Improved Generative Adversarial Behavioral Learning Method for Demand Response and Its Application in Hourly Electricity Price Optimization

Junhao Lin, Yan Zhang, and Shuangdie Xu

***Abstract*—In response to the imbalance between power generation and demand, demand response (DR) projects are vigorously promoted. However, customers' DR behaviors are still difficult to be simulated accurately and objectively. To tackle this challenge, we propose a new DR behavioral learning method based on a generative adversary network to learn customers' DR habits. The proposed method is also extended to maximize the economic revenues of generated DR policies on the premise of obeying customers' DR habits, which is hard to be realized simultaneously by existing model-based methods and traditional learning-based methods. To further consider customers' time-varying DR patterns and trace the changes dynamically, we define customers' DR participation positivity as an indicator of their DR pattern and propose a condition regulation approach improving the natural generative adversary framework to generate DR policies conforming to customers' current DR patterns. The proposed method is applied to hourly electricity price optimization to reduce the fluctuation of system aggregate loads. An online parameter updating method is also utilized to train the proposed behavioral learning model in continuous DR simulations during electricity price optimization. Finally, numerical simulations are conducted to verify the effectiveness and superiority of the proposed method.**

***Index Terms*—Demand response, behavioral learning, reinforcement learning, generative adversarial network, electricity price optimization.**

I. Introduction

With the growth of the social economy, the demand for electricity is increasing rapidly and energy shortages remain a severe problem. In response to the imbalance between power demand and supply, countries and regions are not only vigorously developing renewable energy but also launching various active management projects on the power demand side, among which, demand response (DR) is an effective way.

DR projects can relieve the staggering peaks of power consumption by introducing customers by electricity prices [1] or incentives [2]. Therefore, the fluctuation of system loads can be reduced, and the security level of power grids can be improved [3], making DR projects be worth promoting.

Accurately describing the uncertainties of customer's DR behavior is one of the key issues in DR analysis. Studies have shown that electricity price is one of the essential factors encouraging customers to reschedule their electricity consumption plans, which is the basics of price-based DR projects. Therefore, the correlations between economic profits and customers' DR behaviors have been investigated by some studies. Based on the Cobb-Douglas function, variations in customers' electricity consumption can be estimated by variations in the electricity price via the customers elasticity coefficient models [4], [5]. However, the investigated elasticity coefficient can hardly describe the relation between electricity price and DR behavior for various customer individuals in a different time and will eventually cause the deviations in further analysis considering customers' DR behaviors.

Besides, other model-based methods have also been proposed to describe the uncertainty of DR behaviors, including multi-scenario analysis [6], the probability model method [7], and robust optimization method [8]. Moreover, [9] formulates a customer's demand function as a linear function of electricity price with a random variable with a o-mean random variable. In [10], a two-stage stochastic DR method is proposed. The first stage of the model aims to optimize the electricity price, and in the second stage, the appliance schedule is optimized to reduce electricity costs. The shiftable appliance loads are modeled by a group of manually extracted features and the uncertainty of appliance usage is analyzed by the multi-scenario method and scenario reduction technique. In [11], Monte Carlo simulation is employed for a day-ahead DR problem considering the uncertainty of renewable energy outages and system security. A robust optimization method is used in [12] to solve the uncertainty of wind power output and the price-elastic-based demand curves by taking the worst case into account. In general, model-based methods rely on accurate physical or mathematical models to describe customers' DR behaviors. However, they are usu-

Manuscript received: March 16, 2020; revised: September 25, 2020; accepted: February 19, 2021. Date of crosscheck: February 19, 2021. Date of online publication: July 8, 2022.

This work was supported by the National Key Research and Development Program of China (No. 2015AA050203) and the State Grid Corporation of China (No. SGDK0000NYJS1807745).

J. Lin, Y. Zhang (corresponding author), and S. Xu are with Shanghai Jiao Tong University, Shanghai, China, and J. Lin is also with Shibei Electricity Supply Company, State Grid Shanghai Municipal Electric Power Company, Shanghai, China (e-mail: linjunhao@sjtu.edu.cn; zhang_yan@sjtu.edu.cn; xushangdie@126.com).

DOI: 10.35833/MPCE.2020.000152

ally hard to be obtained and usually contain prior knowledge or hypotheses.

Apart from model-based DR analysis methods, machine learning (ML) methods, especially supervised learning (SL) algorithms, have also been introduced into the prediction of customers' DR behaviors. Learning-based methods can predict DR behaviors by data-driven training methods with few human interventions. Reference [13] sets up a neural network to learn energy consumption modes of home heating, ventilation, and air conditioning. Reference [14] directly predicts customers' DR behaviors by a long-short-term memory (LSTM) network, with temporal state vectors and price information as its input. However, traditional learning-based DR analysis methods, which mainly adopt SL algorithms, can only rigidly copy customers' past response policies and will probably fail when the "domain drift" occurs, in which data distribution of a test data set is different from a training data set. Besides, traditional learning-based methods are hard to be directly applied in multi-objective optimization scenarios.

Some studies convert a DR problem to a Markov decision process (MDP) or an optimization model. In this case, the reward of executing a specific DR action in a decision or optimization process is needed to be defined. The reward of DR defined in [15] mainly focuses on gaining economic profits and improving grid reliability, but without taking customers' electricity consumption habits into account. Some studies establish a DR reward function containing discomfort assessments to measure the compromise degree of the generated response policies. Reference [16] considers customers' comfort requirements based on empirical factors, including indoor temperature and relative humidity, in the optimization process of DR policies and inserts these constraints in the defined reward function. In [17], [18], the discomfort coefficient is defined based on the difference between customers' real response behaviors and their original electricity consumption plan. The discomfort coefficient is then added to the total electricity consumption costs. From the current studies, the reward functions are mainly defined manually and empirically. Thus, the individual differences of customers' DR behavior modes are hardly completely considered.

Reinforcement learning (RL) algorithms are widely used in solving DR models and are generally divided into two categories: value-based and policy-based. Value-based RL algorithms such as Q learning [19] and deep Q learning (DQN) [20] are popular for their convenient implementation, but value-based RL cannot directly deal with continuous variables. The discretization of continuous decision variables would dramatically enlarge the action space, making the solution process too time-consuming. Variable discretization also limits the accuracy of the solutions. In contrast, policy-based RL algorithms directly optimize the generated policies by a policy gradient descent (PGD) [21] approach, enabling them to deal with continuous variables directly. Some studies further fuse a policy-based algorithm with a value-based algorithm and propose the Actor-Critic (AC) algorithm [22], which demonstrates higher accuracy and greater convenience in solving optimization problems with continuous variables. However, the natural AC algorithm requires a large number of sampling calculations, and thus the computational efficiency will be reduced.

The literature review above shows that learning-based methods offer an objective way to learn customers' DR habits while model-based methods are more flexible in considering multiple DR objectives. However, some inherent drawbacks also exist in both model-based methods and traditional learning-based methods. On one hand, when considering customers' electricity consumption habits, model-based methods need to define the electricity consumption model manually with prior knowledge or hypotheses, and will thus increase the subjectivity and inaccuracy of the generated DR policies. On the other hand, traditional learning-based DR behavioral learning methods are difficult to be combined with other methods to consider multiple DR objectives due to their end-to-end framework. Therefore, traditional learning-based methods can only rigidly copy customers' past response behaviors, and thus DR agents on the customer side cannot offer customers' DR policies with higher economic profits. To solve the problems in existing methods, threefold major contributions are proposed in this paper.

1) We propose a new DR policy generation method that simultaneously considers customers' electricity consumption habits in a learning way and maximizes economic revenues. The proposed method learns customers' DR behaviors via a generated adversary network and realizes a multi-objective optimization by an RL algorithm. Therefore, the aforementioned drawbacks of model-based and traditional learning-based methods can be improved and DR policies can be generated in an objective and flexible way.

2) An electricity price optimization model is proposed to reduce the system aggregate load fluctuations and enlarge electricity selling profits. The proposed behavioral learning method for DR is conducted by multiple agents and an iteration framework between the power utility company (PUC) and DR agents is built for the price optimization.

3) We consider the dynamic changes of customers' DR behavior patterns and offer a learning-based tracing method. We define a DR participation positivity index (PPI) to indicate customers' current DR patterns and constrain the generated DR policies conforming to current PPI by proposing a regulated condition generative adversarial imitation learning (RCGAIL) method. This combined measurement increases the effectiveness and accuracy of customers' DR behavioral learning results in the scenarios with various DR behavior patterns.

The rest of this paper is structured as follows. In Section II, we explore the electricity price optimization model of the PUC and the DR analysis model on the customer side. The customers' DR behavioral learning method, its improvement for dynamic response patterns tracing, and the online parameter updating method are presented in Section III. Section IV conducts case studies and evaluates the performance of the proposed model. Finally, conclusions are drawn in Section V.

II. Electricity Price Optimization and DR Model

In this paper, we propose a DR policy generation model based on a new DR behavioral learning method and apply it to an electricity price optimization problem. A sketch of an

integrated system containing the electricity price optimization and customer-side DR behavior learning is shown in Fig. 1.

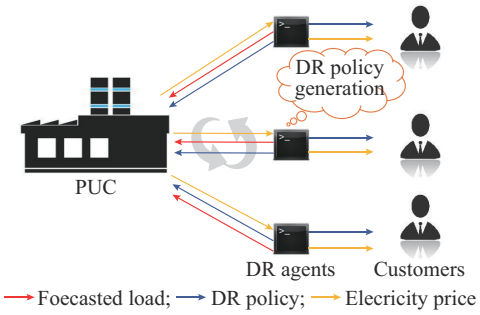

Fig. 1. Electricity price optimization and customer-side DR behavior learning system.

For price optimization containing a price-based DR project, the PUC first initializes a preliminary hourly electricity price. PUC then sends the price information to the DR agents and anticipates to obtain customers' probable response to this preliminary price. The corresponding DR policies are generated by DR agents (like the customer-side energy management systems), which can be deployed in a centralized or distributed way. Figure 1 offers an example of the distribution implement method, which is also adopted in [15], [23]-[25]. In this way, DR simulation results are uploaded back to the PUC for further electricity price adjustment. This process may last several rounds for PUC to gradually update the preliminary price, during which customers won't be notified of the price information until the price optimization has been completed. Till then, DR agents generate the recommended DR policies corresponding to the optimized electricity price and send the formal price information and the corresponding recommended DR policies to the customers. Customers' actual response behaviors will be recorded for the further training of the DR analysis model.

To reduce the computational and communication burdens of the proposed system, some techniques, like parallel-distributed computing and cloud-edge computing, have been applied in the power DR domain [24], [25]. For example, in the power communication network, edge nodes can be built to gather and process customer data. The results are also packaged in the message passed to the PUCs. Since fewer channels are required in this way, the communication burden can be greatly reduced.

A. Objective Function of Electricity Price Optimization Model for PUC

In this paper, we set a twofold-objective for the PUC's electricity price optimization model. The first objective is to reduce the fluctuation of daily aggregate loads (F_1), which brings the long-term profits of PUC by improving the system security level and reducing the operation costs [26], [27]. The second objective is to maximize the electricity selling profits of PUC (F_2).

The fluctuation of the system aggregate loads in different hours of a single day can be described by the coefficient of variation (C_v) index. Therefore, the objectives of the PUC's price optimization model can be formulated as:

$$F_1 = \min C_v(\boldsymbol{L}_{SDR}) = \min \frac{1}{E_L}\sqrt{\sum_{h=1}^{24}\left(\sum_{i=1}^{N_c} L_{DR,i,h} - E_L\right)^2} \tag{1}$$

$$F_2 = \max P_{DR} = \max \sum_{h=1}^{24}\left[(p_{DR,h} - c_h)\sum_{i=1}^{N_c} L_{DR,i,h}\right] \tag{2}$$

where N_c is the number of customers in the system; $L_{DR,i,h}$ is the actual load schedule after DR of the customer i in hour h; $\boldsymbol{L}_{SDR}$ is the system daily load after DR; E_L is the expectation of $\boldsymbol{L}_{SDR}$; P_{DR} is the profit after price optimization and customer DR; $p_{DR,h}$ is the electricity price of hour h to be optimized; and c_h is the PUC's cost purchasing of electricity of hour h from the market.

$L_{DR,i,h}$ is related to the $p_{DR,h}$ in a price-based DR project and this relation is difficult to be accurately described with a static and certain mathematic model. Therefore, a learning-based approach is proposed in Section III to determine this correlation.

B. Constraints of Electricity Price Optimization Model

1) Lower Bound of Daily Profits After Price Optimization

Although a PUC tries to reduce the fluctuation of system loads via DR projects, earning profits is also important. Therefore, a lower limit is needed for the daily profits of PUC, which can be expressed as:

$$\begin{cases} P_{DR} > d_p P_I \\ P_I = \sum_{h=1}^{24}\left[(p_{I,h} - c_h)\sum_{i=1}^{N_c} L_{I,i,h}\right] \end{cases} \tag{3}$$

where P_I is the original total electricity selling profits without price optimization and DR; d_p is the lower bound coefficient of the system daily profits; and $L_{I,i,h}$ and $p_{I,h}$ are the load demands of customer i and the original electricity selling price of hour h, respectively.

d_p can be set according to actual profit requirements of PUC. For the case of further optimizing an existing pricing result, $p_{I,h}$ and $L_{I,i,h}$ can be set as the price and the load schedule in this existing result, respectively. For an original pricing scenario, $p_{I,h}$ is set as c_h, and (3) means that the PUC must be profitable by the price optimization and customer DR. $L_{I,i,h}$ for DR simulation, in this case, can be obtained by the load forecasting.

2) Upper and Lower Bounds of Hourly Electricity Price

The hourly electricity prices should have upper and lower limits. This constraint can be expressed as:

$$\kappa_{\min} p_{h,\min} \le p_{DR,h} \le \kappa_{\max} p_{h,\max} \tag{4}$$

where $p_{h,\max}$ and $p_{h,\min}$ are the upper and lower bounds of the hourly electricity price in hour h, respectively; and $\kappa_{\max}$ and $\kappa_{\min}$ are the adjustment coefficients for $p_{h,\max}$ and $p_{h,\min}$, respectively.

$p_{h,\max}$ and $p_{h,\min}$ can be set as the maximum and minimum of the recent historical hourly electricity price, respectively, and we set $\kappa_{\max}$ and $\kappa_{\min}$ as 1.2 and 0.6 in the simulations of

this paper, respectively.

C. Model of DR Analysis Agents

For a higher adoption probability, customers' electricity consumption habits need to be considered in a DR project. Some studies limit the deviation between generated DR policies and customers' original load schedules to reduce customer dissatisfaction. However, the deviations caused by DR may not inevitably result in customer dissatisfaction, because conducting DR can also bring them extra economic revenues. Some customers may be attracted by such rewards and are willing to partly reschedule their electricity consumption plans. Therefore, we need to find an appropriate model to describe the complex correlations among electricity prices, customers' original electricity consumption plans, and their actual response behaviors with various and variable individual electricity consumption habits.

1) Objective Function of Customer DR Model

In this paper, the DR analysis agents generate DR policies to both minimize customers' daily electricity charges and conform to customers' electricity consumption habits. The objective function of the DR policy generation process focuses on minimizing a customer's electricity charge P_i and can be expressed as:

$$\min P_i = \min \sum_{h=1}^{24} p_{DR,h} L_{DR,i,h} \tag{5}$$

Then, to increase the adoption probability of generated DR policies, we will consider customers' electricity consumption habits in the constraints of the DR model in a data-driven way.

2) Constraints of DR Policy Generation Process

First, a lower bound of electricity consumption schedules after DR needs to be set. Considering customers may have different DR patterns, we set the lower bound of generated DR policies by defining a discount coefficient according to customers' historical DR behaviors as $b_L = \min\{Q_{DR,1}/Q_{I,1}, Q_{DR,2}/Q_{I,2}, \ldots, Q_{DR,j}/Q_{I,j}\}$, where j is the number of the training data; $Q_{I,j}$ and $Q_{DR,j}$ are the quantities of daily electricity consumption before and after DR of the j^{th} data, respectively. b_L needs to be updated when the new DR behavior data are collected. Then a constraint for the DR policies for customer i can be formulated as:

$$\sum_{h=1}^{24} L_{DR,i,h} \geq b_L \sum_{h=1}^{24} L_{I,i,h} \tag{6}$$

Second, to take the customers' electricity consumption habits into account and increase the adoption possibility, DR analysis agents need to limit the deviation between the generated daily DR policies ($\boldsymbol{L}_{DR,i}$) and probable real response behaviors ($\boldsymbol{L}_{C,i}$) of customer i to a certain range ε (ε>0), which can be described as:

$$d_i(\boldsymbol{L}_{DR,i}, \boldsymbol{L}_{C,i}) \leq \varepsilon \tag{7}$$

where $d_i(\cdot)$ is a distance measurement function for DR behaviors of customer i.

However, $d_i(\cdot)$, $\boldsymbol{L}_{C,i}$, and ε in (7) are all hard to be acquired in the simulation process without detailed prior knowledge of model-based methods. To overcome this difficulty, a data-driven DR behavioral learning method is presented in the following section.

III. Customers' DR Behavioral Learning Method and Policy Generation Method

A. Framework of Overall System

In this subsection, we will learn the rules of customers' DR behaviors. Figure 2 shows an overall framework of the electricity price optimization model for PUC and the proposed customers' DR behavioral learning model.

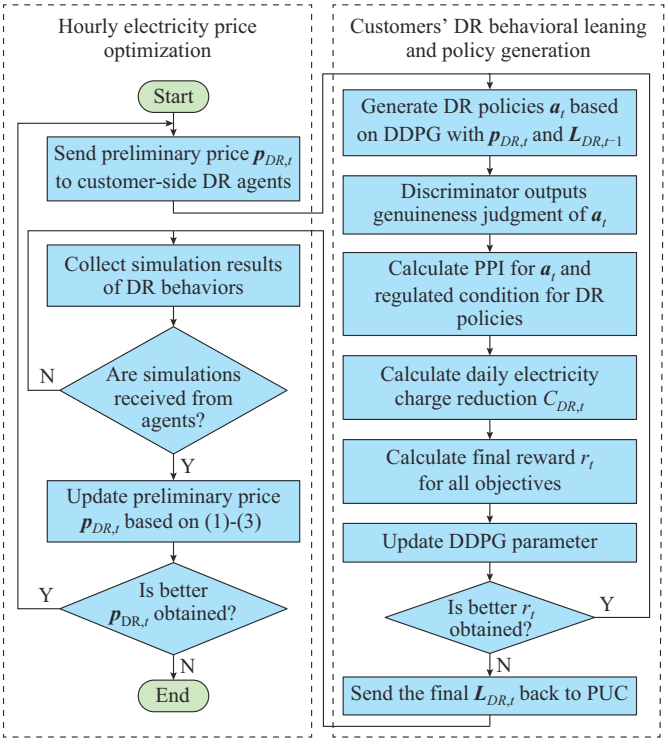

Fig. 2. Framework of electricity price optimization model for PUC and customers' DR behavioral learning model.

The whole system consists of two submodules: an hourly electricity price optimization module and a customers' DR behavior learning and policy generation module. The main process of the proposed system can be implemented with the following steps.

Step 1: the PUC initializes a preliminary hourly electricity price $p_{I,h}$ and sends it to customer-side DR agents. For an original pricing scenario, DR agents conduct load forecasting. The forecasting result $L_{I,i,h}$ is then sent back to the PUC. For a further price optimization scenario, the initial value of $L_{DR,i,h}$ can be inherited from $L_{I,i,h}$. PPI is also predicted once for this day.

Step 2: the PUC updates the daily preliminary price $\boldsymbol{p}_{DR,t}$ ($\boldsymbol{p}_{DR,t} = [p_{DR,1,t}, p_{DR,2,t}, \ldots, p_{DR,24,t}]$), where $p_{DR,h,t}$ is the electricity price of hour h during the price optimization (round t), which is sent to the customer-side DR agents.

Step 3: the DR agent simulates and updates the DR behaviors of customer i using the RCGAIL algorithm based on the current $\boldsymbol{p}_{DR,t}$, customer i's load demands $\boldsymbol{L}_{DR,i,t-1}$ ($\boldsymbol{L}_{DR,i,t-1}$ =

$[L_{DR,i,1,t-1}, L_{DR,i,2,t-1}, \ldots, L_{DR,i,24,t-1}]$, where $L_{DR,i,h,t-1}$ is the behavior of customer i of hour h in optimization round t), and the predicted DR PPI. The simulated DR policies are sent back to the PUC.

Step 4: the PUC aggregates the simulation results of DR behaviors from DR analysis agents and optimizes $\boldsymbol{p}_{DR,t}$ using the electricity price optimization model in Section II. Then go back to *Step 2* and repeat the above procedures until the optimal hourly electricity price is obtained.

Step 5: the PUC officially broadcasts the optimal hourly electricity price to customers and DR agents offer their customers recommended DR policies corresponding to this hourly electricity price. Then, a day-ahead electricity price optimization process is accomplished.

The rest of this subsection illustrates the detailed principles of each module of the proposed DR behavioral learning and DR policy generation method, the definition and prediction approach of the customers' DR PPI values, and an online model parameter updating method. A summarized numerical relation between electricity price and DR behaviors is presented in the part 4 of Section III-D.

B. Customers' DR Behavioral Learning Method Based on Improved Generative Adversary Network

Though some studies formulate the DR analysis process as a prediction problem, and SL algorithms have been employed in DR behavioral learning, a drawback still exists. In this prediction model, DR agents can only copy customers' existing DR patterns due to the end-to-end framework of SL algorithms, and can hardly be extended for other objectives (like reducing customers' electricity charges). Actually, the DR behaviors of some customers are elastic to some extent if more profits can be gained. Therefore, the DR policies don't need to strictly follow customers' DR habits. PUCs can try to actively induce customers to accept the recommended DR policies, which will also contribute to learning and reducing their DR behavioral uncertainties. In this way, multiple DR objectives are needed to be considered, and formulating the DR analysis problem as a predicting process with a single objective is not appropriate.

In this paper, we formulate this problem as an optimization-prediction combined model and apply the generative adversarial imitation learning (GAIL) algorithm to solve this problem. In this way, the function of learning-based methods can be covered, while the merit of the flexibility of model-based methods is also maintained.

1) GAIL-based DR Behavioral Learning Model

The GAIL works based on two networks shown in Fig. 3. The generator network tries to confuse the discriminator by generating DR policies similar to customers' real behaviors, while the discriminator network executes genuineness judgment. The GAIL trains the two networks in an adversarial way. The equilibrium reaches when the discriminator cannot tell the differences between generated DR policies and customers' real behaviors, indicating that the generator can control the deviation of the behavioral learning results within a small limit, so that (7) is also satisfied in a learning-based way.

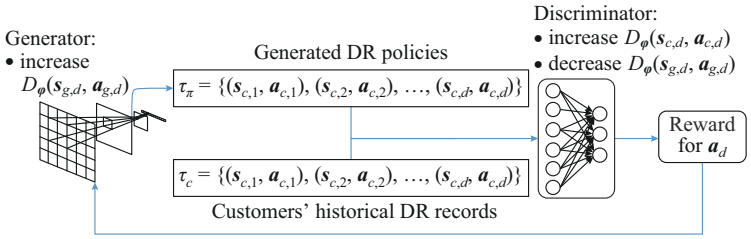

Fig. 3. Framework of natural GAIL algorithm.

The implementation of GAIL training in a DR analysis can be illustrated as follows. Denote $\boldsymbol{s}_{c,d}$ as the observed and environmental states for DR behaviors on day d in history of generating DR policies, including daily load schedule $\boldsymbol{L}_{DR,i,t}$, daily preliminary price $\boldsymbol{p}_{DR,t}$, and customers' DR pattern indicator I_E. The complete and detailed definition of I_E will be presented in Section III-C. Customers' DR behavior on day d in history is denoted as $\boldsymbol{a}_{c,d}$. Then the DR process can be simulated as a decision process from $\boldsymbol{s}_{c,d}$ to $\boldsymbol{a}_{c,d}$. In this way, a customer's historical DR records can be expressed as $\tau_c = \{(\boldsymbol{s}_{c,1}, \boldsymbol{a}_{c,1}), (\boldsymbol{s}_{c,2}, \boldsymbol{a}_{c,2}), \ldots, (\boldsymbol{s}_{c,d}, \boldsymbol{a}_{c,d})\}$. The corresponding generated DR policies by the GAIL can be expressed as $\tau_{\pi} = \{(\boldsymbol{s}_{c,1}, \boldsymbol{a}_{g,1}), (\boldsymbol{s}_{c,2}, \boldsymbol{a}_{g,2}), \ldots, (\boldsymbol{s}_{c,d}, \boldsymbol{a}_{g,d})\}$, where $\boldsymbol{a}_{g,1}, \boldsymbol{a}_{g,2}, \ldots, \boldsymbol{a}_{g,d}$ are the finally optimized DR policies for day 1, 2, …, and d.

In the training process, by given τ_c and observed $\boldsymbol{s}_t = \boldsymbol{s}_{c,d}$, GAIL tries to find DR policy $\boldsymbol{a}_{g,d}$ (at the optimization step t, $\boldsymbol{a}_t = \boldsymbol{a}_{g,d}$) corresponding to $\boldsymbol{s}_t$ and ensure the generated DR policy, that is $\tau_{\pi} = \{(\boldsymbol{s}_t, \boldsymbol{a}_t)\}$, conforms to the same rule as τ_c. During this process, the output of the discriminator at the optimization step t ($D_{GAIL,t}$) serves as guidance to optimize the generated policies. The adversarial training process of GAIL [28] with an objective as $V(\boldsymbol{\varphi}, \boldsymbol{\theta})$ can be described as:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} V(\boldsymbol{\varphi}, \boldsymbol{\theta}) = \mathop{\mathbb{E}}_{(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \tau_{\pi}} (\log_2 D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)) + \mathop{\mathbb{E}}_{(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \tau_c} (\log_2 (1 - D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t))) \quad (8)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are the parameters of the generator and discriminator networks, respectively; and $D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is the output function of the discriminator according to $\boldsymbol{s}_t$ and $\boldsymbol{a}_t$.

2) Improved GAIL for Customers' DR Pattern Tracing

Since natural GAIL does not impose any constraints on the generator, the training process is uncontrollable, so natural GAIL is sometimes hard to converge or converges slowly.

Moreover, in some specific cases, we know empirically that certain additional information has an inner correlation with customers' DR behaviors, for example, customers' current DR participation positivity. Such inner correlations are sometimes difficult to be described with mathematical func-

tions or physical models, so we hope that they can be learned by the generator to make the DR analysis results more reasonable. However, the learning process of a neural network is hard to intervene and the additional information may be ignored by the neural network. If we use SL algorithms or a natural GAIL with additional information as generation conditions [29] directly in this scenario, it would be hard to know whether the final output is related to the additional information.

To solve the aforementioned problems, we propose a regulated conditional method to improve the natural GAIL for dynamic DR patterns tracing and call it RCGAIL. RCGAIL uses a conditional generator to generate DR policies conforming to the expected additional information y, which is the generation condition. To ensure that y will be learned by the generator, RCGAIL inserts a regular term $L(\boldsymbol{a}_t, y)$ for the generation condition y to the objective function $V(\boldsymbol{\varphi}, \boldsymbol{\theta})$ in natural GAIL as:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} V(\boldsymbol{\varphi}, \boldsymbol{\theta}) = \mathop{\mathbb{E}}_{(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \tau_\pi} (\log_2 D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t | y)) + \mathop{\mathbb{E}}_{(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \tau_c} (\log_2 (1 - D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t | y))) + \lambda_L L(\boldsymbol{a}_t, y) \tag{9}$$

where λ_L is the regulation coefficient.

Here we set $\lambda_L = 0.9$ and $\boldsymbol{a}_t$ is influenced by both $\boldsymbol{s}_t$ and additional information y.

In dynamic DR simulations, we expect to constrain the generated DR policies conforming to customers' current DR behavior patterns. Thus, the additional information y in $L(\boldsymbol{a}_t, y)$ represents customers' DR behavior pattern and is related to the probable environment states and action. Thus, we can calculate $L(\boldsymbol{a}_t, y)$ as follows:

$$L(\boldsymbol{a}_t, y) = \left| y_{real} - y_{gen} \right| = \left| \hat{I}_E(y_1, y_2, \ldots, y_{d-1}) - I_E(\boldsymbol{s}_t, \boldsymbol{a}_t) \right| \tag{10}$$

where y_{real} is the real additional information about customers' current DR behaviors; y_{gen} is the additional information calculated from the generated policy $\boldsymbol{a}_t$; $y_1, y_2, \ldots, y_{d-1}$ are the real additional information (DR patterns) before day d; $\hat{I}_E$ is the estimation of y_{real} based on $y_1, y_2, \ldots, y_{d-1}$; and I_E is an indicator of customers' DR behavior pattern related to $\boldsymbol{s}_t$ and $\boldsymbol{a}_t$ in optimization step t.

We set the PPI as I_E and will give a detailed definition of I_E in Section III-C. Since customers have not executed current $\boldsymbol{a}_t$ during the DR behavioral learning process, we provide a prediction result $\hat{I}_E$ to estimate y_{real} based on customers' historical DR patterns $y_1, y_2, \ldots, y_{d-1}$ in Section III-C.

Since $L(\boldsymbol{a}_t, y)$ is always nonnegative, to optimize (9), we should reduce $L(\boldsymbol{a}_t, y)$, so that the generated policy $\boldsymbol{a}_t$, whose y_{gen} have low deviations to the y_{real}, are more likely to be retained and the DR policies generated eventually will conform the customers' current DR patterns. By being given the predicted additional information continuously on different days, RCGAIL can dynamically trace customers' latest DR behavior patterns. Formula (9) will be converted to rewards for the generator of RCGAIL in the training process and will be explained later. To increase the robustness of the algorithm, if $L(\boldsymbol{a}_t, y)$ is smaller than a limit δ, we can regard $L(\boldsymbol{a}_t, y)$ as 0. We set δ as 0.1 in this paper.

The training process of the RCGAIL-based customers' DR behavioral learning model can be realized by the following steps in Algorithm 1. The DR policy generation process will be presented in Section III-D.

Algorithm 1: customers' DR behavioral learning method based on RCGAIL

1. **Input**:
2. τ_π: generated DR policies and state variables
3. τ_c: customer's historical DR records
4. **Output**:
5. $D_{GAIL,t}$: genuineness judgments of generated DR policies at step t
6. Initialize generator parameter $\boldsymbol{\theta}$ and discriminator parameter $\boldsymbol{\varphi}$
7. Initialize the customer DR record buffer B
8. **for** each iteration **do**
9. Sample policy trajectories τ_π
10. Update the buffer B using τ_π
11. Sample customer's DR trajectories $\tau_c \sim B$
12. Update the discriminator parameter $\boldsymbol{\varphi}$ based on $\boldsymbol{a}_t$ and $\boldsymbol{s}_t$ via gradient ascent method as: $\nabla_{\boldsymbol{\varphi}} l_D = \mathbb{E}_{\tau_\pi}(\nabla_{\boldsymbol{\varphi}} \log_2 D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)) + \mathbb{E}_{\tau_c}(\nabla_{\boldsymbol{\varphi}} (1 - \log_2 D_{\boldsymbol{\varphi}}(\boldsymbol{s}_{c,d}, \boldsymbol{a}_t)))$
13. Calculate $\dot{y}_{gen}$ and calculate regular term via (10)
14. Output the genuineness judgment of τ_π as $D_{GAIL,t} = -\log_2(1 - D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)) - L(\boldsymbol{a}_t, y)$ and update the generator parameter $\boldsymbol{\theta}$, which is also the $\boldsymbol{\varpi}$ in (14)
15. **end for**

C. Definition of Customers' DR Participation Positivity and Its Prediction Method

During the training process of the proposed behavioral learning model, customers' historical DR records may contain multiple response patterns, e.g., customers may have different sensitivities to electricity prices on different days and thus their DR participation positivity may be various. Such training data containing multiple DR patterns will confuse behavioral learning algorithms and make those algorithms hard or even unable to learn the real rules of the customers' response behaviors. To solve this problem, we define an index to indicate the customers' current DR pattern.

In this passage, we define customers' DR PPI (I_E) to be the latent code y in RCGAIL as the ratio of electricity charge changes and distance between electricity consumptions before and after DR:

$$I_E = sig\left(\frac{C_{r,opt}}{D_{JS}(\boldsymbol{L}_I, \boldsymbol{L}_{DR})} \right) = sig\left(\frac{\sum_{h=1}^{24} p_{I,h} L_{I,i,h} - \sum_{h=1}^{24} p_{DR,h,t} L_{DR,i,h,t}}{D_{JS}(\boldsymbol{L}_I, \boldsymbol{L}_{DR})} \right) \tag{11}$$

where $C_{r,opt}$ is the change of daily electricity charges before and after DR; $\boldsymbol{L}_I$ and $\boldsymbol{L}_{DR}$ are the daily electricity consumption schedules before and after DR, respectively; $D_{JS}(\cdot)$ is the Jensen-Shannon (JS) divergence between $\boldsymbol{L}_I$ and $\boldsymbol{L}_{DR}$; and $sig(\cdot)$ is the sigmoid function.

In the case that $D_{JS}(\boldsymbol{L}_I, \boldsymbol{L}_{DR}) = 0$, we set I_E as 0. The data for calculating I_E are all available for PUCs by recording daily electricity prices and collecting DR policies uploaded from DR analysis agents.

Since $D_{JS}(\cdot)$ is nonnegative, $C_{r,opt}$ controls the sign of I_E. When $\boldsymbol{L}_{DR}$ has a higher daily charge than $\boldsymbol{L}_I$, $C_{r,opt}$ will be less than zero, which means that the customer is less sensi-

tive to the electricity price and prefers to behave according to his own scheduled electricity consumption plan in an uneconomical way. Conversely, if $C_{r,opt}$ is greater than 0, it means that the customer's rescheduled power consumption plan has the same inclination as that advocated by the DR project. In this case, a profitable electricity price may be attractive to the customer.

In this passage, we use a gated recurrent unit (GRU) algorithm to predict customers' current DR PPI values and its detailed model is presented in [30]. GRU is a kind of recurrent neural network and also works as an SL and backpropagation way. The loss function of GRU (l_{GRU}) in this paper is defined as a mean squared error (MSE), which is defined as:

$$l_{GRU}=\frac{1}{N_b}\sum_{j=1}^{N_b}(I_{E,real,j}-\hat{I}_{E,j})^2 \tag{12}$$

where $I_{E,real,j}$ and $\hat{I}_{E,j}$ are the real DR PPI calculated by (11) and the predicted results of the GRU for data j, respectively; and N_b is the size of the data batch in the training process.

D. DR Behavior Generation Based on RL

We apply the deep deterministic policy gradient (DDPG) [31] algorithm as the DR policy generator. DDPG uses two deep neural networks (DNNs), i.e., a main network and a target network, in both the Actor and Critic networks.

1) Mathematical Model of Actor Network in DDPG

The policy generation process $\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ can be realized with a DNN with parameter $\boldsymbol{\varpi}_t$. To apply DDPG in the DR analysis problem, the input state $\boldsymbol{s}_t$ of the Actor in step t consists of the price information $\boldsymbol{p}_{DR,t}$, current policy $\boldsymbol{L}_{DR,t}$, and predicted pattern indicator $\hat{I}_E$. $\boldsymbol{L}_{DR,t}$ is initialized with the forecasted loads $\boldsymbol{L}_{DR,0}$. $\boldsymbol{L}_{DR,0}$ will be replaced with newly generated policies $\boldsymbol{a}_t$ in the optimization steps. Suppose $J(\cdot)$ is the performance function of a policy from $\boldsymbol{s}_t$ to $\boldsymbol{a}_t$ ($\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)$), its gradient can be formulated as.

$$\nabla_{\boldsymbol{\varpi}_t}J(\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t))=\mathbb{E}(\nabla_{\boldsymbol{\varpi}_t}\log_2\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)\nabla_{\boldsymbol{a}_t}Q(\boldsymbol{a}_t,\boldsymbol{s}_t)) \tag{13}$$

where $Q(\boldsymbol{s}_t,\boldsymbol{a}_t)$ is the Q value for the policy measured by the Critic network.

In DDPG, $\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ is determined by a certain function $\mu(\cdot)$ from the main Actor network as $\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)=\mu(\boldsymbol{s}_t)$, and thus accelerates the learning process compared with the natural AC algorithm. $\boldsymbol{\varpi}_t$ can be updated as follows.

$$\boldsymbol{\varpi}_{t+1}=\boldsymbol{\varpi}_t+\alpha_{\boldsymbol{\varpi}}\nabla_{\boldsymbol{\varpi}_t}\mu(\boldsymbol{s}_t)\nabla_{\boldsymbol{a}_t}Q(\boldsymbol{s}_t,\boldsymbol{a}_t)|_{\boldsymbol{a}_t=\mu(\boldsymbol{s}_t)} \tag{14}$$

where $\alpha_{\boldsymbol{\varpi}}$ is the learning rate of the Actor network.

2) Mathematical Model of Critic Network in DDPG

The main Critic network comments on the generated policies by Q values with parameter $\boldsymbol{w}_t$. The loss function l_c of the Critic network is as:

$$l_c=\frac{1}{m}\sum_{t=1}^{m}(y_{c,t}-Q(\boldsymbol{s}_t,\boldsymbol{a}_t))^2 \tag{15}$$

where m is the mini-batch size of records sampled from replay memory; and $y_{c,t}$ can be determined as follows.

$$y_{c,t}=r_t+\gamma Q^*(\boldsymbol{s}_{t+1},\boldsymbol{a}'_{t+1})=r_t+\gamma Q^*(\boldsymbol{s}_{t+1},\mu^*(\boldsymbol{s}_{t+1})) \tag{16}$$

where $Q^*(\boldsymbol{s}_{t+1},\boldsymbol{a}'_{t+1})$ and $\mu^*(\boldsymbol{s}_{t+1})$ are the outputs of the target Critic and Actor networks with parameters $\boldsymbol{w}'_t$ and $\boldsymbol{\varpi}'_t$, respectively; $\boldsymbol{a}'_{t+1}$ is the action output by the target Actor network; and $\gamma\in[0,1]$ is the probability to choose the action with a high reward in each step.

Algorithm 2 will show the parameter updating method. r_t will be determined in the following part, but if the model constraints are violated, like (6) or upper and lower limits of loads in each hour, r_t will directly minus a negative number to prevent further optimization in this way.

Algorithm 2: electricity price optimization containing DDPG and RCGAIL based DR behavioral learning

Input:
  $\boldsymbol{L}_{L0}$: initial load schedule
  $\hat{I}_E$: predicted customers' PPI
  τ_c: customer's historical DR records
  m: mini-batch size
Output:
  $\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)$: DR policies for the customers
  $\boldsymbol{p}_{DR,t}$: hourly electricity price weighting to be optimized
Initialize the parameters of the Actor network, Critic network, and the discriminator network
Initialize the replay memory B
for $i=1,2,\ldots,T_{\max}$, **do** the following until (1) is satisfied
  Update preliminary hourly price $\boldsymbol{p}_{DR,t}$
  For each DR analysis agent, do the following steps
Receive $\boldsymbol{p}_{DR,0}$ from PUC and initialize the environment state $s_0=[\boldsymbol{p}_{DR,0},\boldsymbol{L}_{L0},\hat{I}_E]$
Initialize a random process $\hbar_t$ for DR behavior exploration in step t
for $t=1,2,\ldots,$ **do** the following until (4)-(6) is satisfied
    Repeat generating a DR policy until (5) is satisfied by the Actor with randomness $\hbar_t$ as:
      $\boldsymbol{a}_t=\mu_{\boldsymbol{\varpi}}(\boldsymbol{s}_t,I_E)+\hbar_t$
    Input $\tau_\pi=\{\boldsymbol{s}_t,\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)\}$ and τ_c into the discriminator to obtain the genuineness judgment $D_{GAIL,t}$ and update $\boldsymbol{\varphi}$ of the discriminator via Algorithm 1
    Calculate r_t for $\boldsymbol{a}_t$ according to (19)
    Update the state variable $\boldsymbol{s}_{t+1}=[\boldsymbol{p}_{DR,t+1},\ \boldsymbol{a}_t,\ \hat{I}_E]$ and store the transition $\{\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1}\}$ into B
    If $|B|>m$:
      Sample a mini-batch of the transition
      Train the main networks of DDPG via (14) and (17)
      Update target networks with a running average method
        $\boldsymbol{\varpi}'\leftarrow\upsilon\boldsymbol{\varpi}+(1-\upsilon)\boldsymbol{\varpi}'$, $\boldsymbol{w}'\leftarrow\upsilon\boldsymbol{w}'+(1-\upsilon)\boldsymbol{w}'$
    else
      continue
  end for
  Return optimal DR policies $\pi_{DR}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ for current $\boldsymbol{p}_{DR,t}$
end for
Output the optimized electricity price $\boldsymbol{p}_{DR,t}$

With (16), the parameters of the main Critic network ($\boldsymbol{w}_t$) can be updated with the following equation.

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_{\boldsymbol{w}}\nabla_{\boldsymbol{w}_t}l_c \tag{17}$$

where $\alpha_{\boldsymbol{w}}$ is the learning rate of the Critic network.

3) Reward Function of DDPG-Based DR Policy Generation Model

In RCGAIL, the rewards of the generated policies are determined by the discriminator $D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t,\boldsymbol{a}_t)$ to make the generated policies conforming to the distribution of the real data. Meanwhile, the generated DR policy should also conform to the customer's current DR behavior pattern, that is $L(\boldsymbol{a}_t,y)=L(\boldsymbol{a}_t,I_E)$. Therefore, the reward function of the RL module in RCGAIL can be expressed as:

$$r_t=D_{GAIL,t}=-\log_2(1-D_{\boldsymbol{\varphi}}(\boldsymbol{s}_t,\boldsymbol{a}_t))-\lambda_L L(\boldsymbol{a}_t,I_E) \tag{18}$$

Formula (18) indicates that the DR policy generation process is controlled by the discriminator in GAIL without human intervention. In this case, the DR behavioral learning has also become DR behavior imitation, and the subjectivity of the generated DR policies will be greatly reduced.

Since the generated DR policies will be recommended to customers before their electricity consumption, if the economic rewards of the generated DR policies are satisfactory considering their DR habits, customers may further adjust their original electricity consumption behaviors. PUCs' purpose of promoting DR can also be achieved in this way. Therefore, we improve the reward function in (18) by defining it from two aspects, i.e., an economic profit aspect and a customer behavior genuineness judgment aspect. The reward in step t is presented as:

$$\begin{cases} r_t = D_{GAIL,t} C_{DR,t} \\ C_{DR,t} = \dfrac{\sum_{h=1}^{24} p_{DR,t}(L_{I,h,t} - L_{DR,h,t})}{\sum_{h=1}^{24} p_{DR,h,t} L_{I,h,t}} \\ D_{GAIL,t} = -\log_2(1 - D_{\varphi}(\boldsymbol{s}_t, \boldsymbol{a}_t)) - \lambda_L L(\boldsymbol{a}_t, \hat{I}_E) \end{cases} \tag{19}$$

where $C_{DR,t}$ is the electricity charge reduction rate representing the economic profits by DR in step t.

In (19), $D_{GAIL,t}$ represents a genuineness judgment of the generated DR policy $\boldsymbol{L}_{DR,h,t}$ compared with the customers' real DR records. Therefore, $D_{GAIL,t}$ can be regarded as the estimation of the DR policies' adoption probability, and $C_{DR,t}$ can be regarded as the potential economic revenues. Formula (18) indicates that the DR analysis agents try to generate DR policies with more economic revenues on the premise of conforming to customers' electricity consumption habits.

Since the RL-based GAIL algorithm also has an open framework that policy generation, behavioral learning, and policy comment are conducted by separated networks, it can realize a comprehensive consideration of objective behavioral learning and further profit obtaining. Therefore, the merits of model-based methods and SL-based methods can be both kept.

4) Numerical Formulation of Relation Between Price and DR Behaviors

With the learned reward function r_t, the complete DR behavioral learning method can be summarized with the following equations, in which the relation between electricity price $\boldsymbol{p}_{DR,t}$ and DR behaviors $\boldsymbol{L}_{DR,t}$ can be numerically modeled.

$$\boldsymbol{L}_{DR} = \boldsymbol{a}_t \quad t \geq T_{\max} \tag{20}$$

$$\boldsymbol{a}_t = \mu(\boldsymbol{s}_t) = \mu([\boldsymbol{p}_{DR,t}, \boldsymbol{a}_{t-1}, \hat{I}_E]) \quad t = 1, 2, \ldots \tag{21}$$

$$\boldsymbol{\varpi}_{t+1} = \boldsymbol{\varpi}_t + \alpha_{\varpi} \nabla_{\varpi_t} \mu(\boldsymbol{s}_t) \nabla_{\boldsymbol{a}_t} Q(\boldsymbol{s}_t, \boldsymbol{a}_t)|_{\boldsymbol{a}_t = \mu_{\varpi}(\boldsymbol{s}_t)} \tag{22}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha_{\boldsymbol{w}} \nabla_{\boldsymbol{w}_t} \left(\frac{1}{m} \sum_{t=1}^{m} (y_t - Q(\boldsymbol{s}_t, \boldsymbol{a}_t))^2 \right) \tag{23}$$

$$y_t = r_t + \gamma Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}'_{t+1}) \tag{24}$$

$$r_t \leftarrow (18) \tag{25}$$

$$\boldsymbol{\varphi}_{t+1} = \boldsymbol{\varphi}_t + \alpha_{\varphi} [\mathbb{E}(\log_2 D_{\varphi}(\boldsymbol{s}_t, \boldsymbol{a}_t)) + \mathbb{E}(\log_2(1 - D_{\varphi}(\boldsymbol{s}_c, \boldsymbol{a}_c)))] \tag{26}$$

where $\boldsymbol{L}_{DR}$ is the final DR policy; $T_{\max}$ is the maximum iterations of the policy generation process; $\mu(\cdot)$, $Q(\cdot)$ and $D_{\varphi}(\cdot)$ are the policy generation network, policy comment network, and DR behavioral learning network with parameters $\boldsymbol{\varpi}$, $\boldsymbol{w}$, and $\boldsymbol{\varphi}$, respectively; and $\boldsymbol{s}_c$ and $\boldsymbol{a}_c$ are the agent's observation states (including original load schedules and price information) and the corresponding DR behaviors of customers' historical records, respectively.

The final action $\boldsymbol{a}_t$ ($t \geq T_{\max}$) is output as the DR policy under price $\boldsymbol{p}_{DR,t}$. Parameters of all the DNNs are trained with the collected historical data set $\{\boldsymbol{s}_c, \boldsymbol{a}_c\}$ and generated data set $\{\boldsymbol{s}_t, \boldsymbol{a}_t\}$.

Formulae (20) and (21) present a price-DR relation with parameters. Formula (22) updates the parameters for the policy generation network. The Q value in (22) is given by the critic network. Formulae (23) and (24) update the critic network, with the learned reward function r_t. r_t is generated by the discriminator of the GAIL model as (25). The reward learning is present in Algorithm 1. In this way, even if customers' DR behaviors are hardly formulated with mathematical models, a numerical estimation method is proposed.

Based on the reward function, the proposed DR behavioral learning and electricity price optimization process are summarized in Algorithm 2. The running average rate υ for target network updating is set as 0.001 in this paper.

E. Model Parameter Updating Method for Dynamic and Online DR Analysis

Since customers' DR behavior patterns may change for various reasons like daily electricity prices, customers' personal willingness and short-term daily life arrangements, it leads to large amounts of uncertainties for DR behavioral learning. Thus, to trace customers' DR behavioral patterns, beside the proposed condition regulation method, the behavioral learning model also has to update its network parameters dynamically and in time with continuously collected customers' actual response data and daily electricity price. However, in traditional off-line training processes, the training data need to be obtained once all together and are usually input to the models in batch forms to update the model parameters. Therefore, an off-line optimization method is not appropriate for this paper and an online learning optimization algorithm is necessary.

To realize the dynamic and online DR policy generation, a follow-the-regularized-leader (FTRL) algorithm [32] is applied. Supposing t is the optimization step and $\boldsymbol{W}_{i,t}$ is the parameter i of the DNN in step t in the proposed DR behavioral learning model, the iteration formula to update $W_{i,t}$ using FTRL can be expressed as:

$$\begin{cases} \boldsymbol{W}_{i,t+1} = \begin{cases} \boldsymbol{0} & |\boldsymbol{Z}_{i,t}| < \lambda_1 \\ -\left(\lambda_2 + \sum_{s=1}^{t} \sigma_s\right)^{-1} [\boldsymbol{Z}_{i,t} - \lambda_1 sign(\boldsymbol{Z}_{i,t})] & |\boldsymbol{Z}_{i,t}| \geq \lambda_1 \end{cases} \\ \boldsymbol{Z}_{i,t} = \sum_{s=1}^{t} \boldsymbol{g}_s - \sum_{s=1}^{t} \sigma_s \boldsymbol{W}_{i,s} \\ \sigma_s = \dfrac{1}{\eta_{i,s}} - \dfrac{1}{\eta_{i,s-1}} \end{cases} \tag{27}$$

where $\boldsymbol{g}_s$ is the gradient of a loss function in iteration s; λ_1 and λ_2 are the regularization coefficients; and $\eta_{i,s}$ is the per-coordinate learning rate for DNN i in the s^{th} iteration.

For step t, $\eta_{i,s}$ can be calculated as:

$$\eta_{i,t}=\frac{\alpha}{\beta+\sqrt{\sum_{s=1}^{t} g_{s,i}^2}} \tag{28}$$

where α and β are the hyper-parameters; and $g_{s,i}$ is the i^{th} coordinate of $\boldsymbol{g}_s$.

Reference [29] offers the empirical values for α, β, λ_1 and λ_2.

IV. Case Study

In this section, we illustrate the results of the proposed DR behavioral learning model and assess its performance. The electricity price information and customers' actual load data are derived from [33] and [34]. The day-ahead forecasted loads are generated with the collected data by an LSTM network. We regard customers' actual loads as their DR behaviors since customers' DR behaviors are induced by the electricity price in a price-based DR project. Therefore, customers' actual electricity consumptions also represent their DR behaviors corresponding to the electricity price. Randomness is added to the used data for data augmentation and increasing their diversity, which also contributes to improving model robustness for noise-contained data. Considering a data set with an excessive period may lead to inaccuracy of the results, since customers' electricity consumption habits a long time ago may be greatly different from their recent DR behaviors. The testing system contains 280 customers and the used data have a period of 6 months.

In this paper, we set up a two-layer neural network as the Actor network of DDPG with 36 and 24 hidden units, respectively. The two-layer neural network containing 10 neurons each layer consists of the Critic network. Networks in DDPG are activated by the tanh function. In the discriminator of RCGAIL, we build a 4-layer neural network using 30, 20, 10, 1 hidden neurons, with leakly Relu active function for the first three layers and sigmoid function for the last layer. A dropout technique is applied in this paper with a dropout rate as 0.1.

A. Customers' DR PPI Prediction

In this subsection, we verify the effectiveness of the DR PPI prediction model. We introduce a support vector regression (SVR) algorithm to compare with GRU. We use electricity prices and customers' actual daily loads in 3 months for pre-training and data in the next 3 months for testing. The testing data are offered to the PPI prediction models successively for online parameter tuning after daily PPI has been predicted. Figure 4 shows the prediction results of the two algorithms for one of the customers.

To evaluate the performance of the prediction models, we apply the mean absolute error (MAE), mean absolute percentage error (MAPE), and R^2 score. Performance comparisons between GRU and SVR are shown in Table I.

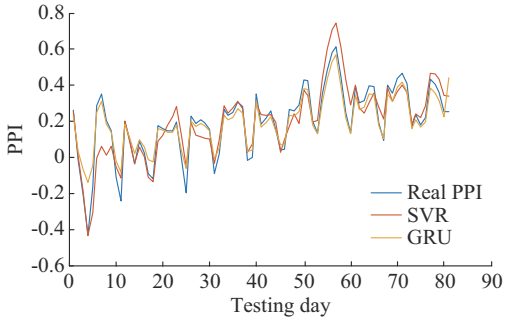

Fig. 4. Results of customer's DR PPI prediction.

TABLE I
Performance of PPI Prediction Algorithms

Algorithm	MAE	MAPE	R^2 score
GRU	0.0354	0.0942	0.9189
SVR	0.0476	0.1240	0.8436

Table I indicates that the GRU has a better performance than SVR in prediction accuracy. This is because deep learning algorithms have a more powerful ability in nonlinear mapping by deepening the network layers and adding nonlinear active functions. These measures help DNNs extract the features of the input data automatically and effectively, while traditional ML algorithms usually extract data features based on prior assumptions or feature engineering. Although SVR has improved its learning ability by introducing a kernel function, the chosen kernel function still has difficulty in distinguishing the data following arbitrary distributions, and thus its mapping ability is still limited.

B. Performance of DR Behavioral Learning Model

Before the behavioral learning, we conduct the load forecasting. The used LSTM network contains 24 units for output and 4 layers. Historical load and price data are the input of the model. Actually, forecasted loads serve as an auxiliary indicating feature of customers' current demands. Therefore, another well-performed forecasting can also be applied. Figure 5 shows the load forecasting results on 3 different days.

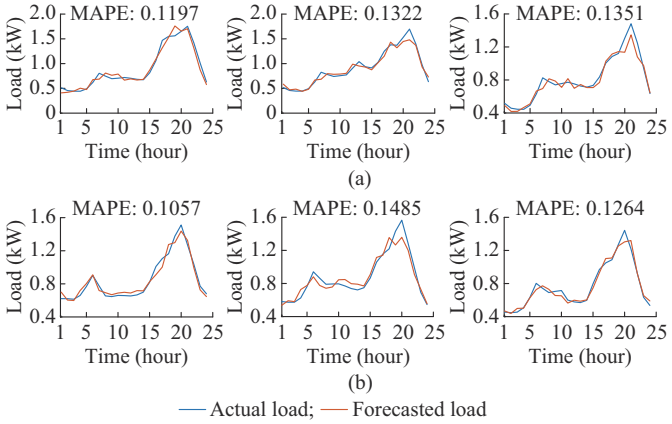

Fig. 5. Load forecasting results. (a) Demand forecasting results for customer 1. (b) Demand forecasting results for customer 2.

Using loads before day j and price before day $j+1$ for load forecasting on day j, we train the load forecasting model before starting the behavioral learning simulations. When training the GAIL model and conducting subsequent simulations, we average prices from day $j-3$ to $j-1$ as a moving average method. The averaged prices are regarded as the initial preliminary prices on day j. The forecasted loads for the GAIL model are generated with such preliminary prices.

The simulation results show that the forecasting errors are acceptable by using an independent forecasting model for each customer. Researches in [35] and [36] also present the methods to improve the accuracy of single customer load forecasting. Actually, the forecasted loads are not required to be completely accurate. This is because that the consistency of DR policies and customers' real DR habits is judged by the behavioral learning network. Forecasted loads serve as an auxiliary indicator to the state recognition for the judgment, whose features will be extracted by the discriminator network. In this way, detailed information, also containing errors, is all probable to be filtered. Therefore, limited errors may not affect behavioral learning results.

Besides, we add random noises to the training data for the behavioral learning model, which helps the model adapt to the fluctuations of the input data, and the model robustness will also be improved. Combined with the generalization ability of neural networks, the forecasting errors for load forecasting may have few negative effects on the behavioral learning results.

As shown in Fig. 1, customer load forecasting is conducted by the customer-side DR agents. Different well-trained load forecasting methods may still cause some limited deviations in different results. In our proposed method, due to the independence of each customer's DR analysis agent, different agents can use different forecasting methods. However, we still suggest a certain agent to avoid changing the used forecasting method, to keep the consistency of distributions of forecasted loads and DR behaviors from slight changes. The added noise in the training data also contributes to freeing the simulation results of the proposed method from being affected by different forecasting results with limited deviations. Also, the stacked auto-encoder model can also be applied to filter the deviations from different load forecasting methods [37].

Then, we will verify the effectiveness of the proposed DR behavioral learning model. The initial preliminary price is also set by the aforementioned moving average methods. The collected price data are used as the optimal preliminary price in this simulation. The presented results are the responses to these optimal preliminary prices. We set the reward function of the RL algorithm in the RCGAIL based on both (18) and (19) and name them as case 1 and case 2, respectively. By setting the two cases about RCGAIL, we can compare the DR analysis results purely considering customers' electricity consumption habits and comprehensively considering consumption habits and economic revenues, so that the flexibility of the proposed DR behavioral learning model can be verified. We take the DR analysis model in [17] as control group 1 and the model based on LSTM adopted in [14] as control group 2. In control group 1, we set the weight of discomfort cost $\omega_{a,t}$ as 0.6, and the reward function r_t of this model can be expressed as $r_t = P_t + \omega_{a,t}\left|\boldsymbol{L}_{DR,t} - \boldsymbol{L}_I\right|$, where P_t is the daily electricity charge; $\boldsymbol{L}_I$ is the daily original load schedule; and $\boldsymbol{L}_{DR,t}$ is the DR policy. Besides, we also make a comparison with an additional linear-program-based DR model, which only focuses on minimizing the daily charge P_t, and is labeled as control group 2.

Figure 6 shows the behavioral learning results for a customer in 24 hours according to day-ahead load forecasting results and the real hourly electricity price. The daily electricity charge and error indices (MAPE and R^2 score) between generated DR policies and the customers' real response behaviors are presented in Fig. 7.

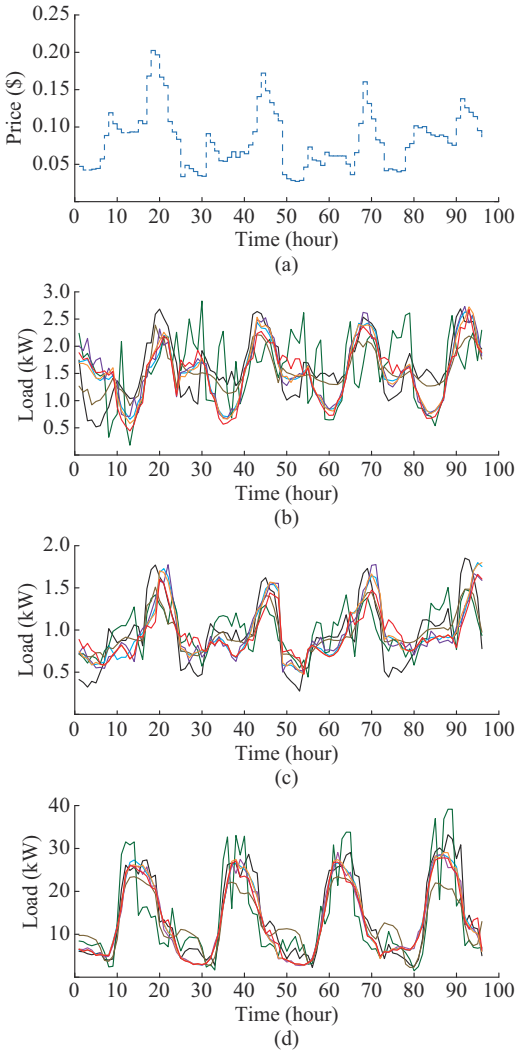

Fig. 6. DR results based on different DR analysis models. (a) Hourly price. (b) Load curves of customer 1. (c) Load curves of customer 2. (d) Load curves of customer 3.

In this simulation, we present the results of customers with three different types of DR behaviors in four days. The

daily load curves of customer 1 have two peaks in the mornings and evenings, respectively. The highest load consumptions appear in the evenings, and day-time loads are relatively lower. For customer 2, the highest loads also appear in the evenings, but the day-time loads are relatively flattened. The daily loads of customer 3 are almost two-side. The loads are low in the evenings and late at nights, but they dramatically increase from the late mornings. Then the loads stay at high levels and last a long time to the late evenings.

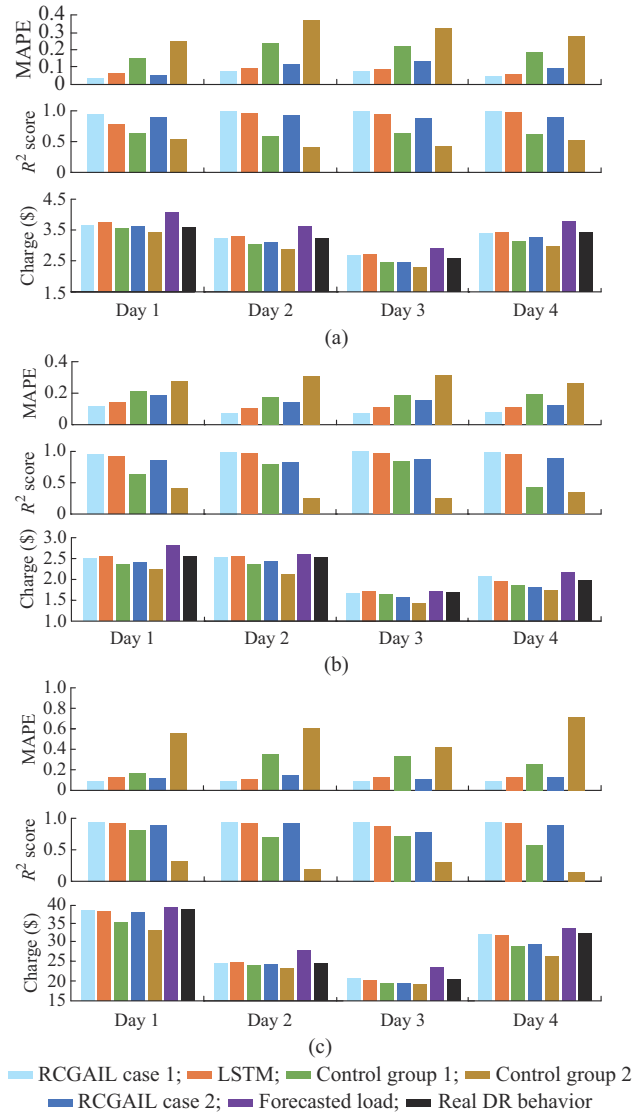

Fig. 7. MAPE, R^2 score, and daily charges of DR analysis methods. (a) Indices of customer 1. (b) Indices of customer 2. (c) Indices of customer 3.

Different response behaviors can be observed in customers' actual electricity consumption curves (regarded as real DR behaviors). Compared with customer 3, customers 1 and 2 are more positive in participating in DR. In the high-price hours in the evenings, customer 2 responds as delaying some peak loads and transfering parts of loads to the deep nights. While customer 1 makes a slight reduction of loads in evening high-price hours rather than delaying the consumptions. The loads in the daytime are obviously reduced. The reduced loads are also transferred to the late nights with low prices. Customer 3 responds as reducing part of loads in the afternoon, which shows a lower elasticity for load rescheduling. Therefore, we can conclude that customers 1 and 2 are relatively more price-sensitive but the sensitivity is low during the evening high-price hours, and their DR behaviors may be more likely to be induced to some extent.

The simulation results also indicate that customers' DR behaviors may be various even with similar electricity prices. A universal model to formulate different DR behaviors is hard to be obtained. Therefore, a learning-based method extracting personalized DR habits from customers' historical records is appropriate for this situation.

The indices in Fig. 7 indicate that the proposed method satisfactorily performs in DR behavioral learning for all the present customer classes. With historical DR records containing multiple response patterns, the result of RCGAIL case 1 with a high R^2 score and low MAPE may be attributed to the PPI prediction and condition regulation methods, which offers the behavioral learning method the customer's current response pattern to follow. The proposed models are trained with historical data containing noises. We can see that customers' behaviors can still be learned with the forecasted loads, which are not completely accurate for the aforementioned lack of price information during the forecasting process.

Figure 6 shows that control group 2 has similar response principles for different customers who usually reduce more loads in high price hours preferentially. Such policies can reduce more charges, and may match the needs of customers having the desire for maximizing profits. However, large deviations are also found in these policies and thus they are not suitable for customers having other response modes. Customers may also ignore the recommended DR policies with huge deviations to their habits. That's why DR agents are recommended to conduct DR behavioral learning for the policy generation.

According to the reward function, control group 1 offers DR policies trying to minimize electricity charges with the least modifications to the original load schedules. In this way, loads in high-price hours are also firstly reduced like control group 2. Even though the control group limits the adjustment amount to the original load curves, its principle still doesn't conform to every customers' electricity consumption habits, like customers 1 and 2. As a result, the DR policies may be more likely to be ignored by the customer due to their great deviations from the customer's actual DR behaviors despite that they have lower daily electricity charges than customers' original electricity consumption plans.

The DR results generated by LSTM mainly imitate the customer's response behavior and also perform well in the MAPE and R^2 score. The only objective of LSTM is to accurately predict customers' actual DR behaviors, and thus it will not actively try to reduce the charges. Simulation results show that RCGAIL case 2 can generate the policies with lower charges while the load curve deviations have not dramatically increased. This performance contributes to attracting customers to accept the recommended policies so that PUCs can actively induce customers' DR behaviors, which

is hardly realized with the SL-based methods.

To further verify the effectiveness of the proposed method, we conduct a test in a real-time DR case. In this case, the final price signal is a real-time price (RTP), and the RTP is set to be published hourly. The response time is the future rest hours of a day, and the DR process is conducted in a rolling form [38], [39]. RTP prediction for the rest of response hours can be applied [38], [40] and is also conducted in a rolling form. Since the dimensions of the state variables are dynamic, we use a 1-D convolution neural network (CNN) to extract the features from the data with variable length and use 1×1 convolution to replace the final full-connected layer. Simulation results are shown in Fig. 8.

Similar to the previous simulation, simulation results show that the drawbacks of model-based methods in DR behavioral learning still exist. Even though they can reduce more daily charges, their policies still have large deviations from customers' real DR behaviors. While the testing data-driven methods still have satisfactory performances in learning customers' DR habits with relatively lower MAPE values, so that the adoption probability of the generated policies may be larger than control groups 1 and 2. Besides, compared with LSTM-based method, the proposed RCGAIL-based method can further reduce the charges with a limited deviation increase. Therefore, the effectiveness and flexibility of our proposed method can be verified in the real-time DR.

From the above cases, we can notice that the proposed method is more comprehensive for taking both the high profitability and customers' electricity consumption habits into account. Moreover, the reward function in (19) can be conveniently modified for more optimization objectives, so the proposed DR analysis method is more flexible than SL-based DR behavior learning methods in generating DR policies with multiple objectives.

C. Performance of RCGAIL in Tracing Different DR Patterns

In this subsection, we analyze the effect of PPI and verify the capability of RCGAIL to dynamically trace the customers' DR patterns. To make the simulation results easy to be compared, we set the reward function in the DDPG as formula (18) in this case.

First, we analyze the effect of PPI. Figure 9 shows the DR behaviors of two customers on different days. Figure 9 indicates that customers may have different DR behaviors with similar state situations (price and forecasted loads). In Fig. 9(a), (c), (d) and (e), (g), (h), the two customers' response behaviors show their desires for reducing the charges, as they reduce part of load consumptions in the high-price hours compared with original schedules. In these cases, PPI is positive. The amplitude of PPI represents the efficiency of profit gaining with load rescheduling. Figure 9(b) and (f) show negative PPI values, in which, loads from the afternoon to the evening are larger than original schedules. These behaviors indicate that they respond in uneconomical ways, which may present customers' personal preferences.

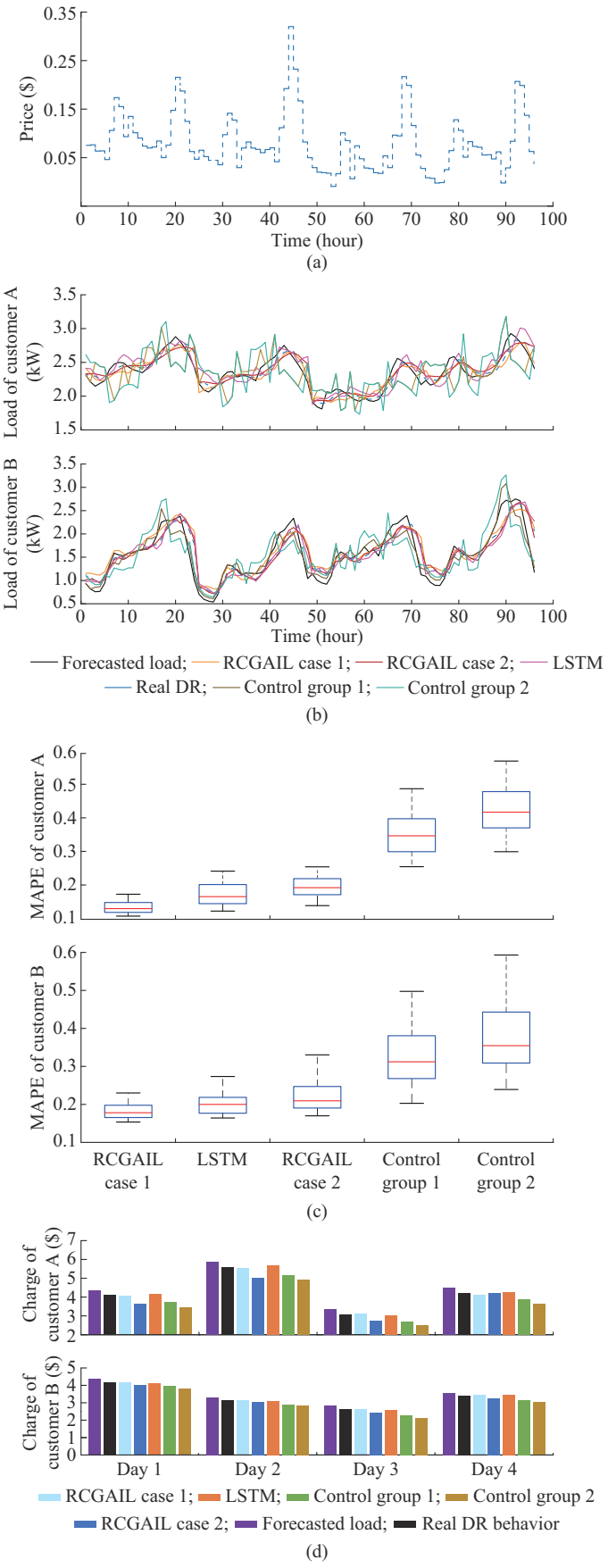

Fig. 8. DR simulation results for RTP. (a) RTP. (b) Loads and DR policies. (c) Error distribution for all testing days. (d) Daily charges.

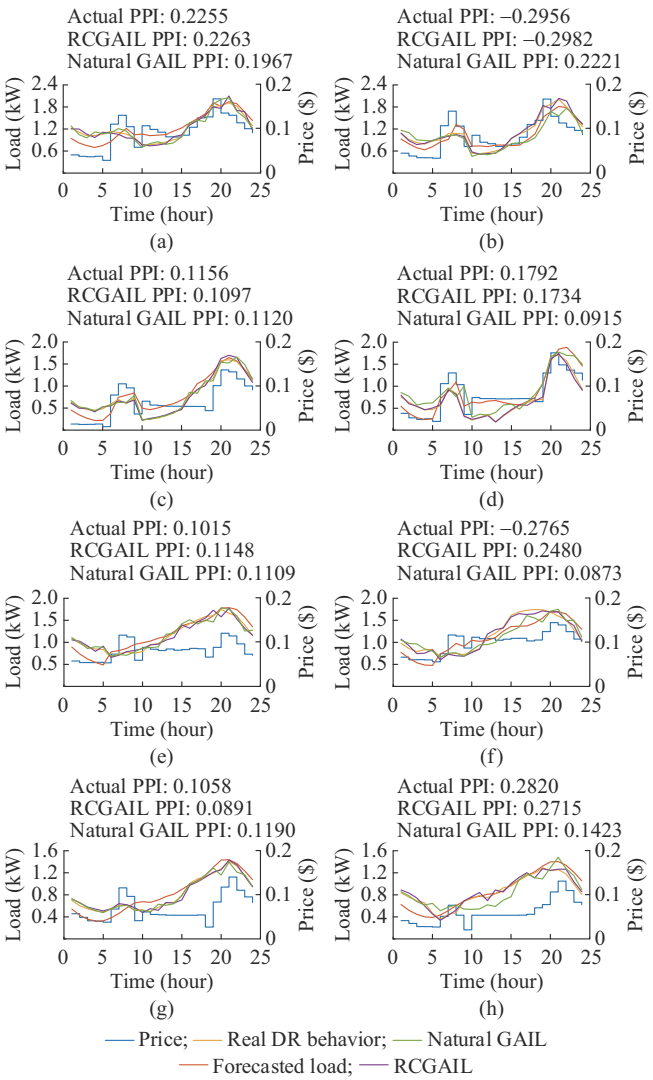

Fig. 9. DR behaviors and simulation results with different PPI values. (a) Load curves of customer 1 in case 1. (b) Load curves of customer 1 in case 2. (c) Load curves of customer 1 in case 3. (d) Load curves of customer 1 in case 4. (e) Load curves of customer 2 in case 1. (f) Load curves of customer 2 in case 2. (g) Load curves of customer 2 in case 3. (h) Load curves of customer 2 in case 4.

The simulation results with and without matching to customers' actual PPI are also presented in Fig. 9. Without the assistance of PPI, natural GAIL may generate different policies under similar price and original load schedules, since customers may respond in various ways existing in history. PPI serves as a further indicator for similar input states and guides the behavioral learning model to choose a certain pattern among DR records. During the policy generation process, PPI works by generating an additional gradient to the generator network, making the actual PPI of the generated policy closer to the predicted one.

In this way, the penalty item in (10) can be eliminated. Therefore, the PPI-RCGAIL model performs better than natural GAIL in customer dynamic behavioral learning and inferring.

Second, since PPI needs to be combined with the proposed regulated condition methods in actual implements, the next case will show the simulation results in dynamic behavioral learning situations. In that simulation, we will analyze also the efforts of the combination of PPI, regulated condition method, and the FTRL algorithm.

Figure 10 shows dynamic behavioral learning results of the natural GAIL-based methods and RCGAIL-based methods on 9 days. On these testing days, we can see that the customer shows limited participation positivity to the DR project in the evening but may be interested in participating in DR in the daytime and after midnight on the first two days. During the high-price hours in the evening, only a small fraction of loads is moved to low-price hours. The majority of the loads in the evening are still non-shiftable. The profile of the customer's actual response changes greatly on days 3 and 6, which means that the customer changes his DR patterns on those days. On day 3, for example, the customer reduces the electricity consumption from midnight to morning and repeats this adjustment on the following 2 days. On day 6, the customer returns to early electricity consumption habit at these hours.

We can note from the simulation results of the first 2 days shown in Fig. 10 that natural GAIL without the condition regulation method can keep the behavioral learning errors within a relatively low limit before the customer changes his response pattern, but natural GAIL fails to trace the dynamic changes after the DR pattern changes on day 3. This is because that after the change occurs, the newly collected customers' actual DR behaviors can only expand the experience pool of the GAIL, but can't offer extra constraints about this new pattern to the generator. Therefore, the generator will continue to follow one of the existing DR patterns, not tracing the new change in the data distribution. Natural GAIL may regard all the generated policies obeying one of the existing distribution rules as real.

RCGAIL also performs well in DR behavioral imitation on the first 2 days and the prediction errors are within reasonable limits. On the third day, when the customer greatly changes the electricity consumption plan for the first time during the testing, a great imitation deviation occurs. This is because no signs have been discovered in advance. RCGAIL still believes that the customer will repeat his previous response pattern. However, RCGAIL predicts the customers' DR PPI with the latest DR data and generates DR policies according to the customers' newly predicted DR PPI, and constrains the generated policies to conform to this PPI by the condition regulation method. The FTRL algorithm also helps the RCGAIL update the parameters of deep networks by using the newly collected data effectively and efficiently. Therefore, on day 4 and day 5, the RCGAIL model succeeds in generating DR policies following the customers' changed behavior pattern. Although the customer adjusts the response pattern again on day 6, RCGAIL continues to learn this adjustment. The prediction deviation is reduced to a low level again on day 7 with a swift model parameter adjustment. Therefore, the effectiveness of RCGAIL can be verified.

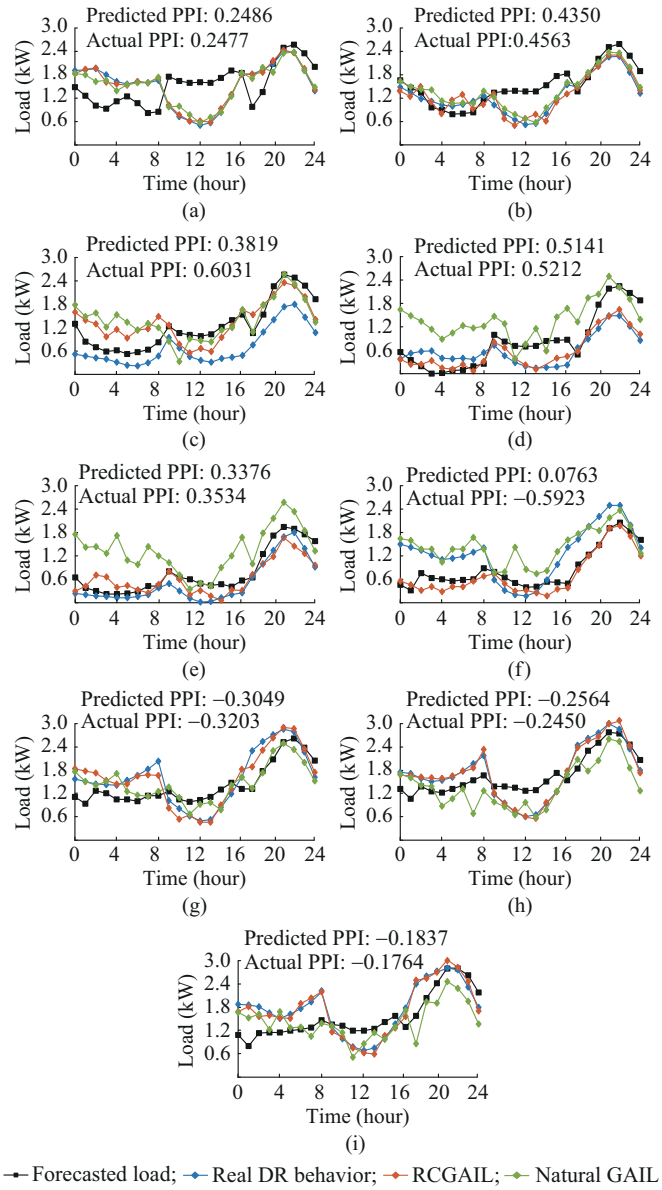

Fig. 10. DR behavior learning and generation results for 9 days. (a) Day 1. (b) Day 2. (c) Day 3. (d) Day 4. (e) Day 5. (f) Day 6. (g) Day 7. (h) Day 8. (i) Day 9.

D. Electricity Price Optimization Results

Finally, we apply the proposed RCGAIL model to the electricity price optimization process. In this case, DR analysis agents need to consider both customers' electricity consumption habits and economic revenues, so we apply the reward function of the RL model defined in (19) for DR policy generation.

In this simulation, we conduct a price optimization for an existing data record to further adjust the system load schedule and enlarge PUC's profits. Hence, $\boldsymbol{p}_I$ and $\boldsymbol{L}_I$ are the original price and load in this case, respectively. Referring to [15], [41], the cost c is estimated as the wholesale market price. We set the d_p in (3) as 1 to ensure that the optimized price will bring PUC a higher profit than that without optimization.

Some results of the price optimization problem in (1) and (2) locating at the Pareto front are presented in Fig. 11. With these results, PUC can find optimal pricing policies according to their preference for profit desire and load fluctuation reduction. In actual implementation, a weighting method can also be applied to find a certain solution to this optimization, when the PUC has specific preferences of the objectives or combined with the non-dimension operation and entropy-weighting method, like [42].

The optimized price and detailed indices of a result (marked red in Fig. 11) are presented in Table II and Fig. 12.

The results show that the PUC receives higher profits after the price optimization. Meanwhile, the system load fluctuation, measured by the C_V index, is effectively reduced, which flattens the daily load profile and will thus contribute to reducing the grid operation costs. By inducing customers to participate in DR, the system daily peak loads are also shaded, so that the grid security level in the heavy hours can be improved. Therefore, the PUC's purposes of promoting DR are realized.

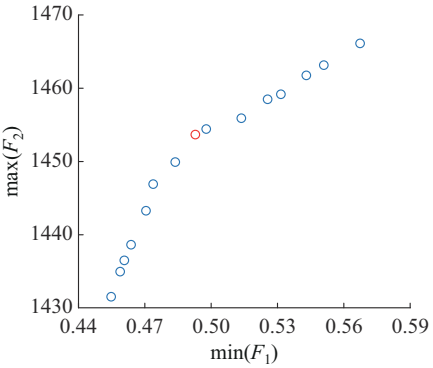

Fig. 11. Tradeoff between F_1 and F_2.

TABLE II
INDICES FOR ELECTRICITY PRICE OPTIMIZATION

State	C_V	System daily profit ($)	Daily maximum load (MW)
Before price optimization and DR	0.7375	1372.93	1.4984
After price optimization and DR	0.4937	1453.74	1.4091

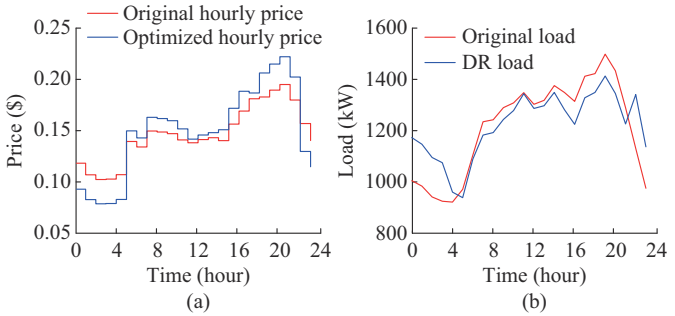

Fig. 12. Simulation results of electricity price optimization and DR module. (a) Electricity price optimization results with DR behavior simulation. (b) Aggregate system loads with and without electricity price optimization and DR.

V. Conclusion

In this paper, we propose a new DR behavioral learning method that overcomes inherent drawbacks in model-based and traditional learning-based methods. With the generated adversarial training and RL-based framework, the proposed method can comprehensively consider customers' electricity consumption habits and economic revenues. We define and predict the PPI to indicate customers' current DR patterns and propose a condition regulation method to improve the performance of the natural GAIL in tracing the DR pattern dynamics. The proposed DR behavioral learning method is applied in a price optimization problem for load fluctuation reduction and profit-maximizing. Besides, the FTRL algorithm is utilized to dynamically update the parameters of the proposed DR behavioral learning method with continuously collected data. From the case study, we can conclude as follows.

1) The numerical results show that the proposed DR behavioral learning method has lower deviations to the customers' real DR behaviors than model-based methods, and reduces more electricity charge with limited deviation increase than LSTM based methods. Lower deviations and electricity charges contribute to increasing the DR policies' adoption probability. Therefore, the effectiveness of the proposed method and its flexibility in considering multiple DR objectives can be verified.

2) The proposed RCGAIL method greatly improves the performance of natural GAIL in tracing customers' time-varying DR patterns. With the FTRL algorithm, the adjustments of RCGAIL parameters are completed rapidly when customers change their DR patterns, which ensures the effectiveness of the DR pattern tracing results.

3) By using the proposed DR behavioral learning method and electricity price optimization model, PUC succeeds in enlarging daily profits, and the fluctuation and peak of system aggregate loads are also reduced effectively.

References

[1] F. Zeng, Z. Bie, S. Liu *et al*., "Trading model combining electricity, heating, and cooling under multi-energy demand response," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 1, pp. 133-141, Mar. 2020.

[2] Y. Chai, Y. Xiang, J. Liu *et al*., "Incentive-based demand response model for maximizing benefits of electricity retailers," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 6, pp. 1644-1650, Nov. 2019.

[3] Y. Wang, I. R. Pordanjani, and W. Xu, "An event-driven demand response scheme for power system security enhancement," *IEEE Transactions on Smart Grid*, vol. 2, no. 1, pp. 23-29, Mar. 2011.

[4] D.S. Kirschen, G. Strbac, P. Cumperayot *et al*., "Factoring the elasticity of demand in electricity prices," *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 612-617, May 2000.

[5] C. Dou, X. Zhou, T. Zhang *et al*., "Economic optimization dispatching strategy of microgrid for promoting photoelectric consumption considering cogeneration and demand response," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 3, pp. 557-563, Jun. 2020.

[6] G. Samuel and S. Krumdieck, "Scenario analysis of residential demand response at network peak periods," *Electric Power Systems Research*, vol. 93, pp. 32-38, Dec. 2012.

[7] S. P. Roukerd, A. Abdollahi, and M. Rashidinejad, "Probabilistic-possibilistic flexibility-based unit commitment with uncertain negawatt demand response resources considering Z-number method," *International Journal of Electrical Power & Energy Systems*, vol. 113, pp. 71-89, Dec. 2019.

[8] D. Gao, Y. Sun, and Y. Lu, "A robust demand response control of commercial buildings for smart grid under load prediction uncertainty," *Energy*, vol. 93, pp. 275-283, Dec. 2015.

[9] K. Kia and E. Bitar, "Risk-sensitive learning and pricing for demand response," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6000-6007, Nov. 2018.

[10] Y. Wang, H. Liang, and V. Dinavahi, "Two-stage stochastic demand response in smart grid considering random appliance usage patterns," *IET Generation Transmission & Distribution*, vol. 12, no. 18, pp. 4163-4171, Oct. 2018.

[11] H. Wu, M. Shahidehpour, and A. Al-bdulwahab, "Hourly demand response in day-ahead scheduling for managing the variability of renewable energy," *IET Generation Transmission & Distribution*, vol. 7, no. 3, pp. 226-234, Mar. 2013.

[12] C. Zhao, J. Wang, J. Watson *et al*., "Multi-stage robust unit commitment considering wind and demand response uncertainties," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2708-2717, Aug. 2013.

[13] D. Zhang, S. Li, M. Sun *et al*., "An optimal and learning-based demand response and home energy management system," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1790-1801, Jul. 2016.

[14] H. Xu, H. Sun, D. Nikovski *et al*., "Learning dynamical demand response model in real-time pricing program," in *Proceedings of Energy Society Innovative Smart Grid Technologies Conference*, Washington D.C., USA, Feb. 2019, pp. 1-8.

[15] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach," *Applied Energy*, vol. 220, no. 15, pp. 220-230, Jun. 2018.

[16] L. A. Hurtado, E. Mocanu, P. H. Nguyen *et al*., "Enabling cooperative behavior for building demand response based on extended joint action learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 1, pp. 127-136, Jan. 2018.

[17] B. Shahab, V. W. S. Wong, and J. Huang, "An online learning algorithm for demand response in smart grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4712-4725, Sept. 2018.

[18] B. G. Kim, Y. Zhang, M. V. D. Schaar *et al*., "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2187-2198, Sept. 2016.

[19] V. Hasselt, H. A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, USA, Feb. 2016, pp. 1-13.

[20] V. R. Mnih, K. Kavukcuoglu, D. Silver *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-541, Feb. 2015.

[21] S. M. Kakade, "A natural policy gradient," in *Proceedings of NIPS2001 14th Neural Information Processing Systems*, Vancouver, Canada, Dec. 2001, pp. 1-6.

[22] J. Peters, S. Vijayakumar, and S. Schaal, "Natural Actor-Critic," in *Proceedings of ECML 2005, 16th European Conference on Machine Learning*, Porto, Portugal, Oct. 2005, pp. 280-291.

[23] R. Deng, Z. Yang, F. Hou *et al.*, "Distributed real-time demand response in multiseller-multibuyer smart distribution grid," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2364-2374, Sept. 2015.

[24] M. H. Yaghmaee, A. Leon-Garcia, and M. Moghaddassian, "On the performance of distributed and cloud-based demand response in smart grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5403-5417, Sept. 2018.

[25] A. Jiang, H. Wei, J. Deng *et al.*, "Cloud-edge cooperative model and closed-loop control strategy for the price response of large-scale air conditioners considering data packet dropouts," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4201-4211, Sept. 2020.

[26] Z. Tan, P. Yang, and A. Nehorai, "An optimal and distributed demand response strategy with electric vehicles in the smart grid," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 861-869, Mar. 2014.

[27] S. Khemakhem, M. Rekik, and L. Krichen, "Double layer home energy supervision strategies based on demand response and plug-in electric vehicle control for flattening power load curves in a smart grid," *Energy*, vol. 167, no. 15, pp. 312-324, Jan. 2019.

[28] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Proceedings of Advances in Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, Dec. 2016, pp. 4565-4573.

[29] M. Mirza and S. Osindero. (2014, Nov.). Conditional generative adversarial nets. [Online]. Available: https://arxiv.org/abs/1411.1784

[30] K. Cho, B. V. Merrienboer, C. Gulcehre *et al.* (2014, Sept.). Learning phrase representations using RNN encoder-decoder for statistical machine translation. [Online]. Available: https://arxiv.org/abs/1406.1078

[31] D. Silver, G. Lever, N. Heess *et al*., "Deterministic policy gradient al-

gorithms," in *Proceedings of the International Conference on Machine Learning*, Beijing, China, Jun. 2014, pp.1-9.
[32] H. B. McMahan, G. Holt, D. Sculley *et al*., "Ad click prediction: a view from the trenches," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago，USA, Aug. 2013, pp. 1-9.
[33] PJM INT. (2013, Jan.). Data base of energy market. [Online]. Available: https://www.pjm.com/markets-and-operations/energy.aspx
[34] National Renewable Energy Laboratory of the U. S. (2013, Jan.). Data base of hourly loads. [Online]. Available: https://openei. org/datasets/files/961/pub/
[35] F. L. Quilumba, W. Lee, H. Huang *et al*., "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918, Mar. 2015.
[36] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting-a novel pooling deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271-5280, Sept. 2018.
[37] X. Wang, Y. Wang, J. Wang *et al*., "Residential customer baseline load estimation using stacked autoencoder with pseudo-load selection," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 1, pp. 61-70, Jan. 2020.
[38] R. Lu, R. Bai, Y. Huang *et al.*, "Data-driven real-time price-based demand response for industrial facilities energy management," *Applied Energy*, vol. 283, p. 116291, Feb. 2021.
[39] J. Wang, H. Zhong, X. Lai *et al.*, "Distributed real-time demand response based on Lagrangian multiplier optimal selection approach," *Applied Energy*, vol. 190, no. 15, pp. 949-959, Mar. 2017.
[40] X. Huang, S. H. Hong, and Y. Li, "Hour-ahead price based energy management scheme for industrial facilities," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2886-2898, Dec. 2017.
[41] H. Yang, J. Zhang, J. Qiu *et al.*, "A practical pricing approach to smart grid demand response based on load classification," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 179-190, Jan. 2018.
[42] S. Yang, Z. Tan, Z. Liu *et al.*, "A multi-objective stochastic optimization model for electricity retailers with energy storage system considering uncertainty and demand response," *Journal of Cleaner Production*, vol. 277, no. 20, pp. 1-17, Dec. 2020.

Junhao Lin received the B.S. degree in electrical engineering from Southeast University, Nanjing, China, in 2015. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering of Shanghai Jiao Tong University, Shanghai, China. His research interests include application of artificial intelligence in power systems.

Yan Zhang received the B.E. degree in power plants and electric power systems from Hefei University of Technology, Hefei, China, in 1982, the M.E. degree in high-voltage engineering from China Electric Power Research Institute, Beijing, China, in 1987, and the Ph.D. degree in electric power systems and their automation from Shanghai Jiao Tong University, Shanghai, China, in 1998. She is currently a Professor with the Department of Electrical Engineering, Shanghai Jiao Tong University. Her research interests include power system stability, planning and reliability.

Shuangdie Xu received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2017 and 2020. Her research interests include smart distribution networks and power system reliability.