A Three-Stage Optimal Operation Strategy of Interconnected Microgrids With Rule-Based Deep Deterministic Policy Gradient Algorithm

Huifeng Zhang, *Member, IEEE*, Dong Yue, *Fellow, IEEE*, Chunxia Dou, *Senior Member, IEEE*, and Gerhard P. Hancke, *Life Fellow, IEEE*

***Abstract*—The ever-increasing requirements of demand response dynamics, competition among different stakeholders, and information privacy protection intensify the challenge of the optimal operation of microgrids. To tackle the above problems, this article proposes a three-stage optimization strategy with a deep reinforcement learning (DRL)-based distributed privacy optimization. In the upper layer of the model, the rule-based deep deterministic policy gradient (DDPG) algorithm is proposed to optimize the load migration problem with demand response, which enhances dynamic characteristics with the interaction between electricity prices and consumer behavior. Due to the competition among different stakeholders and the information privacy requirement in the middle layer of the model, a potential game-based distributed privacy optimization algorithm is improved to seek Nash equilibriums (NEs) with encoded exchange information by a distributed privacy-preserving optimization algorithm, which can ensure the convergence as well as protect privacy information of each stakeholder. In the lower layer of the model of each stakeholder, economic cost and emission rate are both taken as operation objectives, and a gradient descent-based multiobjective optimization method is employed to approach this objective. The simulation results confirm that the proposed three-stage optimization strategy can be a viable and efficient way for the optimal operation of microgrids.**

***Index Terms*—Energy management, optimal operation, potential game, privacy information, stakeholders.**

I. Introduction

A. Motivation and Incitement

THE increasing renewable energy resources and power-supply quality requirements have introduced significant changes to interconnected microgrids. These have become dynamic complex systems with multiple stakeholders' economic, reliability, and information protection requirements [1], [2], which can be a major challenge for the optimal operation of interconnected microgrids. As different stakeholders of microgrids enter the electricity market, they compete with one another to seek economic benefit for themselves, which results in a bargaining situation with multiple participants [3]–[9], which attracts emerging researches on the optimal operation of microgrids with different stakeholders.

B. Literature Review

Wu *et al.* [3] investigated the energy scheduling of energy consumers and sellers to maximize their benefit in response to two types of local trading centers (LTCs), which include nonprofit- and profit-oriented LTCs. In [4], each distributed generator is seen as a player, and a distributed locational marginal pricing-based unified energy management system model is proposed to solve the loss reduction problem with the Shapley value method of game theory. Dou *et al.* [6] designed an agent with a decision-making process of price bidding strategies, where power market participants pursue maximum profit by bidding day-ahead electricity price to achieve Nash equilibriums (NEs). Mediwaththe *et al.* [8] propose a decentralized approach of a dynamic noncooperative repeated game with Pareto-efficient pure strategies to determine day-ahead optimal energy trading scheduling. Du *et al.* [9] utilize potential games to solve economic power dispatch problems with practical operation constraints, where each generator is taken as an independent player. However, those game-based approaches are centralized/decentralized and the defined player lacks information protection.

Since it exists some unknown or model-free parts in power system operations, deep reinforcement learning (DRL) has widely been used due to its knowledge learning mechanism, which can be a data-driven feedback optimization approach [10]–[16]. Literature [10] learns a map from states to optimal actions of wind energy conversion systems with a model-free Q-learning algorithm, which keeps maximum power point tracking of wind energy resources. According to paper [12], the improved reinforcement learning algorithm

Manuscript received 27 January 2022; revised 22 May 2022; accepted 19 June 2022. This work was supported in part by the National Natural Science Fund under Grant 61973171, in part by the Basic Research Project of Leading Technology of Jiangsu Province under Grant BK20202011, in part by the National Natural Science Fund of Jiangsu Province under Grant BK20211276, and in part by the National Natural Science Key Fund under Grant 61833008. *(Corresponding author: Dong Yue.)*

Huifeng Zhang, Dong Yue, and Chunxia Dou are with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: zhanghuifeng_520@163.com; medongy@vip.163.com; cxdou@ysu.edu.cn).

Gerhard P. Hancke is with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, and also with the University of Pretoria, Pretoria 0028, South Africa (e-mail: g.hancke@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3185211.

Digital Object Identifier 10.1109/TNNLS.2022.3185211

can achieve an optimal scheme of generation resources, distributed storage, and customers without prior information about the microgrid system. Huang *et al.* [14] propose a double-Q learning-based power management approach to scale operating frequency, which reduces the overestimation and enhances prediction accuracy. In [15], a distributed system operator learns the multimicrogrid response with deep neural networks without direct access to the user's information, which decreases the demand-side peak-to-average ratio and maximizes the profit from selling energy. However, those existing DRL methods may suffer great computational complexity especially when decision variables are high-dimensional. Here, this article proposes a rule-based DRL approach to tackle this problem.

C. Contribution and Article Organization

Considering the complexity issue, a multiple-stage optimization strategy can be a good choice. Literature [17] proposes a hierarchical distributed energy management of microgrids with energy routing, which can provide good dynamic performance. In [18], it focuses on the energy management of autonomous microgrids, and a three-level hierarchical coordination strategy is proposed to tackle this problem. However, the existing literature focuses merely on one or two optimal operation problems of microgrids, which lacks a systematic viewpoint to solve the optimization problem from both the load demand side and power generation side. In this article, a three-stage optimization strategy is proposed with a reinforcement learning-based potential game approach to systematically tackle with optimal operation of interconnected microgrids with different stakeholders. The main contribution of the proposed strategy can be summarized as follows.

1) Considering the unknown process of consumers' behavior on the demand side in the upper-layer model, a deep deterministic policy gradient (DDPG) approach is developed to learn the load-price function with three proposed rules to achieve an optimal load-shifting scheme under price-incentive-based demand response, which reduces computational complexity to improve learning efficiency.
2) Due to privacy protection requirements of different stakeholders in the middle layer model, a potential game-based distributed privacy optimization is improved to deal with competition issues of stakeholders, the exchange information is coded with designed noise, which can ensure the safety of private information exchange as well as optimization convergence performance.
3) In the lower layer of the model, a gradient descent-based multiobjective cultural differential evolution (GD-MOCDE) algorithm is employed to optimize economic cost and emission rate simultaneously with a two-step embedded constraint handling technique, which can generate a set of Pareto-optimal solutions for assisting power operator's decision making.

This article structure is structured as follows. The three-layered model is presented in Section II, the proposed method is described in Section III, and the simulation results are shown in Section IV.

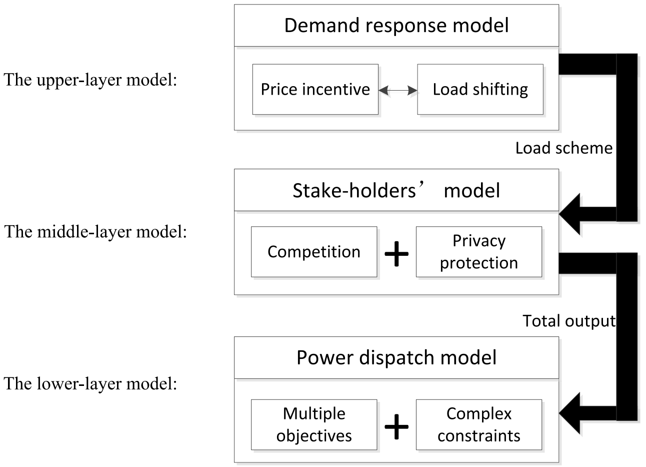

Fig. 1. Structure diagram of a three-layered optimal operation model.

II. Three-Layered Optimal Operation Model of Interconnected Microgrids

The aim of this article is to solve the optimal operation problem of interconnected microgrids, which consists of power generation and load demand in each microgrid. The optimal operation of microgrids mainly consists of three issues: load management under the electricity market, coordinated optimization of stakeholders' microgrids, and power dispatch of the inner microgrid. To tackle these problems, load demand is first considered before power generation since all the power generations of microgrids must meet the load demand; here the price incentive-based demand response is taken into consideration. On the basis of load demand, the power generation scheme of both the microgrid system and each microgrid must be obtained. Hence, a three-layer model is created as an upper-layer model with price incentive-based demand response, a middle-layer model with microgrids with different stakeholders, and a lower-layer model with power dispatch within each microgrid, the structure of the three-layered optimal operation model is shown in Fig. 1. In the upper layer of the model, consumers can adjust their load consumption scheme as electricity price dynamically changes. In the middle layer of the model, it also exists microgrid owners (or stakeholders) with privacy protection requirements, and these stakeholders compete with one another to seek maximum profit. In the lower layer of the model, economic cost and emission issues can be two major objectives of power dispatch within each microgrid.

A. Upper Layer of Demand Response With Load Shifting

As electricity prices can dynamically change, the consumers of each microgrid can make an electricity consumption scheme to maximize load-shifting benefits. Consumers in each microgrid can adjust their consumption behavior, and load shifting occurs among those consumers under incentive-based electricity prices in interconnected microgrids. The system load L_{q,m,t_i} is classified into fixed load $\overline{L_{q,m,t_i}}$ and controllable load $\widetilde{L_{q,m,t_i}}$, then the load shifting model can be formulated as follows:

$$\begin{cases} \max\ C_1 = \sum_{q=1}^{N_G}\sum_{m=1}^{N_q}\sum_{t_i\in T}\sum_{t_j\in T}(\gamma_{t_i}-\gamma_{t_j})L_{q,m,t_i,t_j} \\ L_{q,m,t_i,\min} \le L_{q.m,t_i,t_j} \le L_{q,m,t_i,\max} \\ L_{q,m,\min,t_j} \le L_{q,m,t_i,t_j} \le L_{q,m,\max,t_j} \\ \sum_{t_i\in T} L_{q,m,t_i} = M_{q,m} \\ L_{q,m,t_i} = \overline{L_{q,m,t_i}} + \widetilde{L_{q,m,t_i}} \\ \widetilde{L_{q,m,t_i}} = \sum_{j\in T} L_{q,m,t_j,t_i} - \sum_{j\in T} L_{q,m,t_i,t_j} \ge 0 \\ \gamma_{t_i} = g\left(\sum_{q\in N_G}\sum_{m=1}^{N_q} L_{q,m,t_i}\right) \end{cases} \tag{1}$$

where N_G and q represent the microgrid number and the microgrid index, respectively, N_q is the number of consumers at the qth microgrid, m is the consumer index, t_i and T denote the time period and total time length, respectively, γ_{t_i} and L_{q,m,t_i,t_j} represent the electricity price and the consumer's load migrating from time period t_i to t_j, respectively, $L_{q,m,t_i,\min}$ and $L_{q,m,t_i,\max}$ represent the minimum and maximum emigration load, respectively, $L_{q,m,\min,t_j}$ and $L_{q,m,\max,t_j}$ are the minimum and maximum immigration load, respectively, $M_{q,m}$ is a certain real number, and $g(\cdot)$ denotes the differentiable monotonic decreasing function, which describes the relationship between total load demand and electricity price.

B. Middle Layer of Multiple Stakeholders With a Privacy Issue

On the basis of obtained load demand, all stakeholders will compete to satisfy load demand requirements to gain maximum profit or minimum economic cost as follows:

$$\begin{cases} \min\ C_2 = \sum_{q\in N_G} f_q \\ f_q = B_{q2}P_q^2 + B_{q1}P_q + B_{q0} \end{cases} \tag{2}$$

where f_q represents the stakeholder's economic cost, each stakeholder competes with others to minimize this cost function. B_{q2}, B_{q1}, and B_{q0} are the coefficients of economic cost. And all stakeholders must assign the output to their power generators for satisfying system load on the demand side as follows:

$$\sum_{q\in N_G} P_q(t) = \sum_{q\in N_G} L_{q,t} \tag{3}$$

where $P_q(t)$ denotes the total output of the qth stakeholder. Due to the competition characteristics, the communications among stakeholders and their neighbors must protect their private information. Simultaneously, that output also has some constraint limits as follows:

$$\begin{cases} P_{q,\min} \le P_q(t) \le P_{q,\max} \\ \mathrm{Ram}_{\mathrm{dowm},q} \le P_q(t) - P_q(t-1) \le \mathrm{Ram}_{\mathrm{up},q} \end{cases} \tag{4}$$

where $P_{q,\min}$ and $P_{q,\max}$ represent the minimum and maximum output limits, respectively, and $\mathrm{Ram}_{\mathrm{dowm},q}$ and $\mathrm{Ram}_{\mathrm{up},q}$ denote the ramp down and ramp up limits, respectively. Simultaneously, the power flow constraint can also be taken into consideration as follows:

$$\begin{aligned} h(U_q, P_{n,q}, Q_{n,q}) &= (U_q(t))^2 - (U_n(t))^2 \\ &\quad + 2(R_{n,q}P_{n,q}(t) + X_{n,q}Q_{n,q}(t)) \\ &\quad - [(R_{n,q})^2 + (X_{n,q})^2]\frac{(P_{n,q}(t))^2 + (Q_{n,q}(t))^2}{(U_n(t))^2} = 0 \end{aligned} \tag{5}$$

where $h(\cdot)$ represents the nonlinear function, $P_{n,q}(t)$ and $Q_{n,q}(t)$ represent active power and reactive power flowing from microgrid n to microgrid q, respectively, which satisfies that $P_q(t) - L_q(t) = \sum_{n\in\Xi_{q,t}} P_{n,q}(t) - \sum_{m\in\Xi'_{q,t}} P_{q,m}(t)$. It is also the same between reactive power $Q_q(t)$ of the qth microgrid node and $Q_{n,q}(t)$. $\Xi_{q,t}$ represents the set of microgrids' power flow to the qth microgrid, and power flow from the qth microgrid to microgrid set $\Xi'_{q,t}$. For simplicity, active/reactive power cannot flow from other microgrids to the qth microgrid while active/reactive power of the qth microgrid flows simultaneously to other microgrids, which also means that if $P_{i,q}(t) \ne 0 (i \in \Xi_{q,t})$, then $P_{q,j}(t) = 0 (j \in \Xi'_{q,t})$, and vice versa. $R_{n,q}$ and $X_{n,q}$ denote the resistance and reactance between the nth microgrid and the qth microgrid, respectively, $U_q(t)$ is the voltage of the qth microgrid, and those following constraint limits should also be satisfied:

$$\begin{cases} U_q^{\min} \le U_q(t) \le U_q^{\max} \\ Q_q(t) = P_q(t)\tan\varphi_q \\ Q_q^{\min} \le Q_q(t) \le Q_q^{\max} \end{cases} \tag{6}$$

where $U_n^{\min}$ and $U_n^{\max}$ represent the minimum and maximum voltage limits, respectively, and $Q_q^{\min}$ and $Q_q^{\max}$ denote the minimum and maximum reactive power limits, respectively, and φ_q is the power factor angle.

C. Lower Layer of Power Dispatch Within Each Stakeholder's Power Generation System

Each stakeholder owns an independent power generation system, which consists of stable power generators and intermittent energy resources energy storage units in the microgrid. On the basis of the middle layer model, the obtained total output can further guide the power generation in the lower layer, and the following power balance must be satisfied:

$$P_q = \sum_{i\in N_c} P_{ci} + \sum_{j\in N_I} P_{Ij} + \sum_{k\in N_e} U_{ek}P_{ek} \tag{7}$$

where N_c, N_I, and N_e are the number of stable power generators, intermittent energy resources, and energy storages, respectively, U_{ek} denotes the ON/OFF state of energy storage, and P_{ck}, P_{Ij}, and P_{ek} represent the output of stable power, intermittent energy, and energy storage, respectively. In this independent system, the main goal can be achieved by minimizing economic cost and emission rate simultaneously, the economic cost can be represented as

$$\begin{aligned}\min C_3 = \sum_{i\in N_c}&(\alpha_{1i}P_{ci}^2 + \alpha_{2i}P_{ci} + \alpha_{3i}\\ &+ |\alpha_{4i}\sin(\alpha_{5i}(P_{ci,\min} - P_{ci}))|)\\ &+ \sum_{k\in N_e}|U_{ek}(t) - U_{ek}(t-1)|\gamma_{e,k}\end{aligned} \tag{8}$$

where α_{1i}, α_{2i}, α_{3i}, α_{4i}, and α_{5i} are the coefficients of stable power generation cost, $P_{ci,\min}$ represents the minimum output limit, and γ_{ek} denotes the switching price of energy storage. The emission rate can also be expressed as follows:

$$\min C_4 = \sum_{i\in N_c}\left(\beta_{1i}P_{ci}^2 + \beta_{2i}P_{ci} + \beta_{3i} + \beta_{4i}e^{\beta_{5i}P_{ci}}\right) \tag{9}$$

where β_{1i}, β_{2i}, β_{3i}, β_{4i}, and β_{5i} are the coefficients of emission rate. The total output P_q can also be classified into two parts: stable output $\overline{P_q}$ and uncertain output $\widetilde{P_q}$. Obviously, it satisfies $\overline{P_q} = \sum_{i\in N_c}P_{ci} + \sum_{k\in N_e}U_{ek}P_{ek} + \sum_{j\in N_I}\overline{P_{Ij}}$ and $\widetilde{P_q} = \sum_{j\in N_I}\widetilde{P_{Ij}}$, where $\overline{P_{Ij}}$ and $\widetilde{P_{Ij}}$ represent the stable part and uncertain part of intermittent energy resources, respectively. Those stable power can be controlled without uncertainty, and intermittent energy resources cannot be controlled while causing uncertainty to the system. The stable power required to satisfy some constraints is as follows.

1) *Output limits:* The stable output can be adjusted within the minimum and maximum limits as well as the ramp-up and ramp-down limits, which can be expressed as follows:

$$\begin{cases}P_{ci,\min} \le P_{ci} \le P_{ci,\max}\\ \mathrm{Ram}_{\mathrm{down},i} \le P_{ci}(t) - P_{ci}(t-1) \le \mathrm{Ram}_{\mathrm{up},i}\end{cases} \tag{10}$$

where $P_{ci,\max}$ represents the maximum output of stable power, and $\mathrm{Ram}_{\mathrm{down},i}$ and $\mathrm{Ram}_{\mathrm{up},i}$ denote the ramp down and ramp up limits, respectively.

2) *Minimum* ON/OFF *time constraints:* For protecting the power generator devices, the stable power generator also requires to obey the ON/OFF time limits as follows:

$$\begin{cases}\left(T_{ci,t-1}^{\mathrm{ON}} - T_{ci,\min}^{\mathrm{ON}}\right)(\tau_{ci,t-1} - \tau_{ci,t}) \ge 0\\ \left(T_{ci,t-1}^{\mathrm{OFF}} - T_{ci,\min}^{\mathrm{OFF}}\right)(\tau_{ci,t} - \tau_{ci,t-1}) \ge 0\end{cases} \tag{11}$$

where $T_{ci,t-1}^{\mathrm{ON}}$ and $T_{ci,t-1}^{\mathrm{OFF}}$ represent the continuous online and offline time of power generator i until period $t-1$, $T_{ci,\min}^{\mathrm{ON}}$ and $T_{ci,\min}^{\mathrm{OFF}}$ denote the minimum online and offline time, respectively, and $\tau_{ci,t}$ is the binary decision variable for online state of power generator.

3) *Charging/discharging limits:* The energy storage can be a supplementary energy resource by charging or discharging to keep the stability of the independent system, while the charging and discharging processes must satisfy the following limits:

$$\begin{cases}V_k^{\mathrm{store}}(t+1) = V_k^{\mathrm{store}}(t) + P_{ek}(t) * \Delta T\\ P_{ek} = \eta_l P_{ek}^{\mathrm{store}}\\ V_{k,\min}^{\mathrm{store}} \le V_k^{\mathrm{store}} \le V_{k,\max}^{\mathrm{store}}\\ P_{ek}^{\mathrm{store}} = P_{ek}^{\mathrm{cha}}, \quad \text{if } P_{ek}^{\mathrm{store}} \ge 0\\ P_{ek}^{\mathrm{store}} = -P_{ek}^{\mathrm{dis}}, \quad \text{if } P_{ek}^{\mathrm{store}} < 0\\ 0 \le P_{ek}^{\mathrm{dis}} \le P_{ek,\max}^{\mathrm{dis}}\\ 0 \le P_{ek}^{\mathrm{cha}} \le P_{ek,\max}^{\mathrm{cha}}\\ V_{ek}^{\mathrm{store}}(0) = V_{ek,\mathrm{initial}}^{\mathrm{store}}\end{cases} \tag{12}$$

where V_k^{store} is the storage of the kth energy storage, ΔT is the time period length, $V_{k,\min}^{\mathrm{store}}$ and $V_{k,\max}^{\mathrm{store}}$ are the minimum and maximum storage of the kth energy storage, respectively, P_{ek}^{store} denotes the power discharge/charging of kth energy storage, P_{ek}^{dis}, and P_{ek}^{cha} are the output of discharging and charging state, $P_{ek,\max}^{\mathrm{dis}}$ and $P_{ek,\max}^{\mathrm{cha}}$ are the maximum discharging and charging output at $k \in N_e$th energy storage, respectively, and $\eta_k \in (0, 1]$ is the efficiency factor of charging or discharging state.

4) *Spinning reserve constraint:* For ensuring the safety of the power system, additional stable power is required for avoiding potential risk, which means it needs to satisfy the following constraint:

$$\begin{aligned}\sum_{i\in N_c}\left(P_{ci,\max} - P_{ci}\right) + \sum_{k\in N_{eu}}\left(P_{ek,\max}^{\mathrm{dis}} - P_{ek}^{\mathrm{dis}}\right)\\ + \sum_{k\notin N_{eu}}P_{ek,\max}^{\mathrm{dis}} \ge \sum_{j\in N_I}r_j\left(\widetilde{P_{Ij,\max}} - \widetilde{P_{Ij,\min}}\right)\end{aligned} \tag{13}$$

where the set N_{eu} can be expressed as $\{k|k \in N_e$ AND $U_{ek} = 1\}$, r_j denotes controllable parameter of uncertainty degree, and $\widetilde{P_{Ij,\max}}$ and $\widetilde{P_{Ij,\min}}$ represent the maximum and minimum limits of the uncertain part of the intermittent energy resources, respectively.

III. Proposed Optimization Strategy for Three-Layer Optimal Operation of an Isolated Power System

A. Rule-Based Deep Reinforcement Learning for Load-Shifting-Based Demand Response in the Upper Layer Model

With consideration of load consumers' intelligent activity, system load on the demand side can be dynamic and further affects the output process on the power generation side. Once electricity price has been determined, load consumers can dynamically change their load consuming scheme by moving controllable loads with the high cost to those with low cost. For properly dealing with this dynamic optimal operation problem, it adopts a DRL algorithm to predict the upcoming load according to the historical consumers' activities and makes the best shifting scheme for the lowest economic cost. The load scheme can be taken as a Markov decision process (MDP), which contains the following elements.

1) *State Set:* The state set of system load can be expressed as $V = \{V_1, V_2, \ldots, V_T\}$, and its arbitrary element

$V(t) \triangleq \sum_{q \in N_G} \sum_{m=1}^{N_q} L_{q,m,t}$, which represents the state element.

2) *Action set:* The action set can be presented as $A(t) = \{A_1(t), A_2(t), \ldots, A_{N_G}(t)\}$, and the action $A_q(t)$ of each system load at each time slot can be defined as $A_q(t) \triangleq \{\{L_{q,m,t,1}\}_{m \in N_q}, \{L_{q,m,t,2}\}_{m \in N_q}, \ldots, \{L_{q,m,t,T}\}_{m \in N_q}\}$, where $L_{q,m,t,t} = 0$ for simplicity.
3) *Transition model:* $T(V', A, V) \sim Prob(V'|V, A)$ represents the transition model from the current state V to the next state V' after the current action set A, where $Prob(\cdot)$ denotes the transition probability.
4) *Reward set:* The reward $R(V(t), A(t)) = E(R_{t+1}|V(t), A(t))$ can be provided after the current action $A(t)$, and $E(\cdot)$ denotes the expected value. Since constraint limits must be satisfied, the reward $R(V(t), A(t))$ should also include the penalty term or negative term. Here, its positive term can be expressed as $E(C_1)$, and the negative part can be described as the total constraint violation $\xi_R E(\text{Vio}_{\text{con}})$, where ξ_R is the positive penalty factor, and then the reward $R(V(t), A(t))$ can be defined as $E(C_1 - \xi_R \text{Vio}_{\text{con}})$.
5) *Optimal state-value function:* The optimal state-value function $Q^*(V, A)$ represents the optimal value after all policies, which can also be expressed as $\max_A Q_A(V, A)$. Combined with the Bellman theory, the optimal value of the next state $Q^*(V', A')$ can be achieved with the following iteration:

$$\begin{aligned} &Q^*(V', A') \\ &= \max_{A'} Q_{A'}(V', A') \\ &= \max_{A'(t)} \left[R(V', A') + \xi_Q \sum_{A} \text{Prob}(V'|V, A) \max_{A} Q_A(V, A) \right] \end{aligned} \tag{14}$$

where $\xi_Q \in (0, 1]$ represents the discount factor. For ensuring the feasibility of constraint limits, the total constraint violation Vio_{con} can be expressed as follows:

$$\begin{aligned} &\text{Vio}_{\text{con}} \\ &= \sum_{t_i \in T} \sum_{t_j \in T, j \neq i} \sum_{m=1}^{N_q} [\max(L_{q,m,t_i,\min} - L_{q,m,t_i,t_j}, 0) \\ &\qquad + \max(L_{q,m,t_i,t_j} - L_{q,m,t_i,\max}, 0)] \\ &\quad + \sum_{t_j \in T} \sum_{t_i \in T, i \neq j} \sum_{m=1}^{N_q} [\max(L_{q,m,\min,t_j} - L_{q,m,t_i,t_j}, 0) \\ &\qquad + \max(L_{q,m,t_i,t_j} - L_{q,m,\max,t_j}, 0)] \\ &\quad + |\sum_{t_i \in T} \sum_{m=1}^{N_q} L_{q,m,t_i} - M_{q,m}| \\ &\quad + \sum_{t_i \in T} \sum_{m=1}^{N_q} \max(-\widetilde{L_{q,m,t_i}}, 0). \end{aligned} \tag{15}$$

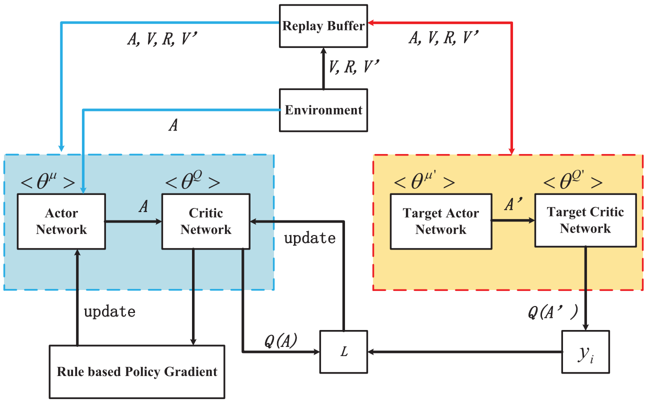

Fig. 2. Structure of rule-based DDPG in the load migration model.

Once the violation Vio_{con} is smaller than the permitted deviation ϵ_{tot}, the feasibility can be assured. During the iteration process, if the state V exceeds the permitted bounds, then force it to the nearest bound [19]. With the consideration of the requirement of the transition probability information, the expected value can be difficult to obtain, deep Q-learning methods can approximate the transition process by training historical data. On the basis of [20], the off-policies learning named DDPG of continuous action is improved to deal with the above problem. Four neural networks including the critic network, the actor network, and two target networks are employed to enhance the learning efficiency, the structure has been presented in Fig. 2. In each iteration, the transition is sampled from $Environment$ to be stored in $ReplayBuffer$ with V, R, V' and generated A, which are taken as input to two networks and their "copy" target networks, which mainly trains optimal action at a certain state. With consideration of the state vector and action vector's high-dimensional issue. After training on actor-critic network with weights θ^μ and θ^Q, where $\mu = \text{argmax}_A Q(V, A)$. Simultaneously, two target networks are copied to calculate the target value $y_i = R(V^{\{i\}}, A^{\{i\}}) + \xi_Q Q(V'^{\{i\}}, A'^{\{i\}})$ ($[\cdot]^{\{i\}}$ denotes the ith sample), then it can deduce the online approximation loss $L = 1/N \sum_{i \in N} (Q(V^{\{i\}}, A^{\{i\}}) - y_i)^2$ (N denotes the number of samples), the actor policy is updated with the sampled policy gradient as follows:

$$\nabla J = 1/N \sum_{i \in N} \nabla_A Q|_{V=V^{\{i\}}, A=A^{\{i\}}} \nabla_{\theta^\mu} \mu(V)|_{V=V^{\{i\}}}. \tag{16}$$

Due to the dynamic time-related load-shifting scheme, state and action variables can be high-dimensional, which can bring high computational complexity for obtaining an optimal load-shifting strategy. Hence, this article proposes a rule-based reinforcement learning approach to tackle this problem. For reducing optimization complexity, three rules are proposed in policy gradient for leading the rapid search to the optimal scheme as follows:

1) *Rule 1:* The load shifting occurs from time period $t_i = \arg\max_{t \in T} \gamma_t$ with highest electricity price γ_{t_i} to the period $t_j = \arg\min_{t \in T} \gamma_t$ with lowest electricity price γ_{t_j}.
2) *Rule 2:* For arbitrary consumer m, load emigration and load immigration cannot occur simultaneously, which means that $L_{q,m,t_i,t_j} = 0$ if $L_{q,m,t_j,t_i} > 0$, and vice versa.

3) *Rule 3:* Since load shifting can dynamic change electricity price, it must stop when L_{q,m,t_j,t_i} satisfies $g(\sum_{q=1}^{N_G}\sum_{m=1}^{N_q}(\sum_{t_j \in T} L_{q,m,t_j,t_i} + \overline{L_{q,m,t_i}})) \le \min_{t_j \in T} \gamma_{t_j}$.

Algorithm 1 Rule Based DDPG for Load Shifting

```
procedure Rule based DDPG for load shifting
    Initialization: Critic network $Q$ and weights $\theta^Q$, actor network $\mu$ and weights $\theta^\mu$, $ReplayBuffer = \{\phi\}$, two target network $Q'$, $\mu'$ and their weights $\theta^{Q'}$, $\theta^{\mu'}$, initial state $V$ and action $A$, $k = 0$;
    while $Episode < maxcount$ do
        Execute action $A$ and observe $R$ and $V'$;
        Select minibatch from $N$ transition samples;
        Update critic network by minimizing loss $L$;
        while $k < maxcount1$ or $g(V^{\{i\}} - A^{\{i\}^{(k)}}) \le g(V^{\{i\}} + A^{\{i\}^{(k)}})$ do
            For $t_j = 1 : T$
            Find current minimum price $\min_{t_j \in T} \gamma_{t_j}$;
            Training with Rule 1 and Rule 2;
            Update actor network with rule based policy gradient $A^{\{i\}^{(k+1)}} = A^{\{i\}^{(k)}} + \eta_{A^{\{i\}}}\delta^{(k)}$;
            Update network weights $\theta^{\mu'}$ and $\theta^{Q'}$;
            end
            $k = k + 1$;
        end while
        Store $(V, A, R, V')$ in $ReplayBuffer$;
        $Episode = Episode + 1$;
    end while
end procedure
```

Combined with above three rules, the solution strategy can be improved as: For a given state vector V, check electricity price γ_t of each time period, suppose current peak price and current valley price are $\gamma_{\max}$ and $\gamma_{\min}$, and its corresponding periods are $t_{\max}$ and $t_{\min}$, system load of each consumer can shift from $t_{\max}$ period to $t_{\min}$ period, which means that $L_{q,m,t_{\min},t_j} = 0$ ($t_j \in T$ and $t_j \neq t_{\min}$). Since electricity price $\gamma_{t_{\min}}$ can rise as load shifting into $t_{\min}$ period and electricity price $\gamma_{t_{\max}}$ can decrease with load emigration to other periods. With consideration of neuron function (sigmoid function) $\mu(\cdot)$, the proposed iteration for actor policy gradient can be taken as

$$\begin{cases} \theta^{\mu(k+1)} = \theta^{\mu(k)} - \eta_\theta \delta_\theta^{(k)} \\ A^{\{i\}^{(k+1)}} = \mu(\theta^{\mu(k+1)}) \end{cases} \tag{17}$$

where $\theta^{\mu(k)}$ and $A^{\{i\}^{(k)}}$ represent the actor network weight and action vector at the kth iteration, η_θ denotes the control parameter, and $\delta_\theta^{(k)}$ represents an adaptive factor. The factor $\delta_\theta^{(k)}$ must adaptively adjust action vector with consideration of dynamic change of electricity price $\gamma_{t_{\min}}$ and $\gamma_{t_{\max}}$, then it can obtain

$$\begin{cases} \delta_\theta^{(k)} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mu(\psi_\mu^{(k)})(1 - \mu(\psi_\mu^{(k)}))V^{\{i\}^{(k+1)^T}} \\ \psi_\mu^{(k)} = \theta^{\mu(k)^T} V^{\{i\}^{(k+1)}} + b \end{cases} \tag{18}$$

where $\psi_\mu^{(k)}$ denotes the input vector of neuron function and b is the constant valve value. The above iteration can stop when $g(V^{\{i\}} - A^{\{i\}^{(k)}}) \le g(V^{\{i\}} + A^{\{i\}^{(k)}})$, or maximum count number $maxcount1$ is achieved. In the target network, $\tau \in [0, 1)$ denotes the updating parameter, and $\theta^{\mu'}$ and $\theta^{Q'}$ represent the weights of actor network and critic network, respectively, which can be updated as follows:

$$\begin{cases} \theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1 - \tau)\theta^{\mu'} \\ \theta^{Q'} \leftarrow \tau\theta^{Q} + (1 - \tau)\theta^{Q'}. \end{cases} \tag{19}$$

Then, the obtained information (V, A, R, V') can be stored in $ReplayBuffer$ and enhance the training efficiency for the next round. The algorithm flowchart can be implemented as it is shown in **Algorithm 1**. According to the above rule-based learning procedures, the computational complexity can be greatly reduced. The dimension of load shifting instant can be decreased from $T - 1$ to 1 with **Rule 1**, dimension variables of load shifting can be reduced from $T \times (T - 1)$ to $(T - 1)$ with **Rule 2**, and it can also be reduced further with **Rule 3** to enhance learning efficiency.

B. State-Based Potential Game With Distributed Optimization for the Middle-Layer Model

After load-shifting procedures, total system load $\sum_{q \in N_G} L_{q,t}$ can be properly deduced, and the next task is to balance system load with cooperating different stakeholders. Here, it is assumed that each stakeholder owns one microgrid, and stakeholder seeks their own maximum profit/minimum cost, which generates competition among these stakeholders, while each stakeholder can protect their own privacy, so the coordination of different stakeholders with privacy issue can be the challenging issue. With consideration of the above issues, a distributed potential game with privacy issues is proposed to solve the middle-level problem. The Lagrangian function can be expressed as follows:

$$\begin{aligned} L(P_q) = &\sum_{q \in N_G} \left(B_{q2}P_q^2 + B_{q1}P_q + B_{q0}\right) \\ &+ \lambda_1 \left(\sum_{q \in N_G} P_q(t) - \sum_{s \in N_G} L_{q,t}\right) \\ &+ \lambda_{2,q}^{+}\left(P_{q,\min} + d_{q2}^{+} - P_q\right) + \lambda_{2,q}^{-}\left(P_q + d_{q2}^{-} - P_{\max}\right) \\ &+ \lambda_{3,q}^{+}\left(\mathrm{Ram}_{\mathrm{down},q} + d_{q3}^{+} - P_q(t) + P_q(t-1)\right) \\ &+ \lambda_{3,q}^{-}\left(P_q(t) - P_q(t-1) + d_{q3}^{-} - \mathrm{Ram}_{\mathrm{up},q}\right) \\ &+ \lambda_{4,q}h(U_q, P_{n,q}, Q_{n,q}) + \lambda_{5,q}^{+}\left(U_q^{\min} - U_q(t) + d_{q5}^{+}\right) \\ &+ \lambda_{5,q}^{-}\left(U_q(t) + d_{q5}^{-} - U_q^{\max}\right) \\ &+ \lambda_{6,q}\left(Q_q(t) - P_q(t)\tan\varphi_q\right) \\ &+ \lambda_{7,q}^{+}\left(Q_q^{\min} + d_{q7}^{+} - Q_q(t)\right) \\ &+ \lambda_{7,q}^{-}\left(Q_q(t) + d_{q7}^{-} - Q_q^{\max}\right) \end{aligned} \tag{20}$$

where λ_1, $\lambda_{2,q}^{+}$, $\lambda_{2,q}^{-}$, $\lambda_{3,q}^{+}$, $\lambda_{3,q}^{-}$, $\lambda_{4,q}$, $\lambda_{5,q}^{+}$, $\lambda_{5,q}^{-}$, $\lambda_{6,q}$, $\lambda_{7,q}^{+}$, and $\lambda_{7,1}^{-}$ represent the Lagrangian multipliers, d_{q2}^{+}, d_{q2}^{-}, d_{q3}^{+}, d_{q3}^{-}, d_{q5}^{+}, d_{q5}^{-}, d_{q7}^{+}, and $d_{q7}^{-} > 0$ denote the penalty factors. After the reinforcement learning at system load on the demand side, each stakeholder can make a decision to minimize economic cost. With consideration of competition issue, the state-based potential game is utilized to model this problem, the state variable is defined as follows:

$$\boldsymbol{x_q} = \begin{pmatrix} P_q, U_q, Q_q, \{P_{n,q}\}_{n\in\Xi_q}, \{Q_{n,q}\}_{n\in\Xi_q}, d_{q2}^{+} \\ d_{q2}^{-}, d_{q3}^{+}, d_{q3}^{-}, d_{q5}^{+}, d_{q5}^{-}, d_{q7}^{+}, d_{q7}^{-}, \lambda_1, \lambda_{2,q}^{+} \\ \lambda_{2,q}^{-}, \lambda_{3,q}^{+}, \lambda_{3,q}^{-}, \lambda_{4,q}, \lambda_{5,q}^{+}, \lambda_{5,q}^{-}, \lambda_{6,q}, \lambda_{7,q}^{+}, \lambda_{7,1}^{-} \end{pmatrix}. \quad (21)$$

The ensuing state $\hat{\boldsymbol{x_q}}$ can be estimated with obtained optimal state $\boldsymbol{x_q^*}$, and the action variable $\boldsymbol{a_q}$ can be expressed with $\boldsymbol{x_q^*} - \hat{\boldsymbol{x_q}}$, then the Lagrangian function can also be rewritten as $L(\boldsymbol{x_q}, \boldsymbol{a_q})$. The scalar function $\Phi_q(\boldsymbol{x_q}, \boldsymbol{a_q})$ can be defined as follows:

$$\Phi_q(\boldsymbol{x_q}, \boldsymbol{a_q}) = L(\boldsymbol{x_q}, \boldsymbol{a_q}) - L(\boldsymbol{x_q}(\boldsymbol{0}), \boldsymbol{a_q}(\boldsymbol{0})) \quad (22)$$

where $\boldsymbol{x_q}(\boldsymbol{0})$ and $\boldsymbol{a_q}(\boldsymbol{0})$ are the initial state of $\boldsymbol{x_q}$ and $\boldsymbol{a_q}$. The above game model can be solved with distributed optimization approach, the increment cost $\lambda_q = \lambda_{2,q}^{+} - \lambda_1 - \lambda_{2,q}^{-} + \lambda_{3,q}^{+} - \lambda_{3,q}^{-} - \lambda_{6,q}\tan\varphi_q$ can be defined as

$$\lambda_q = 2B_{q2}P_q + B_{q1}. \quad (23)$$

As it is known that coordination optimal solution is mainly achieved by exchanging information between an agent and its neighbors, while stakeholders cannot share true information with their competitors. With consideration of each stakeholder's privacy, the designed noise is added into the coordination process in the distributed optimization algorithm, simultaneously it can also ensure the convergence ability. During the coordination process, it exchanges information with added noise for privacy protection while updating itself [21]. Here, each stakeholder updates its own information with true value $\lambda_q(k)$ and $\xi_q(k)$, where $\xi_q(t)$ denotes the deviation parameter and broadcasts noisy information $\lambda_q^{+}(k)$ and $\xi_q^{+}(k)$ to its neighbors.

1) *Added noise for privacy protection:*

$$\begin{cases} \lambda_q^{+}(k) = \lambda_q(k) + \phi_q(k) \\ \xi_q^{+}(k) = \xi_q(k) + \zeta_q(k) \end{cases} \quad (24)$$

where $\phi_q(k)$ and $\zeta_q(k)$ represent the added noise, which can ensure the privacy of each stakeholder as well as the convergence of distributed optimization.

2) *Update of exchanged information:*

$$\lambda_q(k+1) = \sum_{j\in N_G} \left(\omega_{qq}\lambda_j^{+}(k) + \omega_{qj}\lambda_q(k)\right) + \chi_q\xi_q(k) \quad (25)$$

where ω_{qj} denotes the weights between agent q and agent j. For arbitrary q, it satisfies $\sum_{j=1}^{N_G}\omega_{qj} = 1$, and $\chi_q \in (0, 1)$ represents control parameter.

3) *Update of self-information:*

$$\begin{cases} P_q(k+1) = P_q(k) - \eta_P(k)\dfrac{\partial\Phi_q(\boldsymbol{x_q}, \boldsymbol{a_q})}{\partial P_q}\Big|_{P_q=P_q(k),\lambda_q=\lambda_q(k+1)} \\ \eta_P(k) = \dfrac{h_P}{\sqrt{\sum_{j=1}^{k}\dfrac{\partial\Phi_q(\boldsymbol{x_q}, \boldsymbol{a_q})}{\partial P_q}} + \epsilon_P}\Big|_{P_q=P_q(j),\lambda_q=\lambda_q(j+1)} \\ \xi_q(k+1) = \displaystyle\sum_{j\in N_G}\omega_{qj}\xi_q^{+}(k) + \omega_{qq}\xi_q(k) \\ \qquad\qquad + P_q(k) - P_q(k+1) \end{cases} \quad (26)$$

where $\eta_P(k)$ represents the adaptive control parameter, and $h_P > 0$ and $\epsilon_P > 0$ denote iteration step and magnitude parameter, respectively. The adaptive cumulative control parameter can improve gradient decent optimization efficiency. The above iteration stops when it converges, it stops when it achieves maximum iteration number $Maxcount2$ or it satisfies $|\lambda_q(k) - \lambda_q(k-1)| > \epsilon_q$, where $\epsilon_q > 0$ denotes permitted convergence accuracy. Due to the nonconvex characteristic of power flow limits, it is simplified by setting $X_{n,q} = 0$ and $U_n(t) = 1.0$ p.u. at the tth instant, then it can be considered as a convex quadratic function. With consideration of constraint limits, the iteration process must be taken with feasible domain $\boldsymbol{\Omega}$, it can be forced to its upper bound when the iteration exceeds the upper bound $\overline{\boldsymbol{\Omega}}$, and it is forced to $\underline{\boldsymbol{\Omega}}$ when it exceeds the lower-bound $\underline{\boldsymbol{\Omega}}$. The designed noise ϕ_q and ζ_q can ensure asymptotic convergence of distributed optimization, if it satisfies two conditions: One is that $\xi_q^{+}(0) = \xi_q(0)$ and the following condition:

$$\begin{cases} \displaystyle\sum_{k=0}^{\infty}|\phi_q(k)| \le H \\ \displaystyle\sum_{k=0}^{\infty}|\zeta_q(k)| \le H \end{cases} \quad (27)$$

where $H > 0$ denotes a upper bound of designed noise. The other is that the graph of distributed optimization is strongly connected (since all microgrids are interconnected), and the initial equation $\sum_{q\in N_G}\xi_q(0) = \sum_{q\in N_G}P_q(0)$ is satisfied. The algorithm flowchart of potential game-based distributed optimization is presented in **Algorithm 2**.

Algorithm 2 Potential Game Based Distributed Algorithm

```
procedure Potential game based distributed algorithm
    Initialization: State variable $x_q$, $\lambda_q(0)$, $\xi_q(0)$ and $\epsilon_q$;
    Initialize $\lambda_q(0) = 2B_{q2}P_q(0) + B_{q1}$, $P_q(0) = 0$, $\xi_q(0) = 0$, $q = 1$ and $0 < \chi_q < 1$;
    while $q < N_G$ do
        $k = 0$;
        while $(|\lambda_q(k) - \lambda_q(k-1)| > \epsilon_q) or (k < Maxcount2)$ do
            Add designed noise for privacy;
            Update exchange information;
            Update local information;
            if $P_q(k+1) > \overline{\Omega}$ then
                $P_q(k+1) = \overline{\Omega}$ End;
            end if
            if $P_q(k+1) < \underline{\Omega}$ then
                $P_q(k+1) = \underline{\Omega}$ End;
            end if
            $k = k + 1$;
        end while
        $q = q + 1$;
    end while
end procedure
```

C. Gradient Descent-Based Multiobjective Optimization for the Lower-Layer Model

Once the NE of each stakeholder has been achieved, the remaining problem is the economic emission dispatch within the microgrid system. Since the economic cost and emission rate must be optimized simultaneously, a GD-MOCDE algorithm is employed to take care of this problem. Considering two gradient directions: positive space H^+ and negative space H^-, the element $z \in R^n$ can be expressed as follows:

$$\begin{cases} H^+ = \{z \in R^n | \Delta F z > 0\} \\ H^- = \{z \in R^n | \Delta F z < 0\} \end{cases} \tag{28}$$

where F denotes the objective function vector. Then the deviation between two variables $X_{G+1} - X_G$ can be described as $\Delta F(X_{G+1} - X_G) = \Delta F z$, if the mutation operator adopts the following formation:

$$X^j_{G+1} = X^j_G + \gamma^j_{G,1}\left(X^j_{r2,G} - X^j_{r3,G}\right) + \gamma^j_{G,2}\left(X^j_{r4,G} - X^j_{r5,G}\right) \tag{29}$$

where $X^j_{r2,G}$, $X^j_{r3,G}$, $X^j_{r4,G}$, and $X^j_{r5,G}$ represent the jth variable of individuals in archive set $(X^j_{r2,G} \neq X^j_{r3,G} \neq X^j_{r3,G} \neq X^j_{r4,G} \neq X^j_{r5,G})$. The control parameters $\gamma^j_{G,1}$ and $\gamma^j_{G,2}$ can be updated as follows:

$$\begin{cases} \gamma^j_{G,1} = \dfrac{-\eta_G \nu_1 \operatorname{sgn}\left(F_1\left(X_{r2,G} - F_1(X_{r3,G})\right)\right)}{\left(X^j_{r2,G} - X^j_{r3,G}\right)^2 \sqrt{\sum_{j \in n} \frac{1}{\left(X^j_{r2,G} - X^j_{r3,G}\right)^2}}} \\ \gamma^j_{G,2} = \dfrac{-\eta_G \nu_2 \operatorname{sgn}\left(F_2\left(X_{r4,G} - F_2(X_{r5,G})\right)\right)}{\left(X^j_{r4,G} - X^j_{r5,G}\right)^2 \sqrt{\sum_{j \in n} \frac{1}{\left(X^j_{r4,G} - X^j_{r5,G}\right)^2}}} \\ \eta_G = \eta_0[(G_{\max} - G + 1)/G_{\max}]^p \end{cases} \tag{30}$$

where $\eta_0, \eta_G \in R^+$ represent the scaling parameters, ν_1 and ν_2 are the weighted parameters, sgn(·) denotes the sign function, p is a positive integer, and $G_{\max}$ is the maximum generation. For properly dealing with these constraint limits, the constraint handling technique in [22] is employed.

IV. Case Study

For properly verifying the optimization efficiency, five stakeholders compete to balance the dynamic system load for maximum profit while considering privacy protection. Each stakeholder owns a microgrid, which consists of three traditional generator units, two energy storage, one wind farm, and one consumer system, the related details can be found in [19] and [23]. To verify the learning efficiency, five-consumer systems, ten-consumer systems, and 20-consumer systems are taken for implementing the proposed learning method. The topology of the five stakeholders is a complete graph, stakeholders can exchange information with each other to seek maximum profit, and the information exchange process can be coded with designed noise for privacy protection.

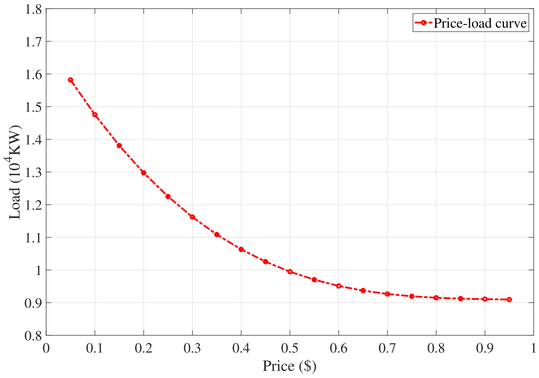

Fig. 3. Dynamic relationship between price and total system load.

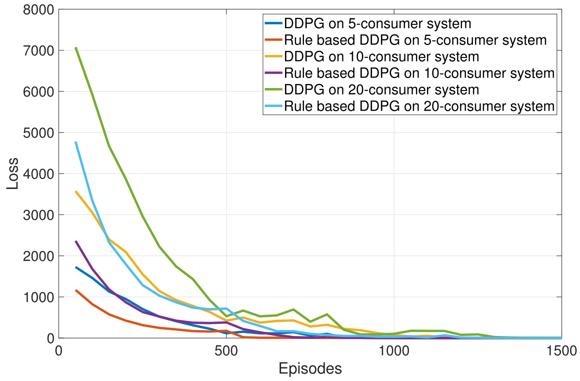

Fig. 4. Training process in comparison with DDPG on different consumer systems.

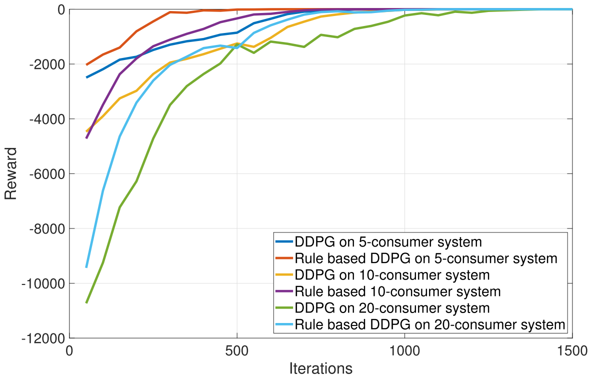

Fig. 5. Reward process in comparison with DDPG on different consumer systems.

A. Upper-Layer Optimization With Rule-Based DDPG Approach

In the electricity market, system load can affect electricity price, which can also change the consumers' behavior and affect the system load in turn. Here, the dynamic relationship between electricity price and system load is presented in Fig. 3, which describes that system load in each interval (1 h) decreases as electricity price increases. For verifying the learning efficiency of different consumer systems, the proposed learning method is implemented on a five-consumer system, ten-consumer system, and 20-consumer system in comparison to traditional DDPG. The DRL training consists of 1500 episodes, and the loss process and reward process have been presented in Figs. 4 and 5. It can be seen in Figs. 4 and 5

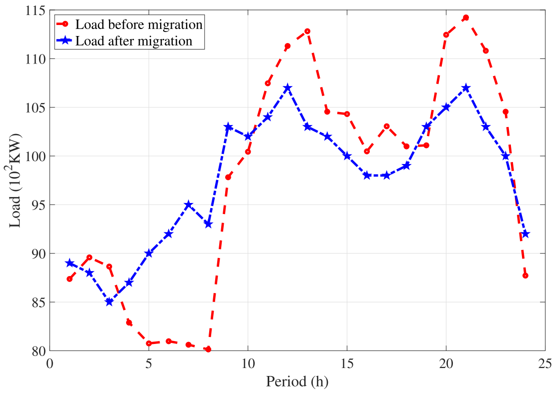

Fig. 6. Load before migration and after migration.

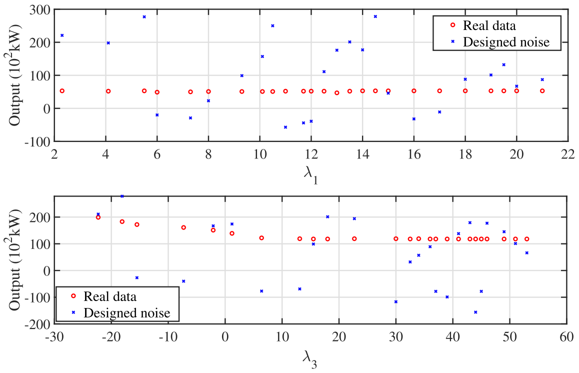

Fig. 7. Real data and coded data with designed noise.

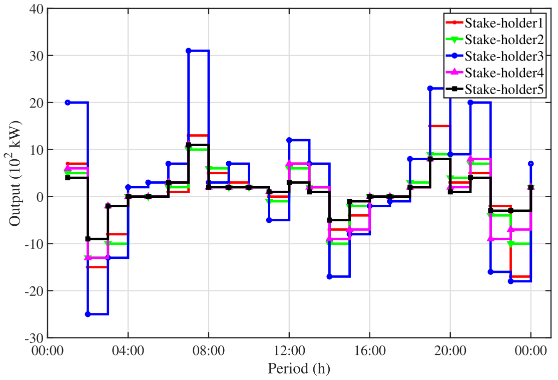

Fig. 8. Action process of five stakeholders.

that the proposed rule-based DDPG has better convergence performance than DDPG, the computational time of three test systems by the proposed learning method are 156, 353, and 732 s, while DDPG requires 177, 412, and 885 s, the priority is more obvious with increase of consumers scale, which reveals that proposed approach has better learning efficiency than DDPG. For further analysis, the five-consumer system results are taken as a typical case. The load-shifting strategy can alleviate the deviation of peak load and valley load, which can be seen in Fig. 6. The rule-based DDPG algorithm learns the consumer's load-shifting strategy to achieve peak shaving and valley filling, then adjusted load can be more stable and convenient to track with power generation, which can also save the economic cost of load consumption.

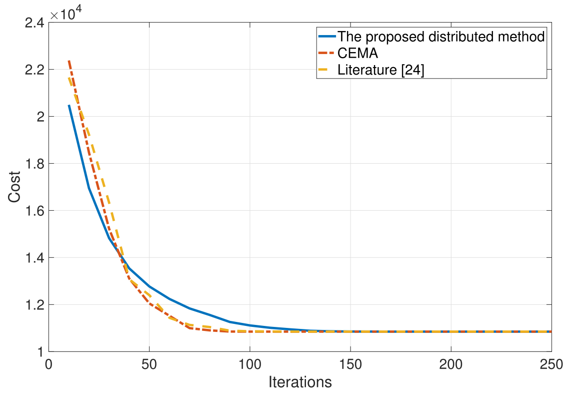

Fig. 9. Convergence performance in comparison with other methods.

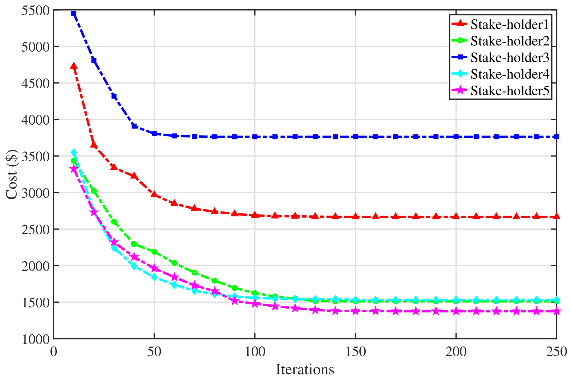

Fig. 10. Convergence process of five agent-based HESs.

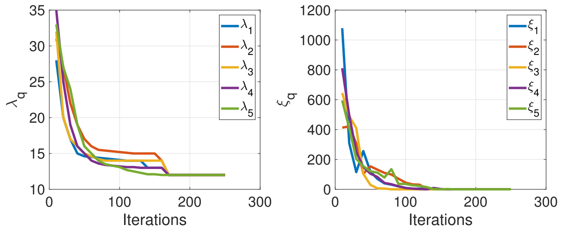

Fig. 11. Convergence process of control parameters λ_q and ξ_q.

B. Middle-Layer With Potential Game-Based Distributed Privacy Optimization

After load shifting on the demand side, stakeholders achieve their maximum profit by exchanging information with designed noise in Fig. 7, where a stable sequence can be coded as disorderly distributed data. The game-based action process of each stakeholder is shown in Fig. 8, where its amplitude is in the range $[-30, 30]$, which cannot exceed 20% of the state value. The voltage of each microgrid range in 1.0 ± 0.005 p.u., and frequency is controlled at 50 ± 0.02 Hz. The convergence of economic cost by the proposed distributed algorithm in comparison with the consensus-based energy management algorithm (CEMA) and literature [24] is presented in Fig. 9, where it can be seen that the proposed distributed method can still converge with those added noises for protecting stakeholders' privacy. The convergence process of five stakeholders is presented in Fig. 10, and it can be seen

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

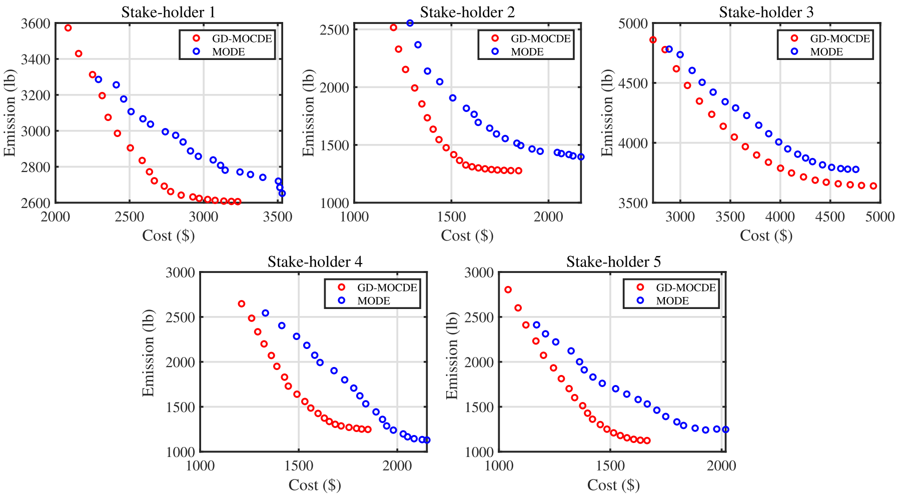

Fig. 12. Comparison of Pareto fronts of five stakeholders.

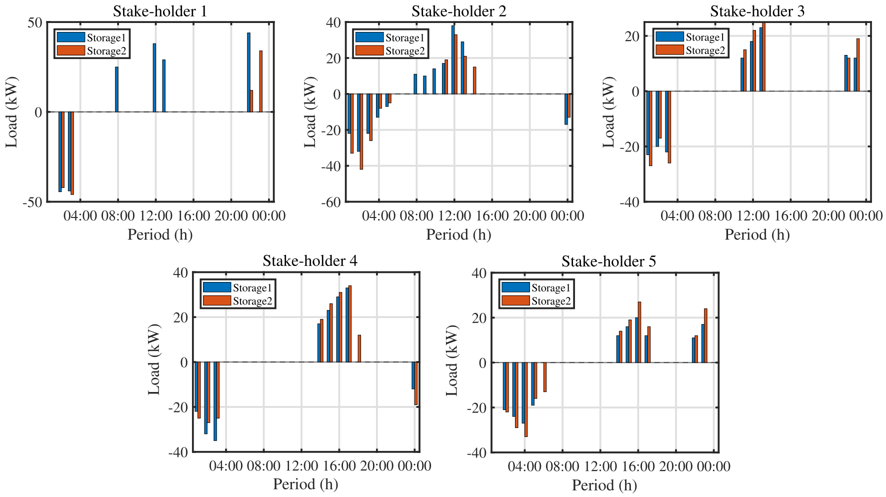

Fig. 13. Charging/discharging process of energy storage.

that all optimization processes converge within 150 iterations. Moreover, the convergence process of control parameters λ_q and ξ_q is also presented in Fig. 11, and the coordinate control parameters of all stakeholders converge to 12 within 150 iterations, and the limits control parameter ξ_q of each stakeholder also converges to 0, which also means that constraint limits are properly satisfied. Combined with the above results, it can reveal that the proposed method can have good convergence performance as well as protect each stakeholder's private information.

C. Lower-Layer Optimization With GD-MOCDE

The total output of stakeholders' energy resources can be deduced with middle-layer model optimization, and the remaining problem is to minimize economic issues and emission rates simultaneously while satisfying various constraint limits. Those obtained Pareto fronts of five power systems are presented in Fig. 12, where the comparison with multi-objective differential evolution (MODE) can reveal that those Pareto fronts obtained by GD-MOCDE can dominate that of MODE, and MODE produces those nondominated schemes disorderly distributed while Pareto fronts of GD-MOCDE has better diversity distribution. For further analysis of the operation process, the tenth scheme is chosen as the compromise scheme, and its optimal operation strategy is presented in Fig. 13, where it can see the charging and discharging process of energy storage.

TABLE I
COMPARISON OF OBTAINED OPTIMIZATION RESULTS AND EFFICIENCY

Index	The proposed method	Literature [24]
Benefit ($)	2177	2279
Switching cost ($)	3223	3435
Total cost($)	13457	15014
Emission (lb)	11512	12593
Average Voltage (p.u.)	0.996	0.98
Average Frequency (HZ)	50.01	50.02
Privacy degree	Privacy	Public

D. Results Analysis of Optimization of the Three-Layered Model

The three-layered hierarchical optimization has less computational complexity in comparison to direct integrated optimization, which means that it can also finish the optimization task in less computational time. To better testify the efficiency of the proposed strategy, the comparison of the proposed method in paper [24] is presented in Table I, where load shifting benefit, total cost, emission rate, privacy degree, and computational time are listed. It can be seen that the proposed method can have better benefit/cost and emission rate in less computational time with a high degree of privacy. The obtained results can help each stakeholder to schedule the power generation with a high privacy degree while considering dynamic changes on the demand side.

V. Conclusion

The existence of stakeholders introduces a major challenge to optimal operation of interconnected microgrid systems in the future electricity market, some merits of this article can be concluded as follows.

1) With consideration of dynamic load demand, reinforcement learning with a rule-based DDPG approach can adjust the consumers' load consumption scheduling by load migration as electricity price changes, which can also shave load peak and fill load valley to maximize economic benefit.
2) Since each stakeholder has a privacy requirement, information exchange cannot be public. Potential game-based distributed privacy optimization can deal with the competition problem as well as the privacy issue.
3) In each stakeholders' microgrid system, multiple objectives are required to be optimized simultaneously. GD-MOCDE can optimize the economic emission problem well with a two-step constraint-handling technique.

References

[1] G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi, "Dynamic energy management system for a smart microgrid," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1643–1656, Aug. 2016.

[2] B. Zhao *et al.*, "Energy management of multiple microgrids based on a system of systems architecture," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6410–6421, Nov. 2018.

[3] Y. Wu, X. Tan, L. Qian, D.-K. Tsang, W.-Z. Song, and L. Yu, "Optimal pricing and energy scheduling for hybrid energy trading market in future smart grid," *IEEE Trans. Ind. Informat.*, vol. 11, no. 6, pp. 1585–1596, Dec. 2015.

[4] K. Wang, Z. Ouyang, R. Krishnan, L. Shu, and L. He, "A game theory-based energy management system using price elasticity for smart grids," *IEEE Trans. Ind. Informat.*, vol. 11, no. 6, pp. 1607–1616, Dec. 2015.

[5] Y. Liang, F. Liu, and S. Mei, "Distributed real-time economic dispatch in smart grids: A state-based potential game approach," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4194–4208, Sep. 2018.

[6] C. Dou, D. Yue, X. Li, and Y. Xue, "MAS-based management and control strategies for integrated hybrid energy system," *IEEE Trans. Ind. Informat.*, vol. 12, no. 4, pp. 1332–1349, Aug. 2016.

[7] A. Belgana, B. P. Rimal, and M. Maier, "Open energy market strategies in microgrids: A Stackelberg game approach based on a hybrid multiobjective evolutionary algorithm," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1243–1252, May 2015.

[8] C. P. Mediwaththe, E. R. Stephens, D. B. Smith, and A. Mahanti, "A dynamic game for electricity load management in neighborhood area networks," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1329–1336, May 2016.

[9] L. Du, S. Grijalva, and R. G. Harley, "Game-theoretic formulation of power dispatch with guaranteed convergence and prioritized bestresponse," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 51–59, Jan. 2015.

[10] C. Wei, Z. Zhang, W. Qiao, and L. Qu, "Reinforcement-learning-based intelligent maximum power point tracking control for wind energy conversion systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6360–6370, Oct. 2015.

[11] Y. Wang and M. Pedram, "Model-free reinforcement learning and Bayesian classification in system-level power management," *IEEE Trans. Comput.*, vol. 65, no. 12, pp. 3713–3726, Dec. 2016.

[12] E. Foruzan, L.-K. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5749–5758, Sep. 2018.

[13] H. Zhang, D. Yue, C. Dou, X. Xie, K. Li, and G. P. Hancke, "Resilient optimal defensive strategy of TSK fuzzy-model-based microgrids' system via a novel reinforcement learning approach," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 31, 2021, doi: 10.1109/TNNLS.2021.3105668.

[14] H. Huang, M. Lin, L. T. Yang, and Q. Zhang, "Autonomous power management with double-Q reinforcement learning method," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1938–1946, Mar. 2020.

[15] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1066–1076, Mar. 2020.

[16] T. Liu, X. Hu, W. Hu, and Y. Zou, "A heuristic planning reinforcement learning-based energy management for power-split plug-in hybrid electric vehicles," *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6436–6445, Dec. 2019.

[17] J. Ahmad, M. Tahir, and S. K. Mazumder, "Improved dynamic performance and hierarchical energy management of microgrids with energy routing," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3218–3229, Jun. 2019.

[18] Y. Du, J. Wu, S. Li, C. Long, and S. Onori, "Hierarchical coordination of two-time scale microgrids with supply-demand imbalance," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 3726–3736, Sep. 2020.

[19] H. Zhang, D. Yue, X. Xie, C. Dou, and F. Sun, "Gradient decent based multi-objective cultural differential evolution for short-term hydrothermal optimal scheduling of economic emission with integrating wind power and photovoltaic power," *Energy*, vol. 122, pp. 748–766, Mar. 2017.

[20] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: 10.1038/nature14236.

[21] C. Zhao, J. Chen, J. He, and P. Cheng, "Privacy-preserving consensus-based energy management in smart grids," *IEEE Trans. Signal Process.*, vol. 66, no. 23, pp. 6162–6176, Dec. 2018.

[22] H. Zhang, D. Yue, W. Yue, K. Li, and M. Yin, "MOEA/D-based probabilistic PBI approach for risk-based optimal operation of hybrid energy system with intermittent power uncertainty," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 4, pp. 2080–2090, Apr. 2021.

[23] J. Aghaei, T. Niknam, R. Azizipanah-Abarghooee, and J. M. Arroyo, "Scenario-based dynamic economic emission dispatch considering load and wind power uncertainties," *Int. J. Electr. Power Energy Syst.*, vol. 47, pp. 351–367, May 2013.

[24] H. Zhang, D. Yue, C. Dou, K. Li, and X. Xie, "Event-triggered multiagent optimization for two-layered model of hybrid energy system with price bidding-based demand response," *IEEE Trans. Cybern.*, vol. 54, no. 4, pp. 2068–2079, Apr. 2021.

Huifeng Zhang (Member, IEEE) received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2013.

From 2014 to 2016, he was a Post-Doctoral Fellow with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, where he is currently an Associate Professor. From 2017 to 2018, he was granted as a Visiting Research Fellow by the China Scholarship Council to study in Queen's University Belfast, Belfast, U.K., and the University of Leeds, Leeds, U.K. His current research interests include electrical power management, optimal operation of power system, distributed optimization, and multiobjective optimization.

Dong Yue (Fellow, IEEE) received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1995.

He is currently a Professor and the Dean of the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, and also a Changjiang Professor with the Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China. He has authored over 100 articles in international journals, domestic journals, and international conferences. His current research interests include the analysis and synthesis of networked control systems, multiagent systems, optimal control of power systems, and the Internet of Things.

Dr. Yue is currently an Associate Editor of the IEEE Control Systems Society Conference Editorial Board and the *International Journal of Systems Science*.

Chunxia Dou (Senior Member, IEEE) received the B.S. and M.S. degrees in automation from the Northeast Heavy Machinery Institute, Qiqihaer, China, in 1989 and 1994, respectively, and the Ph.D. degree from the Institute of Electrical Engineering, Yanshan University, Qinhuangdao, China, in 2005.

In 2010, she joined the Department of Engineering, Peking University, Beijing, China, where she was a Post-Doctoral Fellow for two years. From 2005 to 2016, she was a Professor with the Institute of Electrical Engineering, Yanshan University. Since 2016, she has been a Professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. Her current research interests include multiagent-based control, event-triggered hybrid control, distributed coordinated control, and multimode switching control, and their applications in power systems, microgrids, and smart grids.

Gerhard P. Hancke (Life Fellow, IEEE) received the B.Sc. and B.Eng. degrees and the M.Eng. degree in electronic engineering from the University of Stellenbosch, Stellenbosch, South Africa, in 1970 and 1973, respectively, and the Ph.D. degree from the University of Pretoria, Pretoria, South Africa, in 1983.

He is currently a Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China, and also with the University of Pretoria. He was recognized internationally as a pioneer and leading scholar in industrial wireless sensor networks research. He initiated and coedited the first Special Section on Industrial Wireless Sensor Networks in the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS in 2009 and the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS in 2013. He coedited a textbook *Industrial Wireless Sensor Networks: Applications, Protocols and Standards* (2013), the first on the topic.

Dr. Hancke has been serving as an Associate Editor and a Guest Editor for the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE ACCESS, and previously the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS. Currently, he is a Coeditor-in-Chief of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS.