Letters

Artificial Intelligence-Aided Minimum Reactive Power Control for the DAB Converter Based on Harmonic Analysis Method

Yuanhong Tang, *Student Member, IEEE*, Weihao Hu, *Senior Member, IEEE*, Di Cao, Nie Hou, *Student Member, IEEE*, Yunwei Li, *Fellow, IEEE*, Zhe Chen, *Fellow, IEEE*, and Frede Blaabjerg, *Fellow, IEEE*

***Abstract*—With the aim of reducing the reactive power for the dual-active-bridge (DAB) converter, this letter proposes an artificial intelligence (AI) aided minimum reactive power control scheme based on the harmonic analysis method. Specifically, as an advanced algorithm of the deep reinforcement learning (DRL), the deep deterministic policy gradient (DDPG) is used to train an agent off-line. During the training of DDPG algorithm, the three-phase-shift (TPS) modulation is adopted and the zero-voltage-switching (ZVS) constraints are considered. Thus, the trained agent of the DDPG which likes an implicit function, can provide optimal control strategies for the DAB converter in real-time with the minimum reactive power and soft switching performance in the continuous operation range. Finally, experimental results validate the feasibility and correctness of the proposed AI based optimized method.**

***Index Terms*—Artificial intelligence (AI), DAB converter, deep deterministic policy gradient (DDPG), harmonic analysis, reactive power.**

I. Introduction

DUE to the benefit of a simple circuit structure, high power efficiency, and high power density, the dual-active-bridge (DAB) converter has become one of the popular topologies of the isolated bidirectional converter family [1]. As a typical modulation strategy, the single-phase-shift (SPS) is widely employed in the DAB converter for its simple realization, while this simple method suffers from high reactive power and a narrow zero-voltage-switching (ZVS) range [2]. Aiming to address these issues, the triple-phase-shift (TPS) modulation is widely adopted to promote the control flexibility and the operation efficiency [3].

In general, a linear piecewise time-domain (LPTD) model is usually adopted to make the theoretical analysis with the TPS modulation. However, the piecewise expressions are required for different time intervals and operation modes, which is complicated and inconvenient for solving optimal control variables [4]. Aiming to overcome the disadvantages of the LPTD expression, the harmonic analysis method can be used to simplify the waveform expressions of the TPS modulation as the harmonic series forms in the frequency domain [5], [6].

Furthermore, many advanced schemes have been proposed to solve the optimized phase shift angles for the TPS modulation, such as the constrained numerical optimization method [7], iterative methodology [8], and heuristic algorithm [5]. More specifically, a genetic-algorithm (GA) [9], an ant-colony-optimization (ACO) [10], and a particle-swarm-optimization (PSO) [5] as familiar heuristic algorithms have been widely employed to solve the optimization control problems. However, these solutions suffer from obvious limitations such as time consuming, and poor accuracy.

Today, the artificial intelligence (AI) technique plays a large role in solving the optimization control problems [11]. In our previous work [12], an efficiency optimization scheme for the DAB converter is proposed by using the reinforcement learning (RL). However, this RL method may not be applicable if the training range is widened, due to the states and actions need to be discretized, which will make the volume of the generated lookup table very huge.

In order to overcome the weaknesses of the discretized RL proposed in [12] and the LPTD expression, this letter proposes a deep reinforcement learning (DRL) based reactive power optimization scheme in continuous domain. The novel aspects are summarized as follows: 1) as an advanced algorithm of the DRL, the deep-deterministic-policy-gradient (DDPG) is utilized to train an agent based on the harmonic analysis and TPS modulation method; 2) during the training of the DDPG, the ZVS constraints are considered to guarantee the soft switching performance. Based on this, the trained agent of the DDPG, which likes an implicit function, can provide optimal phase shift angles in real time for the continuous operation range

Manuscript received December 21, 2020; revised January 21, 2021; accepted February 13, 2021. Date of publication February 18, 2021; date of current version June 1, 2021. This work was supported by the Sichuan Science and Technology Program under Grants 2020JDJQ0037 and 2020YFG0312. *(Corresponding author: Weihao Hu.)*

Yuanhong Tang, Weihao Hu, and Di Cao are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: yhtang@std.uestc.edu.cn; whu@uestc.edu.cn; caodi@std.uestc.edu.cn).

Nie Hou and Yunwei Li are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada (e-mail: nhou@ualberta.ca; yunwei.li@ualberta.ca).

Zhe Chen and Frede Blaabjerg are with the Department of Energy Technology, Aalborg University, 9220 Aalborg, Denmark (e-mail: zch@et.aau.dk; fbl@et.aau.dk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPEL.2021.3059750.

Digital Object Identifier 10.1109/TPEL.2021.3059750

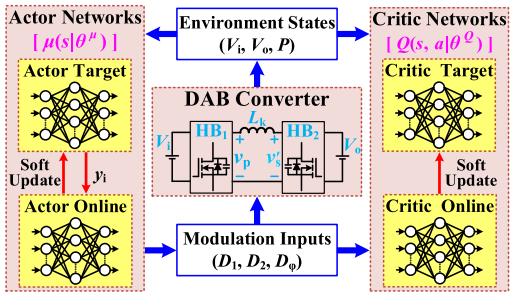

Fig. 1. Structure diagram of the proposed DDPG method for DAB converter.

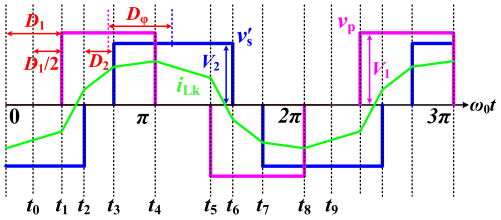

Fig. 2. Typical operating waveforms of the DAB converter based on the TPS.

with the desired minimum reactive power and soft switching performance.

II. Proposed Methodology

This letter is aiming to obtain the optimal control variables for reducing the reactive power based on the DRL method. The principle of the proposed DRL method for the DAB converter is shown in Fig. 1. In this section, the harmonics analysis method based on TPS modulation is given first. Then, the DDPG algorithm is introduced to solve the optimization control problem with minimum reactive power.

A. Harmonic Analysis of the TPS Modulation Method

The equivalent circuit of the DAB converter is depicted in Fig. 1, and the corresponding operation waveforms of the TPS modulation is illustrated in Fig. 2. Specifically, $v_{\rm p}$ denotes the ac voltage of the primary side, $v'_{\rm s}$ is the equivalent ac voltage of the secondary side, $L_{\rm k}$ is the equivalent inductance of the auxiliary inductance and leakage of the transformer, and the duty ratio of the transformer is N: 1. The amplitude value of $v_{\rm p}$ and $v'_{\rm s}$ are V_1 and V_2, where $V_1 = V_{\rm i}$ and $V_2 = N{\cdot}V_{\rm o}$. D_1 is the inner phase shift of the full bridge $\rm HB_1$, D_2 denotes the inner phase shift of the full bridge $\rm HB_2$, and D_φ is the phase shift between the center point of $v_{\rm p}$ and $v'_{\rm s}$. Moreover, $i_{\rm Lk}$ is the current flowing through $L_{\rm k}$. In this letter, we assume the power flows from primary to secondary, and $V_1 \geq V_2$, thus the range of all phase shift angles (D_1, D_2, D_φ) are limited in $[0\sim\pi]$.

Based on the harmonic analysis method, the periodic function expression of the TPS can be made up with the combination of series odd-order harmonics components, which expressed by simple sine and cosine functions [5], [6]. Based on this, the active power $P_{\rm o}$ can be expressed as

$$P_o = \sum_{n=1,3,5,..} \frac{8V_1V_2}{n^3\pi^2\omega_0 L_k}\cos\left(n\frac{D_1}{2}\right)\cos\left(n\frac{D_2}{2}\right)\sin(nD_\varphi) \tag{1}$$

where $\omega_0 = 2\pi f_{\rm s}$ and $f_{\rm s}$ denotes the switching frequency. The reactive power Q can be calculated as follows:

$$Q_{n=1,3,5,..} = \sum_{n=1,3,5,..} \frac{8V_1\sqrt{A^2+B^2}}{n^3\pi^2\omega_0 L_k}\cos\left(n\frac{D_1}{2}\right) \times \sin\left(-\arctan\frac{A}{B}\right) \tag{2}$$

$$Q_{m\neq n=1,3,5,..} = \sum_{m\neq n=1,3,5,..} \frac{8V_1\cos(m\frac{D_1}{2})}{mn^2\pi^2\omega_0 L_k}\sqrt{A^2+B^2} \tag{3}$$

where A and B can be expressed as

$$\begin{cases} A = V_2\cos\left(n\frac{D_2}{2}\right)\cos(nD_\varphi) - V_1\cos\left(n\frac{D_1}{2}\right) \\ B = V_2\cos\left(n\frac{D_2}{2}\right)\sin(nD_\varphi). \end{cases} \tag{4}$$

Thus, the reactive power is given by $Q = Q_{n\ =\ 1,3,5,..} + Q_{m\neq n\ =\ 1,3,5,...}$ Moreover, the current flowing through $L_{\rm k}$ is described as

$$i_{Lk}(t) = \sum_{n=1,3,5,..} \frac{4}{n^2\pi\omega_0 L_k}\sqrt{A^2+B^2} \sin\left(n\omega_0 t + \arctan\frac{A}{B}\right). \tag{5}$$

For ensuring the ZVS performance for all power switches, the current direction should contrary to the terminal voltage at the switching-ON instants. Thus, the ZVS boundary conditions for the switches of each bridge legs can be expressed as follows [11]:

$$\begin{cases} i_{Lk}\left(\omega_0 t = \frac{D_1}{2}\right) \leq 0 \;\; i_{Lk}\left(\omega_0 t = \pi - \frac{D_1}{2}\right) \geq 0 \\ i_{Lk}\left(\omega_0 t = D_\varphi + \frac{D_2}{2}\right) \geq 0 \;\; i_{Lk}\left(\omega_0 t = D_\varphi - \frac{D_2}{2}\right) \geq 0. \end{cases} \tag{6}$$

In this letter, the harmonic components are chosen as m, n = 1, 3, 5, 7 to simplify the calculations as the higher order shows small influence on the calculation accuracy [5].

B. DRL-Based Optimization

As an advanced algorithm of the DRL method, DDPG is suitable for solving the complex and multidimensional optimization problems in the continuous action space [13], [14]. The DDPG algorithm is used in this letter as a policy network to find the control variables for the DAB converter, which corresponds to minimum reactive power.

Based on the RL paradigm, an agent interacts with the environment sequentially to learn the optimal policy. Thus, the optimization control problem of the minimum reactive power can be modeled as a Markov decision process (MDP) under stochastic environments [15]. Moreover, the MDP is usually composed of a four-tuple (S, A, P, R), where S is the state space, A denotes the action space, P is the state transition probability function, and R is the reward function. At each step k, an RL agent

perceives a state $s_k \in S$, and selects an action $a_k \in S$ according to the policy $\pi(a_k|s_k)$, where $\pi(a_k|s_k)$ indicates the policy function, which maps states s to actions a. After that, the RL agent will receive a corresponding reward $r_k \sim R(s_k, a_k)$ and devise next state s_{k+1}. The cumulative discounted reward G_t is used, and it is defined as

$$G_t = \sum_{k=0}^{\infty} \gamma^k \cdot R_{k+t} \tag{7}$$

where γ denotes the discount factor and is limited to [0∼1].

Based on this, the aim of maximizing the value of the reward function R is further adjusted to maximize the discounted reward G_t. Moreover, the action-value function $Q^\pi(s_t, a_t)$ is defined in DDPG algorithm as follows:

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi \left[G(s_t, a_t) + \gamma \mathbb{E}_{a_{t+1} \sim \pi} [Q^\pi(s_{t+1}, a_{t+1})] \right]. \tag{8}$$

In the DAB converter, the environment features consist of the input voltage V_i, the output voltage V_o, and the objective active power P_o'. The active power and operation performance are dependent on the phase-shift angles (D_1, D_2, D_φ). Thus, in the proposed DDPG algorithm, the state space S is defined as $S = [V_i, V_o, P_o]$, the action space A is defined as $A = [D_1, D_2, D_\varphi]$.

As the policy agent is improved with trial and error after the explorations, it is important to design an appropriate reward function. The aim of this letter is to acquire the optimal actions with respect to the minimum reactive power and the minimum active power error, the reward function R is set as

$$R(D_1, D_2, D_\varphi) = -[\delta \cdot |Q| + \xi \cdot (P_o - P_o')^2] \tag{9}$$

where P_o' is the expected output active power and ξ is penalty factor of the active power. Remarkably, for the training of the DDPG algorithm, a large value of ξ will cause poor reactive power performance, while a small value of ξ will cause huge active power error. Through the trial and error during the simulation, ξ is chosen as 200 in this letter. δ is set to 1 when the ZVS constraints is satisfied as mentioned in inequality (6). However, a large value of δ should be given once the ZVS is lost which indicates a small reward value should be given, in order to avoid such poor action being learned by the agent of the DDPG algorithm. Here, δ is chosen as 10 when inequality (6) is not satisfied to ensure a ZVS performance during the training process. Thus, the ZVS can be guaranteed, and the minimum active power error and the minimum reactive power can be obtained by maximizing the reward function R.

As illustrated in Fig. 1, the DDPG algorithm uses an actor–critic based framework, which contains two eponymous ingredients (actor-network and critic-network), where each ingredient is made up with two networks (main network and target network). Moreover, the actor network is used to adjust the weight θ^μ in the policy function $\mu(s|\theta^\mu)$ by fitting a state (V_i, V_o, P_o') to the corresponding action (D_1, D_2, D_φ), while the critic network is used to adjust the weight θ^Q in the action-value function $Q(s, a|\theta^Q)$ [13].

TABLE I
TRAINING PROCESS OF THE DDPG ALGORITHM

Algorithm: Offline Learning with the DDPG algorithm
Input: Environments [V_i, V_o, P_o]
Output: Phase shift angles (D_1, D_2, D_φ)

1: Randomly initialize the weights of actor network θ^μ and critic network θ^Q
2: Initialize target network μ' and Q' by: $\theta^{\mu'} \leftarrow \theta^\mu$, $\theta^{Q'} \leftarrow \theta^Q$
3: **for** episode=1 to max-episode **do**
4: Observe current state s_1
5: **for** actor =1, 2, …10 **do**
6: Select action $a_t = \mu(s_t|\theta^\mu) + n_t$ according to the deterministic policy μ and exploration rate
7: Execute action a_t, observe the reward r_t and observe new state s_{t+1}
8: Store transition (s_t, a_t, r_t, s_{t+1}) in replay buffer R
9: Sample a random mini-batch of transitions (s_i, a_i, r_i, s_{i+1}) from replay buffer R
10: $y_t = r(s_t, a_t) + \gamma Q\left(s_{t+1}, \mu(s_{t+1})\middle|\theta^Q\right)$
11: Update critic net by using the loss function $\mathcal{L}(\theta^Q)$:
$$\mathcal{L}(\theta^Q) = E_{(s,a)}[(Q(s_t, a_t|\theta^Q) - y_t)^2]$$
12: Update actor net by using the policy gradient:
$$\nabla_{\theta^\mu} J^{\theta^\mu} \approx \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{a=\mu(s|\theta^\mu)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)]$$
$$= \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_a Q(s, a|\theta^Q)|_{a=\mu^\beta(s)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)]$$
13: Update target net by using the policy gradient:
$$\nabla_{\theta^\mu} J^{\theta^\mu} \approx \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{a=\mu(s|\theta^\mu)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)]$$
$$= \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_a Q(s, a|\theta^Q)|_{a=\mu^\beta(s)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)]$$
15: end for
16: end for

The weight θ^Q of the critic network is updated by minimizing the loss function $\mathcal{L}(\theta^Q)$, which is given by

$$\mathcal{L}(\theta^Q) = E_{(s,a)}[(Q(s_t, a_t|\theta^Q) - y_t)^2] \tag{10}$$

where $y_t = r_t(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_t|\theta^\mu)|\theta^Q)$. The parameter θ^μ of the actor network is updated by using the policy gradient below

$$\nabla_{\theta^\mu} J^{\theta^\mu} \approx \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{a=\mu(s|\theta^\mu)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)]$$
$$= \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_a Q(s, a|\theta^Q)|_{a=\mu^\theta(s)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)] \tag{11}$$

where ρ represents the discounted distribution and β denotes the specific policy of the current policy π.

Aiming to enhance the stability and reliability during the learning process of the DDPG algorithm, two distinct target networks, which are named as the actor target network $\mu'(s|\theta^{\mu'})$ and the critic target network $Q'(s, a|\theta^{Q'})$ are added to the actor network and critic network, respectively, as depicted in Fig. 1. In each iteration, these soft weights ($\theta^{\mu'}$ and $\theta^{Q'}$) are updated according to the following formulas:

$$Softupdate \begin{cases} \theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \\ \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \end{cases} \tag{12}$$

where $\tau << 1$ and τ represents the soft update coefficient.

Moreover, Table I shows the detailed training process of the DDPG algorithm. For the DDPG algorithm, the environments

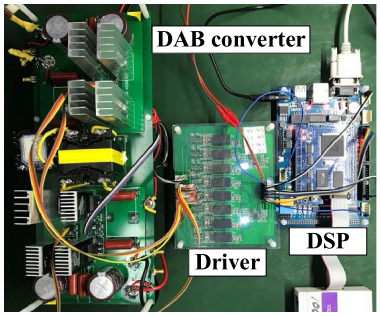

Fig. 3. Experimental prototype for 200 W nominal power.

TABLE II
SYSTEM SPECIFICATIONS OF THE DAB CONVERTER

Parameter	value
Input voltage V_i (V)	100~140
Output voltage V_o (V)	40
Switching frequency f_s (kHz)	50
Inductor L_k (μH)	41
Turns ratio of the transformer (N:1)	1:1

$[V_i, V_o, P_o]$ is used for the training process. During the training of the DDPG algorithm, the training data (Sate S) are sampled randomly in the range of the environments mentioned above. Besides, the strategy of continuous action space in the DDPG algorithm is learned by a deep neural network. Based on this, after the training of the DDPG algorithm is completed, the trained agent will be downloaded to a digital signal processor (DSP). In practice, once the environments $[V_i, V_o, P_o]$ is detected, such agent which likes an implicit function will map these input parameters into the corresponding phase shift angles (D_1, D_2, D_φ) instantly. Therefore, such trained agent can dramatically reduce the memory allocation and the computational time in comparison to the RL method [12], because the lookup table is not needed in this letter. Moreover, as the strategy of continuous action space in the DDPG algorithm is learned by a deep neural network, the trained agent can provide optimal control strategy under whole continuous operation conditions. Based on this, the proposed DRL method is more suitable for the online real-time control. It is worth noting that the trained agent can still provide corresponding phase shift angles (D_1, D_2, D_φ), even if the environments $[V_i, V_o, P_o]$ is out of the previous training range,while such circumstance is not a very likely scenario in practical application.

III. EXPERIMENTAL ANALYSIS

Aiming to verify the feasibility and correctness of the proposed AI-based optimized method (AIO), a 200-W experimental prototype is built, where Fig. 3 shows a photograph of the experimental prototype. The system specifications are summarized in Table II. The specific structure of the DAB converter is illustrated in Fig. 1.

The main hyperparameters of the DDPG algorithm are the number of neural layers, the cell of each hidden layer, and the learning rate. Aiming to find the appropriate hyperparameters for the DDPG algorithm, a *Grid Search* method is adopted in this letter [16]. Where the number of neural layers is chosen from 0 to 5 and the step size defaults to 1, the number of hidden layer is chosen from 10 to 150 and the step size defaults to 10, and the learning rate in the range of $0.1\sim10^{-8}$ and reduces ten times at each tuning time. Similarly, other hyperparameters are also set to a general range. Based on this, two hidden layers with 100 and 100 neurons, respectively, are designed for both the critic and actor networks. And, other detailed hyperparameters of the DDPG algorithm are summarized in Table III.

TABLE III
TRAINING PARAMETERS ADOPTED IN DDPG ALGORITHM

Hyper Parameter	value
Learning rate of critic network (λ_c)	0.002
Learning rate for actor network (λ_a)	0.001
Soft update coefficient (τ)	0.001
Replay buffer size (M)	40000
Max Episode (N)	10000
Step size of each episode	10
Mini-batch size (m)	32

Moreover, the activation functions used in the critic networks are chosen as the rectified linear for both the hidden layers and the output layer [17], [18]. The tanh activation unit and the softplus activation unit are chosen as the activation functions in the actor network for the output layer.

According to the training parameters adopted in the DDGPG algorithm as shown in Table III, the Max Episode N is chosen as 10 000, and the step size of each episode is chosen as 10. Based on the hyperparameters shown in Table III, the whole training process can be accomplished in approximately 30 min, where the training of the ANN is running on the *Windows 10* operating system powered by an Intel(R) Core(TM) i7-8700 CPU @3.20GHz 3.19 GHz.

Based on this, detailed experimental results and corresponding analysis are described as follows.

Fig. 4 illustrates the measured experimental waveforms for the forward power flow under different operation conditions when $V_o = 40$ V, where purple circles indicate the switching performance of the leading leg in the primary side, purple dotted circles indicate the switching performance of the lagging leg in the primary side, blue circles indicate the switching performance of the leading leg in the secondary side, and blue dotted circles indicate the switching performance of the lagging leg in the secondary side. According to Fig. 4(a) and (c), all of the power switches can obtain the ZVS performance. Fig. 4(b) and (d) suggests that the leading leg of the primary side can obtain the ZVS turn-ON, while the other three legs can obtain the ZCS turn-OFF. Fig. 5 shows the measured experimental waveforms for the backward power flow under different operation conditions when $V_i = 100$ V and $V_o = 40$ V. Remarkably, the soft switching conditions in Fig. 5 is analogous to Fig. 4. The experimental results shown in Figs. 4 and 5 indicate that the soft switching performance can be guaranteed under different operation conditions.

Fig. 6 illustrates the dynamic performance for the power transition and the input voltage variation when $V_o = 40$ V. It can be seen from Fig. 6(a), the output voltage v_{out} and the current

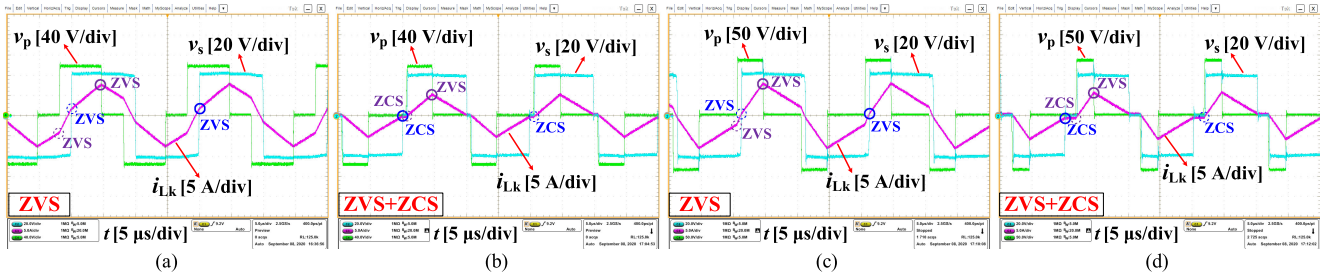

Fig. 4. Experimental waveforms for the forward power flow of the primary side voltage v_{p}, secondary side voltage v_{s}, and current i_{Lk} under different operations with $V_{\mathrm{o}} = 40$ V. (a) $V_{\mathrm{i}} = 100$ V, $P_{\mathrm{o}} = 180$ W. (b) $V_{\mathrm{i}} = 100$ V, $P_{\mathrm{o}} = 100$ W. (c) $V_{\mathrm{i}} = 140$ V, $P_{\mathrm{o}} = 180$ W. (d) $V_{\mathrm{i}} = 140$ V, $P_{\mathrm{o}} = 100$ W.

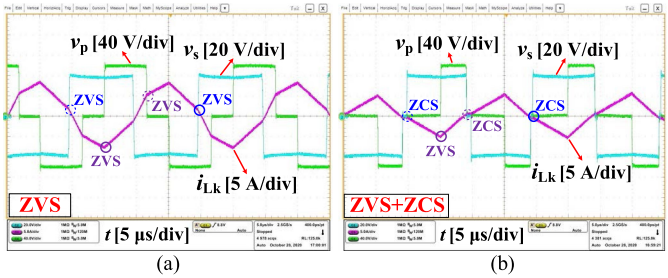

Fig. 5. Experimental waveforms for the backward power flow of the primary side voltage v_{p}, secondary side voltage v_{s}, and current i_{Lk} under different operations with $V_{\mathrm{i}} = 100$ V and $V_{\mathrm{o}} = 40$ V. (a) $P_{\mathrm{o}} = 180$ W. (b) $P_{\mathrm{o}} = 100$ W.

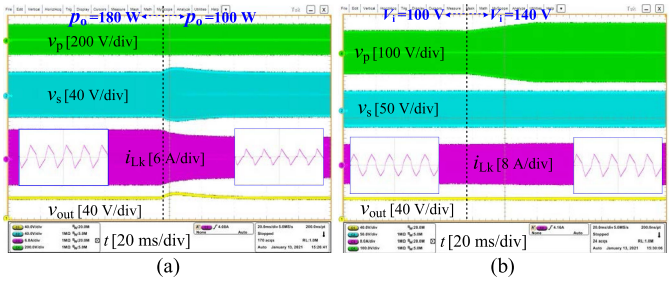

Fig. 6. Dynamic performance for the power transition and the input voltage variation. (a) Curves of the power transition with $V_{\mathrm{i}} = 140$ V and $V_{\mathrm{o}} = 40$ V. (b) Curves of the input voltage variation with $V_{\mathrm{o}} = 40$ V and load $= 8.889\ \Omega$.

TABLE IV
REACTIVE POWER COMPARISON BETWEEN DIFFERENT OPTIMIZATION METHODS

Operating conditions	Algorithms	Reactive power (Var) P_o=100 W	Reactive power (Var) P_o=100 W
V_i=100 V V_o=40 V	GA	115.1	325.1
	ACO	110.3	330.3
	PSO	103.2	320.2
	MPS	98.1	312.4
	QTPS	95.2	310.1
	AIO	101.1	302.1
V_i=140 V V_o=40 V	GA	115.3	330.1
	ACO	120.5	320.5
	PSO	105.4	299.3
	MPS	100.1	293.2
	QTPS	98.4	295.4
	AIO	96.0	282.8

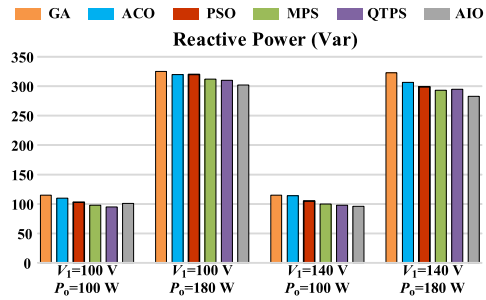

Fig. 7. Reactive power comparison between different optimization methods with $V_{\mathrm{o}} = 40$ V.

i_{Lk} can quickly return to steady state after P_{o} is changed from 180 to 100 W when $V_{\mathrm{i}} = 140$ V and $V_{\mathrm{o}} = 40$ V. Fig. 6(b) shows also a good dynamic performance when the V_{i} is changed from 100 to 140 V when $V_{\mathrm{o}} = 40$ V and load $= 8.889\ \Omega$.

Table IV shows the quantitative reactive power comparison between the GA [9], ACO [10], PSO [5], minimum current stress phase shift control (MPS) [8], Q-learning optimized triple-phase-shift control (QTPS) [12], and the proposed AIO scheme under different operation conditions when $V_{\mathrm{o}} = 40$ V. Corresponding histogram of the reactive power comparison between different algorithms is illustrated in Fig. 7. According to Table IV and Fig. 7, such six optimization methods have similar reactive power, while the proposed AIO scheme shows lower reactive power under most circumstances.

According to the system specifications of the DAB converter mentioned in Table II, the rated power is 200 W, and the input voltage V_{i} is varied from 100 to 140 V. In this letter's application scenarios, supposing that the default intervals of the V_{i} and P_{o} are set to 0.5 V and 0.5 W respectively, in the heuristic algorithm (GA, ACO, PSO) utilization. Therefore, at least 32 000 optimization processes must be required in these heuristic algorithms, which will cause heavy calculation burden. Compared to these heuristic algorithms, the QTPS method can complete the training process without such complicated optimization processes, while at least 32 000 training results have to be saved in the lookup table, as similar to the heuristic algorithm. Thus, the memory size of such a lookup table is at least 4 MB for the heuristic algorithm (GA, ACO, PSO) and the QTPS method, and such a large memory will also slow the query speed down. Remarkably, the generated lookup table is discrete, which cannot provide continuous control strategy for the DAB converter. In practice,

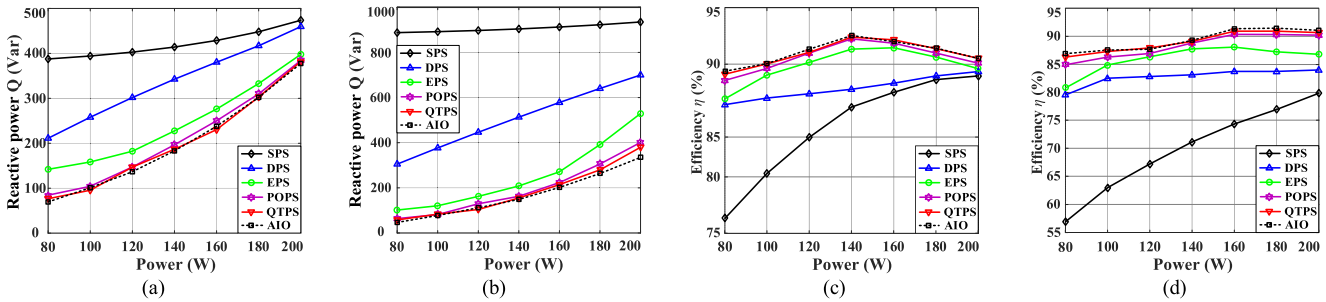

Fig. 8. Curve of the reactive power and corresponding efficiency with respect to the active power with $V_o = 40$ V. (a) Reactive power when $V_i = 100$ V. (b) Reactive power when $V_i = 140$ V. (c) Efficiency when $V_i = 100$ V. (d) Efficiency when $V_i = 140$ V.

if the detected environments $[V_i, V_o, P_o]$ is not found in such a lookup table directly, the closest environments $[V_i, V_o, P_o]$ and corresponding phase shift angles (D_1, D_2, D_φ) will be selected, which will cause certain power error and degrade the operation performance for the DAB converter.

However, in the proposed AIO scheme, the whole training process is completed at once, and the trained agent which likes an implicit function can map the environments $[V_i, V_o, P_o]$ under whole continuous operation range into the corresponding phase shift angles (D_1, D_2, D_φ) instantly. Moreover, the memory size of the trained agent is less than 200 KB. Based on this, the proposed AIO scheme is more suitable for the online real-time control compare to the heuristic algorithm (GA, ACO, PSO) and the QTPS method.

Furthermore, the LPTD model is used in the MPS and the QTPS method. According to [8], five different switching modes are adopted in the MPS method, and each switching mode should be participated in calculation for specific operating conditions to find the global optimum. Therefore, the multiple partial differential calculations will be required for different time intervals and operation modes, which is complexity and inconvenient. According to [12], eight different operation modes are considered during the training of the Q-learning algorithm in this QTPS method, while a unified harmonic analysis model is used during the training of the DDPG algorithm in the proposed AIO scheme. Thus, the proposed AIO scheme can lessen the calculative complexity and calculative burden during the training process compared to the QTPS method.

Fig. 8 illustrates the curve of the reactive power and efficiency with respect to the active power between SPS, dual-phase-shift (DPS) [1], extended-phase-shift (EPS) [11], PSO-optimized phase-shift control (POPS) [5], QTPS [12], and the proposed AIO scheme. According to Fig. 8, the reactive power and the efficiency in the proposed AIO scheme and the QTPS method are obviously better than the other four modulation methods. Besides, the curves of the proposed AIO scheme are similar to the QTPS method, while the reactive power and the efficiency in the proposed AIO scheme is better than the QTPS method for most operating conditions. More specifically, the maximum efficiency in the proposed AIO scheme is up to 92.6% when $P_o = 100$ W according to Fig. 8(c), and up to 90.7% when $P_o = 200$ W according to Fig. 8(d).

VII. Conclusion

This letter proposes an AI-aided minimum reactive power control for the DAB converter based on harmonic analysis method and TPS modulation. By using the state-of-the-art algorithm (DDPG) of the DRL method to train an agent offline, an adaptive control strategy can be obtained to reduce the reactive power. Furthermore, by adding the ZVS constraints to the reward function, the soft switching can be acquired. Compared with the heuristic algorithm and RL, the proposed AIO scheme is more suitable for the online real-time control, without a large lookup table and complicated optimization processes. Moreover, as the unified harmonic analysis model is adopted in the proposed AIO scheme, the calculative complexity and calculative burden for training of the DDPG algorithm is reduced. Experimental results show that the proposed AIO scheme can promote the operation performance. Specifically, the proposed AIO scheme reaches to the maximum efficiency in about 92.7% when $V_i = 100$ V and $V_o = 40$ V, $P_o = 140$ W. Remarkably, the proposed AIO scheme can also be easily used to solve the optimization control problem for other application scenarios and circuit topologies. In the future, a DRL method will be used to improve the operation performance and dynamic response when internal fault or external fault occurs.

References

[1] N. Hou and Y. Li, "Overview and comparison of modulation and control strategies for non-resonant single-phase dual-active-bridge dc-dc converter," *IEEE Trans. Power Electron.*, vol. 35, no. 3, pp. 3148–3172, Mar. 2020.

[2] M. N. Kheraluwala, R. W. Gascoigne, D. M. Divan, and E. D. Baumann, "Performance characterization of a high-power dual active bridge DC-to-DC converter," *IEEE Trans. Ind. Appl.*, vol. 28, no. 6, pp. 1294–1301, Nov./Dec. 1992.

[3] A. K. Bhattacharjee and I. Batarseh, "Optimum hybrid modulation for improvement of efficiency over wide operating range for triple-phase-shift dual-active-bridge converter," *IEEE Trans. Power Electron.*, vol. 35, no. 5, pp. 4804–4818, May 2020.

[4] A. Tong, L. Hang, G. Li, X. Jiang, and S. Gao, "Modeling and analysis of a dual-active-bridge-isolated bidirectional DC/DC converter to minimize RMS current with whole operating range," *IEEE Trans. Power Electron.*, vol. 33, no. 6, pp. 5302–5316, Jun. 2018.

[5] H. Shi, H. Wen, Y. Hu, and L. Jiang, "Reactive power minimization in bidirectional DC–DC converters using a unified-phasor-based particle swarm optimization," *IEEE Trans. Power Electron.*, vol. 33, no. 12, pp. 10990–11006, Dec. 2018.

[6] B. Zhao, Q. Song, W. Liu, G. Liu, and Y. Zhao, "Universal high-frequency-link characterization and practical fundamental-optimal strategy for dual-active-bridge DC-DC converter under PWM plus phase-shift control," *IEEE Trans. Power Electron.*, vol. 30, no. 12, pp. 6488–6494, Dec. 2015.

[7] J. Everts, F. Krismer, J. Van den Keybus, J. Driesen, and J. W. Kolar, "Optimal ZVS modulation of single-phase single-stage bidirectional DAB AC–DC converters," *IEEE Trans. Power Electron.*, vol. 29, no. 8, pp. 3954–3970, Aug. 2014.

[8] J. Huang, Y. Wang, Z. Li, and W. Lei, "Unified triple-phase-shift control to minimize current stress and achieve full soft-switching of isolated bidirectional DC–DC converter," *IEEE Trans. Ind. Electron.*, vol. 63, no. 7, pp. 4169–4179, Jul. 2016.

[9] L. Meng, T. Dragicevic, J. C. Vasquez, and J. M. Guerrero, "Tertiary and secondary control levels for efficiency optimization and system damping in droop controlled DC–DC converters," *IEEE Trans. Smart Grid*, vol. 6, no. 6, pp. 2615–2626, Nov. 2015.

[10] Z. Yin, C. Du, J. Liu, X. Sun, and Y. Zhong, "Research on autodisturbance-rejection control of induction motors based on an ant colony optimization algorithm," *IEEE Trans. Ind. Electron.*, vol. 65, no. 4, pp. 3077–3094, Apr. 2018.

[11] S. Zhao, F. Blaabjerg, and H. Wang, "An overview of artificial intelligence applications for power electronics," *IEEE Trans. Power Electron.*, vol. 36, no. 4, pp. 4633–4658, Apr. 2021.

[12] Y. Tang *et al.*, "Reinforcement learning based efficiency optimization scheme for the DAB DC-DC converter with triple-phase-shift modulation," *IEEE Trans. Ind. Electron.*, to be published, doi: 10.1109/TIE.2020.3007113.

[13] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[14] D. Cao *et al.*, "Reinforcement learning and its applications in modern power and energy systems: A review," *J. Modern Power Syst. Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020.

[15] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 4120–4123, Sep. 2020.

[16] H. A. Fayed and A. F. Atiya, "Speed up grid-search for parameter selection of support vector machines," *Appl. Soft Comput.*, vol. 80, pp. 202–210, Jul. 2019.

[17] G. Zhang *et al.*, "Deep reinforcement learning based approach for proportional resonance power system stabilizer to prevent ultra-low-frequency oscillations," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5260–5272, Nov. 2020.

[18] M. Gheisarnejad, H. Farsizadeh, and M. H. Khooban, "A novel non-linear deep reinforcement learning controller for DC/DC power buck converters," *IEEE Trans. Ind. Electron.*, to be published, doi: 10.1109/TIE.2020.3005071.