Letters

DC/DC Power Converter Control-Based Deep Machine Learning Techniques: Real-Time Implementation

Mojtaba Hajihosseini, Milad Andalibi, Meysam Gheisarnejad, Hamed Farsizadeh, and Mohammad-Hassan Khooban, *Senior Member, IEEE*

***Abstract*—The recent advances in power plants and energy resources have extended the applications of buck-boost converters in the context of dc microgrids (MGs). However, the implementation of such interface systems in the MG applications is seriously threatened with instability issues imposed by the constant power loads (CPLs). The objective is that without the accurate modeling information of a dc MG system, to develop a new adaptive control methodology for voltage stabilization of the dc–dc converters feeding CPLs with low ripples. To achieve this goal, in this letter, the deep reinforcement learning (DRL) technique with the Actor–Critic architecture is incorporated into an ultralocal model (ULM) control scheme to address the destabilization effect of the CPLs under the reference voltage variations. In the suggested control approach, the feedback controller gains of the ULM controller are considered as the adjustable controller coefficients, which will be adaptively designed by the DRL technique through online learning of its neural networks (NNs). It is proved that the suggested scheme will ensure the rigorous stability of the power electronic system, for simultaneous effects of CPL and reference voltage changes, by adaptively adjusting the ULM controller gains. To appraise the merits and usefulness of the suggested adaptive methodology, some dSPACE MicroLabBox outcomes on a real-time testbed of the dc–dc converter feeding a CPL are presented.**

***Index Terms*—Constant power load (CPL), dc–dc buck-boost converter, deep reinforcement learning (DRL), ultralocal model (ULM).**

I. Introduction

RECENTLY, the usage of dc microgrid (MG) has widened in numerous industrial applications due to its more advantages than the ac MG [1]. Despite the advantages of dc MGs, they are faced with an instability problem of the dc–dc converters feeding constant power loads (CPLs) [2]–[4] that can lead to large oscillations in the voltage and frequency terms. The advances in the hardware technologies with high computing power facilitated the practical implementation of advanced control mythologies, e.g., nonfragile controller [1], sliding mode controller (SMC) [5], and backstepping scheme [6], to ameliorate the control performance of dc converters feeding CPLs. By combining a nonlinear disturbance observer and backstepping technique, a composite nonlinear controller is developed in [6] to mitigate the instability imposed by CPLs on the dc MGs. In [7], a systematic and simple state feedback controller has been extended to stabilize the dc MGs with multiple CPLs. To meet the stability and efficient performance, the authors of [7] stated the nonlinear dc MG with some CPLs in a Takagi–Sugeno fuzzy model combined with a quadratic D-stability theory.

In the abovementioned works, the stabilization of the converters is satisfied in the presence of ideal CPLs, however, due to the inevitable uncertainties (e.g., unmodeled dynamics) in practical applications, the robust model-based strategies fail to effectively suppress the CPL's nonlinearity. Moreover, the need for accurate modeling to design the model-based control strategies limits their applicability to handle the process with high nonlinearities. These difficulties motivated researchers to develop their control techniques based on the input–output (I/O) measurements, referred to as data-driven strategies [8], which disappear the modeling procedure and unknown dynamics. The model-independent schemes are one of the most popular data-driven techniques, which are also known as model-independent adjusting or intelligent controllers such as intelligent proportional integral derivative [9] and model-independent nonsingular terminal sliding-mode control (MINTSMC) [10]. Based on the ultralocal model (ULM) concept, the model-independent schemes adopt a quick observer (e.g., extended state observer, sliding mode (SM) observer, etc. [10], [11]) to estimate the unknown terms of the process model. To achieve the optimal performance of the intelligent controllers, the evolutionary algorithms (e.g., genetic algorithm) are often adopted to adjust the design coefficients of the intelligent controllers in a heuristic manner. However, the implementation of such approaches can guarantee the optimal performance of the system only for a specific cycle period and suffer the lack of capability to learn from the observed process data and restricted generalization capability.

Manuscript received January 3, 2020; revised February 9, 2020; accepted February 24, 2020. Date of publication March 2, 2020; date of current version June 23, 2020. *(Corresponding author: Mohammad-Hassan Khooban.)*

Mojtaba Hajihosseini and Milad Andalibi are with the School of Electrical and Computer Engineering, Shiraz University, Shiraz 71946-84471, Iran (e-mail: mojtaba. hajihosseini.ir@gmail.com; m.andalibi@shirazu.ac.ir).

Meysam Gheisarnejad is with the Department of Electrical Engineering, Islamic Azad University, Najafabad Branch, Isfahan 85141-43131, Iran (e-mail: me.gheisarnejad@gmail.com).

Hamed Farsizadeh is with the Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz 71946-84471, Iran (e-mail: hthfarsi@gmail.com).

Mohammad-Hassan Khooban is with the DIGIT, Department of Engineering, Aarhus University, 8200 Aarhus, Denmark (e-mail: khooban@ieee.org).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPEL.2020.2977765

The recent advances in the algorithmic procedure of the reinforcement learning (RL), decreasing the variance in coefficient updates, have enabled the deployment of deep neural networks (NNs) in RL techniques, creating the context of deep RL (DRL) [12]. However, for the complex environments with high-dimensional inputs, the deep deterministic policy gradient (DDPG) [13]–[15] is preferred by performing control actions of an Actor-network to the system. In spite of the DDPG can learn the complex specification of continuous problems, the strong sensitivity of this algorithm to its hyperparameters makes it hard to adjust. With the aim of developing a DRL algorithm with high robustness in training, proximal policy optimization (PPO) [16] was introduced to solve the problem of learning rate selection that has made significant progress for policy search in the RL domains.

This letter explores the potential of the PPO algorithm to tune the ULM control scheme in stabilizing the voltage term of dc–dc buck-boost converters. As one of the majors threatens in the stability of the converter systems, a time-varying CPL is applied to the dc–dc converter to verify the robustness of the suggested adaptive data-driven scheme to deal with the serious stability issue of such systems.

Consequently, the major contribution of this letter is mentioned in the following.

1) The CPLs imposes a destabilizing nonlinear impact on the dc power electronic converters by an inverse voltage, which results in remarkable fluctuations in the voltage term of the main bus or even its collapse. The stability threats of dc converters feedings CPLs is further intensified when the voltage reference is varied during the simulation. In this letter, the simultaneous impact of both the CPL and voltage reference variations in the dc–dc buck-boost converter is investigated to consider the worst condition for such systems from the stability perspective.
2) A new ULM control scheme based on SM observer that can estimate the unknown dc–dc converter dynamics is designed, instead of to developed the global mathematical model of the system.
3) The PPO algorithm with Actor–Critic architecture has been adopted for online adjusting of ULM control coefficients in an adaptive manner. The suggested PPO mechanism, naturally, incorporates the feedback controller gains of the ULM controller into the design goal and offers the ULM controller with online gain designing by employing the learning ability of Actor–Critic NNs.
4) The suggested scheme was experimentally tested and compared with that of a conventional state-of-the-art model-independent scheme, on a laboratory prototype of the buck-boost converter.

II. Model of the Buck-Boost Converter With CPL

The buck-boost converter feeding a time-varying CPL circuit topology is shown in Fig. 1. The average state-space model of the system under certain assumptions is given by [17]

$$L\frac{di}{dt} = -(1-u)\,v + uE \tag{1}$$

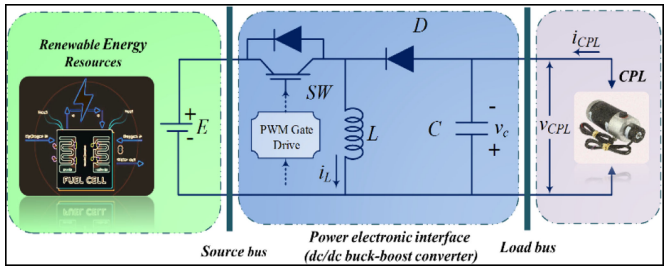

Fig. 1. Circuit representation of the dc-dc buck-boost converter with a CPL.

$$C\frac{dv}{dt} = (1-u)\,i - \frac{P}{v} \tag{2}$$

where i and $v \in \mathbb{R} > 0$ denote the inductor current and the output voltage, respectively, $P \in \mathbb{R} > 0$ denotes the CPL's power, $E \in \mathbb{R} > 0$ denotes the input voltage, and $u \in [0,\ 1]$ denotes the duty ratio of the control switch signal. The certain equilibrium of the dc/dc buck-boost converter feeding a CPL is given by [17]

$$\mathcal{E} := \left\{ (i,v) \in \mathbb{R}^2 > 0 \mid i - P\left(\frac{1}{v} + \frac{1}{E}\right) = 0 \right\}. \tag{3}$$

III. PPO Designed Ultralocal Model Controller-Based SM Observer

For control of dc–dc buck-boost converters feeding CPLs, the voltage stabilization performance of conventional methodologies like fuzzy logic, SMC, and model predictive controller (MPC) are restricted from the following regulatory aspects.

1) Since the time intervals in the simulation of dc–dc buck-boost converters are in the microseconds, the computational time for designing the model-based schemes is too exhaustive to be solved in real time.
2) Due to the destabilization properties and nonlinearities imposed by the CPLs to the converters, the control approaches fail to ameliorate the settling time and overshoot terms simultaneously. This necessitates further efforts to mitigate the CPL destructive effects in an optimal manner and to ensure stability requirements.
3) For a power electronic system, like a buck-boost converter, after a change in operating condition, the control performance of the deterministic techniques deteriorates due to lack of adaptive ability.

Owing to the deficiency of the existing control methodologies, a PPO-based ULM scheme is proposed in this letter, which is designed to address the aforesaid issues. First, a simple model-independent feedback controller-based ULM scheme (i.e., intelligent proportional-integral (PI) controller) is established to reduce the dependency on the converter model and obtain the initial desired control specifications. Then, an SM observer is incorporated into the ULM scheme to estimate the unknown dynamics of the converter and apply its feedback to the controller. Lastly, an adaptive based controller coefficient tuning approach is developed for the optimal setting of the predesigned feedback controller, employing the learning ability of PPO.

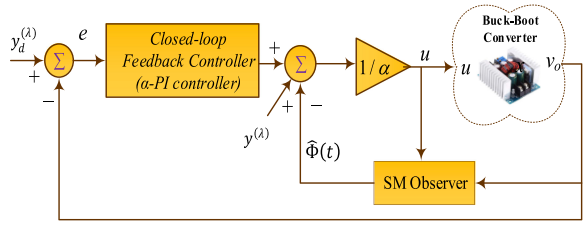

Fig. 2. Structure of the feedback controller with SM observer.

A. Ultralocal Model Control Based on SM Observer

1) Design of Feedback Controller: Based on the knowledge of I/O measurements of the dc–dc converter, some nonlinearities, uncertain parameters, and mathematical complexity of the system can be replaced by the ULM scheme. A numerical model of the ULM mechanism, over a short lapse of time, can be described as [9], [11]

$$y^{(\lambda)}(t) = \Phi + \alpha u(t) \tag{4}$$

where $y^{(\lambda)}(t)$ denotes the λth derivative order of the system output $y(t)$. The total unknown phenomena are represented by Φ; $\alpha \in \mathbb{R}$ denotes a nonphysical constant factor.

Establishing the PI-based factor $\alpha(\alpha - PI)$ controller as the feedback controller, the loop will be closed with the control law of

$$u(t) = \frac{\hat{\Phi}}{\alpha} - \frac{\left(y_d^{(\lambda)}(t) + k_{sp}e_1^{(\lambda)}(t) + k_{si}\int e_1^{(\lambda)}(t)\right)}{\alpha} \tag{5}$$

where k_{sp} and k_{si} are the feedback controller coefficients, which are manually adjusted; $y_d^{(\lambda)}(t)$ is the desired trajectory of the $y(t)$; $e_1(t)$ is the tracking error, which represents the difference between $y_d(t)$ and $y(t)$; and $\hat{\Phi}$ denotes a real-time estimation of $\Phi(t)$.

2) Design of SM Observer: The estimation of Φ is a crucial factor to cancel the influence of the disturbances and unmodeled dynamics included in the practical implementation. For the estimation of Φ, an SM observer that enjoys great robustness in view of the uncertain plant is introduced into the ULM control scheme. The structure of the ULM controller based on the SM observer is depicted in Fig. 2.

According to (4), the term Φ can be calculated using the SM observer, given as [10]

$$\dot{\hat{y}}(t) = \sigma \mathrm{sgn}\left(y(t) - \hat{y}(t)\right) + \alpha u(t) \tag{6}$$

where $\hat{y}$ denotes the estimated value of y; σ denotes the designed variable. The SM observer error is defined as $e_2(t) = y(t) - \hat{y}(t)$. By subtracting (4) from (6), one can obtain the following:

$$\dot{e}_2(t) = \Phi - \sigma sgn\left(y(t) - \hat{y}(t)\right). \tag{7}$$

Theorem I (stability analysis) [10]: After defining the SM manifold $S(t) = e_2$, the estimation observer error will converge to zero if the term σ be properly set.

Proof: As usual, the following term is considered as the Lyapunov function

$$V = \frac{1}{2}S(t)^T S(t). \tag{8}$$

The time derivative of (8) is yielded as

$$\begin{aligned}\dot{V} &= S(t)^T\dot{S}(t) = e_2^T(t)\dot{e}_2(t)\\ &= e_2^T(t)\left(\Phi - \sigma \mathrm{sgn}\left(e_2(t)\right)\right) \le |e_2(t)|\left(|\Phi| - \sigma\right).\end{aligned} \tag{9}$$

Now, assume σ meets the condition of $|\Phi| + \eta < \sigma$, where $\eta > 0$, then one can obtain $\dot{V} \le -\eta|e_2(t)|$ which ensures the observer is stable asymptotically. ■

B. The Proximal Policy Optimization Algorithm

In RL framework, a task can be described by Markov decision process (MDP) characterized by a quintuple $\{S, A, r, p, \gamma\}$, where $S \in \mathbb{R}^n$ denotes the state space, $A \in \mathbb{R}^m$ denotes the action space, $r : S \times A \to \mathbb{R}$ denotes the reward function, $p : S \times A \times S \to [0, 1]$ denotes the transition function, which expresses the probability of transferring a new state s_{t+1}, emitting a reward r under executing action a_t on the state s_t, and $\gamma \in [0\ 1]$ denotes the discount factor. With an initial state s_t, which is arbitrary set, the RL is aimed to maximize the obtained rewards $\epsilon[\sum_{t=0}^{\infty}\gamma^t r_t]$.

PPO is an Actor–Critic and on-policy-based RL algorithm that follows an optimal policy π to act on an environment described as an MDP. Compared with the existing Actor–Critic-based algorithms [13], [14], in many cases, the hyperparameters of PPO are more robust for a large number of tasks and relatively converges quicker [16]. By considering Kullback–Leibler divergence of the policy updates in the optimization process, PPO can guarantee the optimum convergence. The Monte Carlo (MC) method is used to approximate the samples of the policy loss function and the gradients of the loss function are as follows:

$$J(\theta) = \mathbb{E}_{\mathcal{T}\sim\pi_\theta(\tau)}\left[\sum_t R\left(s_t, a_t\right)\right] = \mathbb{E}_{\mathcal{T}\sim\pi_\theta(\tau)}\left[R\left(\tau\right)\right] \tag{10}$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\mathcal{T}\sim\pi_\theta(\tau)}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log\pi_\theta\left(a_t|s_t\right)R\left(\tau\right)\right]\right. . \tag{11}$$

The main object in policy gradient methods is trying to reduce the variance of the gradient estimations toward better policies causing consistent progress. The Actor–Critic architecture by representing a new definition of value function makes a significant impact in this approach:

$$Q^\pi(s,a) = \sum_t \mathbb{E}_{\pi_\theta}\left[R\left(s_t, a_t\right)|s, a\right] \tag{12}$$

$$V^\pi(s) = \sum_t \mathbb{E}_{\pi_\theta}\left[R\left(s_t, a_t\right)|s\right] \tag{13}$$

$$A^\pi(s,a) = Q^\pi\left(s,a\right) - V^\pi\left(s\right). \tag{14}$$

The advantage function $A^\pi(s,a)$ measures how good an action is compared to the others available in that state. The value function $V(s)$ measures how good it is to be in that state. The Critic network separately is trained to predict the value function by analyzing the cumulative receiving rewards. The PPO, as one of the most efficient Actor–Critic methods, aims to maximize the

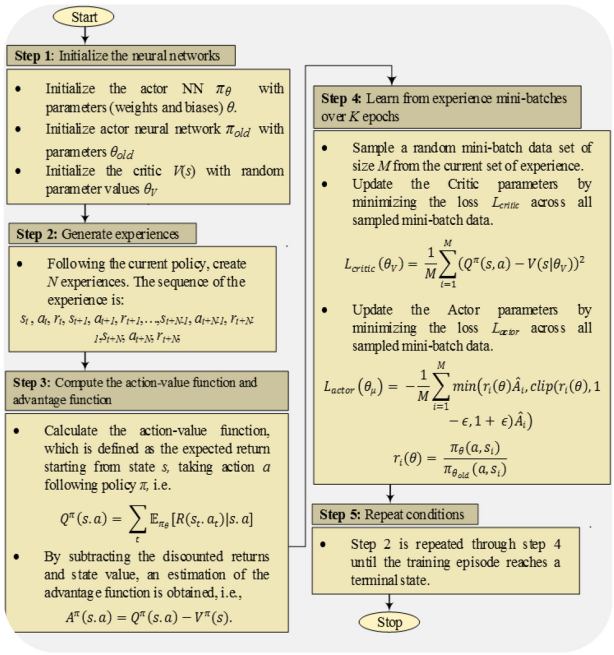

Fig. 3. Flowchart of the PPO algorithm with Actor–Critic architecture.

objective function formulating as follows:

$$L(\theta) = \widehat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}\left(r_t(\theta), 1-\epsilon,\ 1+\epsilon\right)\hat{A}_t\right)\right] \tag{15}$$

where $\hat{A}$ and $\widehat{\mathbb{E}}$ denotes the estimation of the advantage function and expectation, respectively, and $r_t(\theta)$ is the probability ratio defined as follows:

$$r_t(\theta) = \frac{\pi_\theta(a_t, s_t)}{\pi_{\theta_{\text{old}}}(a_t, s_t)}. \tag{16}$$

Vanilla policy gradients require examples of optimal policy making that cannot be applied to the modified policy after one more optimization step. PPO uses the importance of sampling to obtain the expectation of samples from an old policy under the new policy. For this purpose, each sample can be used for several gradient ascent steps. When the new policy is refined, both old and new policies will diverge and result in an increased variance of the estimation, also the old policy would be updated to the new policy. To achieve this goal, a similar state transition function has to exist, which is ensured by clipping the probability ratio to the region $[1-\epsilon,\ 1+\epsilon]$. The algorithmic steps of the PPO scheme are depicted in Fig. 3. (For more details about the learning procedure of the PPO NNs, readers are referring to [18]).

C. PPO-Based Feedback Controller Coefficient Tuner

In this letter, the PPO algorithm, as an adaptive mechanism tuner, is adopted to adjust the feedback controller coefficients of the ULM control scheme by extracting the advantages of the online learning and model-independent property of RL. According to the suggested strategy, the feedback controller (i.e., $\alpha - PI$) coefficients are considered as the design control objective and the PPO tuner adjusts these coefficients through online learning

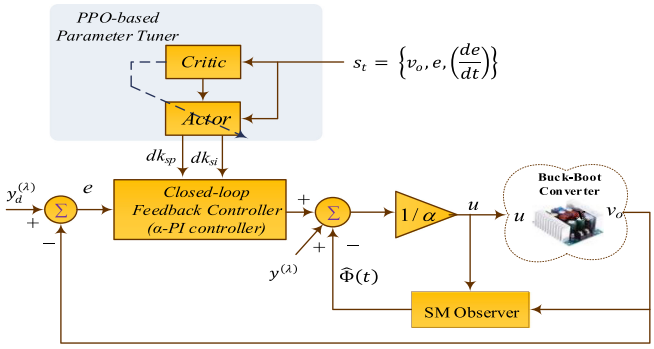

Fig. 4. Self-adaptive ULM control scheme based on the PPO algorithm.

of Actor–Critic NNs. The structure of the proposed adaptive ULM control scheme based on the PPO tuner is illustrated in Fig. 4.

According to Fig. 4, by employing the Actor and Critic NNs, the PPO produces the regulatory commands $[dk_{sp}(t)\ \ dk_{si}(t)]$ to tune the feedback controller coefficients. Since, usually, these control coefficients are nonzero, a feedback controller with $\alpha - PI$ structure is designed as $k_{sp}(t+1) = k_{sp}(t) + dk_{sp}(t)$ and $k_{si}(t+1) = k_{si}(t) + dk_{si}(t)$.

Remark 1: It is noted that the feedback controller in this application is not necessarily a $\alpha - PI$ controller, that is, the suggested adaptive scheme can be extended to other types of adjustable controller coefficients (e.g., MINTSMC controller).

The PPO agent aims to train the coefficients of Actor and Critic NNs in such a way that reduces the output voltage error e, i.e., the error between the desired voltage of the converter v_{ref} with its actual value v_o. The feedbacks in time step t from the buck-boost converter are selected as the voltage output v_o, the output-voltage error e and their derivative $\dot{e}$, i.e., state $= \{v_o,\ e,\ (\frac{de}{dt})\}$. To compensate for the output voltage, the reward signal r_t in the PPO algorithm is set as $r_t = \frac{1}{abs(e_t)}$. Based on the immediate reward r_t, the PPO agent evaluates how good the undertaken actions $[dk_{sp}(t)\ dk_{si}(t)]$. The action (regulatory signals) is applied to the environment (i.e., converter) to eliminate the CPL disturbances occurred during the operation of the power electronic system, that causes the system experiences the transition to a new state and to release a feedback signal r_t. At each time step, the transition vector $(s_t,\ s_{t+1},\ a_t,\ r_t)$ is gathered and each episode is finished when the terminal condition is reached, i.e., the maximum number of steps for each episode is met. Then, the gathered information vector is adopted to train the RL agent by adopting a PPO tuner, training the policy coefficients (weight coefficient and bias of the Actor and Critic NNs). Thus, an updated regulatory signal is produced to adaptively mitigate the influence of the nonideal CPL.

In this letter, two hidden layers (HLs) with 189 and 30 neurons are used in the Actor NN while the Critic NN is built with two completely connected HLs with 30 neurons. The *rectified linear unit* is adopted as a nonlinear mapping function for all HLs in the NNs. With the defined parameters of the PPO algorithm and the NNs configured in Fig. 5, the rest of the parameters for the algorithm configuration are furnished in Table I.

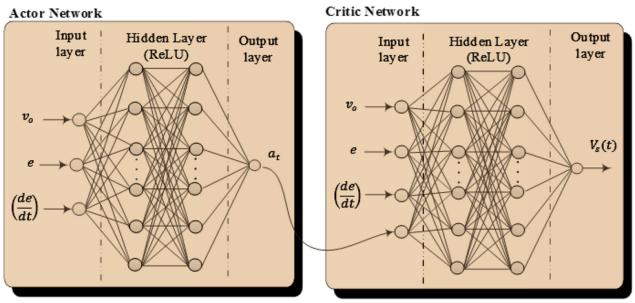

Fig. 5. Feed-Forward NNs as couple Actor–Critic.

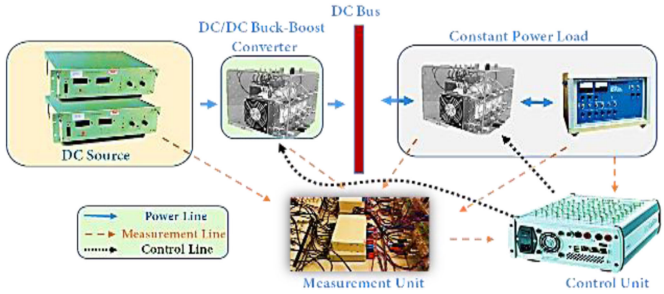

Fig. 6. Photograph of laboratory prototype adopted in the experiment.

TABLE I
PARAMETERS OF THE PPO TUNER

Parameter	Value	Parameter	Value
PPO training episode length	200 ts	Discount factor	0.95
Batch size	1024 eps	Learning rate	0.008
Learning rate of Actor NN	0.008	Number of MC cycles	1000
Learning rate of Critic NN	0.008		

TABLE II
SPECIFICATIONS OF THE DC–DC BUCK-BOOST CONVERTER

Parameter	Value	Parameter	Value
Input voltage, E	48	Capacitance, C	40mF
Reference output voltage, V_{ref}	30, 80, 110	Inductance, L	10μH

Remark 2: The merits of the suggested control scheme are highlighted as follows.

1) Comparing with the conventional ULM control schemes (e.g., MINTSMC controller [10]), the suggested controller is more robust due to its online learning ability to deal with the CPL's instability.
2) Comparing with the nonfragile, MPC, and backstepping controllers that can guarantee the stability of systems only in a particular cycle period [1], [6], the suggested controller can guarantee the stability of systems in various periods during the experiment.
3) Comparing with DRL and DDPG algorithms [12]–[14], training of the PPO is more robust with less sensitivity to its hyperparameters.

IV. EXPERIMENTAL RESULTS

The dc–dc buck-boost converter model explained in Fig. 1 is experimentally tested to appraise the applicability of real-time implementation of the suggested adaptive controller and verify its performance. Applying a CPL under input voltage variations in the buck-boost converter, the worst scenario situation that can be imposed on such power electronic systems is investigated from the stability point of view. A comparative study is accomplished in two typical scenarios to validate the superior transient performance of the intelligent feedback controller-based PPO tuner with the SM observer in the stabilization of the output voltage than that of the MINTSMC scheme [10]. The configuration of the experimental testbed is illustrated in Fig. 6 and the parameters corresponding to the circuit components of this setup are presented in Table II. In the experimental setup of Fig. 6, the output voltage of the buck-boost converter is controlled by a dSPACE MicroLabBox with DS1202 PowerPC Dual Core 2 GHz processor board and DS1302 I/O board. Besides, the real-time simulation executes on Windows 10, Core i7, 2.6 GHz, 8 GB of RAM. The expected design process and fast design performance are realized by the real-time interface (RTI) based on MicroLabBox. The MATLAB C code generator Simulink Coder (formerly Real-Time Workshop) is developed for the automatic execution and seamless implementation of Simulink models on the real-time testbed. The model can be connected to the dSPACE I/O board by, simply, drag the I/O module from the RTI block library onto the model and connect it to the Simulink blocks. By clicking the suitable blocks, any set out or change in the parameters is achievable. To prepare the real-time simulation, the collaboration of Simulink Coder and the RTI produce, respectively, the model code and blocks to implement the I/O capabilities of dSPACE in Simulink models.

Scenario I: In the first scenario, the adaptive model-independent capability of the PPO optimized ULM controller with respect to the ideal CPL and reference voltage variations is studied. The CPL's power is applied as a constant value of $P = 150$ W throughout the real-time simulation while the reference voltage is set as 80 V for $t \in [0, 0.3)$ s, 110 V for $t \in [0.3, 0.7)$ s, and 30 V for $t \in [0.7, 1]$ s. With the concerned scenario, the experimental outcomes of CPL's power (blue line), bus voltage (red line), and CPL's current (green line), respectively, for the suggested controller and the MINTSMC scheme are depicted in Figs. 7 and 8. As shown in Figs. 7 and 8, in response to the constant CPL and reference voltage variations, the output voltage of the suggested controller tracks its references during the experimental analysis while the voltage responses of the MINTSMC scheme experience a small oscillation as the voltage reference varies suddenly, i.e., at $t = 0.3$ s and $t = 0.7$ s. Moreover, the suggested controller provides a quicker and better outcome to restore the power and current of CPL as compared with the MINTSMC scheme.

Scenario II: Here, a nonideal CPL is applied to the converter at a heavy load situation with the initial power of $P = 150$ W.

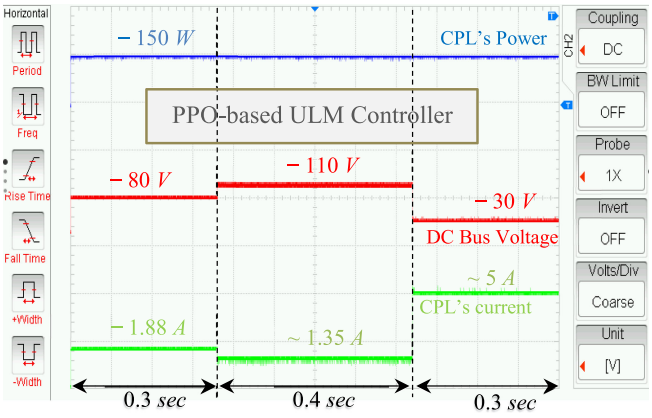

Fig. 7. Transient outcomes of the PPO based-ULM control scheme with SM observer according to Scenario I.

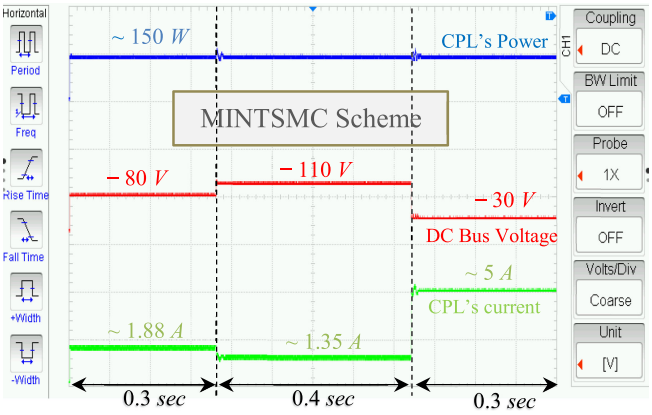

Fig. 8. Transient outcomes of the MINTSMC scheme according to Scenario I.

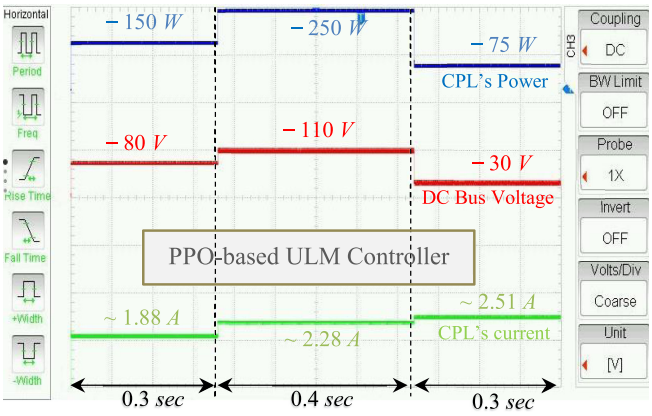

Fig. 9. Transient outcomes of the PPO based-ULM control scheme with SM observer according to Scenario II.

At $t = 0.3$ s, the CPL's power is increased from its initial power to 250 W; at $t = 0.7$ s, the CPL's power is reduced from 250 to 75 W. The experimental outcomes of Figs. 9 and 10 including CPL power (blue line), bus voltage (red line), and CPL's current (green line), respectively, illustrate how the suggested controller and the MINTSMC scheme stabilize the buck-boost converter feeding the nonideal CPL under the reference voltage variations. By comparing the results of Figs. 9 and 10, one can

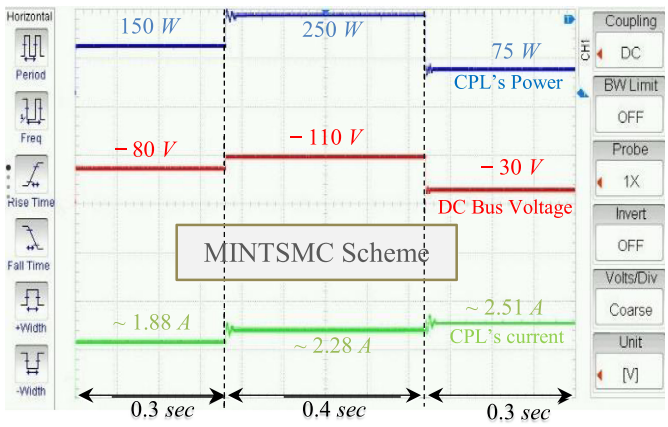

Fig. 10. Transient outcomes of the MINTSMC scheme according to Scenario II.

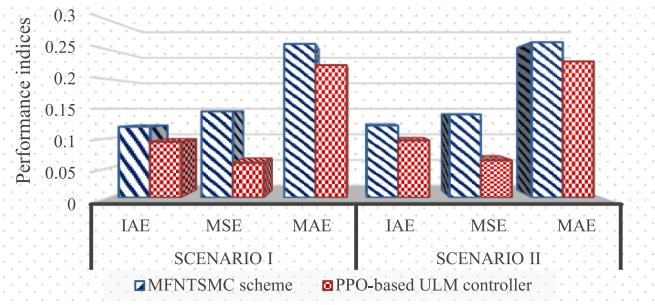

Fig. 11. Bar chart comparison of different performance indices.

observe that the experimental outcomes of the suggested ULM control scheme (realized based on the PPO agent) experience minor control degradation than that of the MINTSMC scheme; specifically, within the range [0.3 s 0.7 s] of operation, where the highest CPL's power is applied to the converter, the set points of the voltage term are precisely tracked while simultaneously the power and current of CPL experience smaller deviation from their nominal values. Such improvement in the converter stabilization performance is valuable in the power electronic engineering, to protect the CPL in which it is connected to, from possible great deviations of system responses during a long transient.

For quantitative analysis of the suggested controller, several error measurement criteria including integral absolute error, mean square error, and mean absolute error corresponding to the Scenario I and Scenario II are compared using bar charts, as depicted in Fig. 11.

V. Conclusion

The robust stabilization problem of a class of power electronic systems exposed to the dynamic loads has been studied in this letter. Particularly, by employing the adaptive capability of DRL, a novel adaptive model-independent ULM controller-based SM observer has been developed to suppress the destructive effects of CPLs when the system is subjected to the reference voltage changes. This control strategy can achieve promising outcomes

due to the following two reasons. With the aim of practical implementation, an SM observer is incorporated into the ULM control scheme that ensures good compatibility with the unmodeled system dynamics and the learning ability of the PPO agent keeps driving the feedback controller to its optimal point, where the coefficients in the Actor and Critic NNs are trained under the CPL and the reference voltage changes.

The experimental outcomes of the prototype confirm an excellent transient behavior in the voltage responses of the dc–dc buck-boost converter with the use of suggested adaptive methodology than the MINTSMC controller.

REFERENCES

[1] N. Vafamand, M. H. Khooban, T. Dragicevic, F. Blaabjerg, and J. Boudjadar, "Robust non-fragile fuzzy control of uncertain DC microgrids feeding constant power loads," *IEEE Trans. Power Electron.*, vol. 34, no. 11, pp. 11300–11308, Nov. 2019.

[2] S. R. Huddy and J. D. Skufca, "Amplitude death solutions for stabilization of DC microgrids with instantaneous constant-power loads," *IEEE Trans. Power Electron.*, vol. 28, no. 1, pp. 247–253, Jan. 2012.

[3] M. H. Khooban, M. Gheisarnejad, H. Farsizadeh, A. Masoudian, and J. Boudjadar, "A new intelligent hybrid control approach for DC/DC converters in zero-emission ferry ships," *IEEE Trans. Power Electron.*, vol. 35, no. 6, pp. 5832–5841, Jun. 2020.

[4] H. Farsizadeh, M. Gheisarnejad, M. Mosayebi, M. Rafiei, and M. H. Khooban, "An intelligent and fast controller for DC/DC converter feeding CPL in a DC microgrid," *IEEE Trans. Circuits Syst. II: Express Briefs*, 2019.

[5] S. Singh, D. Fulwani, and V. Kumar, "Robust sliding-mode control of DC/DC boost converter feeding a constant power load," *IET Power Electron.*, vol. 8, no. 7, pp. 1230–1237, Jul. 2015.

[6] Q. Xu, C. Zhang, C. Wen, and P. Wang, "A novel composite nonlinear controller for stabilization of constant power load in DC microgrid," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 752–761, Jan. 2019.

[7] M. M. Mardani, N. Vafamand, M. H. Khooban, T. Dragičević, and F. Blaabjerg, "Design of quadratic D-stable fuzzy controller for DC microgrids with multiple CPLs," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4805–4812, Jun. 2019.

[8] J. Sun, J. Yang, W. X. Zheng, and S. Li, "GPIO-based robust control of nonlinear uncertain systems under time-varying disturbance with application to DC–DC converter," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 63, no. 11, pp. 1074–1078, Nov. 2016.

[9] H. Abouaïssa and S. Chouraqui, "On the control of robot manipulator: A model-free approach," *J. Comput. Sci.*, vol. 31, pp. 6–16, 2019.

[10] K.-H. Zhao *et al.*, "Robust model-free nonsingular terminal sliding mode control for PMSM demagnetization fault," *IEEE Access*, vol. 7, pp. 15737–15748, 2019.

[11] H. P. Wang, G. I. Y. Mustafa, and Y. Tian, "Model-free fractional-order sliding mode control for an active vehicle suspension system," *Advances Eng. Softw.*, vol. 115, pp. 452–461, 2018.

[12] L. Huang, X. Feng, C. Zhang, L. Qian, and Y. Wu, "Deep reinforcement learning-based joint task offloading and bandwidth allocation for multi-user mobile edge computing," *Digit. Commun. Netw.*, vol. 5, pp. 10–17, 2019.

[13] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2124–2136, Mar. 2019.

[14] Y. Wang, J. Sun, H. He, and C. Sun, "Deterministic policy gradient with integral compensator for robust quadrotor control," *IEEE Trans. Syst., Man, Cybern., Syst.*, 2019.

[15] M. Gheisarnejad, J. Boudjadar, and M.-H. Khooban, "A new adaptive type-II fuzzy-based deep reinforcement learning control: Fuel cell air-feed sensors control," *IEEE Sens. J.*, vol. 19, no. 20, pp. 9081–9089, Oct. 2019.

[16] X. Wang, T. Li, and Y. Cheng, "Proximal parameter distribution optimization," *IEEE Trans. Syst., Man, Cybern., Syst.*, 2019.

[17] W. He and R. Ortega, "Voltage regulation in buck–boost coniverters feeding an unknown constant power load: An adaptive passivity-based control," 2019, *arXiv:1909.04438*.

[18] Y. Zhang, Z. Deng, and Y. Gao, "Angle of arrival passive location algorithm based on proximal policy optimization," *Electronics*, vol. 8, p. 1558, 2019.