Journal of Control and Decision

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjcd20

Robustness enhancement of DRL controller for DC–DC buck convertersfusing ESO

Tianxiao Yang, Chengang Cui, Chuanlin Zhang & Jun Yang

To cite this article: Tianxiao Yang, Chengang Cui, Chuanlin Zhang & Jun Yang (2023): Robustness enhancement of DRL controller for DC–DC buck convertersfusing ESO, Journal of Control and Decision, DOI: 10.1080/23307706.2023.2201587

To link to this article: https://doi.org/10.1080/23307706.2023.2201587

Published online: 25 Apr 2023.

JOURNAL OF CONTROL AND DECISION
https://doi.org/10.1080/23307706.2023.2201587

Taylor & Francis
Taylor & Francis Group

Robustness enhancement of DRL controller for DC–DC buck converters fusing ESO

Tianxiao Yang [a], Chengang Cui [a], Chuanlin Zhang [a] and Jun Yang [b]

[a]Intelligent Autonomous Systems Lab, Shanghai University of Electric Power, Shanghai, People's Republic of China; [b]Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough, UK

ABSTRACT

Recent application studies of deep reinforcement learning (DRL) in power electronic systems have successfully demonstrated its superiority over conventional model-based control design methods, stemming from its adaption and self-optimisation capabilities. However, the inevitable gap between offline training and real-life application presents a significant challenge for practical implementation, owing to its insufficient robustness. With this in mind, this paper proposes a novel robust DRL controller by fusing an extended state observer (ESO) for the DC–DC buck converter system feeding constant power loads (CPLs). To be specific, the mismatched lumped terms are reconstructed by an ESO in real time, and then fed forward into the agent's action, aiming to improve the adaptability to parameter variations of the real-life converter systems. By carefully conducting simulation and experimental tests, the robustness enhancement ability of the proposed framework compared with model-free DRL and conventional PI controllers are clearly verified.

ARTICLE HISTORY
Received 9 November 2022
Accepted 7 April 2023

KEYWORDS
Deep reinforcement learning; extended state observer; DC–DC buck converter; robustness enhancement

1. Introduction

Reinforcement learning (RL), inspired by exploration behaviour in nature, is a kind of learning procedure to facilitate intelligent systems with human-like learning and reasoning capabilities by interacting with the environment (Kaelbling et al., 1996). Combining traditional optimal control with adaptive control algorithms, RL has numerous advantages due to its effectiveness in obtaining optimal performance, features of partly/totally data-driven and adaptability to uncertain systems. Benefiting from the decision-making ability of RL, new developments have been seen in the power electronics industry (Wang et al., 2019), see for instance, self-tuning maximum power point tracking scheme in PV power systems (Lin et al., 2021), anomaly detection for inverter (Bandyopadhyay et al., 2018), remaining useful life prediction for super-capacitors (El Mejdoubi et al., 2017), etc.

Standard RL algorithm learns how to realise a specific work through 'trial and error' rule, with both exploration and exploitation balanced to fulfil better performance (Coggan, 2004). The core obstruction is that the sampling efficiency is exceedingly low, and it is necessary to repeatedly interact with the environment to muster data to train the agent (Yu, 2018). However, a bulky number of explorations in the interaction progress will cause serious cost losses in many real-world scenarios. Moreover, it is well known that the agent may fail to converge in the early exploration stage. If trained in a real-life power electronics system, problems such as voltage collapse and current runaway may occur, which are fatal to power electronic components and may cause safety problems (Cao et al., 2020). Fortunately, offline RL provides a possibility of realisation that learns from the collected trajectory data instead of interacts directly with the real environment (Kumar et al., 2020). In other words, offline RL researches how to maximise the use of static offline datasets to train intelligent agents. Generally speaking, the idea of obtaining the optimal policy through offline RL and then transferring it to the actual system can effectively handle the problems of standard RL methods mentioned above.

However, power electronic systems have specific challenges and characteristics, such as high switching frequencies in control, complex operating conditions and so on. The implementation of RL in power electronics has its own characteristics different from other engineering fields. Although many scholars have tried to apply RL methods in the field of power electronics, there are certain limitations in stability, safety, robustness, etc. (Peng et al., 2019). To ensure the safety and stability of the system, numerous composite control strategies have been proposed. Typically, the parameters of these controllers are tuned by a linearised model of the system under specific operating conditions whilst the integration of more power electronic interfaces and loads makes it more challenging. Thus

CONTACT Chengang Cui cgcui@shiep.edu.cn

the existing literature on reinforcement learning (RL) in power electronics systems mainly focuses on combining RL with classical controllers to achieve self-tuning ability. For example, studies have combined adaptive deep deterministic policy gradient (DDPG) compensators with iPI controllers (Gheisarnejad et al., 2020), used PPO-based feedback controllers for coefficient tuning (Hajihosseini et al., 2020), employed data-driven PPO model predictive control (Prag et al., 2021) and developed the composite nonrecursive DRL controller (Huangfu et al., 2022). In contrast, there are limited works using RL as a basic controller, and there are still certain limitations. For instance, a model-free DRL controller based on duty ratio mapping (DRM) was proposed in Cui et al. (2022), however, it requires reacquisition of the DRM when the system varies slightly.

Therefore, in a preliminary conference paper (Yang et al., 2022), some exploratory ideas are introduced for a buck converter system with different component parameters by simulation tests. However, the comprehensive experimental evaluations are still left open. In this paper, we propose a robust DRL controller applied to a real-life DC–DC buck converter feeding constant power loads (CPLs). On one hand, the CPLs have negative impedance characteristics, which will negatively impact the system performance (Xu et al., 2017). The system damping will decrease, leading to system instability when the CPLs interact with their respective source converters (Kwasinski & Onwuchekwa, 2010). On the other hand, the adaptability to variable component parameters provides a potential that the agent only needs to be trained once and hence can be applied to different systems with feed-forward compensation. Thanks to the active disturbance rejection control (ADRC) algorithm proposed by Han (2009), an inner-loop controller that partially contains information of system model can be used to enhance the robustness of the offline-trained DRL controller. The extended state observer (ESO) effectively observes, compensates and adjusts for total perturbations with reduced modelling requirements compared to traditional disturbance observers. As a result, it has seen widespread use in complex nonlinear systems such as PMSM motion control (Xu et al., 2019; Zhang et al., 2020) and UAV trajectory control (Li et al., 2022; Qi et al., 2021). These applications often incorporate ESO with advanced control techniques like slide mode control and model predictive control, which compared to RL, tend to have higher modelling demands.

In this paper, the key to compensating the real-time uncertainties and system variations is applying an ESO to the inner loop controller, as shown in Figure 1. To summarise, the main contributions of this paper are listed as follows.

- The proposed method incorporates an ESO in the DRL controller to compensate for mismatched lumped terms, ensuring robustness and nominal performance even for real-life systems with significant deviations from their simulation models.
- The use of ESO is independent of the specific DRL algorithm and can be applied generally to improve the robustness of any DRL controller.

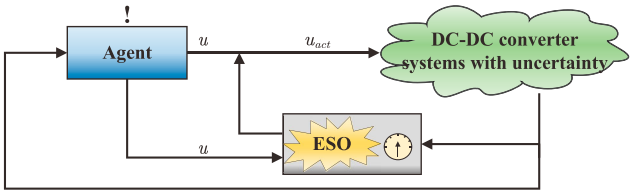


Figure 1. Brief sketch of enhancing the robustness of DRL controller.

The rest of this paper is organised as follows. Section 2 contains a short review of the basic structure of the DC–DC converter buck system. Section 3 proposes a new method combined with DRL and ESO to compensate the mismatched disturbances between the offline-training environment and the real-life DC–DC buck system. Section 4 discusses the simulation and experimental results to verify the validity of the proposed controller. Finally, Section 5 concludes the paper and points the future work.

2. Problem formulation

Normally, the environment of offline training stage is the simplified model expressed in mathematical form of the real-life systems. The system unmodelled dynamic of external disturbances and uncertainties is generally not considered in simulations (Zhao et al., 2020). Therefore, the offline-trained DRL agent performs satisfied control characteristics after training iterations but may cause instability when applied to the real-life system due to the mismatched disturbances between the simulation environment and the platform, which is depicted in Figure 2.

Figure 3 shows the general simplified DC–DC buck converter topology connected to constant power loads (CPLs). The system dynamics can be linearly described as Zhang et al. (2019)

$$i_L = C_0 \frac{\mathrm{d}v_c}{\mathrm{d}t} + i_o + \tau_0(t), \tag{1}$$

where $\tau_0(t)$ denotes the uncertainties between the offline-training environment and real-time external disturbances. For the CPL, i_o is depicted as $i_o = \frac{P_{CPL}}{v_c}$. Meanwhile, the voltage across the diode called switch voltage v_i is depicted as

$$v_i = L_0 \frac{\mathrm{d}i_L}{\mathrm{d}t} + v_c. \tag{2}$$

Specifically, in the constant current mode (CCM) of the buck circuit systems, the switch voltage is approximated

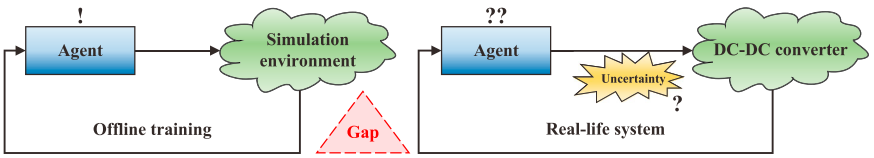


Figure 2. Mismatched terms between offline-trained environment and implementation.

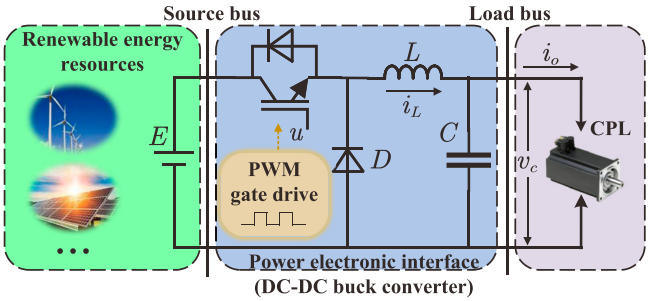


Figure 3. Simplified circuit of the DC–DC buck converter with CPL.

as the product of the duty ratio u and the input voltage E (Davoudi et al., 2006). Then, we can replace v_i with Eu. Define $x_1 = v_c$, the system can be transferred into a controllable canonical form. The derivative of x_1 yields

$$\dot{x}_1 = \frac{1}{C}(i_L - i_o + \tau_0(t)). \tag{3}$$

Setting $\dot{x_1}$ as x_2, the system can be transferred into

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = \dfrac{1}{C}(\dot{i}_L - \dot{i}_o + \dot{\tau}_0(t)) \\ \quad = \dfrac{1}{LC}(Eu - x_1) - \dfrac{1}{C}(\dot{i}_o - \dot{\tau}_0(t)) \\ \quad = b_0 u + f(t, x_1(t)) + \kappa(t), \\ y = x_1, \end{cases} \tag{4}$$

where $b_0 = \frac{E}{LC}$, $f(t, x_1(t)) = -\frac{1}{LC}v_c$. $\kappa(t) = -\frac{1}{C}(\dot{i}_o - \dot{\tau}_0(t))$ is a lumped uncertainty term which consists of inevitable system internal uncertainties, external disturbances and unmodelled dynamics.

3. Main framework

In this section, an extended state observer is utilised to estimate the lumped uncertainty term. The primary observed values are the states of the test bench (i.e. the real-time output voltage and current and the derivative of voltage), the disturbances occurred by unmodelled dynamics (i.e. the mismatched circuit parameters such as resistance, capacitance and inductance, the sampling circuits and protection circuits).

3.1. Controller design based on DRL

The design procedure of the specific DRL controller mainly consists of four parts by using the DQN algorithm (Cui et al., 2021). S_t is the state space, i.e. the state of various information that the agent can obtain from the environment. $r_t = r(s_t, a_t, a_{t+1})$ is the reward function, which is used to judge the merit of training at moment t and feed back to the agent to decide the next action. a_t is the discrete action space, which is the command signal required by the system in this paper and is only related to the previous action.

State Space: For a DC–DC buck converter system, the control purpose is to regulate the bus voltage, which depends on the output voltage, inductor current, as well as circuit parameters including value of resistor, capacitor and inductance, parasitic circuit of switch tube, line impedance, etc. For a specific class of buck converters, state variables other than real-time voltage and current are inherent characteristics of the environment and can be regarded as unknown constants. The design process should take into account the training difficulty, so in order not to increase the computational complexity of the neural network, the proposed DRL controller is designed without considering the inherent parameters of the environment. At last, the state S is defined as $S_t = \{v_o,\ \dot{v}_o,\ v_o^d,\ \Delta v_t,\ \Delta\dot{v}_t,\ \Delta v_t^d\}$, where the superscript d denotes a delayed signal that is used to prevent divergence during the first training episode.

Action Space: The duty cycle required by the system at each moment is given by action a_t chosen through a policy π that provides a distribution of possible actions for each case. The action range and action reference value are derived from the duty ratio by classical double loop PI controller specified as $a_t \in \{a_{\min};\ c;\ a_{\max}\}$. $a_{\min}$, $a_{\max}$ are the upper and lower bounds of 0.45 and 0.55 respectively. c is the minimum incremental of each action set to 0.01.

Reward Function: A reward or a penalty related to the deviation is needed to make the bus voltage close to the reference value. By using the reward function (5), the network can quickly train to the vicinity of the

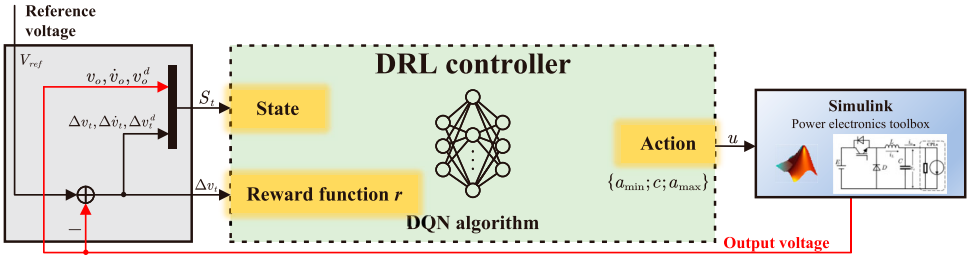


Figure 4. Design of the DRL controller.

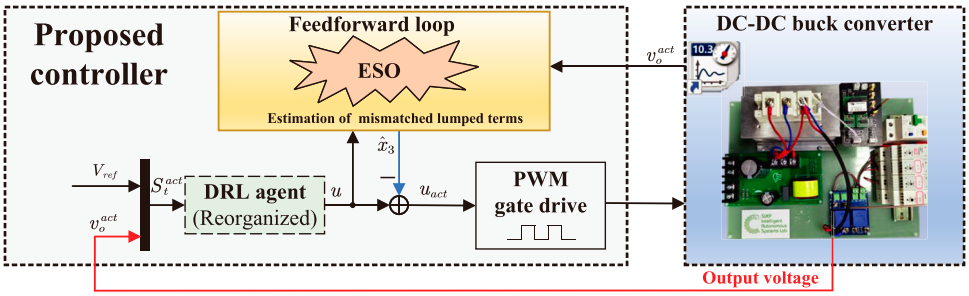


Figure 5. Structure of the proposed composite DRL controller by fusing an extended state observer.

optimal solution.

$$r = \begin{cases} 10 - |\Delta v_t|, & \text{if} |\Delta v_t| < 0.1; \\ 1 - |\Delta v_t|, & \text{if} 0.1 \le |\Delta v_t| \le 1; \\ -10|\Delta v_t|, & \text{else.} \end{cases} \tag{5}$$

Structure of the neural network: The number of hidden layers of the neural network is two, and in each hidden layer, the number of nodes M and N is 64 and 64 respectively. The activation functions of the hidden layer and the output layer are the two ReLU functions and a linear function.

By doing so, the whole structure and offline training procedure of the DRL controller is shown in Figure 4.

3.2. Mismatched lumped terms estimate and rejection via ESO

In mathematical analysis (4), $f(t, x_1(t))$ and $\omega(t)$ do not need to be known explicitly. From the perspective of feedback control, i.e. $F(t) = f(t, x_1(t)) + \kappa(t)$ can be defined as the lumped disturbance needed to be overcome by the control signal. Treat $F(t)$ as an additional state variable artificially, $x_3 = F(t)$. Then letting $\dot{F}(t) = G(t)$, with $G(t)$ unknown, the original plant in (4) can be transferred to

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3 + b_0 u, \\ \dot{x}_3 = G(t), \\ y = x_1. \end{cases} \tag{6}$$

Later on, a linear extended state observer is designed as

$$\begin{cases} \dot{\hat{x}}_1 = \hat{x}_2 + \beta_1 (y - \hat{y}), \\ \dot{\hat{x}}_2 = \hat{x}_3 + b_0 u + \beta_2 (y - \hat{y}), \\ \dot{\hat{x}}_3 = \beta_3 (y - \hat{y}), \end{cases} \tag{7}$$

where $[\beta_1 \quad \beta_2 \quad \beta_3]^T$ is an observer gain vector with its components being corresponding to the coefficients of a Hurwitz polynomial. By choosing the appropriate gain β_1, β_2, β_3, ESO enables real-time tracking of each variable in the system (6), i.e. $\hat{x}_1 \to y$, $\hat{x}_2 \to \dot{y}$, $\hat{x}_3 \to F(t)$.

In addition, considering that the inputs of ESO are the system output y and the control signal u from the DRL controller, and the output of the ESO provides the important information $F(t) = f(t, x_1(t)) + \kappa(t)$. By subtracting $\hat{x}_3$, the original object can be regarded as a control problem of the DRL controller with feedforward compensation for interference signals, then the updated control law is given by

$$u_{act} = u - \frac{\hat{x}_3}{b_0}. \tag{8}$$

Finally, the detailed structure of the proposed method is shown in Figure 5, where $*^{act}$ represents signals in practical applications. The key to the proposed method is the feedforward loop to estimate and compensate the mismatched lumped terms in real-life applications.

3.3. The convergence of extended state observer

The characteristic polynomial of ESO (7) is constructed as

$$\lambda(s) = s^3 + \beta_1 s^2 + \beta_2 s + \beta_3. \tag{9}$$

Table 1. Specifications of theDC–DC buck converter.

Operating parameters	Input voltage E	Bus voltage V_{ref}	Switching frequency f_s	Sampling rate
Value	200 V	100 V	20 kHz	40:1
	Case I	Case II	Case III	Case IV
Component parameters	L, C[a]	L, C	L, C	L, C
Value	0.5 mH, 1.0 mF	1.0 mH, 1.0 mF	1.5 mH, 1.0 mF	2.0 mH, 1.0 mF

[a] represents the abbreviations of inductance and capacitance.

Letting the ideal characteristic polynomial $\lambda(s) = (s+\omega_o)^3$, we have

$$\beta_1 = 3\omega_o,\ \beta_2 = 3\omega_o^2,\ \beta_3 = \omega_o^3. \tag{10}$$

where ω_o is the bandwidth of ESO.

Combining equation (7) and (10), the transfer functions of $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$ can be obtained as

$$\begin{cases} \hat{x}_1 = \dfrac{3\omega_o s^2 + 3\omega_o^2 s + \omega_o^3}{(s+\omega_o)^3} y + \dfrac{b_0 s}{(s+\omega_o)^3} u, \\ \hat{x}_2 = \dfrac{\left(3\omega_o^2 s + \omega_o^3\right) s}{(s+\omega_o)^3} y + \dfrac{b_0 (s+3\omega_o) s}{(s+\omega_o)^3} u, \\ \hat{x}_3 = \dfrac{\omega_o^3 s^2}{(s+\omega_o)^3} y - \dfrac{b_0 \omega_o^3}{(s+\omega_o)^3} u. \end{cases} \tag{11}$$

Letting the tracking error $e_1 = \hat{x}_1 - y$, $e_2 = \hat{x}_2 - \dot{y}$, we get

$$\begin{cases} e_1 = -\dfrac{s^3}{(s+\omega_o)^3} y + \dfrac{s}{(s+\omega_o)^3} b_0 u, \\ e_2 = -\dfrac{(s+3\omega_o) s^3}{(s+\omega_o)^3} y + \dfrac{s}{(s+\omega_o)^3} b_0 u. \end{cases} \tag{12}$$

Denote $e_3 = \hat{x}_3 - F(t)$. Combining with (4), $F(t) = x_3 = \dot{x}_2 - bu = \ddot{y} - bu$, then we have

$$\begin{aligned} e_3 = z_3 - \ddot{y} + b_0 u &= b_0 \left(1 - \frac{\omega_o^3}{(s+\omega_o)^3}\right) u \\ &- \left(1 - \frac{\omega_o^3}{(s+\omega_o)^3}\right) s^2 y. \end{aligned} \tag{13}$$

Taking into account the typicality of the analysis, y, u both take the step signal with the amplitude K, i.e. $y(s) = K/s$, $u(s) = K/s$, then the steady-state error can be obtained

$$\begin{cases} e_{1s} = \lim\limits_{s\to 0} s e_1 = 0, \\ e_{2s} = \lim\limits_{s\to 0} s e_2 = 0, \\ e_{3s} = \lim\limits_{s\to 0} s e_3 = 0, \end{cases} \tag{14}$$

which shows that the ESO has satisfactory convergence and estimation ability and can achieve indifference estimation of system state variables and generalised disturbances (Zheng et al., 2007).

Remark 3.1: Previous literature, such as Fan et al. (2020), has established methods to prove the stability of the DQN algorithm framework used in this study. However, this paper does not focus on the process of proving stability because it is a widely studied issue in the field of RL.

Remark 3.2: This paper has demonstrated the stability of the proposed linear ESO, while the stability of the DQN algorithm has been established in prior literature. The combination of these results enables the verification of the stability of the proposed robust DRL controller.

4. Simulation and experimental results

This section presents the effectiveness of the proposed strategy via simulations and experiments. To verify the robustness of the proposed methods against the changes of component parameters, four cases are considered. The detailed parameters corresponding to the circuit components of this DC–DC buck converter are listed in Table 1. Meanwhile, the proposed controller is compared with a classical double-loop PI controller, where an inner current loop and an outer voltage loop are used to control the DC bus voltage cooperatively.

In the preparation stage, the approximate optimal PI parameters are determined through frequency domain analysis, followed by a trial-and-error process. The hyper-parameters of the DRL controller play a crucial role in the neural network structure and the model training process and cannot be altered during training. These parameters impact the architectural decisions of the model, such as the choice of activation function and the number of neurons in hidden layers, and also determine the efficiency and accuracy of the training process, such as mini batch size and exploration rate (Ripley, 1993). As a result of extensive training, a relatively optimal set of hyper-parameters was obtained, which is documented in Table 2 along with the PI parameters. The detailed parameters used in the experiments are presented in Table 3. The parameter β was determined through trial-and-error, while b_0 was calculated based on E, L, C.

Finally, the proposed method is experimentally tested on a custom-designed DC platform, as shown in Figure 6. The detailed devices consist of a DC power supply (Chroma 62012P-600-8), a DC electronic load (Chroma 63202E-150-200), a DC–DC buck converter and dSPACE 1202. All waveforms are obtained by a digital oscilloscope on the same scale.

Table 2. Details of the PI controller, hyper-parameters of the DRL controller and ESO in the simulation.

PI	Voltage loop		Current loop	
	k_{pv}[b] 3.3	k_{iv}[c] 394	k_{pc} 0.02	k_{ic} 200
DRL	Learning rate 0.001	Discount factor 0.95	Mini batch size 256	Experience buffer length 2×10^5
	Exploration rate 0.1	Training episode length 100	Number of neurons (M) 64	Number of neurons (N) 64
ESO	b_0 2×10^8	β_1 1.8×10^4	β_2 1.08×10^8	β_3 2.16×10^{11}

[b,c] represent the proportional (k_{p*}) and integral (k_{i*}) coefficient.

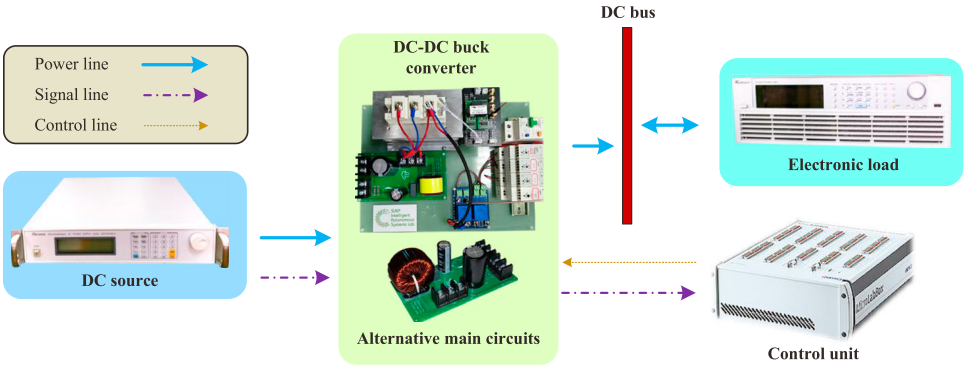


Figure 6. Experimental setup.

Table 3. Details of the PI controller and the proposed controller in the experiments.

PI	Voltage loop		Current loop	
	k_{pv} 2	k_{iv} 83	k_{pc} 0.02	k_{ic} 30
ESO	b_0 2×10^8	β_1 3×10^3	β_2 3×10^6	β_3 1×10^9

Simulation with and without ESO: The simulation was performed with $L = 1$ mH, $C = 1$ mF, the raw parameters used to train the agent. The case involves a load variation from 200 W to 800 W at 0.14 s and a drop from 800 W to 200 W at 0.2 s. An ESO was designed with $\omega_o = 6000$. The voltage response was expected to quickly recover to the reference bus voltage (V_{ref}) with minimal overshoot. The simulation results, shown in Figures 7(c and d), indicate that the proposed method with ESO provides improved voltage recovery performance in terms of settling time and overshoot and reduces fluctuations and peaks in the current response.

Subsequently, we performed simulations with variations in the inductance (L). The results, depicted in Figures 7(a–b, e–h), demonstrate the efficacy of the proposed approach in compensating for real-time disturbances and adapting to system parameter variations. These results are consistent with those obtained in the previous case ($L = 1$ mH, $C = 1$mF). The ESO proves effective in both compensating real-time disturbances, such as those due to CPL variations, and enhancing the adaptability to variations in system parameters.

Simulation comparison results with PI controller: In a comparison between the PI controller and the proposed method Figures 8(a–d), the results show the adaptability of the latter. While the parameters of both controllers were tuned once and remained unaltered throughout the simulation, the PI controller was unable to maintain desirable performance with variations in inductance, while the proposed method remained consistent across different cases. This highlights the enhanced adaptability of the proposed method.

Experimental results: Figure 9 presents the experimental results for the two comparison groups.

- **Columns A and C**: In our study, the adaptability of the proposed control approach is demonstrated through comparisons with a PI controller. The parameters of both controllers are tuned in the same scenario (Case II, with $L = 1$ mH and $C = 1$ mF). Subsequently, the transient-time control performance of the proposed approach is evaluated in scenarios with varying circuit component parameters (Cases I, III and IV). Results show that the proposed controller outperforms the PI controller in terms of maximum overshoot (less than 2 V for the proposed approach, more than 4 V for PI) and settling time (5 ms for the proposed approach, 30 ms for PI). Additionally, the PI controller exhibits voltage oscillation in scenarios with large variations in circuit components as shown in Figures 9(IV-A).
- **Columns B and C**: By incorporating an ESO into the DRL controller, the voltage response fluctuation is reduced, and there is improvement in the transient-time control performance, as especially demonstrated

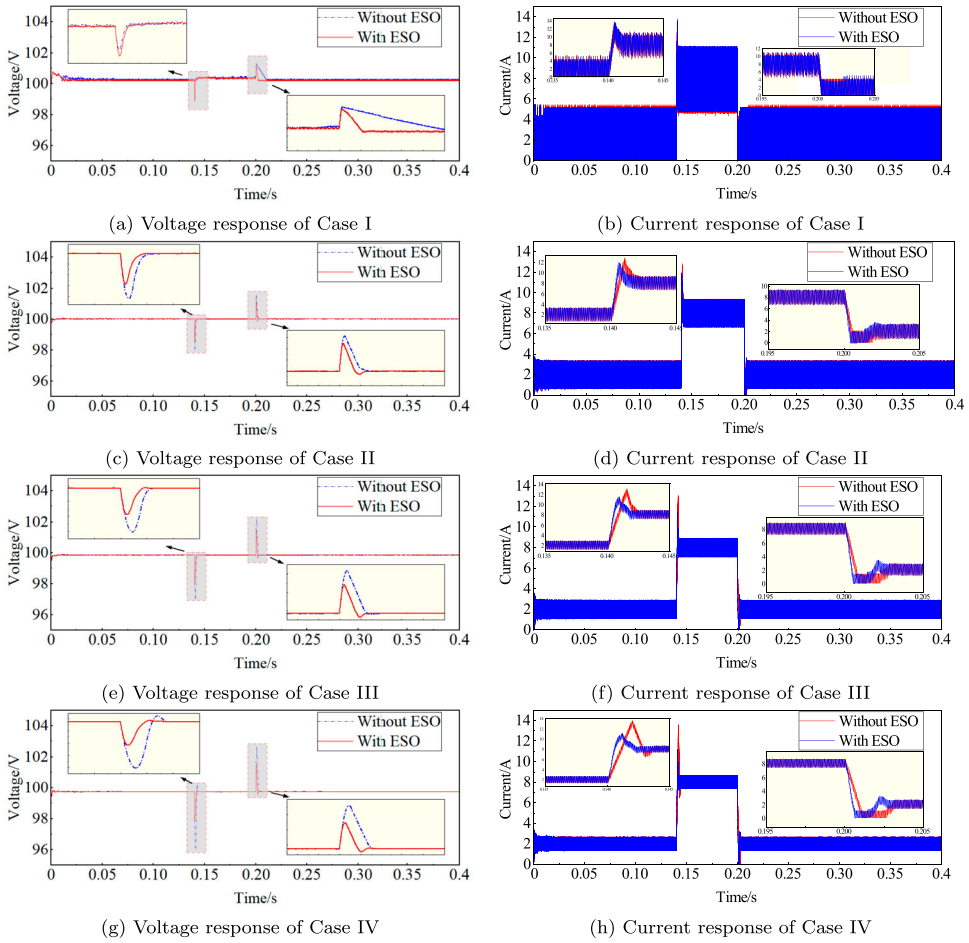


Figure 7. Simulation results of voltage and current response curves in different systems. (a) Voltage response of Case I, (b)current response of Case I, (c)voltage response of Case II, (d)current response of Case II, (e)voltage response of Case III, (f)current response of Case III, (g)voltage response of Case IV, (h)current response of Case IV.

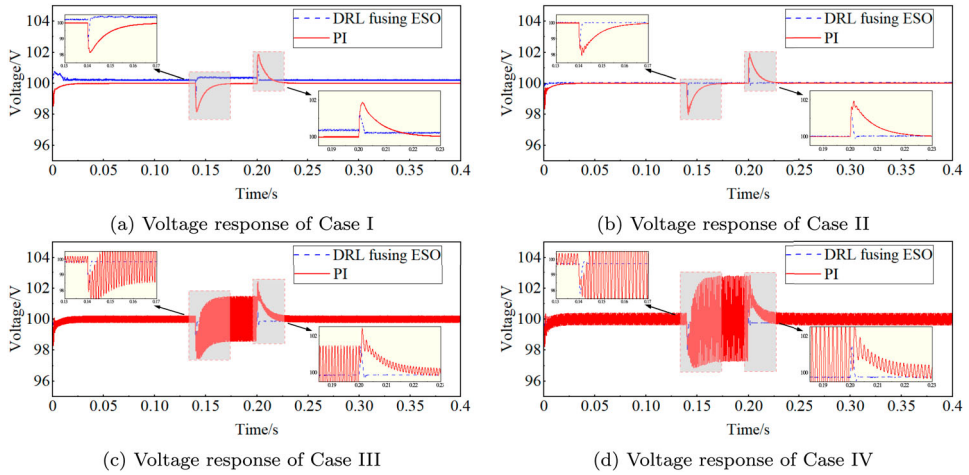


Figure 8. Comparison results of PI controller and proposed controller: (a) voltage response of Case I, (b)voltage response of Case II, (c)voltage response of Case III and (d)voltage response of Case IV.

in Figures 9(I-B) and (I-C). The ESO effectively compensates for mismatched lumped terms, enabling disturbance rejection and DRL implementation.

• **Row I-IV**: The proposed controller demonstrates better performance than the PI controller, even when the inductance parameter L varies by a factor of 4 (from 0.5 mH to 2 mH). The PI controller's control performance remains similar in terms of transients, but may exhibit steady-state oscillations. In contrast, the proposed method not only guarantees steady-state stability but also enables a quick return of the voltage to the reference voltage with less overshoot.

Additionally, Figure 10 depicts the outcome curves of the proposed controller when confronted with large CPL variations. During the pre-training stage, the agent is trained with CPL varying from 200 W to

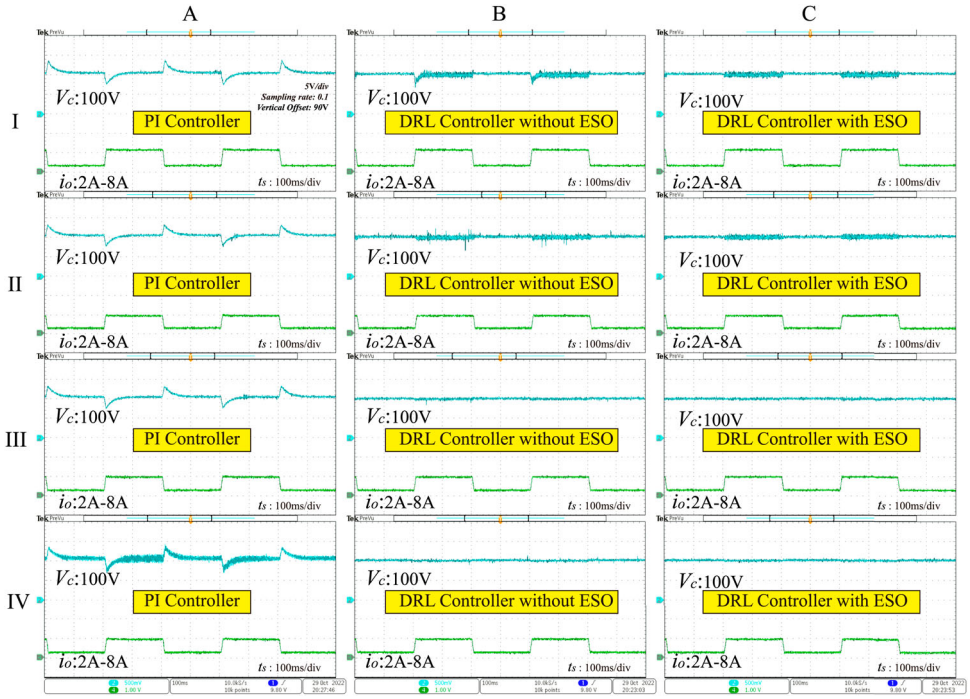


Figure 9. Experimental comparison results with PI controller.

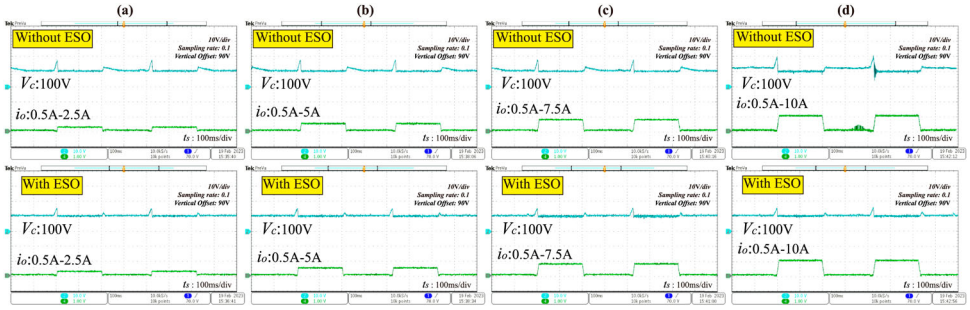


Figure 10. Experiments with large CPL variations: (a) 50 W–250 W–50 W; (b) 50 W–500 W–50 W; (c) 50 W–750 W–50 W; (d) 50 W–1000 W–50 W.

800 W. In the experimental setup, the performance of the proposed controller is evaluated following a CPL modification ranging from 5 to 20 times. Notably, the integration of the ESO effectively mitigates overshoot and reduces transient time during CPL variations. Specifically, as illustrated in Figure 10(d), the proposed method not only minimises fluctuations in steady state but also reduces overshoot by over 50% when the change multiplier is notable. This implies that as the magnitude of the perturbations increases, the proposed algorithm shows an enhanced ability to reject them, resulting in a more robust control system.

5. Conclusion

This work presents a new robust control algorithm that combines a DRL controller with an ESO. The proposed approach is applied to a DC–DC buck converter to regulate the bus voltage to a reference value while compensating for uncertainties and disturbances caused by CPL variations and internal/external sources. The proposed method demonstrates improved transient-time and steady-state performance compared to a traditional PI controller under varying circuit parameters and disturbances. Future research may explore the generalisability of reinforcement learning to a wider range of parameters and applications, and address stability concerns for broader industrial implementation.

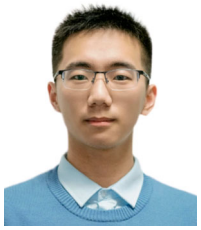
Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported in part by the National Natural Science Foundation of China under Grant 62233006, 62173221 and in part by Shanghai Rising-Star program under Grant 20QA1404000.

Notes on contributors



Tianxiao Yang received the B.E. degree in electrical engineering from Shenyang University of Technology, Shenyang, China, in 2020. He is currently working toward the Master's degree in automatic control from Shanghai University of Electric Power. His research interests include deep reinforcement learning theory and its applications for microgrid and power systems.

Chengang Cui received the B.E. degree in automation engineering from Jilin University, China, in 2004, the Ph.D. degree in control theory and control from Zhejiang University, China, in 2010. He worked at Shanghai Institute for Advanced Studies, Chinese Academy of Sciences, engaged in energy management and optimal scheduling from 2012 to 2015. He has been with the School of Automation, Shanghai University of Electric Power, where he is currently a associate professor. His research interests cover the control and schedule of renewable energy systems and microgrid.

Chuanlin Zhang received the B.S. degree in mathematics and the Ph.D. degree in control theory and control engineering from the School of Automation, Southeast University, Nanjing, China, in 2008 and 2014, respectively. He was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Texas at San Antonio, USA, from 2011 to 2012; a Visiting Scholar with the Energy Research Institute, Nanyang Technological University, Singapore, from 2016 to 2017; a visiting scholar with Advanced Robotics Center, National University of Singapore, from 2017 to 2018. Since 2014, he has been with the College of Automation Engineering, Shanghai University of Electric Power, Shanghai, where he is currently a Professor. His research interests include nonlinear system control theory and applications for power systems.

Jun Yang received the B.Sc. degree in automation from the Department of Automatic Control, Northeastern University, Shenyang, China, in 2006, and the Ph.D. degree in control theory and control engineering from the School of Automation, Southeast University, Nanjing, China, in 2011. He joined the Department of Aeronautical and Automotive Engineering at Loughborough University from 2020 as a senior lecturer. His research interests include disturbance estimation and compensation, and advanced control theory and its application to flight control systems and motion control systems. He is a fellow of IET.

ORCID

Tianxiao Yang http://orcid.org/0000-0003-4244-3208
Chengang Cui http://orcid.org/0000-0002-9463-384X
Chuanlin Zhang http://orcid.org/0000-0001-6052-1682
Jun Yang http://orcid.org/0000-0002-4290-9568

References

Bandyopadhyay, I., Purkait, P., & Koley, C. (2018). Performance of a classifier based on time-domain features for incipient fault detection in inverter drives. *IEEE Transactions on Industrial Informatics*, *15*(1), 3–14. https://doi.org/10.1109/TII.2018.2854885

Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., & Blaabjerg, F. (2020). Reinforcement learning and its applications in modern power and energy systems: A review. *Journal of Modern Power Systems and Clean Energy*, *8*(6), 1029–1042. https://doi.org/10.35833/MPCE.2020.000552

Coggan, M. (2004). Exploration and exploitation in reinforcement learning. Research supervised by Prof. Doina Precup, CRA-W DMP Project at McGill University.

Cui, C., Yan, N., Huangfu, B., Yang, T., & Zhang, C. (2021). Voltage regulation of DC-DC buck converters feeding CPLs via deep reinforcement learning. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *69*(3), 1777–1781. https://doi.org/10.1109/TCSII.2021.3107535

Cui, C., Yang, T., Dai, Y., Zhang, C., & Xu, Q. (2022). Implementation of transferring reinforcement learning for DC-DC buck converter control via duty ratio mapping. *IEEE Transactions on Industrial Electronics*, *70*(6), 1–10. https://doi.org/10.1109/TIE.2022.3192676

Davoudi, A., Jatskevich, J., & De Rybel, T. (2006). Numerical state-space average-value modeling of PWM DC-DC converters operating in DCM and CCM. *IEEE Transactions on Power Electronics*, *21*(4), 1003–1012. https://doi.org/10.1109/TPEL.2006.876848

El Mejdoubi, A., Chaoui, H., Sabor, J., & Gualous, H. (2017). Remaining useful life prognosis of supercapacitors under temperature and voltage aging conditions. *IEEE Transactions on Industrial Electronics*, *65*(5), 4357–4367. https://doi.org/10.1109/TIE.2017.2767550

Fan, J., Wang, Z., Xie, Y., & Yang, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for dynamics and control* (pp. 486–489).

Gheisarnejad, M., Farsizadeh, H., & Khooban, M. H. (2020). A novel nonlinear deep reinforcement learning controller for DC–DC power buck converters. *IEEE Transactions on Industrial Electronics*, *68*(8), 6849–6858. https://doi.org/10.1109/TIE.2020.3005071

Hajihosseini, M., Andalibi, M., Gheisarnejad, M., Farsizadeh, H., & Khooban, M. H. (2020). DC/DC power converter control-Based deep machine learning techniques: Real-Time implementation. *IEEE Transactions on Power Electronics*, *35*(10), 9971–9977. https://doi.org/10.1109/TPEL.63

Han, J. (2009). From PID to active disturbance rejection control. *IEEE Transactions on Industrial Electronics*, *56*(3), 900–906. https://doi.org/10.1109/TIE.2008.2011621

Huangfu, B., Cui, C., Zhang, C., & Xu, L. (2022). Learning-Based optimal large-Signal stabilization for DC/DC boost converters feeding CPLs via deep reinforcement learning. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 1–1. https://doi.org/10.1109/JESTPE.2022.3189078

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285. https://doi.org/10.1613/jair.301

Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, *33*, 1179–1191.

Kwasinski, A., & Onwuchekwa, C. N. (2010). Dynamic behavior and stabilization of DC microgrids with instantaneous constant-power loads. *IEEE Transactions on Power*

Electronics, *26*(3), 822–834. https://doi.org/10.1109/TPEL.2010.2091285

Li, X., Qi, G., Guo, X., Chen, Z., & Zhao, X. (2022). Improved high order differential feedback control of quadrotor UAV based on improved extended state observer. *Journal of the Franklin Institute*, *359*(9), 4233–4259. https://doi.org/10.1016/j.jfranklin.2022.03.019

Lin, D., Li, X., Ding, S., Wen, H., Du, Y., & Xiao, W. (2021). Self-tuning MPPT scheme based on reinforcement learning and beta parameter in photovoltaic power systems. *IEEE Transactions on Power Electronics*, *36*(12), 13826–13838. https://doi.org/10.1109/TPEL.2021.3089707

Peng, Q., Jiang, Q., Yang, Y., Liu, T., Wang, H., & Blaabjerg, F. (2019). On the stability of power electronics-dominated systems: Challenges and potential solutions. *IEEE Transactions on Industry Applications*, *55*(6), 7657–7670. https://doi.org/10.1109/TIA.28

Prag, K., Woolway, M., & Celik, T. (2021). Data-driven model predictive control of DC-to-DC buck-boost converter. *IEEE Access*, *9*, 101902–101915. https://doi.org/10.1109/ACCESS.2021.3098169

Qi, G., Li, X., & Chen, Z. (2021). Problems of extended state observer and proposal of compensation function observer for unknown model and application in UAV. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(5), 2899–2910. https://doi.org/10.1109/TSMC.2021.3054790

Ripley, B. D. (1993). Statistical aspects of neural networks. *Networks and Chaos–statistical and Probabilistic Aspects*, *50*, 40–123. https://doi.org/10.1007/978-1-4899-3099-6

Wang, H., Li, C., Li, J., He, X., & Huang, T. (2019). A survey on distributed optimisation approaches and applications in smart grids. *Journal of Control and Decision*, *6*(1), 41–60. https://doi.org/10.1080/23307706.2018.1549516

Xu, Q., Zhang, C., Wen, C., & Wang, P. (2017). A novel composite nonlinear controller for stabilization of constant power load in DC microgrid. *IEEE Transactions on Smart Grid*, *10*(1), 752–761. https://doi.org/10.1109/TSG.2017.2751755

Xu, W., Junejo, A. K., Liu, Y., & Islam, M. R. (2019). Improved continuous fast terminal sliding mode control with extended state observer for speed regulation of PMSM drive system. *IEEE Transactions on Vehicular Technology*, *68*(11), 10465–10476. https://doi.org/10.1109/TVT.25

Yang, T., Cui, C., & Zhang, C. (2022). On the Robustness Enhancement of DRL Controller for DC-DC Converters in Practical Applications. In *IEEE 17th International Conference on Control & Automation (ICCA)* (pp. 225–230).

Yu, Y. (2018). Towards Sample Efficient Reinforcement Learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* (pp. 5739–5743).

Zhang, C., Wang, X., Lin, P., Liu, P. X., Yan, Y., & Yang, J. (2019). Finite-time feedforward decoupling and precise decentralized control for DC microgrids towards large-signal stability. *IEEE Transactions on Smart Grid*, *11*(1), 391–402. https://doi.org/10.1109/TSG.5165411

Zhang, Y., Jin, J., & Huang, L. (2020). Model-free predictive current control of PMSM drives based on extended state observer using ultralocal model. *IEEE Transactions on Industrial Electronics*, *68*(2), 993–1003. https://doi.org/10.1109/TIE.41

Zhao, W., Queralta, J. P., & Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *IEEE Symposium Series on Computational Intelligence* (pp. 737–744).

Zheng, Q., Gaol, L. Q., & Gao, Z. (2007). On stability analysis of active disturbance rejection control for nonlinear time-varying plants with unknown dynamics. In *46th IEEE Conference on Decision and Control* (pp. 3501–3506).