Energy 205 (2020) 118019

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Modified deep learning and reinforcement learning for an incentive-based demand response model

Lulu Wen [a, b, c], Kaile Zhou [a, b, *], Jun Li [d], Shanyong Wang [e]

[a] School of Management, Hefei University of Technology, Hefei, 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making of Ministry of Education, Hefei University of Technology, Hefei, 230009, China
[c] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, CA, 94720, USA
[d] School of Economics, Hefei University of Technology, Hefei, 230009, China
[e] University of Science and Technology of China, Hefei, 230026, China

ARTICLE INFO

Article history:
Received 13 February 2020
Received in revised form
23 May 2020
Accepted 30 May 2020
Available online 5 June 2020

Keywords:
Demand response
Modified deep learning
Reinforcement learning
Smart grid

ABSTRACT

Incentive-based demand response (DR) program can induce end users (EUs) to reduce electricity demand during peak period through rewards. In this study, an incentive-based DR program with modified deep learning and reinforcement learning is proposed. A modified deep learning model based on recurrent neural network (MDL-RNN) was first proposed to identify the future uncertainties of environment by forecasting day-ahead wholesale electricity price, photovoltaic (PV) power output, and power load. Then, reinforcement learning (RL) was utilized to explore the optimal incentive rates at each hour which can maximize the profits of both energy service providers (ESPs) and EUs. The results showed that the proposed modified deep learning model can achieve more accurate forecasting results compared with some other methods. It can support the development of incentive-based DR programs under uncertain environment. Meanwhile, the optimized incentive rate can increase the total profits of ESPs and EUs while reducing the peak electricity demand. A short-term DR program was developed for peak electricity demand period, and the experimental results show that peak electricity demand can be reduced by 17%. This contributes to mitigating the supply-demand imbalance and enhancing power system security.

1. Introduction

Since the participators of wholesale electricity market usually have difficulty in acquiring detailed market information, end users (EUs) usually face the market uncertainty and the financial risks of buying electricity at real-time prices [1]. Thus, energy service providers (ESPs) which can monitor the market effectively, start to participate into wholesale market on behalf the EUs. As the connectors between grid operator (GO) and EUs, ESPs purchase electricity from GO and sell electricity to EUs for profits. They can help EUs reduce energy expenditure and reduce the risks of market uncertainties, but they always face the revenue risk [2]. In addition, the imbalance between electricity supply and demand can bring high risks for power system operation [3]. Therefore, ESPs have to establish effective measures to manage their financial risks, and they should also take the responsibility of keeping supply-demand balance to improve power system stability and reliability [4].

Demand response (DR) is an important strategy to address these problems. In DR, EUs would change their electricity consumption patterns by reducing or shifting electricity demand when they receive induced signals from ESP [5]. In smart grid, advanced metering infrastructure can support bidirectional communication among GO, ESP and EUs [6]. Besides, smart appliances that can be scheduled to perform their tasks during a time period specified by EUs, have been widely adopted in buildings [7]. This has made DR to be a promising way to promote supply-demand balance, improve power system stability, as well as reduce financial risks of ESP. In general, DR can be divided into price-based DR and incentive-based DR. Price-based DR is to shift EUs' peak electricity demand to off-peak periods by implementing specific tariffs including time-of-use (TOU) pricing, critical peak pricing, and real-time pricing [8]. Rahmani-Andebili and Shen [9] investigated a price-based energy management method to minimize the cost of generation company. To reduce the difference of electricity demand between peak and valley periods, a hybrid price-based DR program was proposed in

* Corresponding author. School of Management, Hefei University of Technology, Hefei, 230009, China.
E-mail address: zhoukaile@hfut.edu.cn (K. Zhou).

https://doi.org/10.1016/j.energy.2020.118019
0360-5442/

Ref. [10]. Srinivasan et al. [11] developed a dynamic pricing strategy based on game theory and achieved peak electricity demand reduction by 10% and 5% for residential and commercial users respectively.

Incentive-based DR programs generally present certain rewards for EUs to reduce their electricity demand during specific periods [12]. They can be dispatched more flexibly than price-based DR programs, and thereby ESP can get the required DR resources more easily. In addition, EUs have high initiative to participate in the incentive-based DR programs since they can get the rewards more directly [13]. For example, in 2017, the peak electricity demand reductions by incentive-based DR programs accounted for 93% of the total peak electricity demand reduction in USA [14]. Up to now, many research efforts have focused on developing different incentive-based DR programs. Rahmani-Andebili established a nonlinear incentive-based DR model and a nonlinear price-based DR model which were implemented in four different power markets [15]. It revealed that the DR programs can shift part of the electricity demand and save energy, while may lead to different demand pattern with different responsive load behavioral model. In Ref. [16], a nonlinear incentive-based DR and TOU-based DR were used in unit commitment problem by considering the nonlinear behaviors of residential customers. Erdinc et al. [17] proposed a credits-based incentive approach for EUs to decrease the critical load demands and maintain the balance between supply and demand during peak electricity demand periods, wherein the uncertainty of the ambient temperature variations were taken into account. Li et al. [18] developed a dynamic coupon incentive-based DR program for distributed energy system with multiple load aggregators, and a fairness function was defined to guarantee that aggregators are rewarded. In Ref. [19], an improved incentive-based DR model was presented, in which the elasticity depends not only on the electricity price, but also on time and customer type. Besides, Yu et al. [20] proposed an incentive-based DR model from the perspective of GO to dispatch DR resources, and then a two-loop Stackelberg game was constructed to capture interactions among GO, ESP, and EUs. In addition, in Ref. [21], the incentives for drivers of plug-in electric vehicles were optimized by fleet management, and the charging demand was transferred to off-peak period from peak period.

However, most of the existing related studies are based on traditional model-based methods, including stochastic programming, game theory, and mixed integer linear programming. It is difficult to choose appropriate models for an actual energy system and identify corresponding parameters that are EUs-dependent. Besides, many assumptions were made in these model-based methods which may be not applicable to real-world situations. In addition, few studies have considered the uncertainty of both electricity price and power load when developing incentive-based DR program. Therefore, some traditional models may be ineffective in volatile market environment. Also, some existing research work just focused on a single EU and ignored the impact of multiple EUs with different characteristics on the incentive-based DR.

In recent years, artificial intelligence (AI) based methods have been widely used in decision making problems [22]. Reinforcement learning (RL) is a typical AI-based method, which has excellent performance in solving the problem of maximizing returns or achieving specific goals through learning strategies in the interaction between agents and environment [23]. RL is a model-free algorithm, and it has been widely used in smart energy management [22,24]. For example, in Ref. [25], RL was applied to control the building climate under dynamic pricing, in which the sequential decision making problem was converted to a Markov decision problem. Arif et al. [26] investigated the load scheduling of plug-in electric vehicles using RL considering different dynamic price schemes, thereby to minimize the total energy costs while meeting the electricity demand of users. Wang et al. [27] proposed a RL-based model to investigate the optimal energy trading among microgrids, where each microgrid chooses a strategy individually and randomly to trade the energy in an independent market and maximizes its average revenue. In addition, Mahapatra et al. [28] developed an improved RL algorithm to manage home appliances, with the aim of reducing the power load during peak period, promoting energy conserving, and reducing the carbon footprint of residential dwellings. In Ref. [29], a multi-agents RL was developed for a real-world smart grid scenario in which the forecasting method was used to identify non-stationary electricity demand of EUs. To deal with the uncertainty of electricity prices, Lu et al. [30] forecasted market price using artificial neural network (ANN), and then a multi-agents RL was used to make optimal decisions for the operation of different home appliances. Similarly, in Ref. [31], the power load and wholesale electricity price were forecasted using ANN, and then the optimal incentives were acquired by RL based on the forecasting results. However, few of above studies have considered the uncertainties of users' load demand and the power purchase cost of ESP at the same time. The forecasting methods in these works also have certain limitation in obtaining more accurate results. Besides, the DERs and their uncertainties were rarely taken into account when developing incentive-based DR models.

Therefore, to bridge these research gaps, this study proposed an incentive-based DR model based on modified deep learning (MDL) and RL, enabling the ESP to get required DR resources with minimum costs and reducing the electricity bills of EUs. As a result, it can promote supply-demand balance and improve power system reliability. To reduce the uncertainties of external environment, a modified deep learning model based on recurrent neural network (MDL-RNN) was developed to forecast day-ahead wholesale electricity price, PV power output, and power load respectively. Then, RL was used to obtain the optimal incentive rate for each EU while maximizing the total profits of ESP and EUs. The contributions of this study are as follows: (1) An incentive-based DR model based on MDL and RL was proposed and the optimal incentive rate at each hour can be autonomously learned by the interactions between the ESP and EUs; (2) The uncertainties of environment were considered by accurate forecasting of the day-ahead wholesale market price, PV power output, and EUs power load with the proposed MDL-RNN model, which achieved better performance; (3) The diversity of EUs was taken into consideration and a theoretical proof on the relationship between the attitudes of EUs towards energy saving and the optimal incentive rates was given; (4) A short-term DR program was presented for peak electricity demand period to enhance the security of power system operation.

The remainder of the study is organized as follows. Section 2 presents the incentive-based DR model. Section 3 introduces the proposed MDL-RNN model for forecasting wholesale market price, PV power output, and EUs' power load. Section 4 provides the RL method to determine the optimal incentive rates at each hour. The experimental results and discussions are given in Section 5. Section 6 presents the conclusions.

2. Demand response model

Fig. 1 shows the schematic of a hierarchical electricity market. There are mainly three kinds of participators in the electricity market, i.e. GO, ESP, and EUs. ESP purchases electricity from GO according to wholesale electricity price, and then sells electricity to EUs at retail electricity price. Meanwhile, ESP has to help GO to ensure electricity supply-demand balance and reduce peak electricity demand while pursuing maximum profit.

As shown in Fig. 1, the incentive-based DR program is

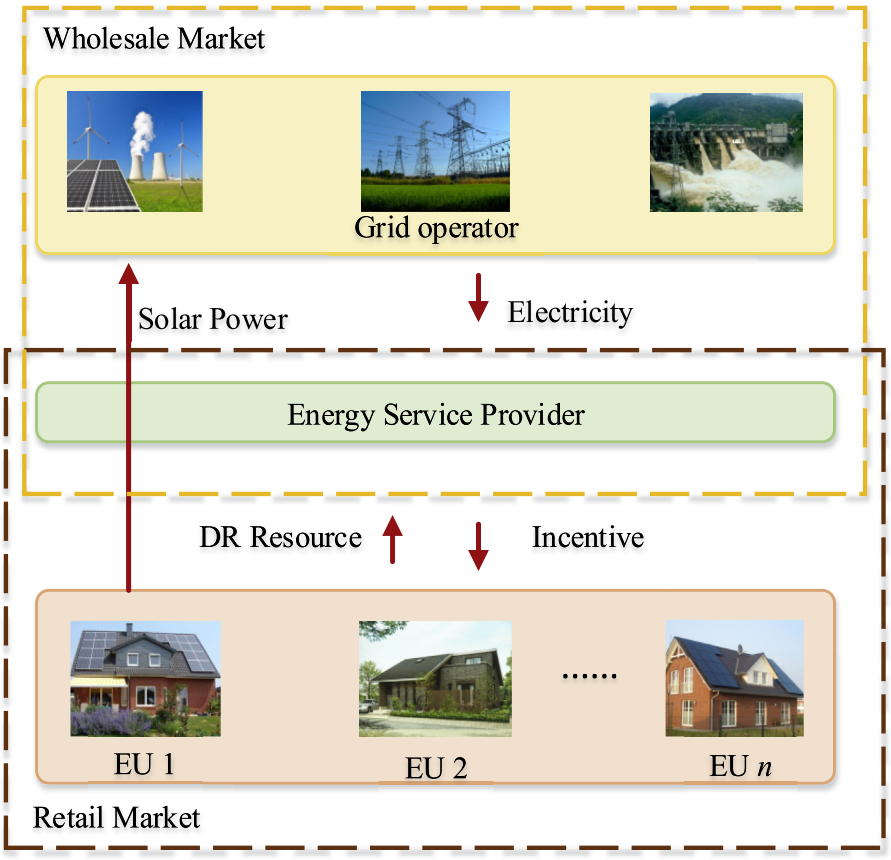


Fig. 1. Schematic of a hierarchical electricity market.

implemented in retail market to reduce peak electricity demand for the better balance of supply and demand, while maximizing the total profits of ESP and EUs. The solar power also was considered in the incentive-based DR program.

2.1. Energy service provider's profits

By giving EUs certain rewards, ESP can procure required DR resources from EUs, and reduce the cost of electricity purchasing from wholesale market. Here, the reduced electricity purchase cost minus the rewards for EUs is regarded as the profit of ESP, as described in Eq. (1). Therefore, ESP can obtain maximum profits by setting optimal incentive rates for EUs.

$$profit_{esp} = \sum_{i=1}^{n}\sum_{j=1}^{h}\left(p_j - \alpha_{ij}\right)\Delta E_{ij} \tag{1}$$

where i represents the i-th EU, n is the total number of EUs, j denotes j-th hour, h is the last hour of a day, p_j is the wholesale electricity price at hour j, ΔE_{ij} is the reduced electricity demand of user i at hour j in response to the incentives, and α_{ij} is the incentive rate for EU i at hour j. In addition, α_{ij} cannot be less than $\alpha_{\min}$ or exceed $\alpha_{\max}$, and this constraint is usually determined by electricity trading market and regulatory authority [32].

2.2. End users' profits

EUs can get rewards when they response to the incentive of ESP to reduce electricity demand, but they will also have to bear the discomfort caused by the reduction of electricity demand. Hence, as shown in Eq. (2), the profits of EUs mainly come from the difference between the electricity reduction rewards and the discomfort cost, and the income that EUs sell their residual solar power in wholesale electricity market. In particular, EUs were assumed to first consume solar power, and the incentive rate is set to 0 when EUs' electricity demand can be met by solar power.

$$profit_{eu} = \sum_{i=1}^{n}\sum_{j=1}^{h}\left[\lambda_i \cdot \alpha_{ij} \cdot \Delta E_{ij} - (1-\lambda_i)\cdot cost_{ij}\left(\Delta E_{ij}\right) + p_j \cdot PV_{ij}\right] \tag{2}$$

$$\Delta E_{ij} = E_{ij} \cdot \xi_j \cdot \frac{\alpha_{ij} - \alpha_{\min}}{\alpha_{\min}} \tag{3}$$

In Eq. (2), λ is the weight parameter which represents the attitude of EUs towards incentive reward and discomfort cost. A small λ_i denotes that the EU i prefer comfort than incentive reward. In contrast, a big λ_i represents that the EU i believe that incentive reward is more important than comfort. $cost_{ij}(\Delta E_{ij})$ is the discomfort cost of EU i at hour j when the electricity demand of EU i is reduced by ΔE_{ij}, and PV_{ij} represents the amount of solar power sold to GO by EU i at hour j. Here, p_j is the tariff of solar power feed-in to power system and is set to be equal to wholesale electricity price in this study. In Eq. (3), E_{ij} is the actual electricity demand of EU i at hour j, ξ_j is the elasticity coefficient that represents the ratio of electricity demand variation to incentive variation [33]. In addition, the electricity demand reduction ΔE_{ij} should be within a range [34]. It cannot exceed the upper bound $\Delta E_{\max}$ and be less than lower

Table 1
Inputs of the MDL-RNN model.

Forecasting targets	Inputs
EUs load	Month, week, day, hour, temperature, humidity, wind speed, historical load
PV power output	Month, day, hour, humid, temperature, global horizontal radiation, diffuse horizontal radiation, historical PV power output
Wholesale market price	Month, week, day, hour, holiday, historical wholesale market price

Table 2
Ranges of hyper-parameters.

Hyper-parameters	Range
Neuron number of layers 2, 3 and 4	[10, 200]
Neurons type of layer 2, 3	[LSTM, GRU, SimpleRNN]
Activation function of layer 4	[Sigmoid, ReLU]
Dropout of layer 2, 3, and 4	[0,1]

bound 0. $\Delta E_{\max}$ is determined by the characteristics of EUs and the capacity of power generation.

The discomfort cost of EU i at hour j is a quadratic function of ΔE_{ij}, which is defined as Eq. (4) [35].

$$cost_{ij}(\Delta E_{ij}) = \frac{\mu_i}{2}(\Delta E_{ij})^2 + \omega_i \cdot \Delta E_{ij} \tag{4}$$

In Eq. (6), μ_i and ω_i are discomfort parameters. They are positive and EUs-dependent [20]. For the same amount of electricity reduction, the discomfort degree of different EUs may be different. This is due to that different EUs have different household appliances and electricity consumption characteristics. Bigger μ_i and ω_i mean that EU i will suffer from more discomfort even when reducing the same electricity demand as EUs who have smaller μ and ω.

Next, the objective of the DR model is to maximize the total profits of ESP and EUs, as shown in Eq. (5).

$$\max(profit_{esp} + profit_{eu}) \tag{5}$$

Then, the day-ahead incentive rates at each hour for every EU were optimized by MDL-RNN and RL.

3. Modified deep learning model for forecasting

To overcome the uncertainties of environment, an MDL-RNN model which has good forecasting performance, was proposed to forecast wholesale electricity price, PV power output, and EUs power load respectively.

3.1. Recurrent neural network

There have been some many research efforts on time series forecasting [36,37]. As a variant of ANN, recurrent neural network (RNN) is a recursive neural network which adopts sequence data as input, and it has been more and more used in time series forecasting [38]. RNN has the property of memory, parameters sharing, and Turing completeness, so it can learn the non-linear characteristics of time series data with high efficiency [39]. However, simple RNN is vulnerable to gradient explosion or vanishing during training. To deal with these problems, some novel cells were proposed to replace original neurons in RNN, such as long short-term memory (LSTM) unit [40] and gated recurrent unit (GRU) etc. [41].

3.2. Modified deep learning model based on RNN (MDL-RNN)

Considering the superior ability of deep learning model in modeling nonlinear relationship and the excellent performance of RNN in learning time dependencies, an MDL-RNN model is proposed to forecast wholesale market price, PV power output, and EUs power load.

3.2.1. Inputs and model structure

The selection of input variables is critical for improving the forecasting accuracy [42]. In this paper, the inputs of the MDL-RNN model were chosen based on the guideline of previous work [43], and they were also limited by the availability of related data. The inputs include three kinds of variables: time variables (month, week, day, hour, and holiday), environment variables (temperature, humidity, wind speed, global horizontal radiation, and diffuse horizontal radiation), and historical variable. The inputs of the MDL-RNN model for forecasting different targets are shown in Table 1.

To obtain more accurate forecasting results, the MDL-RNN adopts a 5-layers network which includes one input layer, two GRU-RNN layers, one simple hidden layer, and one output layer. It has been demonstrated that multi-layers networks usually have better forecasting performance [44]. The input layer has 24 input vectors which once adopts 24-h time series data, and each input

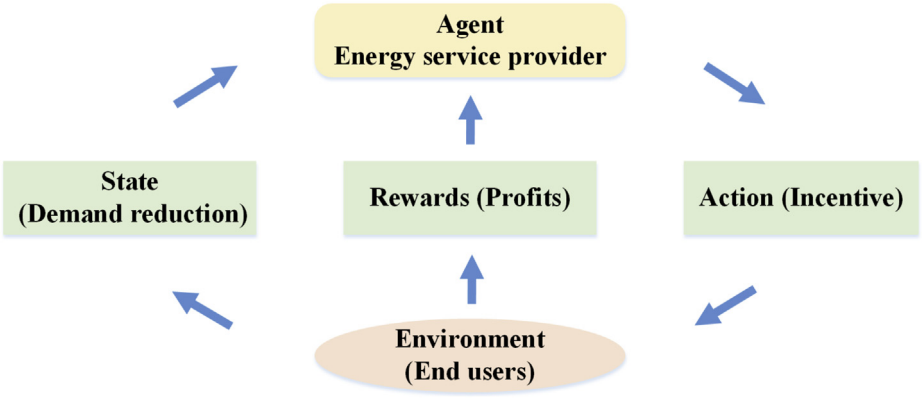


Fig. 2. Schematic of RL to find the optimal incentive rates.

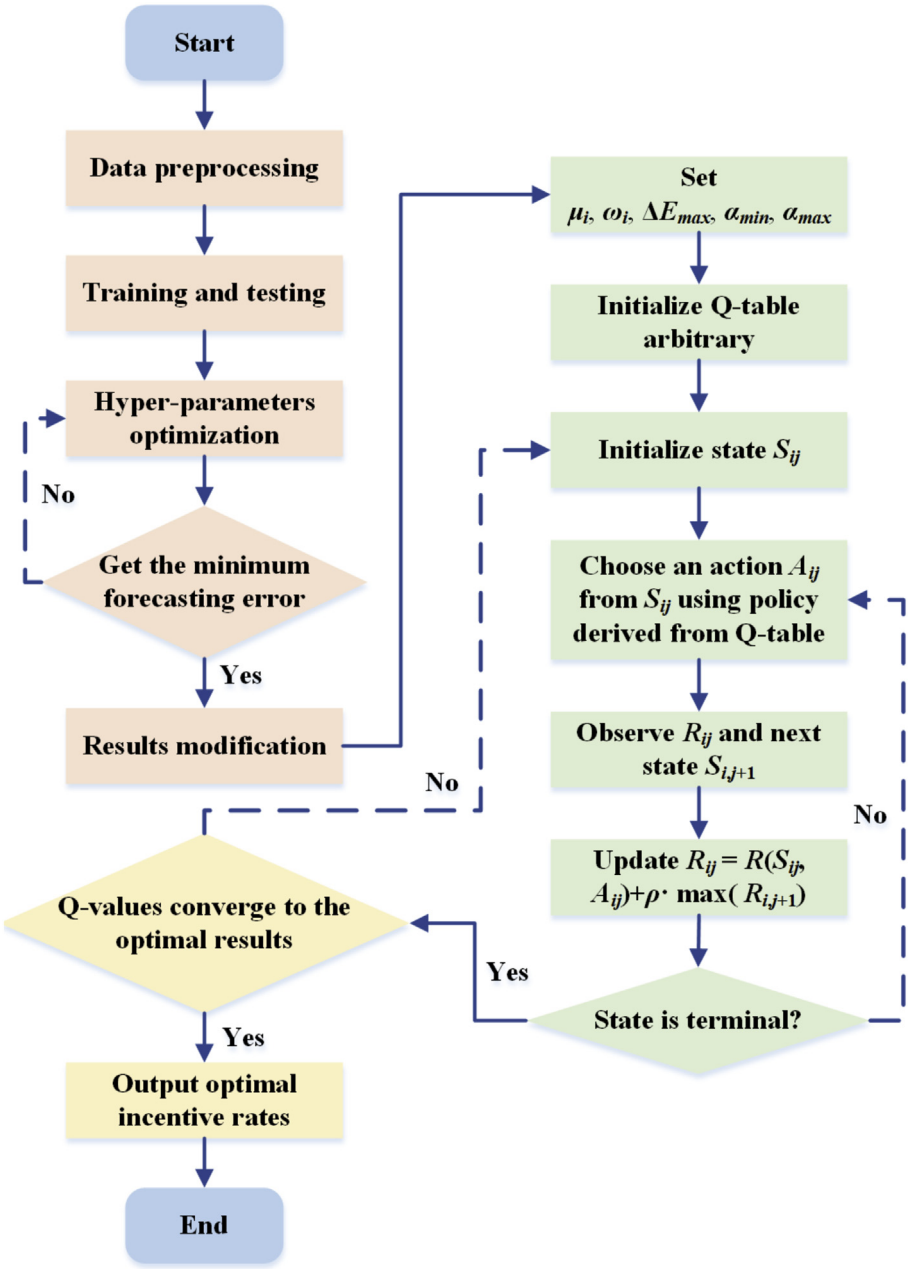


Fig. 3. Flowchart of IDR-MDLRL algorithm.

vector includes time variables, environment variables, and historical variable. This has improved forecasting accuracy because the periods of wholesale electricity price, PV power output, and EUs power load are 24-h.

3.2.2. Model setup

In order to obtain better forecasting results, some setups were made in the training process of MDL-RNN model. First, the training data was preprocessed by min-max normalization to eliminate dimension and improve calculation efficiency [45]. For a given variable x, its normalization can be expressed as

$$x'_k = \frac{x_k - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

where x_k is a value of variable x, x_{min} is the maximum value of x, and x_{max} is the minimum value of x.

Then, weight decay regularization was used to solve the overfitting problem by adding regularization term to cost function [46]. Meanwhile, dropout was also employed in the MDL-RNN model [47]. At last, the Adam gradient descent algorithm was used to train the MDL-RNN model through back-propagation due to its fast convergence speed [48].

In addition, there are some other hyper-parameters that need to be set, such as neuron number of layer 2 and layer 3, the neurons type of layer 2 and layer 3, the activation function of layer 4, and the dropout of layer 2, layer 3, and layer 4. Table 2 shows the range of these hyper-parameters.

Next, a global optimized method Hyperopt was used to determine the most suitable value of hyper-parameters until the optimal MDL-RNN model was obtained.

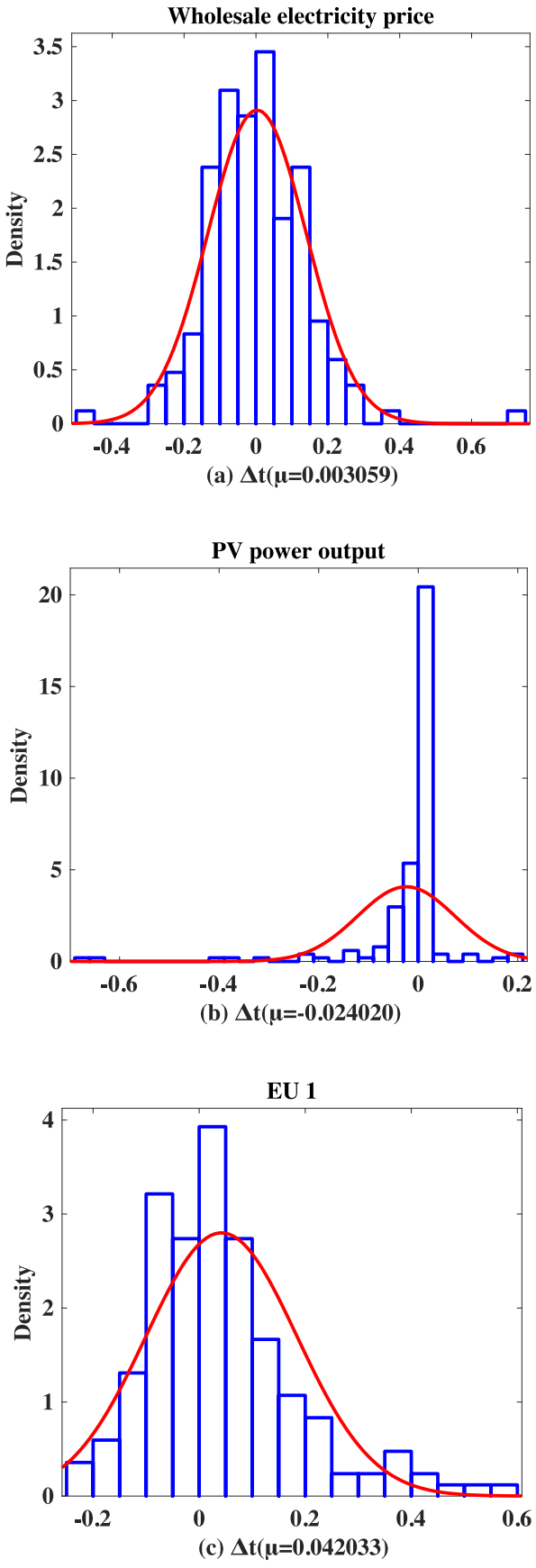


Fig. 4. Distributions of Δ_t in forecasting wholesale electricity price, PV power output, and EU 1's power load.

3.2.3. Modified forecasting results

The forecasting results affect the optimization of incentive rates. However, absolutely accurate forecasting results cannot be obtained. Therefore, a modified method was employed to further improve the forecasting performance. As shown in Eq. (7), Δ_t was assumed to obey normal distribution.

$$\Delta_t = \frac{y_{t,\ true} - y_{t,\ forecast}}{y_{t,\ forecast}} \sim N\left(\mu, \sigma^2\right) \tag{7}$$

Then, the modified forecasting results can be obtained as shown in Eq. (8).

$$y_{t,true} = (1+\Delta_t) \times y_{t,forecast} = E\left[(1+\Delta_t) \times y_{t,forecast}\right] = E\Big[\times (1+\Delta_t) \times y_{t,forecast}\Big] = (1+\mu) \times y_{t,forecast} \tag{8}$$

In real scenarios, the final forecasting model was obtained by learning from historical data. When the related variables are input into the MDL-RNN model, the future targets (i.e. wholesale market price, PV power output, and power load) can be output. However, the actual future targets cannot be acquired in advance, and thus the distribution of Δ_t cannot be identified. Therefore, the distribution of Δ_t in training process was used to replace it.

3.2.4. Evaluation metrics

The forecasting results were evaluated by two common metrics, i.e. mean absolute error (MAE) and mean absolute percentage error (MAPE). Their definitions are shown in Eq. (9) and Eq. (10) respectively.

$$\mathrm{MAE} = \frac{\sum_{t=1}^{T}\left|y_{t,true} - y_{t,forecast}\right|}{T} \tag{9}$$

$$\mathrm{MAPE} = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{y_{t,true} - y_{t,forecast}}{y_{t,true}}\right| \tag{10}$$

In Eq. (9) and Eq. (10), t represents the time step, T is the total time step, $y_{t,true}$ is the true value at time step t, and $y_{t,forecast}$ is the forecasting value at time step t.

4. Reinforcement learning for incentive rates optimization

As an agent-based machine learning method, RL can learn the optimal actions (i.e. optimal policy) by the iterations between agent and environment. With the optimal policy, the agent can get the biggest reward. In this paper, RL was used to explore the optimal incentive rates at each hour of a day to obtain the maximal total profits of ESP and EUs. Fig. 2 shows the schematic of RL to find the optimal incentive rates.

In Fig. 2, the agent represents ESP and the environment represents EUs. When ESP gives EUs incentive (i.e. action), EUs response to the incentive and reduce their electricity demand. Then, ESP will get the reduced electricity demand from EUs (i.e. state), and the profits of ESP and EUs (i.e. reward) can be obtained. Then, the iterations will continue until reaching the maximal profits of ESP and EUs.

Generally, RL can be formalized as a Markov Decision Process (MDP) which contains three elements: state $S_{ij} \in S(\Delta E_{ij})$, action $A_{ij} \in A(\alpha_{ij})$, and reward $R_{ij}(S_{ij}, A_{ij})$, where i is the i-th EU, j is the j-th hour of a day, S_{ij} is the reduced electricity demand of EU i at hour j, $A_{ij}$is the incentive rate for EU i at hour j, and $R_{ij}(S_{ij}, A_{ij})$ is the current reward of EU i at hour j. In RL model, the state transition only relies on the current state and current action, and thus the profits and the

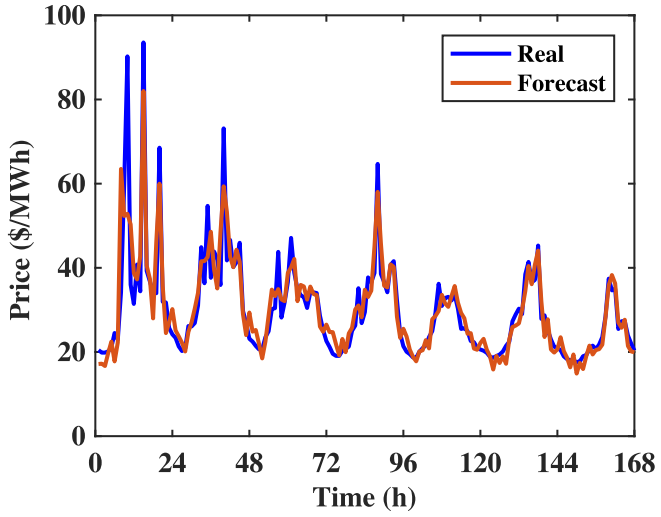


Fig. 5. Forecasting results of wholesale electricity price from July 23, 2018 to July 29, 2018.

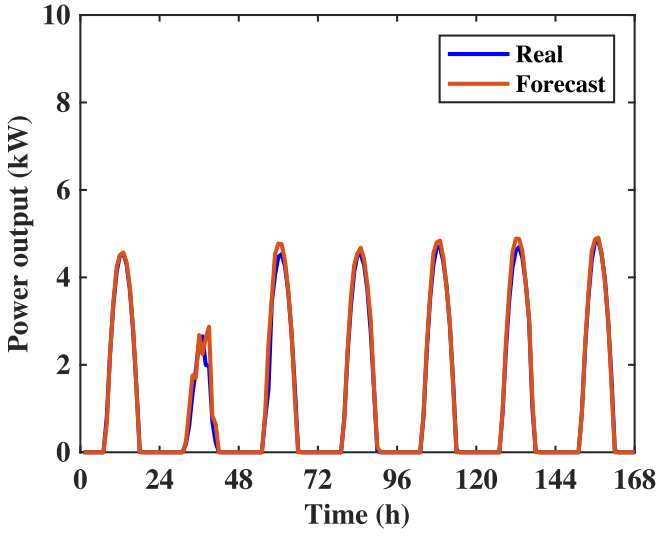


Fig. 6. Forecasting results of PV power output from July 23, 2018 to July 29, 2018.

electricity demand reduction only rely on the current incentive rate. Then, the MDP can be expressed as Eq. (11) in which H denotes the final hour of a day, $R_{i1}(S_{i1}, A_{i1})$ is the reward of EU i at hour 1, and $R_{ih}(S_{ih}, A_{ih})$ is the reward of EU i at the final hour of a day.

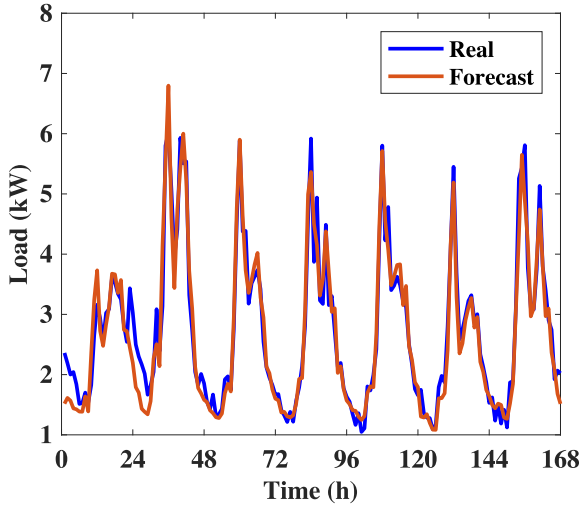
$$S_{i1}, A_{i1}, R_{i1}(S_{i1}, A_{n1}); S_{i2}, A_{i2}, R_{i2}(S_{i2}, A_{i2}) \cdots\cdots S_{i,h}, A_{i,h}, R_{ih}(S_{ih}, A_{ih}). \tag{11}$$

Considering the long-term returns, future reward has to be taken into account besides current reward [24]. The future reward will decay at a discount rate ρ, so the cumulative discounted reward of EU i at the first hour and hour l can be described as Eq. (12) and Eq. (13) respectively.

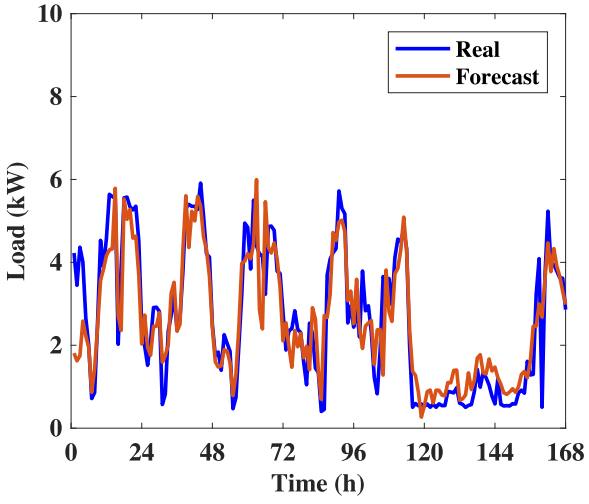
$$R_{i1} = R_{i1}(S_{i1}, A_{i1}) + \rho \cdot R_{i2}(S_{i2}, A_{i2}) + \cdots + \rho^{h-1} \cdot R_{ih}(S_{ih}, A_{ih}) \tag{12}$$

$$R_{il} = R(S_{il}, A_{il}) + \rho \cdot R(S_{i,l+1}, A_{i,l+1}) + \cdots + \rho^{h-l} \cdot R(S_{ih}, A_{ih}) \tag{13}$$

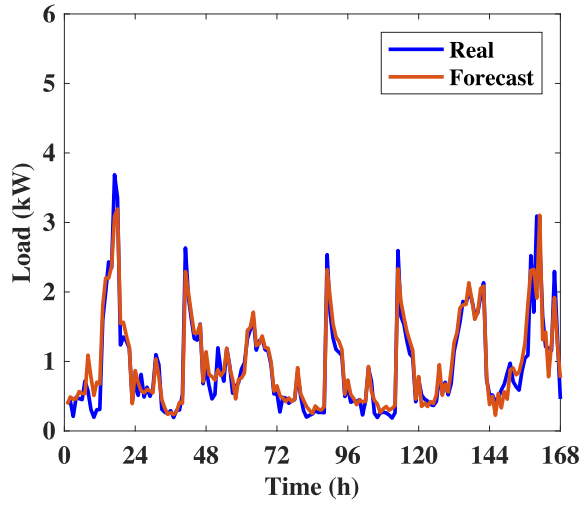
In Eq. (12) and Eq. (13), $\rho \in [0, 1]$ indicates that agent pursues



(a) EU 1



(b) EU 2



(c) EU 3

Fig. 7. Power load forecasting results of three EUs from July 23, 2018 to July 29, 2018.

Table 3
Quantitative evaluation of forecasting results.

	MDL-RNN		ANN		SVM		ELM	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Price	3.31	0.096	5.88	0.153	8.19	0.391	6.28	0.159
PV	0.11	0.039	0.17	0.106	0.16	0.101	0.29	0.249
Load 1	0.25	0.129	0.55	0.173	0.59	0.185	0.36	0.131
Load 2	0.14	0.111	0.69	0.379	0.72	0.385	0.38	0.161
Load 3	0.28	0.137	0.36	0.382	0.38	0.394	0.31	0.154

Table 4
Parameters of RL model.

Parameters	Value
Discomfort parameter $\mu_1/\mu_2/\mu_3$	1/2/3
Discomfort parameter ω_i	1
Maximum electricity reduction$\Delta E_{\max}$	$0.3E_{ij}$
Minimum incentive rate $\alpha_{\min}$	$0.3p_{\min}$
Maximum incentive rate $\alpha_{\max}$	$p_{\min}$

Table 5
Electricity elasticity in different hours.

Period	Elasticity(ξ_j)
0 am–6 am, 22 pm–23 pm	0.5
7 am - 16 pm	0.3
17 pm–21 pm	0.1

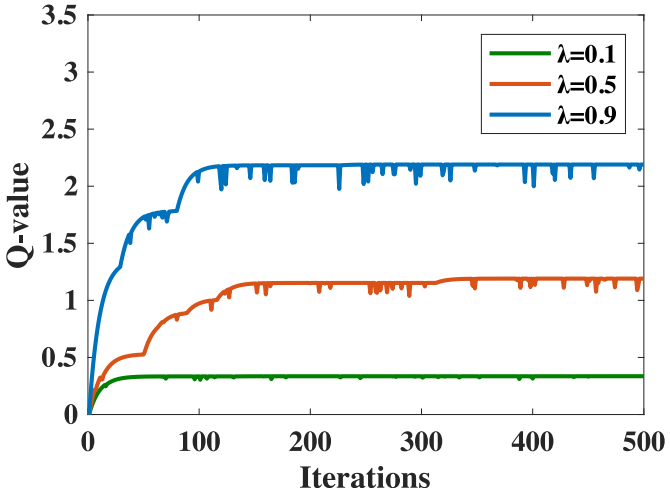


Fig. 8. Convergence of Q-value to acquire optimal incentive rates for EU 2 with different λ.

current reward or strive for future reward. When $\rho = 0$, the agent only consider the current reward, and when $\rho > 0$ the agent will consider both current and future rewards. In this paper, $\rho$is set as 0.9 which is a common value in RL modeling [49]. R_{nl} is the reward of EU i at hour l, S_{il} is the electricity demand reduction of EU i at hour l, A_{il} is the incentive rate for EU i at hour l, and $R(S_{il}, A_{il})$ is the current reward of EU i at hour l.

The solution of MDP is to find the optimal policy which can maximize the cumulative discounted reward, and then the optimal actions and states corresponding to the optimal policy can be obtained. Therefore, Eq. (12) can be transformed into Eq. (14).

$$R_{ij} = R(S_{ij}, A_{ij}) + \rho \cdot \max R_{i,j+1} \tag{14}$$

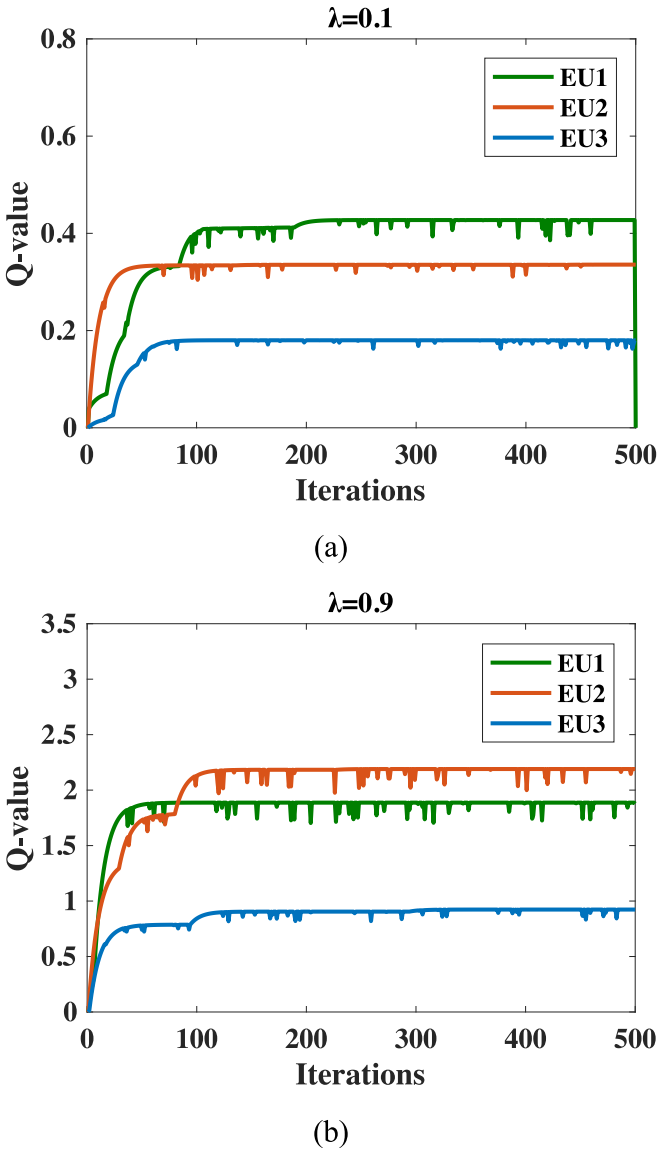
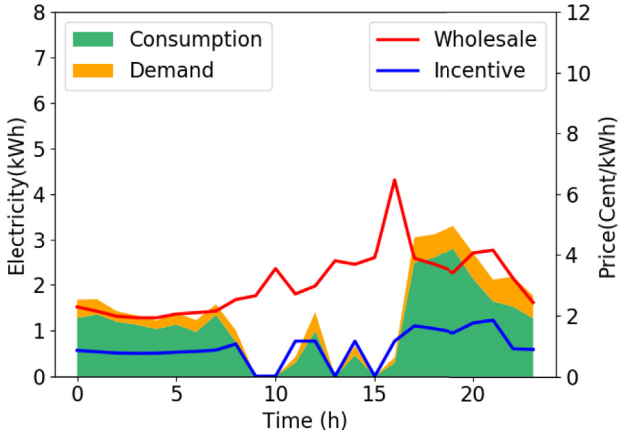


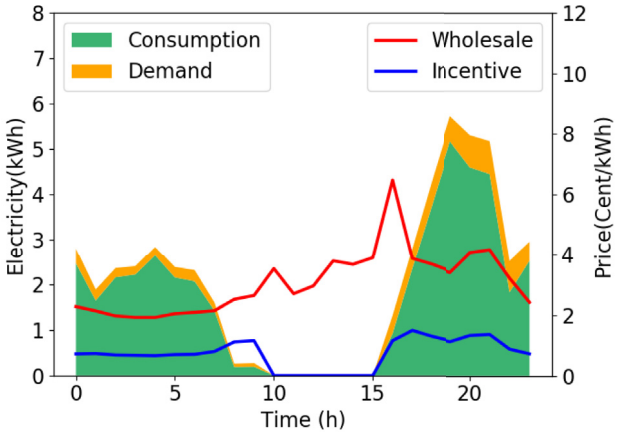
Fig. 9. Convergence of Q-value to acquire optimal incentive rates for different EUs.

Eq. (14) satisfies the Bellman equation which is usually solved by finding approximate solution [50]. The approximate solution can be found through policy based approaches (e.g. policy gradient algorithm) or value based approaches (e.g. Q-learning and Sarsa) [51]. Some other methods combining policy gradient based algorithm and value based algorithm can also solve the Bellman equation (e.g. Actor-Critic) [52]. In addition, these methods can be divided into model-based approach and model-free approach according to whether the dynamics of the system (i.e. transition probabilities between states) are known [53]. In the model-free method, the agent finds the optimal policy without determining transition probabilities between the states, but the model-based method is on the contrary. As for the RL model in this paper, the dynamics of the system are known. Thus, model-free method is more suitable to solve the model.

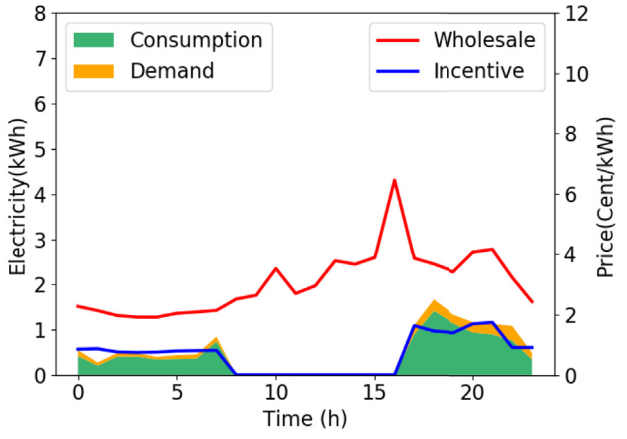
Q-learning is a model-free and off-policy method which has been widely used to solve RL model due to its simplicity [54]. Besides, it can learn from environment directly without knowing the environment, and that is why Q-learning was chosen to solve the proposed RL model. In a simple Q-learning example, a table is



(a) EU 1: λ_1=0.1, μ_1=1



(b) EU 2: λ_2=0.1, μ_2=2



(c) EU 3: λ_3= 0.1, μ_3= 3

Fig. 10. Results of different EUs when $\lambda = 0.1$.

established and the state-action values (i.e. Q-values) are stored in it. For a Q-value $Q(S_{ij}, A_{ij})$, it will be updated at each iteration to optimize the result until acquiring the maximum discounted reward R_{ij}. Then, Eq. (14) can be transformed into Eq. (15).

$$Q_{ij} = Q(S_{ij}, A_{ij}) + \rho \cdot \max Q_{i,j+1} \tag{15}$$

At each hour, the ESP provides an incentive rate for EUs. EUs will reduce their electricity demand for rewards, and then the corresponding Q-value (i.e. action and state) will be updated as Eq. (16).

$$Q_{ij} = (1-\theta)Q_{ij} + \theta\left[Q(s_{ij}, a_{ij}) + \rho \cdot \max Q_{i,j+1}\right] \tag{16}$$

In Eq. (16), $\theta \in [0, 1]$ is the learning rate which represents to what extent the new knowledge overrides old knowledge [24]. If $\theta = 0$, the agent learns nothing. If $\theta = 1$, all previous knowledge is lost. To balance the old knowledge and new knowledge, θ should be set as decimal between 0 and 1. In practice, θ is usually set to 0.1.

After several iterations, the Q-value will converge to the optimal value which corresponds to the optimal policy of MDP. Then, the optimal incentive rates at each hour for every EU can be obtained. The algorithm of the proposed incentive-based DR program based on MDL and RL (IDR-MDLRL) is shown in Fig. 3.

As shown in Fig. 3, the modified forecasting results will be input into RL model to overcome the uncertainties of the environment. In the RL model, the action was chosen by ε-greedy policy which can realize exploration and exploitation mechanism. It is because that the agent cannot acquire maximum reward by only exploiting the already known knowledge. The agent is supposed to explore new knowledge for better reward. ε is generally set to a decimal between 0 and 1 [55]. A big ε denotes that the agent tends to explore new actions, and a small ε means that agent tends to exploit current actions. ε was set to 0.1 in the proposed IDR-MDLRL model as it was a widely used value in RL modeling [56].

5. Results and discussions

5.1. Data

The experimental data were all obtained from public data sources. The power load data comes from Dataport which provides the data that ranges from electricity market operations to appliance-level EUs behavioral research [57]. The building types of EUs include apartment, mobile home, single-family home, sales and town home. The hourly load data (i.e. average load in an hour) from January 1, 2018 00:00 am to July 29, 2018 23:00 pm of three users in Austin, Texas, USA was chosen randomly for experiments. Meanwhile, the environment data was acquired from Mesowest which provides free detailed environment data throughout the USA [58]. The training period of load forecasting is from January 1, 2018 00:00 am to July 22, 2018 23:00 pm, and the testing period is from July 23, 2018 00:00 am to July 29, 2018 23:00 pm.

The hourly wholesale electricity price data (i.e. average price in an hour) was obtained from PJM by Data Miner 2 [59]. PJM is one of the electric industry leaders in reliable operations and efficient wholesale electricity markets in the USA. It provides rich data sources about electric power system operation and wholesale electricity markets. The wholesale electricity price data from January 1, 2018 00:00 am to July 29, 2018 23:00 pm was used in the wholesale electricity price forecasting. The training period and testing period division is the same as that in power load forecasting.

DKA Solar Center is a demonstration facility for commercialized solar technologies operating in the arid solar conditions of Alice Springs, Central Australia [60]. It provides free access to PV power output data and related environment data for researchers all over the world. The hourly solar power output data (i.e. average output in an hour) of a PV panel whose rated output power is 5.5 kW was used in the experiments, and the division of training and testing data is the same as that in the forecasting of load and wholesale

Table 6
Cost benefit analysis of end users.

λ	0.1			0.5			0.9		
users	EU 1	EU 2	EU 3	EU 1	EU 2	EU 3	EU 1	EU 2	User3
μ_n	1	2	3	1	2	3	1	2	3
DR resources (kWh)	7.01	6.44	2.27	9.12	13.12	3.13	9.12	13.19	3.13
Incentive income (¢)	8.34	6.72	2.74	12.03	17.57	4.41	12.03	17.69	4.40
Discomfort cost (¢)	8.49	9.51	2.93	11.53	24.68	4.29	11.53	24.88	4.29
Profit 1 (¢)	−6.81	−7.89	−2.36	0.25	−3.56	0.06	9.67	13.43	3.53
Solar power income (¢)	6.56	49.93	100.22	6.56	49.93	100.22	6.56	49.93	100.22
Profit 2 (¢)	−0.25	42.04	97.86	6.81	46.37	100.28	16.23	63.36	103.75

Table 7
Cost benefit analysis of energy service provider.

λ	0.1	0.5	0.9
Gross income of DR (¢)	50.05	76.05	76.23
Cost of DR (¢)	17.80	34.01	34.12
Profit (¢)	32.25	42.04	42.11

electricity price.

It should be pointed out that the hourly data represents the average value in each hour. Meanwhile, although wholesale electricity price, PV power output, and power load data come from different data sources, the experiment results will not be affected as they are only used to simulate real-world situations.

5.2. Modified deep learning-based forecasting results

In this section, the proposed MDL-RNN model was used to forecast wholesale electricity price, PV power output, and EUs' power load respectively. After training and testing, we obtained the optimal forecasting model which has the minimum forecasting errors. Then, the modified method was used to further improve forecasting accuracy. Fig. 4 (a)-(c) are the distributions of Δ_t in forecasting wholesale electricity price, PV power output, and EU 1's power load.

It can be seen from Fig. 4 that Δ_t generally obey normal distribution. By using more historical data to train the MDL-RNN model, the distribution of Δ_t will be more stable and the forecasting can be further improved.

Fig. 5 shows the forecasting results of wholesale electricity price from July 23, 2018 to July 29, 2018. Fig. 6 shows the forecasting results of PV power output from July 23, 2018 to July 29, 2018. Fig. 7 (a)-(c) presents the load forecasting results of three different EUs from July 23, 2018 to July 29, 2018. Table 3 shows the evaluation results of MDL-RNN in forecasting wholesale electricity price, PV power output, and EUs power load by two common metrics (i.e. MAE and MAPE). It should be noted that we just calculate the MAPE of PV power output at daytime when PV system is on operation.

As shown in Fig. 5 and Table 3, the forecasting errors of wholesale electricity price are small, although there are many fluctuations in the profile. From Fig. 6, it can be found that the PV power output present periodicity. It increases smoothly with the enhancement of illumination and reaches the maximum power output at around 13:00 pm when sunlight is the strongest. Besides, we find from Table 3 that the MDL-RNN model achieved best performance (MAE: 0.11, MAPE: 0.039) in forecasting PV power output as its profiles are smoother and more regular than other profiles.

From Fig. 7, it can be seen that the power loads of these three EUs are of high volatility and show different characteristics. Meanwhile, there are basically two load peaks in the morning and evening over which EUs tend to consume more electricity due to their living needs. We can also find from Table 3 that the forecasting errors are small in forecasting the power load of different EUs. It illustrates that the MDL-RNN model has great capability in forecasting complex and nonlinear power load. In addition, compared with the forecasting in Ref. [31], the proposed MDL-RNN method achieve less MAE value.

Table 3 shows that the MDL-RNN model performs well in forecasting wholesale electricity price, PV power output, and EUs' power load. At the same time, the MDL-RNN model is superior to conventional methods, such as simple ANN, support vector machine (SVM), and extreme learning machine (ELM). It can learn the time dependencies and nonlinearity in time series data. In addition, it can be noted that the MDL-RNN model achieves better performance in forecasting PV power output and wholesale electricity price than EUs' power load. This is due to the fact that EUs' power load is of higher volatility and fluctuation which will increase the difficulty in learning the relationship between the input variables and the output power load.

5.3. Incentive rate optimization based on reinforcement learning

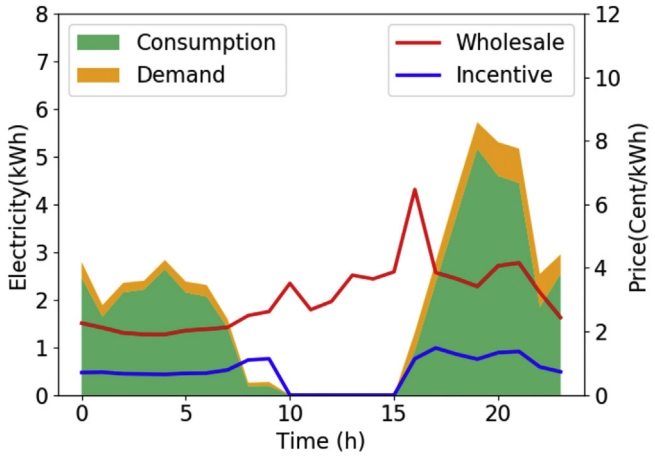
The day-ahead forecasting results were regarded as the inputs of the RL model to reduce the impact of environment uncertainties. But the forecasting hourly PV power output and EUs' power load data have to be first converted to hourly power generation and hourly electricity demand data. In this section, the forecasting results on July 23, 2018 were selected as a detailed case study. The incentive rates at each hour were optimized by the proposed RL model for each EU to maximize the total profits of ESP and EUs.

5.3.1. Parameters setup

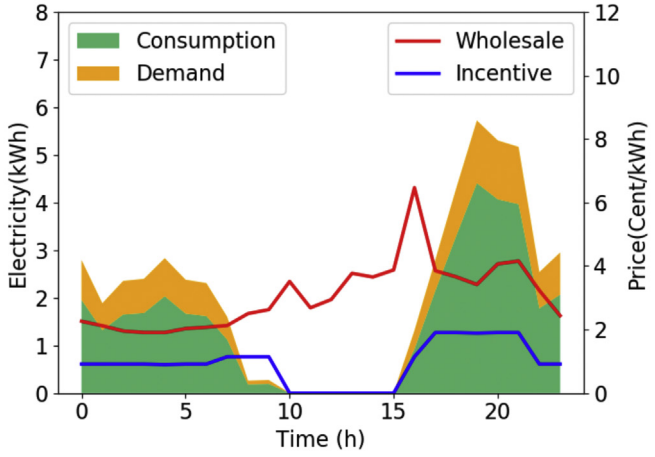
Table 4 shows the parameters of RL model including the discomfort parameters, the upper bound of electricity reduction, the minimum incentive rate, and the maximum incentive rate, which were referenced from Refs. [20]. Table 5 presents the electricity elasticity at different hours [61]. The electricity elasticity reflects the impact of incentive variation on electricity demand variation. The elasticity period can be divided into three parts corresponding to valley, mid-peak, and peak periods respectively. It should be noted that the values of parameters will not affect the mechanism of our RL model in essence.

5.3.2. Convergence of Q-value

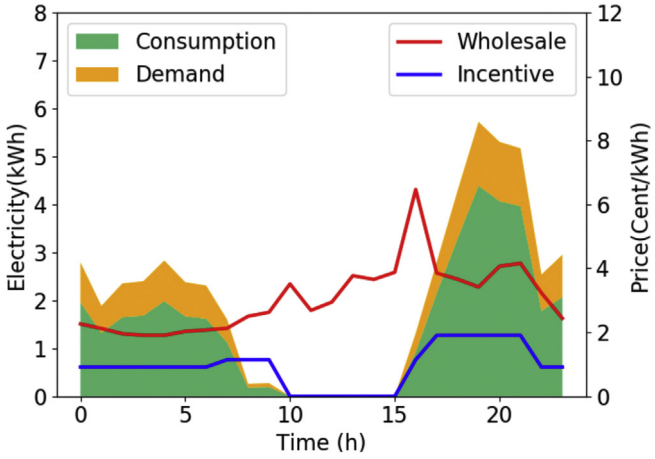
During the iterations of RL, Q-value will gradually converge to the maximum value which corresponds to the maximum profits of ESP and EUs. Fig. 8 shows the Q-value profile of EU 2 to acquire optimal incentive rates under different λ. As shown in Fig. 8, the start Q-value is very small as the agent has very limited knowledge on choosing the best actions which will bring optimal reward. After several iterations, the agent can learn from previous experience to determine the optimal action and the Q-value tends to be stable. It



(a) EU 2: λ_2=0.1, μ_2=2



(b) EU 2: λ_2=0.5, μ_2=2



(c) EU 2: λ_2=0.9, μ_2=2

Fig. 11. Results of EU 2 with different λ_2.

can be also found from Fig. 8 that the optimal Q-value will grow with the increase of λ. It is because that EU places a higher premium on the reward than discomfort cost when λ is larger, and thus the total profits of ESP and EUs increase with λ.

Fig. 9 (a)-(b) shows the convergence of Q-value for different EUs. When λ=0.9, EUs are willing to reduce their electricity demand to obtain rewards and even almost neglect the incurred discomfort costs. It can be seen from Fig. 9 that the maximum Q-value of EU 2 ($\mu_2 = 2$, $\omega_2 = 1$) is larger than that of others. It is because that the discomfort cost paprameter μ_2 is less than that of EU 3 ($\mu_3 = 3$). Meanwhile, the overall electricity consumption of EU 2 is relatively higher than that of others. Hence, EU 2 has more poetntial to response to incentives and reduce their electricity demand for the rewards. When λ=0.1, the EUs place higher premium on discomfort cost than the reward provided by ESP. This leads to less electricity demand reduction and small Q-value as shown in Fig. 9. Moreover, the discomfort cost has more influence on the profits of EUs under this circumstance. Thereby, it can be found that the maximal Q-value of EU 1 is more than that of EU 2.

In addition, from Figs. 8 and 9, it can be seen that Q-value follows a fluctuating ascending trend, and still fluctuates when converging to optimal Q-value. It is caused by the ε-greedy policy which makes the agent explore new knowledge while exploiting acquired konledge during iterations. The convergence of Q-value will be proofed theoretically in Appendix A.1.

5.3.3. Results of different EUs with different λ

First, in order to explore how EUs response to the optimal incentive rate at each hour obtained by the proposed RL model, the incentive rate profiles and electricty demand reduction of different EUs with the same λ value ($\lambda = 0.1$) were presented in Fig. 10.

In Fig. 10, the blue line represents the optimal incentive rates, the red line is the day-ahead wholesale electricity price, the yellow part is the original electricity demand, and the green part denotes actual electricity consumption when the EUs response to the incentive-based DR program. It can be seen that the actual electricity demand is low for EU 1 and even equal to 0 for EU 2 and EU 3 at noon. It is because that each EU is assumed to install the same PV panels which can provide enough electricity for their usage and sell extra solar power in wholesale electricity market over that time period. Meanwhile, ESP will also not provide incentive for EUs when their actual electricity demand is 0. In addition, it can be expected that EU 3 will has less electricity demand reduction when it has same original electricity demand as EU 1 or EU 2. The reason is that EU 3 has larger discomfort parameter $\mu_3 = 3$ which denotes that EU 3 possess more conservative attitude towards incentive reward.

From Fig. 10, what can be also found is that the incentive rates vary with wholesale market price changes. During lower wholesale market price period, ESP will provide less incentive to reduce EUs' electricity demand. Nevertheless, electricity supply is insufficient over higher wholesale market price period, and ESP is more incline to induce EUs to reduce electricity demand by higher incentive rate. This will promote the reliability and stability of power system, and also bring certain profits for ESP and EUs.

The objective of the proposed RL model is to maximize the total profits of ESP and EUs. This can balance the benefits of ESP and EUs and encourage EUs to participate into the incentive-based DR program. The cost benefit analysis of EUs and ESP was carried out as shown in Tables 6 and 7 respectively.

In Table 6, profit 2 represents the total profits of EUs that is calculated by Eq. (2), and profit 1 equals profit 2 subtract the selling income of solar power. Since EUs first consume solar power and ESP provides the unmet electricity demand, the selling income of solar power is fixed for each EU with different λ. It can be seen that EU 3

Table 8
Financial analysis of EU 2 with different λ_2.

λ_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DR resources (kWh)	6.44	7.96	9.82	12.02	13.12	**13.19**	**13.19**	**13.19**	**13.19**
Incentive income (¢)	6.72	8.86	11.88	15.70	17.56	**17.68**	**17.68**	**17.68**	**17.68**
Discomfort cost (¢)	9.51	12.44	16.44	21.82	24.68	**24.88**	**24.88**	**24.88**	**24.88**
Profit 1 (¢)	−7.89	−8.18	−7.95	−6.81	−3.56	0.66	4.91	9.17	13.43
PV income (¢)	49.93	49.93	49.93	49.93	49.93	49.93	49.93	49.93	49.93
Profit 2 (¢)	42.04	41.75	41.98	43.12	46.37	50.59	54.84	59.10	63.36

Table 9
Cost benefit analysis of energy service provider.

λ_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Gross income of DR (¢)	21.41	25.71	30.96	36.93	39.56	39.72	39.72	39.72	39.72
Cost of DR (¢)	6.72	8.86	11.88	15.70	17.56	17.68	17.68	17.68	17.68
Profit (¢)	14.69	16.85	19.08	21.23	22.00	22.04	22.04	22.04	22.04

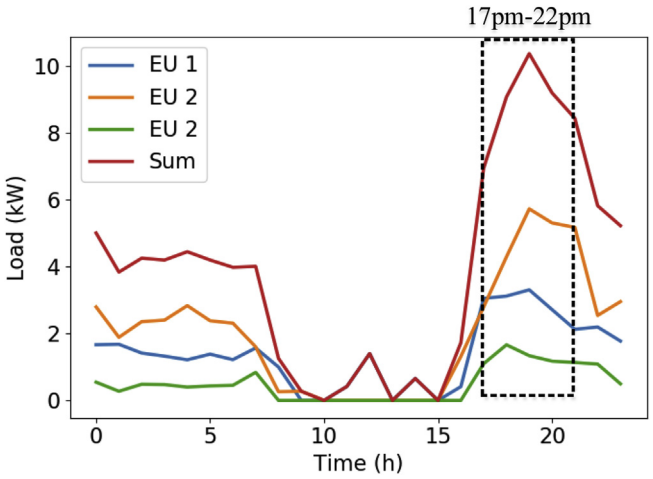


Fig. 12. Real load demand of EUs.

Table 10
Benefits of energy service provider implementing the short-term DR program.

EUs	1	2	3	Total
Cost of DR (¢)	6.898	8.869	4.007	19.774
Cost without DR (¢)	15.599	21.332	8.315	45.246
Peak electricity demand under DR (kWh)	2.726	4.840	1.303	8.869
Peak electricity demand without DR (kWh)	3.304	5.723	1.661	10.688
Total electricity demand reduction (kWh)	4.284	5.776	2.294	12.354

gained the maximum income (100.22 ¢) by selling extra solar power to GO, and EU 1 obtained the minimum income (6.56 ¢) because its electricity demand is relatively higher during daytime. With the increase of λ, EUs prefer incentive reward to comfort, and they are more active to reduce electricity consumption. Therefore, it can be found that profit 1 and profit 2 of each EU are gradually increasing. Here, it can be noticed that the procured DR resources, incentive income, and discomfort cost do not increase when λ>0.5. It is because that the EUs prefer reward than comfort. This also conforms to the actual situation that EUs tend to enjoy better comfort than reward.

Table 7 shows the profits of ESP which was calculated by Eq. (1). The gross income of DR means the reduced electricity purchase cost of ESP, and the cost of DR denotes the rewards for EUs in the incentive-based DR program. It can be seen that cost of DR grows with the increase of λ. However, with the increase of λ, ESP will get more DR resources and profits. At the same time, it can be found that the increase of ESP's profit is not obvious when λ changes from 0.5 to 0.9. In a real scenario, ESP will not raise the incentive rate anymore when the EUs place higher premium on comfort than reward. The detailed explanation will be given in next Section.

5.3.4. Results of same EU with different λ

Fig. 11 shows the results of EU 2 adopting different λ_2 (0.1, 0.5, and 0.9). As we can see form Fig. 11, the profiles of incentive rate show the same trend as wholesale electricity price. A big λ_2 denotes that EU 2 pursues the reward provided by ESP. To improve the profits, ESP would raise the incentive rate to procure more DR resources. However, the ESP cannot further reduce the electricity demand if $\lambda_2 > 0.5$ as the EU obtained the maximal electricity reduction at each time step. At the same time, the ESP will not raise the incentive rate anymore.

To further explore the relationship between responsive resources and λ_2, more values of λ_2 were tested and the cost benefit analysis of EU 2 is shown in Table 9.

In Table 8, it can be found that the incentive income and the discomfort cost of EU 2 are increasing with the increase of λ_2, but do not change when λ_2 is more than 0.6. However, the profit 1 and profit 2 always increase with λ_2. The profit 1 is negative when $\lambda_2 < 0.6$. The reason is that the final discomfort cost is more than the final reward. As for ESP, it can be seen from Table 9 that the gross income of DR, cost, and profits stop increasing when $\lambda_2 \geq 0.6$ as ESP cannot acquire more DR resources. This can be proved theoretically as shown in Appendix A.2.

5.3.5. Short-term incentive-based DR program

Incentive-based DR programs are effective methods to reduce EUs' electricity demand during peak electricity demand period rather than all over a day. This contributes to promoting the balance between supply and demand and ensuring the security of power system operation in real time. Besides, more accurate results of short-term forecasting can be obtained as more accurate environmental variables in the next few hours can be acquired, such as temperature and light intensity. Therefore, a short-term DR program which can also be called an emergency DR program, was developed using the proposed IDR-MDLRL algorithm. Fig. 12 shows the actual load demand profiles of three EUs and their total load demand profile.

As shown in Fig. 12, the load demands in daytime are low as the solar systems deployed in EUs side can provide enough solar power.

It can be seen that the peak load demand occurs at evening corresponding to 17:00 pm - 22:00 pm. The ESP would like to reduce the total electricity demand to ensure the security of power system by providing rewards for EUs. Once receiving the incentive rates form ESP, EUs will response to short-term DR program and sacrifice a part of discomfort for the rewards. Besides, EUs are of great potential to response to DR program at that time due to the high electricity price. To guarantee the necessary electricity demand for normal life, they can shift the peak electricity demand to other periods. For example, EUs can schedule their washing machine to work at deep night rather than during the peak period. What should be noticed is that p_{min} is the minimum wholesale market price over 17:00 pm - 22:00 pm, not the minimum value of the day. The benefits of ESP implementing the short-term incentive-based DR program are show in Table 10.

In the short-term DR program, we just considered the worst scenario, i.e. λ is set to 0.1 and μ is set to 3. In this scenario, EUs prefer comfort than the reward, and the unit discomfort cost is more than the unit reward incurred by the electricity demand reduction. The cost of DR represents the payments to EUs by ESP to obtain required DR resources. The cost without DR is the payments by ESP to GO for purchasing the same resources from GO. As it can be seen from Table 10, the cost for obtaining required resources was reduced by 25.472 ¢. Besides, the peak electricity demand which represents the maximal electricity demand during 17:00 pm - 22:00 pm, decreased to 8.869 kWh, almost a 17% reduction. The total electricity demand reduction also achieved a reduction of 12.354 kWh. It can be expected that higher reduction of peak electricity demand and total electricity demand can be achieved if other better scenarios are taken into consideration.

6. Conclusions

This study proposed an incentive-based DR program based on deep learning and reinforcement learning for smart grid operation. The complexities and uncertainties of environment were considered by forecasting wholesale electricity price, PV power output, and power load with a DL-RNN model using a modified method. Then, a model-free method, RL, was employed to find the day-ahead optimal incentive rates at each hour for each EU. Particularly, a short-term incentive-based DR program was proposed to reduce electricity demand during peak electricity demand periods. The experimental results show that the peak electricity demand was reduced by 17%. This shows that the proposed incentive-based DR program contributes to balancing supply-demand and improving power system reliability. It also provides an effective way to implement incentive-based DR program in uncertain and dynamic environment. In future work, we will collect real-world data to quantify the parameters and investigate the relationships among EU-dependent parameters, socioeconomic backgrounds and weather conditions. In addition, the interactions among multiple ESPs, EUs and the electricity elasticity will be further explored.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Lulu Wen: Conceptualization, Methodology, Data curation, Software, Visualization, Formal analysis, Writing - original draft. **Kaile Zhou:** Conceptualization, Methodology, Resources, Visualization, Writing - review & editing, Project administration, Supervision. **Jun Li:** Resources, Writing - review & editing. **Shanyong Wang:** Writing - review & editing.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 71822104), and the China Scholarship Council.

Appendix A

A. 1 The convergence of Q-value

From Eq. (1), Eq. (3), Eq. (4), Eq. (6), and Eq. (9), the total profits of EUs and ESP can be obtained as follows.

$$profit_{sum} = \sum_{i=1}^{n}\sum_{j=1}^{h}\left[\left(p_j - \alpha_{ij}\right)\Delta E_{ij} + \lambda_i \cdot \alpha_{ij} \cdot \Delta E_{ij} - \left(1-\lambda_i\right)\cdot\left[\frac{\mu_i}{2}\Delta E_{ij}^2 + \omega_i \cdot \Delta E_{ij}\right] + p_j \cdot PV_{ij}\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{h}\left[\left(p_j - \alpha_{ij}\right)\frac{E_{ij}\cdot\xi_j\cdot(\alpha_{ij}-\alpha_{\min})}{\alpha_{\min}} + \lambda_i \cdot \alpha_{ij}\cdot\frac{E_{ij}\cdot\xi_j\cdot(\alpha_{ij}-\alpha_{\min})}{\alpha_{\min}}\right.$$

$$\left. -(1-\lambda_i)\cdot\left[\frac{\mu_i}{2}\cdot\frac{E_{ij}^2\cdot\xi_j^2\cdot(\alpha_{ij}-\alpha_{\min})^2}{\alpha^2{}_{\min}} + \omega_i\frac{E_{ij}\cdot\xi_j\cdot(\alpha_{ij}-\alpha_{\min})}{\alpha_{\min}}\right] + p_j\cdot PV_{ij}\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{h}\left\{-\left[\frac{E_{ij}\cdot\xi_j\cdot(1-\lambda_i)}{\alpha_{\min}} + \frac{\mu_i}{2}\frac{E_{ij}^2\cdot\xi_j^2\cdot(1-\lambda_i)}{\alpha^2{}_{\min}}\right]\alpha_{ij}^2\right.$$

$$+\left[\frac{E_{ij}\cdot\xi_j\cdot p_j}{\alpha_{\min}} + \frac{E_{ij}\cdot\xi_j\cdot\alpha_{\min}\cdot(1-\lambda_i)}{\alpha_{\min}} + \frac{\mu_i\cdot E_{ij}^2\cdot\xi_j^2\cdot(1-\lambda_i)}{\alpha_{\min}} - \frac{\omega_i\cdot E_{ij}\cdot\xi_j\cdot(1-\lambda_i)}{\alpha_{\min}}\right]\alpha_{ij}$$

$$\left. +\left[\omega_i\cdot E_{ij}\cdot\xi_j\cdot(1-\lambda_i) - E_{ij}\cdot\xi_j\cdot p_j - \frac{\mu_i}{2}E_{ij}^2\cdot\xi_j^2\cdot(1-\lambda_i) + p_j\cdot PV_{ij}\right]\right\}$$

Thus, the total profits of EUs and ESP is a quadratic function of incentive rate α_{ij}. Since$1 > \lambda_i > 0, 1 - \lambda > 0, \alpha_{\min} > 0, E_{ij} > 0, \xi_j > 0, p_j > 1, \mu_i > 0, \omega_i = 1$, it can be easily obtained that

$$A = -\left[\frac{E_{ij}\cdot\xi_j\cdot(1-\lambda_i)}{\alpha_{\min}} + \frac{\mu_i}{2}\frac{E_{ij}^2\cdot\xi_j^2\cdot(1-\lambda_i)}{\alpha_{\min}^2}\right] < 0$$

$$B = \left[\frac{E_{ij}\cdot\xi_j\cdot p_j}{\alpha_{\min}} + \frac{E_{ij}\cdot\xi_j\cdot\alpha_{\min}\cdot(1-\lambda_i)}{\alpha_{\min}} + \frac{\mu_i\cdot E_{ij}^2\cdot\xi_j^2\cdot(1-\lambda_i)}{\alpha_{\min}} - \frac{\omega_i\cdot E_{ij}\cdot\xi_j\cdot(1-\lambda_i)}{\alpha_{\min}}\right] > 0$$

$$-\frac{B}{2A} > 0$$

Therefore, the opening of the parabola is down, and the symmetry axis is on the right. From above, it can be found that the total profits of EUs and ESP increase with incentive rates and then decrease with incentive rates. In other words, the total profits can get the maximum value within the range of incentive rates, and thus the Q-value will converge to optimal value.

A. 2 The discomfort cost do not change any more when $\lambda_2 > 0.6$

From Eq. (2), Eq. (4), and Eq. (5), it can be obtained that $0.3p_{\min} \le \alpha_{ij} \le p_{\min}$ and $0 \le \Delta E_{ij} \le 0.3E_{ij}$, thus $0 \le E_{ij}\cdot\xi_j\cdot\frac{\alpha_{ij}-\alpha_{\min}}{\alpha_{\min}} \le 0.3E_{ij}$. Therefore, $0.3p_{\min} \le \alpha_{ij} \le \left(\frac{0.3}{\xi_j}+1\right)*0.3p_{\min}$.

From above, it can be seen that the incentive rate α_{ij} is within certain range. When λ_i is growing, it means that EU i regard incentive reward as more important than comfort. Therefore, in the learning of RL, the agent will improve incentive rate at each hour to acquire more DR resources for rewards. But α_{ij} has reached the maximum value when $\lambda_2 \ge 0.6$, and thus the DR resources, incentive income, and discomfort cost which just correlate with α_{ij}, will not change as shown in Table 8. This has also proved the results in Section 5.3.3 where the increase of ESP's profit is not obvious when λ changes from 0.5 to 0.9.

References

[1] Ghazvini MAF, Faria P, Ramos S, Morais H, Vale Z. Incentive-based demand response programs designed by asset-light retail electricity providers for the day-ahead market. Energy 2015;82:786–99.
[2] Khalili T, Jafari A, Abapour M, Mohammadi-Ivatloo B. Optimal battery technology selection and incentive-based demand response program utilization for reliability improvement of an insular microgrid. Energy 2019;169:92–104.
[3] Feng Z-k, Niu W-j, Cheng X, Wang J-y, Wang S, Song Z-g. An effective three-stage hybrid optimization method for source-network-load power generation of cascade hydropower reservoirs serving multiple interconnected power grids. J Clean Prod 2020;246:119035.
[4] Feng Z-k, Niu W-j, Cheng C-t, Zhou J-z. Peak shaving operation of hydro-thermal-nuclear plants serving multiple power grids by linear programming. Energy 2017;135:210–9.
[5] Haider HT, See OH, Elmenreich W. A review of residential demand response of smart grid. Renew Sustain Energy Rev 2016;59:166–78.
[6] Yu M, Hong SH. Supply–demand balancing for power management in smart grid: a Stackelberg game approach. Appl Energy 2016;164:702–10.
[7] Diekerhof M, Peterssen F, Monti A. Hierarchical distributed robust optimization for demand response services. IEEE Transactions on Smart Grid 2018;9:6018–29.
[8] Yan X, Ozturk Y, Hu Z, Song Y. A review on price-driven residential demand response. Renew Sustain Energy Rev 2018;96:411–9.
[9] Rahmani-Andebili M, Shen H. Energy management of end users modeling their reaction from a GENCO's point of view. In: International conference on computing, networking and communications (ICNC). IEEE; 2017. p. 577–81. 2017.
[10] Monfared HJ, Ghasemi A, Loni A, Marzband M. A hybrid price-based demand response program for the residential micro-grid. Energy 2019;185:274–85.
[11] Srinivasan D, Rajgarhia S, Radhakrishnan BM, Sharma A, Khincha H. Game-Theory based dynamic pricing strategies for demand side management in smart grids. Energy 2017;126:132–43.
[12] Ghasemkhani A, Yang L, Zhang J. Learning-based demand response for privacy-preserving users. IEEE Transactions on Industrial Informatics 2019;15(9):4988–98.
[13] Li Y-C, Hong SH. Real-time demand bidding for energy management in discrete manufacturing facilities. IEEE Trans Ind Electron 2016;64:739–49.
[14] Asadinejad A, Tomsovic K. Optimal use of incentive and price based demand response to reduce costs and price volatility. Elec Power Syst Res 2017;144:215–23.
[15] Rahmani-andebili M. Modeling nonlinear incentive-based and price-based demand response programs and implementing on real power markets. Elec Power Syst Res 2016;132:115–24.
[16] Rahmani-Andebili M. Nonlinear demand response programs for residential customers with nonlinear behavioral models. Energy Build 2016;119:352–62.
[17] Erdinc O, Taşcikaraoğlu A, Paterakis NG, Catalão JP. Novel incentive mechanism for end-users enrolled in DLC-based demand response programs within stochastic planning context. IEEE Trans Ind Electron 2019;66:1476–87.
[18] Li Z, Wang S, Zheng X, De Leon F, Hong T. Dynamic demand response using customer coupons considering multiple load aggregators to simultaneously achieve efficiency and fairness. IEEE Transactions on Smart Grid 2016;9:3112–21.
[19] Shahryari E, Shayeghi H, Mohammadi-Ivatloo B, Moradzadeh M. An improved incentive-based demand response program in day-ahead and intra-day electricity markets. Energy 2018;155:205–14.
[20] Yu M, Hong SH. Incentive-based demand response considering hierarchical electricity market: a Stackelberg game approach. Appl Energy 2017;203:267–79.
[21] Rahmani-Andebili M. Planning and operation of plug-in electric vehicles. Springer; 2019.
[22] Du G, Zou Y, Zhang X, Liu T, Wu J, He D. Deep reinforcement learning based energy management for a hybrid electric vehicle. Energy 2020:117591.
[23] Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT press; 2018.
[24] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: a review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89.
[25] Costanzo GT, Iacovella S, Ruelens F, Leurs T, Claessens BJ. Experimental analysis of data-driven control for a building heating system. Sustainable Energy, Grids and Networks 2016;6:81–90.
[26] Arif A, Babar M, Ahamed TI, Al-Ammar E, Nguyen P, Kamphuis IR, et al. Online scheduling of plug-in vehicles in dynamic pricing schemes. Sustainable Energy, Grids and Networks 2016;7:25–36.
[27] Wang H, Huang T, Liao X, Abu-Rub H, Chen G. Reinforcement learning in energy trading game among smart microgrids. IEEE Trans Ind Electron 2016;63:5109–19.
[28] Mahapatra C, Moharana A, Leung V. Energy management in smart cities based on Internet of Things: peak demand reduction and energy savings. Sensors 2017;17:2812.
[29] Marinescu A, Dusparic I, Clarke S. Prediction-based multi-agent reinforcement learning in inherently non-stationary environments. ACM Trans Autonom Adapt Syst 2017;12:9.
[30] Lu R, Hong SH, Yu M. Demand response for home energy management using reinforcement learning and qrtificial neural network. IEEE Transactions on Smart Grid 2019;10(6):6629–39.
[31] Lu R, Hong SH. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl Energy 2019;236:937–49.
[32] Fang X, Hu Q, Li F, Wang B, Li Y. Coupon-based demand response considering wind power uncertainty: a strategic bidding model for load serving entities. IEEE Trans Power Syst 2015;31:1025–37.
[33] Asadinejad A, Rahimpour A, Tomsovic K, Qi H, Chen C-f. Evaluation of residential customer elasticity for incentive based demand response programs. Elec Power Syst Res 2018;158:26–36.

[34] Wang Y, Ai X, Tan Z, Yan L, Liu S. Interactive dispatch modes and bidding strategy of multiple virtual power plants based on demand response and game theory. IEEE Transactions on Smart Grid 2015;7:510–9.
[35] Yu M, Lu R, Hong SH. A real-time decision model for industrial load management in a smart grid. Appl Energy 2016;183:1488–97.
[36] Feng Z-k, Niu W-j, Tang Z-y, Jiang Z-q, Xu Y, Liu Y, et al. Monthly runoff time series prediction by variational mode decomposition and support vector machine based on quantum-behaved particle swarm optimization. J Hydrol 2020;583:124627.
[37] Niu W-j, Feng Z-k, Chen Y-b, Zhang H-r, Cheng C-t. Annual streamflow time series prediction using extreme learning machine based on gravitational search algorithm and variational mode decomposition. J Hydrol Eng 2020;25:04020008.
[38] Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. Recurrent neural network based language model. In: Eleventh annual conference of the international speech communication association; 2010.
[39] Yan Z, Wang J. Model predictive control of nonlinear systems with unmodeled dynamics based on feedforward and recurrent neural networks. IEEE Transactions on Industrial Informatics 2012;8:746–56.
[40] Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. 1999.
[41] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014. arXiv preprint arXiv:14123555.
[42] Guo Z, Zhou K, Zhang X, Yang S. A deep learning model for short-term power load and probability density forecasting. Energy 2018;160:1186–200.
[43] Rahman A, Srikumar V, Smith AD. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl Energy 2018;212:372–85.
[44] Wang Y, Gan D, Sun M, Zhang N, Lu Z, Kang C. Probabilistic individual load forecasting using pinball loss guided LSTM. Appl Energy 2019;235:10–20.
[45] Patro S, Sahu KK. Normalization: a preprocessing stage. 2015. arXiv preprint arXiv:150306462.
[46] Girosi F, Jones M, Poggio T. Regularization theory and neural networks architectures. Neural Comput 1995;7:219–69.
[47] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
[48] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. arXiv preprint arXiv:14126980.
[49] Lu R, Hong SH, Zhang X. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach. Appl Energy 2018;220:220–30.
[50] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res 1996;4:237–85.
[51] Liu T, Hu X, Hu W, Zou Y. A heuristic planning reinforcement learning-based energy management for power-split plug-in hybrid electric vehicles. IEEE Transactions on Industrial Informatics 2019;15(12):6436–45.
[52] Peters J, Schaal S. Natural actor-critic. Neurocomputing 2008;71:1180–90.
[53] Huys QJ, Cruickshank A, Seriès P. Reward-based learning, model-based and model-free. Encyclopedia of Computational Neuroscience 2015:2634–41.
[54] Peng J, Williams RJ. Incremental multi-step Q-learning. Machine Learning Proceedings 1994:226–32. Elsevier; 1994.
[55] Kumar R, Moseley B, Vassilvitskii S, Vattani A. Fast greedy algorithms in mapreduce and streaming. ACM Transactions on Parallel Computing (TOPC) 2015;2:14.
[56] Schulman J, Chen X, Abbeel P. Equivalence between policy gradients and soft q-learning. 2017. arXiv preprint arXiv:170406440.
[57] Dataport. Pecan Street Inc.. Available: https://dataport.cloud/.
[58] MesoWest. University of Utah Department of Atmospheric Sciences. Available: http://mesowest.utah.edu/.
[59] PJM.. Available, http://dataminer2.pjm.com/list.
[60] DKASC. Available: http://dkasolarcentre.com.au.
[61] Yu M, Hong SH, Kim JB. Incentive-based demand response approach for aggregated demand side participation. In: IEEE international conference on smart grid communications (SmartGridComm). IEEE; 2016. p. 51–6. 2016.