Adaptive Power Control Based on Double-layer *Q*-learning Algorithm for Multi-parallel Power Conversion Systems in Energy Storage Station

Yile Wu, Le Ge, Xiaodong Yuan, Xiangyun Fu, and Mingshen Wang

***Abstract*—An energy storage station (ESS) usually includes multiple battery systems under parallel operation. In each battery system, a power conversion system (PCS) is used to connect the power system with the battery pack. When allocating the ESS power to multi-parallel PCSs in situations with fluctuating operation, the existing power control methods for parallel PCSs have difficulty in achieving the optimal efficiency during a long-term time period. In addition, existing *Q*-learning algorithms for adaptive power allocation suffer from the curse of dimensionality. To overcome these challenges, an adaptive power control method based on the double-layer *Q*-learning algorithm for *n* parallel PCSs of the ESS is proposed in this paper. First, a selection method for the power allocation coefficient is developed to avoid repeated actions. Then, the outer action space is divided into $n+1$ power allocation modes according to the power allocation characteristics of the optimal operation efficiency. The inner layer uses an actor neural network to determine the optimal action strategy of power allocations in the non-steady state. Compared with existing power control methods, the proposed method achieves better performance for both static and dynamic operation efficiency optimization. The proposed method optimizes the overall operation efficiency of PCSs effectively under the fluctuating power outputs of the ESS.**

***Index Terms*—Double-layer *Q*-learning, adaptive power control, energy storage station (ESS), operation efficiency, power conversion system (PCS).**

I. Introduction

In recent years, the energy storage station (ESS) has attracted considerable attention in the generation, transmission, distribution, and power consumption parts of the power system. The ESS provides various services for the power system, such as peak load shaving and valley load filling, alleviating the operation pressure, smoothing the power fluctuation of renewable energy, and improving the reliability of the power supply [1]. With the development of energy-storage technology, the deployment of energy storage in power systems is growing rapidly [2], [3]. For ESSs with a large installed capacity, a parallel structure with multiple battery systems is adopted to improve the operation reliability [4].

In a battery system, the power conversion system (PCS) connects the power system to the battery pack (BP) and realizes bidirectional power exchange. The operation efficiency of a PCS is related to the exchanged power. For an ESS with multi-parallel PCSs, the overall operation efficiency of all the PCSs is determined by the power allocated to each PCS, which is controlled by the energy management system (EMS). In general, the ESS power is allocated via the traditional power sharing and hierarchical switching methods in practical engineering [5]-[7]. For improving the overall operation efficiency of parallel PCSs, these two methods are not applicable. Thus, it is necessary to develop an effective power control method for improving the overall operation efficiency of multi-parallel PCSs in ESSs.

Several methods have been proposed for improving the overall operation efficiency by managing the operation configuration of parallel converters. Reference [8] proposed a method to improve the selection of an arrangement of converters based on efficiency considerations. A tertiary control method was proposed for improving the overall operation efficiency by adjusting the power allocation proportion for parallel DC-DC converters in [9]. Reference [10] proposed a steady-state operation point control method for parallel converter systems, where the power allocation between *n* parallel converters was determined via the forward and backward substitution methods. Reference [11] optimized the efficiency of an ESS with two parallel dual active bridge converters by switching on/off modules. However, when the ESS responds to the stochastic renewable power or power system frequency deviations, the ESS power fluctuates significantly, and the average power allocated to each PCS during a whole day is far lower than 50% of the rated power [12]. The operation efficiency of a PCS under low exchanged power is significantly lower than that under the rated power. In fluctuating operation situations, the number of operating battery systems and the battery charging/discharging status vary dynamically with time. The methods proposed in [8]-[11] are not

Manuscript received: January 9, 2021; revised: June 24, 2021; accepted: November 3, 2021. Date of CrossCheck: November 3, 2021. Date of online publication: April 11, 2022.

This work was supported by the National Natural Science Foundation of China (No. 51707089), the Science and Technology Project of State Grid Corporation of China (No. 5210D0180006), and the Postgraduate Innovation Project of Jiangsu (No. SJCX20_0723).

Y. Wu and L. Ge (corresponding author) are with the School of Electric Power Engineering, Nanjing Institute of Technology, Nanjing, China (e-mail: y00450190414@njit.edu.cn; gele@njit.edu.cn).

X. Yuan, X. Fu, and M. Wang are with State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China (e-mail: 1838658@qq. com; 645014055@qq. com; wangmingshen@tju.edu.cn).

DOI: 10.35833/MPCE.2020.000909

suitable for optimizing the operation efficiency of parallel PCSs in fluctuating operation situations. Thus, the adaptive control method is needed to solve the dynamic operation efficiency optimization problem for parallel PCSs in fluctuating operation situations.

Adaptive power allocation for wind power converters was realized offline via exhaustive calculations and online using a lookup table for two parallel converters [7]. Intelligent algorithms, e.g., particle swarm optimization (PSO) and genetic algorithms (GAs), were used to achieve optimal adaptive control for different types of converters [13]-[15]. However, the complexity of the nonlinear equation for calculating the overall operation efficiency increases with the number of parallel PCSs. For the dynamic nonlinear problem, the adaptive method proposed in [7] led to a long calculation time. The intelligent algorithms proposed in [13]-[15] had low convergence speeds and may find local optimal solutions. In [16], the performance of several state-of-the-art intelligent algorithms for solving nonlinear equations was reviewed. However, with the increasing complexity of the nonlinear equations, the performance of intelligent algorithms will inevitably deteriorate owing to the equation roots.

Considering the aforementioned issues, the objective of this study was to use the *Q*-learning (QL) algorithm to adaptively obtain the optimal power allocation strategies for multi-parallel PCSs. The power allocation problem of the ESS under significant power fluctuations is an optimal decision problem of a dynamic nonlinear system. The Markov decision process (MDP) can be used to describe the dynamic decision-making problem, and the reinforcement learning (RL) employs the MDP as the framework to express the interaction with the environment. QL algorithm is a type of RL and can obtain the optimal solution for the dynamic nonlinear system. The adaptive control realized by the QL algorithm can respond quickly for online control after pre-learning. The QL algorithm has been used for the dynamic optimization of integrated energy systems [17], [18], the automatic generation control of power systems [19], and the auxiliary regulation of power systems [20], [21].

However, the original QL algorithm uses discrete action and state variables, which makes the learning process inefficient and convergence difficult for a dynamic nonlinear system with n parallel PCSs. To increase the learning efficiency, a double-layer QL algorithm for n parallel PCSs is developed to realize dynamic decision-making for adaptive power control, and the following improvements are made.

First, to avoid repeated actions, the selection of power allocation coefficients is simplified according to the system constraints.

Then, when a PCS is controlled to cut in or off, the parallel system will be in the non-steady state and adaptively change the power allocations of all the PCSs to ensure a high operation efficiency. In the steady state, the optimal operation efficiency is achieved for the parallel PCSs with the power sharing method. In fluctuating operation situations, the parallel PCSs may spend more time in the non-steady state than in the steady state. Because of these characteristics, the double-layer QL algorithm is developed as follows.

1) Outer layer: according to the power allocation characteristics, the power sharing method is helpful for achieving the optimal overall operation efficiency of the parallel PCSs in the steady states. To obtain the interval values of non-steady states, the action space of the outer layer is divided into $n+1$ power allocation modes.

2) Inner layer: in the non-steady interval, the inner layer uses the actor neural network to determine the optimal action for the power allocation. Meanwhile, the action reward obtained by the inner layer dynamically revises the non-steady interval values in the outer layer.

In the simulation, a wind power plant is supported by an ESS, which was utilized to smooth the wind power fluctuations. First, after the QL controller is obtained by pre-learning within the range of the ESS power, the QL controller is subjected to real working conditions. Then, the efficiencies of the parallel PCSs are compared between the adaptive power control method and traditional methods. Finally, for the ESS used to respond to the wind power fluctuations, the comparisons of the optimization results between the proposed double-layer QL algorithm and intelligent algorithms are performed.

The remainder of this paper is organized as follows. Section II presents the efficiency model for ESSs. Section III presents the proposed adaptive power control method based on the double-layer QL algorithm. Section IV presents the simulation analysis results obtained using a detailed simulation model. Conclusions are presented in Section V.

II. Efficiency Model for ESSs

A. Structure of ESSs

The structure of an ESS is shown in Fig. 1. The ESS is composed of an EMS and multi-parallel battery systems. Each battery system includes a PCS, a battery management system (BMS), and several BPs [22]. P_{total} is the power output of the ESS and is defined as the ESS power; i is the index for the battery systems; and P_i is the power allocated to the i^{th} PCS. P_{total} is allocated to all the parallel PCSs.

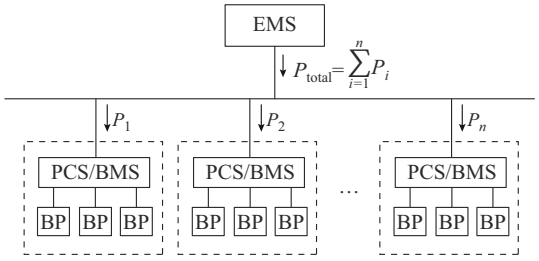

Fig. 1. Structure of an ESS.

After the ESS receives the regulation requirement, the EMS generates the control signal for each PCS to allocate P_{total} to the battery systems according to the state of charge (SOC) values and the rated power of the battery systems. During the charging and discharging processes, the power exchange leads to power loss in the conversion of PCSs. According to the real-time P_{total}, the overall operation efficiency of multi-parallel PCSs can be optimized by adaptively allo-

cating P_i to the i^{th} PCS, which can significantly reduce the power loss.

B. Objective Function

To realize the adaptive power allocations for multi-parallel PCSs in the ESS, the following assumptions are made.

Assumption 1: the values of the battery parameters such as the rated capacity and rated power are assumed to be identical among all the battery systems.

Assumption 2: the power loss models of the two working directions of the PCSs are identical.

Assumption 3: the self-discharge of the battery is ignored.

The objective function is to maximize the overall operation efficiency of the parallel PCSs, which is calculated as:

$$g=\max\{\eta_{\mathrm{T}}(P_1^{\mathrm{b}},P_2^{\mathrm{b}},\ldots,P_i^{\mathrm{b}},\ldots,P_n^{\mathrm{b}})\} \tag{1}$$

where n is the total number of PCSs; η_{T} is the overall operation efficiency of the parallel PCSs; and P_i^{b} is the power allocated to the i^{th} PCS on the grid side.

1) Efficiency Model for PCS

For an individual PCS, the operation efficiency is defined as:

$$\eta(P_{\mathrm{in}})=\frac{P_{\mathrm{out}}}{P_{\mathrm{in}}}=\frac{P_{\mathrm{in}}-P_{\mathrm{loss}}}{P_{\mathrm{in}}} \tag{2}$$

where P_{in} is the input power of the PCS; P_{out} is the output power of the PCS; $\eta(\cdot)$ is the operation efficiency of the PCS; and P_{loss} is the total power loss of the PCS.

According to the power loss model proposed in [23] and actual power loss data obtained from industry, the operation efficiency of a PCS can be calculated. The typical efficiency curve of a PCS is obtained using the multilinear curve fitting method, as shown in Fig. 2. The fitting efficiency function is a piecewise function given as follows:

$$\eta(P)=\begin{cases}0.18(P-0.1)+0.96 & 0<P\le 0.2\\ 0.0275(P-0.2)+0.978 & 0.2<P\le 0.4\\ 0.0125(P-0.4)+0.9835 & 0.4<P\le 0.6\\ -0.004(P-0.6)+0.986 & 0.6<P\le 0.8\\ -0.006(P-0.8)+0.9852 & 0.8<P\le 1.0\end{cases} \tag{3}$$

where P is the power input of PCS.

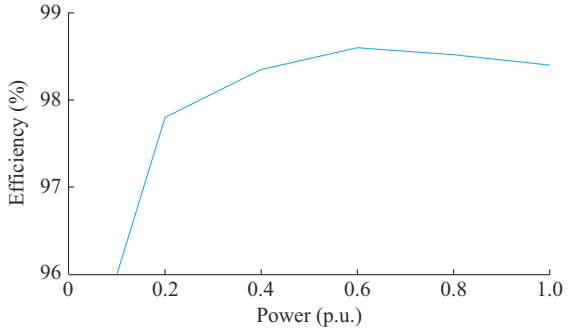

Fig. 2. Typical efficiency curve of a PCS with respect to allocated power.

For a PCS, the linear fitting results describe the relationship between the operation efficiency and the power allocated to the PCS. The operation efficiency increases with the allocated power increase when the allocated power varies within the range of [0, 0.6]p. u., and it increases significantly when the allocated power increases from 0 to 0.2 p.u.. The operation efficiency of the PCS is maximized when the allocated power is 0.6 p.u., and it decreases slowly when the allocated power increases from 0.6 to 1 p.u.. Thus, the power allocated to the PCS significantly affects its operation efficiency.

2) Efficiency Model for Parallel PCSs

In the case of n parallel PCSs, the overall operation efficiency of the PCSs during the discharging and charging processes is calculated as:

$$\eta_{\mathrm{T}}(P_{\mathrm{total}})=\begin{cases}\dfrac{P^{\mathrm{a}}}{P^{\mathrm{b}}}=\dfrac{\sum_{i=1}^{n}P_i^{\mathrm{a}}}{\sum_{i=1}^{n}P_i^{\mathrm{b}}}=\dfrac{\sum_{i=1}^{n}P_i^{\mathrm{b}}\eta_i^{\mathrm{c}}(P_i^{\mathrm{b}})}{\sum_{i=1}^{n}P_i^{\mathrm{b}}} & \text{discharging}\\ \dfrac{P^{\mathrm{b}}}{P^{\mathrm{a}}}=\dfrac{\sum_{i=1}^{n}P_i^{\mathrm{b}}}{\sum_{i=1}^{n}P_i^{\mathrm{a}}}=\dfrac{\sum_{i=1}^{n}P_i^{\mathrm{b}}}{\sum_{i=1}^{n}\dfrac{P_i^{\mathrm{b}}}{\eta_i^{\mathrm{d}}(P_i^{\mathrm{b}})}} & \text{charging}\end{cases} \tag{4}$$

where P_{total} varies within the range of (0, 1)p.u.; $\eta_{\mathrm{T}}(P_{\mathrm{total}})$ is the overall operation efficiency of the parallel PCSs when the ESS power value is P_{total}; P^{a} and P^{b} are the total power outputs on the battery side and grid side, respectively; P_i^{a} and P_i^{b} are the power outputs of the i^{th} PCS on the battery side and grid side, respectively; and η_i^{c} and η_i^{d} are the operation efficiencies of the i^{th} PCS during the charging and discharging processes, respectively.

According to assumptions 1 and 2, the overall operation efficiency of the parallel PCSs is simplified as:

$$\eta_{\mathrm{T}}(P_{\mathrm{total}})=\begin{cases}\dfrac{\sum_{i=1}^{n}P_i^{\mathrm{b}}\eta(P_i^{\mathrm{b}})}{\sum_{i=1}^{n}P_i^{\mathrm{b}}} & \text{discharging}\\ \dfrac{\sum_{i=1}^{n}P_i^{\mathrm{b}}}{\sum_{i=1}^{n}\dfrac{P_i^{\mathrm{b}}}{\eta(P_i^{\mathrm{b}})}} & \text{charging}\end{cases} \tag{5}$$

C. Constraints of ESSs

1) Power Balance Constraint

According to the regulation requirement, the ESS power allocated by the EMS to each PCS is given as [22]:

$$P_{\mathrm{total}}=\sum_{i=1}^{n}P_i \tag{6}$$

2) Battery Power Constraint

According to the conditions of the BP, the battery charging and discharging power should satisfy the constraint of (8), and the ESS power should be allocated to each PCS within the acceptable charging and discharging power range of the BP.

$$-P_i^{\mathrm{c}}\le P\le P_i^{\mathrm{d}} \tag{7}$$

where P_i^{d} and P_i^{c} are the rated discharging and charging power of the i^{th} battery system, respectively.

3) SOC Constraint

The SOC indicates the remaining energy of the BP. In the repeating process of charging and discharging, the SOC of the BP is constrained as [24]:

$$SOC_i^{\min} \le SOC_i \le SOC_i^{\max} \tag{8}$$

where SOC_i is the real-time SOC value of the i^{th} battery system; and $SOC_i^{\max}$ and $SOC_i^{\min}$ are the maximum and minimum SOC values of the i^{th} battery system, respectively. The variation range of the SOC is [0.2, 0.8] in this study.

According to the constraints given by (6) and (7), the selection of the power allocation coefficient can be simplified. For the charging and discharging process, this coefficient is obtained as:

$$\begin{cases} \mu_1 \in [1/n,\ 1/P_{\text{total}}] & P_{\text{total}} > 1 \\ \mu_1 \in [1/n,\ 1] & P_{\text{total}} \le 1 \\ \mu_2 \in [(1-\mu_1)/(n-1),\ \min\{1-\mu_1, \mu_1\}] \\ \mu_3 \in [(1-\mu_1-\mu_2)/(n-2),\ \min\{1-\mu_1-\mu_2, \mu_2\}] \\ \vdots \\ \mu_j \in \left[\dfrac{1}{(n-j+1)}\left(1-\sum_{l=1}^{j-1}\mu_l\right),\ \min\left\{1-\sum_{l=1}^{j-1}\mu_l, \mu_{j-1}\right\}\right] \\ \vdots \\ \mu_n = 1-\sum_{l=1}^{n-1}\mu_l \end{cases} \tag{9}$$

where μ_i is the power allocation coefficient for the i^{th} PCS. The maximum power allocation coefficient μ_1 varies with respect to P_{total}, and $\mu_1 \ge \mu_2 \ge \ldots \ge \mu_n$. Equation (9) satisfies the constraints of (6) and (7). As the parameters are identical among the battery systems, repeated selections are avoided. Therefore, the selection of power allocation coefficients is simplified.

III. Adaptive Power Control Method Based on Double-layer QL Algorithm

In fluctuating operation situations, the EMS adaptively generates the control signal for each PCS according to the significantly fluctuating ESS power. The proposed double-layer QL algorithm adaptively generates the optimal power allocation. In this study, to optimize the overall operation efficiency of the parallel PCSs, the EMS performs online dynamic decision-making for adaptive power allocation with the double-layer QL algorithm.

A. Principle of QL Algorithm

The QL algorithm estimates the *Q*-value dynamically using real-time feedback and the *Q*-value function according to the MDP [25]. The QL algorithm is not a supervised algorithm based on labeled datasets. In the QL algorithm, the agent has no prior knowledge after initialization and takes actions to obtain the reward given by a specific environment. The agent obtains the optimal action strategy in the process of accumulating experience.

In this study, we regard the EMS as an agent to generate the control signal. The original QL algorithm continuously optimizes the action value function $Q_k(s,a)$ in each state in the iteration, reinforces the action probability to maximize the total expected discount reward, and seeks the optimal strategy online. In the k^{th} learning iteration, the agent takes action a in state s and then enters the next state s', obtains the immediate reward r, and updates the *Q*-value of the corresponding position in the $\boldsymbol{Q}$ matrix according to (10).

$$\begin{cases} Q_{k+1}(s,a) = (1-\alpha)Q_k(s,a) + \alpha(r + \gamma \max\limits_{a' \in A} Q_k(s',a')) \\ Q_{k+1}(\tilde{s},\tilde{a}) = Q_k(\tilde{s},\tilde{a}) \quad \forall(\tilde{s},\tilde{a}) \ne (s,a) \end{cases} \tag{10}$$

where a' is the action in state s'; A is the action space; $\tilde{s}$ and $\tilde{a}$ are the state and action in $\boldsymbol{Q}$ matrix that are not equal to s and a, respectively; α is the learning factor ($0 < \alpha < 1$); and γ is the discount rate ($0 < \gamma < 1$). References [26] and [27] presented the general principles for selecting parameters in the QL algorithm, including the discount rate γ and learning factor α.

The selection of the action strategy is a crucial part of the QL algorithm. The goal of the agent is to select the strategy with the maximum reward, i.e., to maximize the *Q*-value in any state. We express the action strategy with the highest *Q*-value as the greedy action strategy π^*. The greedy action strategy π^* that maximizes the *Q*-value in state s in the k^{th} iteration is given by:

$$\pi^*(s) = \arg\max_{a \in A} Q_k(s,a) = Q_k(s,a_g) \tag{11}$$

where a_g is the greedy action.

However, selecting the action with the highest *Q*-value in each iterative learning process makes the agent always perform similar actions and find local optimal solutions, and the agent does not fully perform all the actions in the action set. Therefore, in this study, an action selection strategy based on the probability distribution is adopted [26]. The action selection probability matrix $\boldsymbol{Pr}$ is updated according to (12) with the updating of the $\boldsymbol{Q}$ matrix. In any state, the selection probability of the action with the highest *Q*-value increases, while the selection probabilities of the other actions decrease proportionally (the selection probabilities of all the actions are nonzero). The selection probability of the optimal action is close to 1 after the learning process.

$$\begin{cases} Pr_{k+1}(s,a_g) = Pr_k(s,a_g) - \beta(1 - Pr_k(s,a_g)) \\ Pr_{k+1}(s,a) = Pr_k(s,a)(1-\beta) & \forall a \in A, a \ne a_g \\ Pr_{k+1}(s,a) = Pr_k(s,a) & \forall a \in A, \forall \tilde{s} \in S, \tilde{s} \ne s \end{cases} \tag{12}$$

where β is the probability distribution factor; and $Pr_k(s,a)$ and $Pr_k(s,a_g)$ are the probabilities of conducting actions a and a_g in state s in the k^{th} iteration, respectively. The initial value of $Pr(s,a)$ is $1/|A|$, where $|A|$ is the dimension of action space A, and its range is $Pr(s,a) \in [0, 1]$. The action space A is given as:

$$A = \prod_{i=1}^{n} a_i \tag{13}$$

where a_i is the action space of the i^{th} battery system. There are n battery systems in total. Let the action number of each battery system be M, and $|A| = M^n$. In the power allocation problem, M is related to the rated power of each battery system. $|A|$ increases significantly with the number of PCSs,

which will cause the curse of dimensionality. Therefore, it is necessary to improve the setting of A to resolve the curse of dimensionality.

When the QL algorithm is used to solve the optimal power allocation problem for parallel PCSs, the agents learn the input (i.e., η_T, P_{total}, SOC_i, and r) and the output (i.e., action strategy π), and continuously enhance the cognition of the model and update the ***Pr*** and ***Q*** matrices online. In the QL algorithm, a ***Q*** matrix is constantly reinforced in learning.

B. Actor Neural Network

In the optimal power allocation problem for multi-parallel PCSs, s is a time-series variable. For continuous state variables, the actor neural network $\mu(s|\theta)$ is used to approximate π, and the update method for the strategy function parameters is given as:

$$\theta_{k+1}=\theta_k+\alpha_\mu\nabla_\theta \ln \pi_\theta(s_k,a_k)\delta_k \tag{14}$$

where δ_k is the dominant function in the k^{th} learning iteration; θ_k is the hyperparameter of the actor neural network in the k^{th} learning iteration; $\pi_\theta(s_k,a_k)$ is the action strategy in the k^{th} learning iteration; and α_μ is the learning rate of actor neural network. The input of the actor neural network is state s, and the output is action strategy π. The parameters of the actor neural network are updated according to the learning rate α_μ.

The update method for the target network parameters is given as:

$$\theta_{\mu',k+1}=\tau\theta_{\mu,k}+(1-\tau)\theta_{\mu',k} \tag{15}$$

where τ is the learning rate of the target network parameters and $\tau \ll 1$; μ is the optimal strategy; μ' is the updated optimal strategy; and $\theta_{\mu,k}$ is the hyper parameter of the actor neural network in strategy μ.

The actor neural network intends to achieve the optimal action to maximize the action value of the current state, which is calculated as:

$$\max L_\mu = Q(s_k,a_k|\theta_\mu)\Big|_{a_k=\mu(s_k|\theta_\mu)} \tag{16}$$

where L_μ is the target of the optimal strategy μ; $Q(s_k,a_k|\theta_\mu)$ is the Q-value obtained by action a_k in s_k and θ_μ; and $\mu(s_k|\theta_\mu)$ is the optimal strategy in s_k and θ_μ.

C. Analysis of Characteristics

Figure 3 presents the optimal power allocation strategy for two PCSs, which is obtained using the original QL algorithm. In interval A, P_{total} is less than 0.465 p.u. and only PCS 1 operates. In interval B, when P_{total} varies from 0.465 to 0.5 p.u., PCS 2 operates with a constant power, and the power allocated to PCS 1 increases linearly with respect to the ESS power. When P_{total} varies from 0.5 to 0.6 p.u., PCS 1 operates with a constant power, and the power allocated to PCS 2 increases linearly with respect to P_{total}. In interval C, P_{total} is equally divided between the two PCSs.

Interval B can be regarded as the non-steady state when a new PCS cuts in. In the steady state, the optimal operation efficiency for the parallel PCSs is achieved with the power sharing method [10].

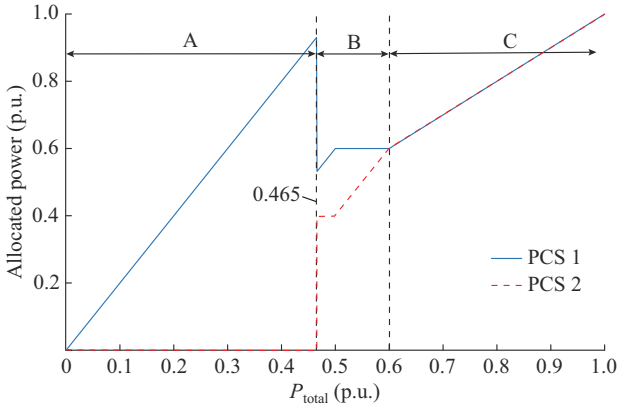

Fig. 3. Optimal power allocation strategy for two PCSs.

D. Power Control Method Based on Double-layer QL Algorithm

According to the characteristic of the optimal power allocation process presented in Section III-C, a double-layer QL algorithm is proposed herein. The agent in the double-layer QL algorithm includes two layers of learning units. The outer layer obtains the interval values of non-steady states, and the inner layer uses an actor neural network to determine the optimal action for power allocation. Meanwhile, the non-steady interval values in the outer layer are dynamically revised according to the reward obtained by the inner layer. The overall structure of the agent is shown in Fig. 4.

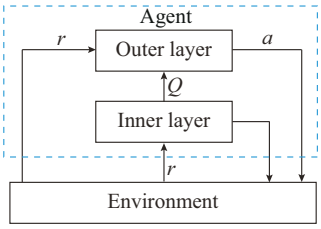

Fig. 4. Overall structure of agent.

1) Action Space

The action space A is formed by $n+1$ working modes, which include n multi-PCS sharing modes and one adaptive allocation mode. It is given as:

$$A=\{MS_1,MS_2,\ldots,MS_n,MA\} \tag{17}$$

where MS_i denotes that the parallel system is operating in the power sharing mode of PCS i, and the power allocated to each PCS in this mode is equal to P_{total}/n; and MA denotes that the system is operating in the adaptive mode, where the actor neural network selects the power allocation proportion within the power constraint of the battery according to (9). In action MA, the actor neural network updates the parameters according to (14), where $\delta(k)$ is used as the new reward and punishment information to determine the updating direction of the action probability. It is calculated as:

$$\delta(k)=r_k(s,MA)-\max\{r(s,MS_1),r(s,MS_2),\ldots,r(s,MS_n)\} \tag{18}$$

where $r_k(s,MA)$ is the reward obtained by action MA in state s in the k^{th} learning iteration; and $r(s,MS_i)$ is the reward which is obtained by action MS_i in the first few learning iter-

ations.

The reward value of the inner layer action is used to obtain $Q_{k+1}(s, MA)$ in the outer layer, and the $\boldsymbol{Q}$ matrix in the outer layer is updated according to (19). The Q-value under $a = MA$ is updated only when $Q_{k+1}(s, MA)$ exceeds Q-value for all sharing modes.

$$Q_{k+1}(s, MA) = \begin{cases} (1-\alpha)Q_k(s,a) + \alpha(r_k + \gamma \max\limits_{a' \in A} Q_k(s', a')) \\ \qquad\qquad Q_{k+1}(s, MA) > \max\limits_{a \in A} Q_k(s,a) \\ Q_k(s, MA) \quad Q_{k+1}(s, MA) \le \max\limits_{a \in A} Q_k(s,a) \end{cases} \tag{19}$$

where r_k is the reward in the k^{th} learning iteration.

2) State Space

One crucial characteristic of RL is judging the action performed in the environment. In the optimal overall operation efficiency problem for parallel PCSs, the action selection is directly related to P_i, and the power allocation strategy is changed according to the ESS power. For the algorithm to adaptively follow the power fluctuations, the ESS power in a time period for power allocation is defined as a state s in the state space S, and ε is defined as the sampling accuracy. The ESS charges when P_{total} is negative and discharges when P_{total} is positive. The dimensions of the $\boldsymbol{Q}$ matrix are $(2P_{\text{total}}/\varepsilon + 1,\ n+1)$.

3) Reward Function

In the optimal efficiency problem, there is no correlation between states, and the Q-value is only related to the corresponding state. In the process of RL, a larger reward is better. Thus, the overall operation efficiency can be set as the reward function r. However, if the overall operation efficiency η_{T} is directly set as the reward function, the differences among the reward values are small, and do not reflect the action performance. Therefore, the efficiency is transformed into the range of 0-100. Then, r is calculated as:

$$r = (\eta_{\text{T}} - \eta_{\text{T}}^{\min}) \frac{100}{\eta_{\text{T}}^{\max} - \eta_{\text{T}}^{\min}} \eta^* \tag{20}$$

where $\eta_{\text{T}}^{\max}$ and $\eta_{\text{T}}^{\min}$ are the maximum and minimum values of η_{T}, respectively; and η^* is the immediate efficiency value.

After A, S, and r are determined, Algorithm 1 is used for the training and application of the double-layer QL controllers.

E. Optimal Power Allocation Process

To achieve SOC balance among the battery systems, the principle of SOC priority is followed in this study, i.e., the battery system with the lowest SOC is charged first, and the battery system with the highest SOC is discharged first. The allocated power is determined by multiplying the power allocation coefficients $\mu_1, \mu_2, \ldots, \mu_n$ of the output of the QL controller from pre-learning by P_{total}. According to $SOC_i(t)$, when discharging, the maximum power is allocated to the battery system with the highest SOC. When charging, the maximum power is allocated to the battery system with the lowest SOC, and so on. The SOC variation is calculated in each time period for power regulation according to (21), and SOC balance between different battery systems is achieved.

$$\Delta SOC = \mu P_{\text{total}} \Delta T / C_0 \tag{21}$$

where ΔT is a time period; and C_0 is the rated capacity of the battery systems.

Algorithm 1: training and application workflow for the proposed control-method based on double-layer QL algorithm

```
Result: optimal power allocation for parallel PCSs learned and utilized
if in the training mode then
  Initialize Q, Pr, μ(s|θ) and k = 0
  while true do
    Enter a current state s randomly, and select the action a(k) according to
      Pr(s)
    Observe the immediate reward given by (20) and correct the next state
    if a = MS_i then
      Update Q and Pr according to (10) and (12), respectively
    else if a = MA then
      Update Q and Pr according to (19) and (12), respectively
      Calculate δ(t) using (18), and then update the network parameter θ ac-
        cording to (14)
    end
    Judge whether the Q matrix is a Q* matrix composed of the optimal Q
      If not, assign the next state to the current state and set k = k + 1
  end
else
  Load the trained network weights and learned control strategy
end
```

If the SOC of a battery system exceeds the constraints given by (8) under the control of the EMS, this battery system stops working, and the EMS reallocates the ESS power. The flowchart of the optimal power allocation in each time period is presented in Fig. 5.

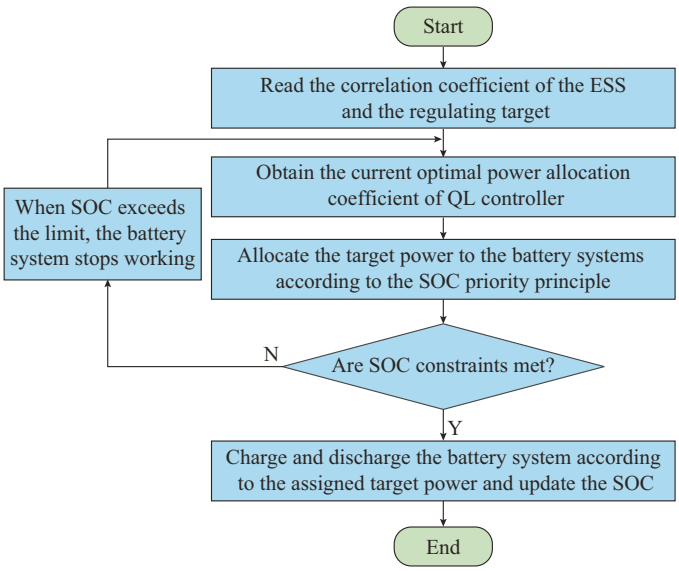

Fig. 5. Flowchart of optimal power allocation in a time period.

IV. Simulation Analysis

In this section, the optimal allocation of the ESS power to four parallel PCSs is taken as an example to validate the feasibility of the adaptive power control method based on the double-layer QL algorithm. First, the adaptive power control method is obtained by the QL controller with pre-learning. Then, the QL controller is applied to a simulated scenario for conducting static and dynamic comparisons with existing methods. In the simulation, an ESS is utilized to smooth the power fluctuations caused by a wind power plant.

A. Simulation Parameters

To validate the double-layer QL algorithm for solving the optimal operation efficiency problem for n parallel PCSs, the simulation is conducted for an ESS with four independent adjustable lithium battery systems. Each battery system contains a PCS. The rated power and the battery capacity of each battery system are 1.5 MW and 1 MWh, respectively. The initial SOCs of the battery systems are 55%, 65%, 60%, and 50%. The parameter values of the PCSs are presented in Table I.

TABLE I
PARAMETER VALUES OF PCSS

Parameter	Value
Rated grid voltage (V)	380
Allowable grid voltage (V)	$380\times(1\pm7\%)$
Rated grid frequency (Hz)	50
Maximum DC power (MW)	1.5
DC voltage range (V)	250-850
Maximum input current (kA)	2.5
Maximum operation efficiency (%)	98.7

For the four parallel PCSs, the action space of the outer layer A_{out} is divided into five action modes in accordance with (22), and $|A_{\text{out}}|$ is equal to 5.

$$a \in A_{\text{out}} = \{MS_1, MS_2, MS_3, MS_4, MA\} \tag{22}$$

P_{total} is standardized within the variation range of [–1, 1]p.u.. When P_{total} is positive, the ESS is in the discharging state. When P_{total} is negative, the ESS is in the charging state. ε is set to be 0.001 p.u.. The dimension of S is 2001, and the dimension of the $\boldsymbol{Q}$ matrix is (2001, 5).

The training parameter values of the double-layer QL algorithm are presented in Table II. These values are obtained via simulation experiments, and they make the algorithm have good performance in solving the problem of optimal power allocation for multi-parallel PCSs.

TABLE II
TRAINING PARAMETER VALUES OF DOUBLE-LAYER QL ALGORITHM

Training parameter	Value
Sampling time	2000
Outer learning factor α	0.99
Inter learning factor α_μ	0.001
Outer discount factor γ	0.0005
Sampling accuracy ε	0.001
Probability distribution factor β	0.01

In Table II, the outer learning factor α represents how much trust is given to the Q-value in the k^{th} learning iteration, and the outer discount factor γ represents how much reservation is taken for the optimal Q-value obtained in the previous iteration. The previous action has little effect on the follow-up actions in the power allocation problem; thus, α is set to be 0.99, and γ is set to be 0.0005 [26], [27].

According to the typical efficiency curve shown in Fig. 2, the overall operation efficiency of multi-parallel PCSs varies within the range of [94%, 99%]. The probability that the efficiency under MA exceeds that under MS_i is low, and the $\boldsymbol{Q}$ matrix is updated only when the efficiency under MA exceeds those under all other sharing modes. Therefore, to reinforce action MA at each effective learning iteration, the reward value is set to be $100r$ when $a = MA$, where r is given by (20). The obtained reward function R is given as:

$$R = \begin{cases} 1724\eta_{\text{T}} - 1624.14 & a = MS_i \\ 100\times(1724\eta_{\text{T}} - 1624.14) & a = MA \end{cases} \tag{23}$$

B. Pre-learning

As the agent of the double-layer QL algorithm has no prior knowledge, a large number of trial-and-error learning iterations are performed in the early stage to interact with the environment. The agent accumulates experience to reinforce the intensity of the action that can yield the maximum reward value in the environment. Without pre-learning, the battery system will be switched on and off unreasonably in the actual environment, which negatively affects the security and stability of the system. The continuous change of state will also make it difficult for the double-layer QL algorithm to learn effectively. Thus, the QL controller can only be applied in the actual environment after pre-learning [28].

The results of a pre-learning comparison between the original QL algorithm and the double-layer QL algorithm are presented in Table III. Herein, "average efficiency" refers to the average of the overall operation efficiency of the four parallel PCSs with different P_{total} values (P_{total} is updated by pre-learning), which is obtained under the optimal power allocation after pre-learning. It can be observed from Table III that the proposed double-layer QL algorithm obtains the optimal solution in a shorter time by reducing the required number of iterations for pre-learning.

TABLE III
PRE-LEARNING COMPARISON BETWEEN ORIGINAL QL ALGORITHM AND DOUBLE-LAYER QL ALGORITHM

Algorithm	Iteration number	Time cost for pre-learning (s)	Average efficiency (%)
Original QL	>10000000	2948.00	98.251
Double-layer QL	658801	150.74	98.379

Figure 6 shows the variations in the average efficiency with different values of n in the pre-learning process of the double-layer QL algorithm. Samples are selected every 2000 learning iterations. The double-layer QL algorithm finally converges to the optimal solutions.

With a specific value of n, the average efficiency increases with the number of iterations until the optimal average efficiency is reached. The optimal average efficiency increases with the number of parallel PCSs, because the available capacity of the system increases. With an increase of n, the required number of iterations for the double-layer QL algorithm increases. The simulation results shown in Fig. 6 confirm that the proposed double-layer QL algorithm completes the training of the controller and obtains the optimal power allocation result within a limited number of iterations. According to the characteristics of optimal power allocation, it

is clear that a larger number of parallel PCSs corresponds to a shorter non-steady interval. Thus, increasing the number of parallel PCSs has a small effect on the convergence of the proposed double-layer QL algorithm.

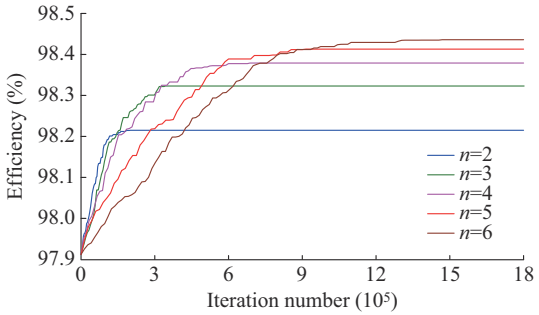

Fig. 6. Average efficiency with different values of n in pre-learning process of double-layer QL algorithm.

C. Comparisons of Static Efficiencies of Parallel PCSs

In practical applications, the EMS generates the control signal to allocate the ESS power to the parallel PCSs according to the power control method. Two traditional power control methods are introduced below.

1) Power sharing method: the ESS power is equally allocated to all parallel PCSs.

2) Hierarchical switching method: according to the target power, the PCSs are cut in or off stepwise. When a new PCS is controlled to cut in, all the other operating PCSs are working with the rated power.

As shown in Fig. 7(a), the four PCSs share the same allocated power with the power sharing method, and the power allocated to each PCS increases linearly with an increase in P_{total}. Figure 7(b) shows the power allocation for the four parallel PCSs with the hierarchical switching method. The power allocated to the cut-in PCS increases linearly with an increase in P_{total} until the rated power is reached. The power allocation for four parallel PCSs obtained via the proposed adaptive power control method using the double-layer QL algorithm is shown in Fig. 7(c), which verifies the characteristics of power allocation described in Section III-C. The optimal power control method is identical for the charging and discharging processes, as the power allocation has similar characteristics. The sum of P_1, P_2, P_3, and P_4 is P_{total}, and $P_1 \geq P_2 \geq P_3 \geq P_4$. When the ESS power varies within the intervals of A, C, E, and G, the parallel PCSs operate under MS_1, MS_2, MS_3, and MS_4, respectively. When the ESS power varies within the intervals of B, D, and F, the parallel PCSs operate under MA.

In Fig. 8, the static operation efficiency of the four parallel PCSs with the proposed method is compared with those for two traditional methods. The static operation efficiency is the overall operation efficiency of the parallel PCSs under each P_{total}.

Among the three methods examined, the proposed method always achieves the highest efficiency. When P_{total} is lower than 0.233 p.u., the efficiency of the proposed method is equal to that of the hierarchical switching method, where only one battery system operates. When P_{total} varies between 0.233 and 0.6 p.u., the efficiency of the proposed method is significantly higher than those of the other methods. When P_{total} is higher than 0.6 p.u., the efficiency of the proposed method is equal to that of the power sharing method. Therefore, with the proposed method, the overall operation efficiency of the four parallel PCSs is optimal with changes in P_{total}.

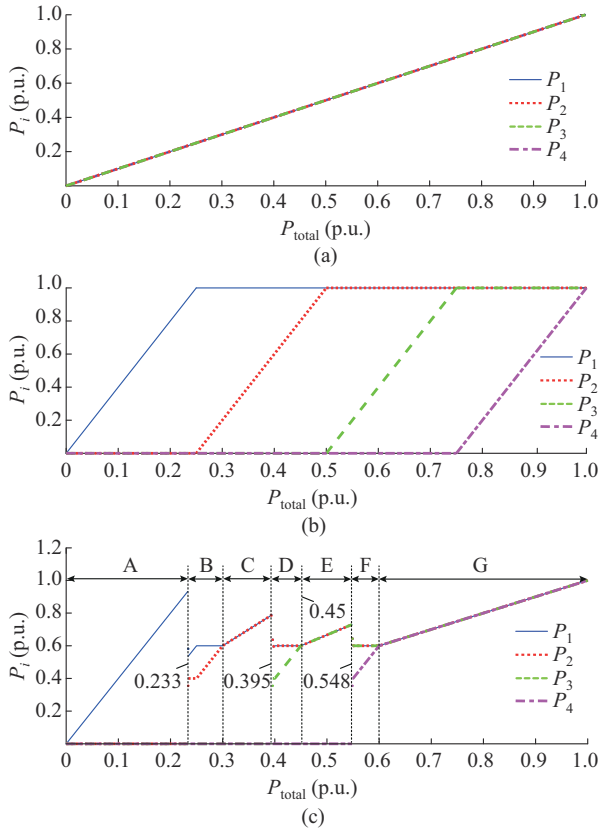

Fig. 7. Power allocation for four parallel PCSs with three different power control methods. (a) Power sharing method. (b) Hierarchical switching method. (c) Proposed method.

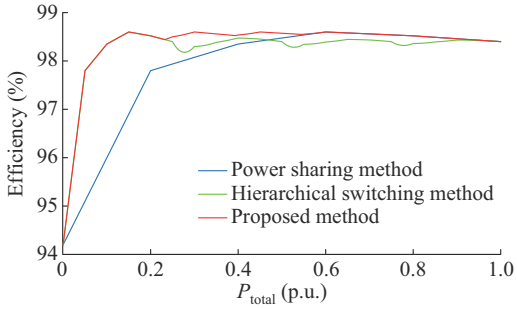

Fig. 8. Static operation efficiency of four parallel PCSs with three different methods.

D. Comparisons of Dynamic Operation Efficiencies of Parallel PCSs

The historical wind power data of 16 wind turbines of a wind farm for 1 day are selected as reference data for the simulation. The rated power of each wind turbine is 1 MW, and the sampling period is 1 min. MATLAB is used to analyze the power output data of the wind farm, and five-layer wavelet packet decomposition is utilized to obtain the expect-

ed power of the wind farm for grid connection [29]. The actual output curve of the wind power for 1 day and the expected output curve are shown in Fig. 9.

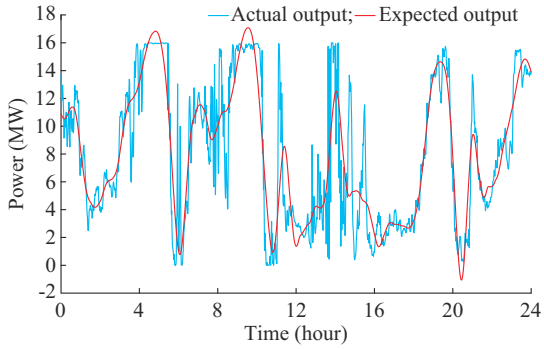

Fig. 9. Actual output curve of wind farm and expected output curve.

To avoid frequent battery charging and discharging, only the low-frequency power component obtained via the wavelet packet decomposition is taken as the expected power output of the ESS [30]. The time period for power allocation is 1 min. The target power curve for ESS charging and discharging is shown in Fig. 10. The positive and negative power values correspond to the charging and discharging states, respectively. Clearly, the wind power induces significant power fluctuations of the ESS. The average target power of charging/discharging is determined by calculating the average of all the positive/negative power values. The average target power for the whole day is –0.616 MW for charging and 0.621 MW for discharging. The average target power of charging/discharging is significantly lower than the maximum charging/discharging power for the whole day.

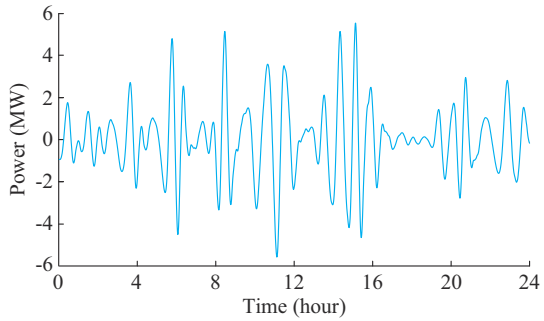

Fig. 10. Target power curve for ESS charging and discharging.

Figure 11 presents the SOC variation curve of the four battery systems with the smoothing of the wind power fluctuations, where BS_i represents the i^{th} battery system. Clearly, the charging and discharging of the four battery systems are controlled according to the SOC priority principle, and the different initial SOC values of the four battery systems initially cause an SOC imbalance. After 21 min, the SOC gradually tends to be balanced, and it always varies within the range of [20%, 80%].

When the charging and discharging target power shown in Fig. 10 is satisfied, the dynamic operation efficiency is the real-time overall operation efficiency of the four PCSs in a whole day. The dynamic operation efficiencies of the four PCSs with the proposed method, the power sharing method, and the hierarchical switching method are compared in Fig. 12, where $\eta_3-\eta_1$ is the dynamic operation efficiency difference between the proposed method and the power sharing method, and $\eta_3-\eta_2$ is the dynamic operation efficiency difference between the proposed method and the hierarchical switching method. When the ESS power is lower than 60% of the total rated power of the four PCSs, the operation efficiency of the parallel PCSs is significantly higher for the proposed method than for the power sharing method. Only when the ESS power is lower than 23% of the total rated power of the four PCSs, the operation efficiency with the hierarchical switching method is equal to that with the proposed method. Therefore, when the ESS power varies between 23% and 60% of the total rated power for a long time, the effectiveness of the proposed method is significantly higher than those of the two traditional methods.

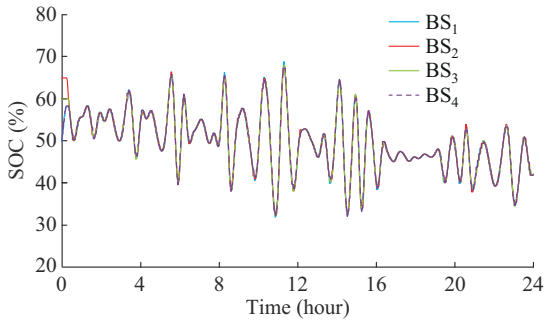

Fig. 11. SOC variation curve of four battery systems.

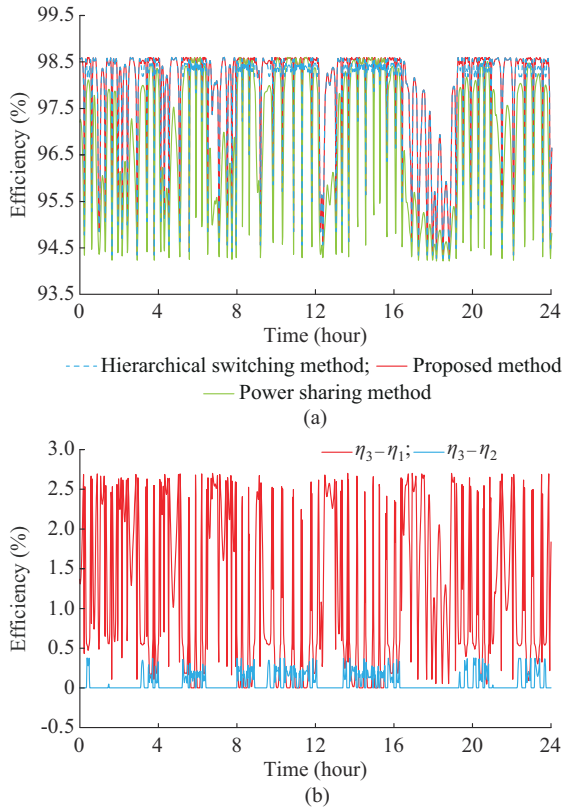

Fig. 12. Dynamic operation efficiencies of four parallel PCSs with three control methods and dynamic operation efficiency differences. (a) Dynamic operation efficiencies. (b) Dynamic operation efficiency differences.

For the ESS with parallel PCSs, the average dynamic operation efficiency and the average execution time with different power control methods are presented in Table IV. The average dynamic operation efficiency is the average value of the overall operation efficiency of the parallel PCSs for a whole day. The average executive time is the average value of the simulation time cost at each time step. Compared with the traditional hierarchical switching and power sharing methods, the adaptive methods using GA and PSO have higher operation efficiencies. However, if the number of parallel PCSs is increased, the average dynamic operation efficiencies of the adaptive methods using GA and PSO are comparable to or worse than that of the hierarchical switching method. For the proposed method, the average execution time is significantly shorter than those of the other adaptive methods, and the average dynamic operation efficiency is higher than those of all the other methods. This is because GA and PSO may find local optimal solutions with the increasing number of parallel PCSs, whereas the proposed method obtains the optimal solution in a relatively short time after pre-learning and updates the power allocations online for real-time control.

TABLE IV
AVERAGE DYNAMIC OPERATION EFFICIENCIES AND AVERAGE EXECUTION TIME WITH DIFFERENT POWER CONTROL METHODS

Power control method	Average dynamic operation efficiency (%)			Average execution time (s)		
	$n=4$	$n=6$	$n=8$	$n=4$	$n=6$	$n=8$
Proposed method	97.947	98.143	98.250	0.001388	0.001512	0.001685
Adaptive method using GA [13]	97.883	98.050	98.143	1.147000	4.198000	4.228000
Adaptive method using PSO [14]	97.879	98.053	98.143	1.512000	5.047000	5.182000
Hierarchical switching	97.683	98.053	98.143			
Power sharing	96.752	96.752	96.752			

The proposed method can optimize the operation efficiency of the PCS parallels at any time under fluctuations in the ESS power. This method reduces the power loss and is economically beneficial for long-term operation of the ESS.

V. CONCLUSION

An adaptive power control method based on the double-layer QL algorithm for the multi-parallel PCSs in an ESS is proposed for achieving the optimal operation efficiency of the PCSs. The proposed method allows the ESS power to be adaptively allocated to parallel PCSs in fluctuating situations. The following conclusions are drawn.

1) The proposed method optimizes the overall operation efficiency of four parallel PCSs in fluctuating operation situations.

2) Compared with the original QL algorithm, the proposed double-layer QL algorithm obtains the optimal solution in a shorter time by reducing the required number of iterations for pre-learning. Moreover, increasing the number of parallel PCSs has a small effect on the convergence of the proposed double-layer QL algorithm.

3) Compared with adaptive power control methods using different intelligent algorithms, the proposed method achieves the best performance without falling into the local optimization.

REFERENCES

[1] R. H. Byrne, T. A. Nguyen, D. A. Copp *et al*., "Energy management and optimization methods for grid energy storage systems," *IEEE Access*, vol. 6, pp. 13231-13260, Aug. 2017.

[2] Y. Yoo, S. Jung, and G. Jang, "Dynamic inertia response support by energy storage system with renewable energy integration substation," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 2, pp. 260-266, Mar. 2020.

[3] K. Wang, Y. Qiao, L. Xie *et al*., "A fuzzy hierarchical strategy for improving frequency regulation of battery energy storage system," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 689-698, Jul. 2021.

[4] A. K. Bae, B. J. Kim, C. J. Lee *et al*., "Multi-parallel operation control method of high efficiency PCS module for ESS," in *Proceedings of 10th International Conference on Power Electronics and ECCE Asia*, Busan, South Korea, May 2019, pp. 1-6.

[5] Y. Han, H. Li, P. Shen *et al*., "Review of active and reactive power sharing strategies in hierarchical controlled microgrids," *IEEE Transactions on Power Electronics*, vol. 32, no. 3, pp. 2427-2451, Mar. 2017.

[6] H. Han, X. Hou, J. Yang *et al*., "Review of power sharing control strategies for islanding operation of AC microgrids," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 200-215, Jan. 2016.

[7] G. Chen and X. Cai, "Adaptive control strategy for improving the efficiency and reliability of parallel wind power converters by optimizing power allocation," *IEEE Access*, vol. 6, pp. 6138-6148, Mar. 2018.

[8] T. Vogt, A. Peters, N. Fröhleke *et al*., "Power profile based selection and operation optimization of parallel-connected power converter combinations," in *Proceedings of 2014 International Power Electronics Conference*, Hiroshima, Japan, May 2014, pp. 2887-2892.

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

[10] S. Wang, J. Liu, Z. Liu *et al*., "Efficiency-based optimization of steady-state operating points for parallel source converters in standalone power system," in *Proceedings of 2016 IEEE 8th International Power Electronics and Motion Control Conference*, Hefei, China, May 2016, pp. 163-170.

[11] M. Rolak, C. Sobol, M. Malinowski *et al*., "Efficiency optimization of two dual active bridge converters operating in parallel," *IEEE Transactions on Power Electronics*, vol. 35, no. 6, pp. 6523-6532, Jun. 2020.

[12] F. Díaz-González, A. Sumper, O. Gomis-Bellmunt *et al*., "A review of energy storage technologies for wind power applications," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 4, pp. 2154-2171, Jan. 2012.

[13] F. H. Dupont, J. Zaragoza, C. Rech *et al*., "A new method to improve the total efficiency of parallel converters," in *Proceedings of Brazilian Power Electronics Conference*, Gramado, Brazil, Apr. 2014, pp. 210-215.

[14] J.-H. Teng, S.-H. Liao, W.-H. Huang *et al*., "Smart control strategy for conversion efficiency enhancement of parallel inverters at light loads," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 12, pp. 7586-7596, Dec. 2016.

[15] J. Yan and L. Mu, "A method to improve the efficiency of asymmetric parallel converters for PV generation," in *Proceedings of 5th IEEE International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*, Changsha, China, Nov. 2015, pp. 1919-1923.

[16] W. Gong, Z. Liao, X. Mi *et al*., "Nonlinear equations solving with intelligent optimization algorithms: a survey," *Complex System Modeling and Simulation*, vol. 1, no. 1, pp. 15-32, Mar. 2021.

[17] Y. Wu, Z. Huang, H. Liao *et al*., "Adaptive power allocation using artificial potential field with compensator for hybrid energy storage systems in electric vehicles," *Applied Energy*, vol. 257, no. 1, pp. 1-14, Jan. 2020.

[18] B. Zhang, W. Hu, D. Cao *et al*., "Deep reinforcement learning-based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy," *Energy Conversion and Management*, vol. 202, no. 1, pp. 1-14, Dec. 2019.
[19] P. Kofinas, A. I. Dounis, and G. A. Vouros, "Fuzzy *Q*-learning for multi-agent decentralized energy management in microgrids," *Applied Energy*, vol. 219, no. 1, pp. 53-67, Jun. 2018.
[20] X. S. Zhang, Q. Li, T. Yu *et al*., "Consensus transfer *Q*-learning for decentralized generation command dispatch based on virtual generation tribe," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2152-2165, May 2018.
[21] Y. Shang, W. Wu, J. Guo *et al*., "Stochastic dispatch of energy storage in microgrids: an augmented reinforcement learning approach," *Applied Energy*, vol. 261, no. 1, pp. 1-11, Mar. 2020.
[22] D. Wang, J. Xue, J. Ye *et al*., "The economical optimal dispatching strategy of energy storage power station based on particle swarm algorithm," *Renewable Energy*, vol. 37, no. 5, pp. 714-719, May 2019.
[23] S. Inoue and H. Akagi. "A bidirectional DC-DC converter for an energy storage system with galvanic isolation," *IEEE Transactions on Power Electronics*, vol. 22, no. 6, pp. 2299-2306, Nov. 2007.
[24] M. Al-Saffar and P. Musilek, "Reinforcement learning-based distributed BESS management for mitigating overvoltage issues in systems with high PV penetration," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 2980-2994, Jul. 2020.
[25] C. J. C. H. Watkins and P. Dayan, "*Q*-learning," *Machine Leaning*, vol. 8, no. 1, pp. 279-292, May 1992.
[26] N. Zhang, *Reinforcement Learning*: *Theory, Algorithms and Its Application*. Harbin, China: Harbin Engineering University Press, 2001, pp. 126-155.
[27] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an Introduction*. Cambridge: MIT Press, 1998, pp. 87-160.
[28] T. P. I. Ahamed, P. S. N. Rao, and P. S. Sastry, "A reinforcement learning approach to automatic generation control," *Electric Power Systems Research*, vol. 63, no. 1, pp. 9-26, Aug. 2002.
[29] A. Meng, J. Ge, H. Yin *et al*., "Wind speed forecasting based on wavelet packet decomposition and artificial neural networks trained by crisscross optimization algorithm," *Energy Conversion and Management*, vol. 114, no. 1, pp. 75-88, Apr. 2016.
[30] N. Altin and S. E. Eyimaya, "A combined energy management algorithm for wind turbine/battery hybrid system," *Journal of Electronic Materials*, vol. 47, no. 8, pp. 4430-4436, Mar. 2018.

Yile Wu received the B.S. degree in electrical engineering from Changchun University of Technology, Changchun, China, in 2019. He is currently pursuing the M.S. degree in electrical engineering in Nanjing Institute of Technology, Nanjing, China. His research interests include application of artificial intelligence in power and energy system.

Le Ge received the B. S. and M. S. degrees in electrical engineering from Southeast University, Nanjing, China, in 2002 and 2008, respectively, the Ph.D. degree in control science and engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. He is currently working as an Associate Professor at the School of Electric Power Engineering, Nanjing Institute of Technology, Nanjing, China. His research interests include smart grid, new energy, and energy storage.

Xiaodong Yuan received the B.S. and M.S. degrees in electrical engineering from Southeast University, Nanjing, China, in 2002 and 2005, respectively. He is currently the Deputy Technology Director at Electric Power Research Institute of State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China, and Convener of IEC TC8 JWG9. His research interests include low-voltage direct current (LVDC), power quality, and energy storage.

Xiangyun Fu received the M.S. degree in electrical engineering from Northeast Electric Power University, Jilin, China, in 2003, and the Ph. D. degree in electrical engineering from Harbin Institute of technology, Harbin, China, in 2007. He is a Professor Level Senior Engineer of State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China. His research interests include planning of power grid and applications of energy storage system.

Mingshen Wang received the B.S., M.S., and Ph.D. degrees in electrical engineering from Tianjin University, Tianjin, China, in 2013, 2016, and 2020, respectively. From 2017 to 2019, he was a joint Ph.D. student with the University of Tennessee, Knoxville, USA. He is currently an Engineer at Electric Power Research Institute of State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China. His research interests include modeling, operation, and control of demand-side resources.