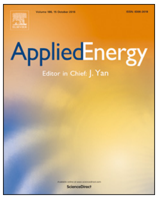
Contents lists available at ScienceDirect

Applied Energy

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



A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach

Renzhi Lu, Seung Ho Hong*, Xiongfeng Zhang

Department of Electronic Engineering, Hanyang University, Ansan 15588, Republic of Korea

HIGHLIGHTS

- Propose an artificial intelligence based dynamic pricing demand response algorithm.
- Reinforcement learning is used to illustrate the decision-making framework.
- Uncertainty of customer's demand and flexibility of wholesale prices are achieved.
- Effects of customers' private preferences in the electricity market are addressed.

ARTICLE INFO

Keywords:
Demand response
Dynamic pricing
Artificial intelligence
Reinforcement learning
Markov decision process
Q-learning

ABSTRACT

With the modern advanced information and communication technologies in smart grid systems, demand response (DR) has become an effective method for improving grid reliability and reducing energy costs due to the ability to react quickly to supply-demand mismatches by adjusting flexible loads on the demand side. This paper proposes a dynamic pricing DR algorithm for energy management in a hierarchical electricity market that considers both service provider's (SP) profit and customers' (CUs) costs. Reinforcement learning (RL) is used to illustrate the hierarchical decision-making framework, in which the dynamic pricing problem is formulated as a discrete finite Markov decision process (MDP), and Q-learning is adopted to solve this decision-making problem. Using RL, the SP can adaptively decide the retail electricity price during the on-line learning process where the uncertainty of CUs' load demand profiles and the flexibility of wholesale electricity prices are addressed. Simulation results show that this proposed DR algorithm, can promote SP profitability, reduce energy costs for CUs, balance energy supply and demand in the electricity market, and improve the reliability of electric power systems, which can be regarded as a win-win strategy for both SP and CUs.

1. Introduction

Owing to the modern advanced information and communication technologies in smart grid systems, demand response (DR) has become an effective method for improving grid reliability and reducing energy costs due to the ability to react quickly to supply-demand mismatches by adjusting flexible loads on the demand side [1,2]. According to the United States Department of Energy, DR refers to "a tariff or program established to motivate changes in the price of electricity over time, or to give incentive payments designed to induce lower electricity usage at times of high market prices or when grid reliability is jeopardized" [3].

The existing literature generally discusses two categories of DR: price-based and incentive-based [4]. Price-based DR motivates customers (CUs) to change their energy usage patterns in response to time-varying electricity prices, while incentive-based DR provides fixed or time-varying incentives to CUs if they reduce their energy consumption during periods of power system stress [5]; both categories have their own benefits and take advantage of different aspects of the potential for flexible demand. This study focuses on price-based DR, whose efficiency has been evaluated in several studies [6–9].

A number of studies have investigated the price-based DR, focusing on directly controlling appliances to maximize the social welfare of the smart grid systems from the CUs' perspective. For example, in [10–12], energy consumption scheduling of residential appliances was studied considering time-of-use (TOU) pricing to reduce CUs' costs and enhance energy efficiency. Similarly, the work in [13] evaluated the impact of a large-scale field deployment of mandatory TOU pricing on the energy use of commercial and industrial CUs. In [14], the authors investigated the DR of commercial and industrial businesses to critical peak pricing plots where the time and duration of the price increase were

* Corresponding author.
E-mail address: shhong@hanyang.ac.kr (S.H. Hong).

https://doi.org/10.1016/j.apenergy.2018.03.072
Received 22 November 2017; Received in revised form 7 February 2018; Accepted 25 March 2018

Nomenclature

Variables

$e_{t,n}$	energy consumption of customer n at time slot t
$E_{t,n}$	energy demand of customer n at time slot t
$e_{t,n}^{curt}$	energy consumption of customer n at time slot t for curtailable load
$E_{t,n}^{curt}$	energy demand of customer n at time slot t for curtailable load
$e_{t,n}^{critic}$	energy consumption of customer n at time slot t for critical load
$E_{t,n}^{critic}$	energy demand of customer n at time slot t for critical load
t	index for time slot
n	index for customer
$\lambda_{t,n}$	retail electricity price for customer n at time slot t
π_t	wholesale electricity price at time slot t
$\varphi_{t,n}$	dissatisfaction cost of customer n at time slot t
i	index for iteration in Q-learning

Parameters

ξ_t	elasticity at time slot t
α_n	customer preference parameter of dissatisfaction cost function
β_n	predetermined paremeter of dissatisfaction cost function
$D_{\min}$	lower bound of demand reduction at time slot t
$D_{\max}$	upper bound of demand reduction at time slot t
$\varkappa_1$	coefficient of lower retail price bound
$\varkappa_2$	coefficient of upper retail price bound
ρ	weighting factor between SP's profit and CUs' costs

Other symbols

DR	demand response
RL	reinforcement learning
SP	service provider
CU	customer
MDP	Markov decision process
GO	grid operator

predetermined. The works in [15–20] described a deterministic DR model with day-ahead prices for CUs, wherein the next-day electricity prices are known in advance and optimal energy consumption scheduling can be predefined though minimizing daily costs. In [21], the authors proposed a day-ahead price-based and real-time incentive-based strategy for large electricity CUs; however, the period and the value of the incentive rate were assumed to be decided ahead of schedule. In [22,23], two further price-based DR schemes were designed for industrial loads, considering both current and future load control in the schedule horizon; however, although the authors modeled the future price uncertainty, the mathematical formulations in these two papers were complex, and real-world implementation would be cumbersome. Thus far, the majority of previous works on energy consumption scheduling based on a given pricing policy, and cannot accommodate uncertainties in the dynamic electricity market environment. Given this, it is imperative to devise an innovative dynamic pricing DR mechanism for smart grid systems.

Dynamic pricing is a business strategy that adjusts the product price in a timely fashion, to allocate the right service to the right CU at the right time [24]. There have been several works on dynamic pricing DR algorithms for smart grids. The study in [25] investigated a dynamic pricing strategy with DR for a microgrid retailer in an integrated energy system, where the retail rates and microgrid dispatch were formulated as a mixed integer quadratic programming problem with the aim of maximizing the retailer's profit. In [26,27], Stackelberg games were used to model energy trading between a retailer and CUs, where the retailer determined the dynamic retail price based on the energy pricing scheme to maximize profit, and then the CUs minimized their payment bill by managing the energy usage of appliances according to the announced prices. More recently, another three works [28–30] proposed the dynamic price-based energy management scheme between the retailer and CUs. However, in these works, the dynamic pricing policies deployed by the retailer were predetermined by abstract models (e.g., linear model) without logical process of determination. To some degree, these studies are still deterministic and cannot react to the flexibility of CUs' demand profiles and wholesale electricity prices in the electric power market.

From the above existing literatures, we can conclude that the energy management system operation still relies on conventional ways such as deterministic rules and abstract models (e.g., mix integer linear programming), which mainly suffer two key criticisms: (a) applying deterministic rules for operating non-stationary system cannot guarantee optimality, any changes of a variable may result in a loss of money and (b) abstract models are usually approximations of the reality and therefore might be unrealistic compared with real energy systems, since the performance of the abstract model is strictly limited by the modeler's skill and experience. In recent years, with the rapid development of artificial intelligence, there has been growing interest in adopting reinforcement learning (RL) to solve the decision-making problem in smart grids. A number of breakthroughs in RL have been reported, in particular, like deep Q-network in Atari[31] and AlphaGo [32]. RL is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in a stochastic environment to maximize some notion of cumulative reward, as shown in Fig. 1. An agent interacts with its environment in discrete time steps; at each time step, the agent chooses an action from the set of available actions, which is subsequently sent to the environment. Then the agent receives a reward and the environment moves to a new state. The goal of an agent is to collect as much reward as possible. In [33–38], RL algorithms were used to schedule energy storage systems and obtain an optimal charging/discharging policy, e.g., a battery or an electric vehicle. This scenario is relatively easy for its limited space of actions and states, and has thus been the focus of a number of papers. The studies in [39–42] used RL to obtain energy scheduling for specific devices in DR, e.g., electric water heaters, thermostatically controlled loads, or others. In [43], the authors considered microgrids as a whole, with each microgrid having the capacity to buy or sell energy to another microgrid; RL was used between the microgrids to choose a buying/selling strategy for energy trading, to maximize the average revenue. Motivated by the dominant and unique features of "no need of expert knowledge" and "model-free", RL is becoming one of the most promising tools to realize optimal operation of energy management system in face of ever-changing ambient factors, e.g., dynamic electricity prices and energy consumptions.

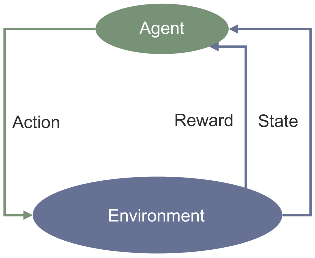


Fig. 1. Reinforcement learning (RL) setup.

In contrast to these existing studies, in this paper, we consider a hierarchical electricity market, as shown in Fig. 2, where the service provider (SP) decides the dynamic pricing strategy according to the CUs' energy demand profiles and dissatisfaction levels, as well as the wholesale electricity prices from the grid operator (GO), to enable more efficient energy consumption. Specifically, the SP (*agent*) determines the retail price (*action*) that is sent to the CUs (*environment*) at each time slot, and charges the CUs an electricity bill as a *reward*; the *states* are denoted by the CUs' energy demands and energy consumptions. RL is used to analyze how the SP can learn and obtain dynamic retail pricing strategies while interacting with different CUs, to maximize the SP's profit and minimize the CUs' costs. Employing RL to address the dynamic pricing decision problem has three main advantages that provide the best solution to the mentioned challenges. First, RL is model free. It does not require a pre-specified model of the environment on which retail price actions are selected. Instead, the relationship between retail price and profit is learned though dynamic interaction with CUs. Second, RL is adaptive. It is capable of responding to a dynamically changing environment though ongoing learning and adaptation, considering the uncertainty and flexibility of the electricity market. Third, RL is concise. The whole computational process of the algorithm is based on a look-up table and an updating mechanism shown in Table 1.

To the best of our knowledge, this is the first paper that investigates a dynamic pricing DR using RL methodology in a hierarchical electricity market for both SP's profit and CUs' costs. The main contributions of this paper are as follows:

(1) Propose an artificial intelligence based dynamic pricing DR algorithm in a hierarchical electricity market.
(2) Reinforcement learning is used to illustrate the hierarchical decision-making framework, in which the dynamic retail pricing problem is formulated as a finite discrete Markov decision process (MDP), and Q-learning is adopted to solve this decision-making problem.
(3) The uncertainty of CUs' load demand profiles and the flexibility of wholesale electricity prices are achieved by on-line learning process.
(4) The effects of CUs' private preferences in the electricity market are addressed, e.g., dissatisfaction cost function.

The rest of this paper is organized as follows. Section 2 introduces the hierarchical electricity market model and describes the mathematical formulations of this system model. In Section 3, the RL methodology is presented in detail, including formulation of the dynamic pricing problem into MDP and adoption of Q-learning to solve this decision-making problem. Section 4 discusses the numerical simulation results. Finally, conclusions and future work are discussed in Section 5.

2. System model

A hierarchical electricity market model is considered, which includes a GO, a SP, and CUs, as shown in Fig. 2. The electricity grid is installed, managed, and maintained by the GO. The GO operates the national high-voltage grid, while low-voltage electricity is transmitted by the SP. The SP buys electricity from the GO at wholesale market prices and sells the electricity to CUs at retail market prices.

This study focuses on the DR algorithm between the SP and CUs. The SP decides what dynamic retail pricing policies to adopt to enable more efficient energy consumption and maximize its profit, and CUs join in this dynamic pricing DR program to balance their energy demand and reduce their energy cost. The SP can adaptively decide retail electricity prices based on CUs' load demand profiles and the cost of buying electricity from the GO.

2.1. Customer model

The load profiles of CUs can be classified as critical or curtailable loads according to their priorities and requirement characteristics [25].

Critical load: it is very important that these load demands of CUs are critically met, for example, electricity usage in data centers.

$$e_{t,n}^{critic} = E_{t,n}^{critic} \tag{1}$$

where $t \in \{1,2,3...T\}$ denotes time slot t,T is the final time slot of a day, i.e., $T = 24$ if the price is updated every one hour, and $n \in \{1,2,3...N\}$ represents CU n. $E_{t,n}$ and $e_{t,n}$ indicate the energy demand and energy consumption of CU n at time slot t, respectively.

Curtailable load: apart from critical loads, electricity demands such as heating, ventilation, and air conditioning (HVAC) of CUs usually decrease as the electricity price increases. Once CU n consumes energy $e_{t,n}^{curt}$ at time slot t, the corresponding energy amount $e_{t,n}^{curt}$ of CU n's load demand is satisfied and the remainder of the load demand $E_{t,n}^{curt}-e_{t,n}^{curt}$ is not satisfied. This reduced energy causes dissatisfaction of CU n at time slot t, which is denoted by a dissatisfaction cost function $\varphi_{t,n}$. This models the degree of discomfort that CUs might experience when reducing their energy demand, and is defined to be convex that will increase dramatically with a larger reduction in energy [44].

The consumed energy of the curtailable load for CU n at time slot t is defined as:

$$e_{t,n}^{curt} = E_{t,n}^{curt} \cdot \left(1 + \xi_t \cdot \frac{\lambda_{t,n} - \pi_t}{\pi_t}\right) \tag{2}$$

$$\xi_t < 0 \tag{3}$$

$$\lambda_{t,n} \geqslant \pi_t \tag{4}$$

where ξ_t is the elasticity coefficient at time slot t, $\lambda_{t,n}$ represents the retail electricity price for CU n at time slot t, and π_t denotes the wholesale electricity price at time slot t.

In economics, elasticity ξ_t is a measurement of how responsive one economic variable is to a change in another. For a specific situation, the price elasticity of demand is a measure used to show the responsiveness, or elasticity, of the quantity demanded of a good or service to a change in its price. More precisely, in a smart grid, this elasticity denotes the change in energy demand for a 1% change in price. Elasticity is generally negative that indicating an inverse relationship between electricity demand and electricity price [25]. Several studies have been done to investigate the price elasticity of demand in smart grids [45–48]. Massimo [46] examined the variation in price elasticity with time of day and planning horizon, and concluded that electricity demand is more elastic during peak hours compared to off-peak hours, and that long-run elasticity is usually greater than short-run elasticity. Miller and Alberini [48] found that the price elasticities of demand

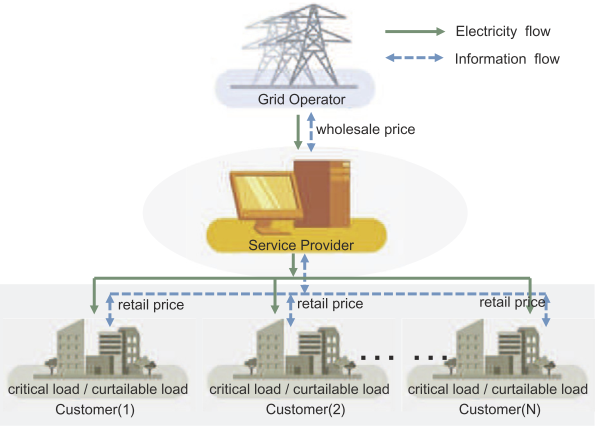


Fig. 2. Hierarchical electricity market model.

Table 1
Q-learning algorithm.

```
Mechanism to obtain the maximum Q-value

Initialize Q-value arbitrarily
Repeat for each iteration $i$
 At each time slot $t$
 Select and execute an action $\lambda_{t,n}$ at state $S(E_{t,n})$
 Observe reward $r(e_{t,n}|E_{t,n},\lambda_{t,n})$ and new state $S(E_{t+1,n})$
 $Q(e_{t,n}|E_{t,n},\lambda_{t,n}) \leftarrow Q(e_{t,n}|E_{t,n},\lambda_{t,n}) + \theta\cdot[r(e_{t,n}|E_{t,n},\lambda_{t,n}) + \gamma\cdot\max Q(e_{t+1,n}|E_{t+1,n},\lambda_{t+1,n})$
                    $-Q(e_{t,n}|E_{t,n},\lambda_{t,n})]$
End
```

ranging from -0.2 to -0.8 though experiments with three nationwide datasets from the U.S. – the American Housing Survey, Forms EIA-861, and the Residential Energy Consumption Survey. In this study, we focus on exploring the feasibility of adopting RL to make decisions for energy management in smart grid systems, thus, the elasticity values are obtained directly from existing papers.

The dissatisfaction cost function of CU n at time slot t is defined as follows:

$$\varphi_{t,n} = \frac{\alpha_n}{2}(E_{t,n}^{curt} - e_{t,n}^{curt})^2 + \beta_n(E_{t,n}^{curt} - e_{t,n}^{curt}) \tag{5}$$

$$\alpha_n > 0 \tag{6}$$

$$\beta_n > 0 \tag{7}$$

$$D_{\min} < E_{t,n}^{curt} - e_{t,n}^{curt} < D_{\max} \tag{8}$$

In (5), α_n and β_n are customer-dependent parameters, where α_n is a CU preference value varying between different CUs [44], and β_n is a predetermined constant [49]. α_n reflects the attitude of a CU with respect to electricity demand reduction: a greater value of α_n indicates that the CUs prefer less demand reduction to improve their satisfaction level, and vice versa. $D_{\min}$ and $D_{\max}$ are the ranges of demand reduction when the retail electricity prices are in effect.

The goal of CU n is to minimize its costs as described below

$$\min \sum_{t=1}^{T} [\lambda_{t,n}\cdot(e_{t,n}^{curt} + e_{t,n}^{critic}) + \varphi_{t,n}] \tag{9}$$

where the first term represents the cost of CU n for buying electricity from the SP, and the second term denotes the dissatisfaction incurred from demand reduction.

2.2. Service provider model

We assume that the SP joins the wholesale electricity market organized by the GO. At each time slot, the SP buys energy from the GO at a wholesale price determined by the GO, then sells the energy to CUs at a retail price determined by itself. Hence, the goal of the SP is to perform dynamic retail pricing that maximizes its profit as follows:

$$\max \sum_{n=1}^{N} \sum_{t=1}^{T} (\lambda_{t,n} - \pi_t)\cdot(e_{t,n}^{curt} + e_{t,n}^{critic}) \tag{10}$$

$$\kappa_1 \pi_{t,\min} \leqslant \lambda_{t,n} \leqslant \kappa_2 \pi_{t,\max} \tag{11}$$

Generally, $\lambda_{t,n}$ is bigger than π_t, but should be bound. In (11), κ_1 and κ_2 are predetermined coefficients of the retail price bounds. This property can be regarded as a regulatory requirements or mutual agreements between SP and CUs, to keep the prices fair and protect each profits [25].

2.3. Objective function

In this paper, we consider both SP's profit and CUs' costs as follows:

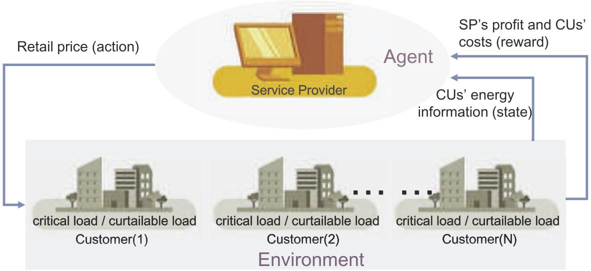

Retail price (action)
Service Provider
Agent
SP's profit and CUs' costs (reward)
CUs' energy information (state)
critical load / curtailable load
Customer(1)
critical load / curtailable load
Customer(2)
critical load / curtailable load
Customer(N)
Environment

Fig. 3. Framework of the electricity market model with RL.

$$\max \sum_{n=1}^{N} \sum_{t=1}^{T} [\rho\cdot(\lambda_{t,n} - \pi_t)\cdot e_{t,n} - (1-\rho)\cdot(\lambda_{t,n}\cdot e_{t,n} + \varphi_{t,n})] \tag{12}$$

$$e_{t,n} = e_{t,n}^{curt} + e_{t,n}^{critic} \tag{13}$$

Here, $\rho \in [0,1]$ denotes the weighting factor that indicates the relative importance between SP's profit and CUs' costs. The value of ρ should be determined by the policy of the SP, and its impact is also discussed in Section 4 of this paper.

3. Reinforcement learning methodology

RL is suitable for illustrating the hierarchical decision-making framework presented in the previous section, as shown in Fig. 3, where the SP serves as the agent, CUs are the environment, the retail price denotes the action that the SP sends to the CUs at each time slot, the energy information (energy demand and consumption) of the CUs represents the state, and the SP's profit and CUs' costs indicate the reward. In this section, we first formulate the dynamic retail pricing problem as a discrete finite horizon MDP. Then by adopting Q-learning, we develop an efficient dynamic pricing DR algorithm that does not require the full knowledge of the system dynamics and uncertainties.

3.1. Formulating system model to Markov decision process

In this paper, the dynamic retail pricing problem is modeled as a discrete finite horizon MDP because it is a decision-making problem in a stochastic environment. In this MDP model, the reward and energy consumption depend only on the energy demand and retail price at the corresponding time slot but not on the historical data. The key components to be modeled in the MDP include: discrete time t, action $A(\lambda_{t,n})$, state $S(E_{t,n},e_{t,n})$, and reward $R(r(e_{t,n}|E_{t,n},\lambda_{t,n}))$.

(1) t is the finite discrete time slot at which retail price actions are executed.
(2) $\lambda_{t,n}$ is a retail price that the SP chooses at time slot t for CU n.
(3) $E_{t,n}$ represents the CU's energy demand before it receives the retail price signal from SP. $e_{t,n}$ indicates the actual CU's energy consumption after it receives the retail price signal from SP.
(4) $r(e_{t,n}|E_{t,n},\lambda_{t,n})$ is the SP's profit and CUs' costs as defined in Section 2, which specifies the expected immediate reward gained by executing retail price $\lambda_{t,n}$ at state $E_{t,n}$.

One episode of the MDP forms a finite sequence of time slots, states, actions and rewards: 1, $E_{1,n}$, $\lambda_{1,n}$, $e_{1,n}$, $r(e_{1,n}|E_{1,n},\lambda_{1,n})$; 2, $E_{2,n}$, $\lambda_{2,n}$, $e_{2,n}$, $r(e_{2,n}|E_{2,n},\lambda_{2,n})$; …$t$, $E_{t,n}$, $\lambda_{t,n}$, $e_{t,n}$, $r(e_{t,n}|E_{t,n},\lambda_{t,n})$; …$T$, $E_{T,n}$, $\lambda_{T,n}$, $e_{T,n}$, $r(e_{T,n}|E_{T,n},\lambda_{T,n})$.

Given one run of the MDP, we can easily calculate the total reward for one episode

$$R = r(e_{1,n}|E_{1,n},\lambda_{1,n}) + r(e_{2,n}|E_{2,n},\lambda_{2,n}) + \cdots + r(e_{T,n}|E_{T,n},\lambda_{T,n}) \tag{14}$$

$$r(e_{t,n}|E_{t,n},\lambda_{t,n}) = \sum_{n=1}^{N} [\rho\cdot(\lambda_{t,n} - \pi_t)\cdot e_{t,n} - (1-\rho)\cdot(\lambda_{t,n}\cdot e_{t,n} + \varphi_{t,n})] \tag{15}$$

Then the total future reward from time slot t onward can be expressed as

$$R_t = r(e_{t,n}|E_{t,n},\lambda_{t,n}) + r(e_{t+1,n}|E_{t+1,n},\lambda_{t+1,n}) + \cdots + r(e_{T,n}|E_{T,n},\lambda_{T,n}) \qquad (16)$$

However, the environment is stochastic, we can never be sure whether we will get the same reward the next time we perform the same action. The further into the future we progress, the more it may diverge. Hence, it is common to use the discounted future reward instead:

$$R_t = r(e_{t,n}|E_{t,n},\lambda_{t,n}) + \gamma\cdot r(e_{t+1,n}|E_{t+1,n},\lambda_{t+1,n}) + \gamma^2\cdot r(e_{t+2,n}|E_{t+2,n}|\lambda_{t+2,n}) + \cdots + \gamma^{T-t}\cdot r(e_{T,n}|E_{T,n},\lambda_{T,n}) \qquad (17)$$

where $\gamma \in [0,1]$ is the discount factor representing the relative importance of future system reward compared with the current system reward. In particular, when γ equals to 0, the system will be short-sighted and only rely on the current reward. If the environment is deterministic and the same action always results in the same reward, γ can be set to 1. If we want to balance the current reward and future rewards, the value of γ should be set to a real decimal, e.g., 0.9. It is easy to see that the discounted future reward at time slot t can be expressed in terms of the identical expression at time slot $t + 1$:

$$R_t = r(e_{t,n}|E_{t,n},\lambda_{t,n}) + \gamma\cdot[r(e_{t+1,n}|E_{t+1,n},\lambda_{t+1,n}) + \gamma^1\cdot r(e_{t+2,n}|E_{t+2,n},\lambda_{t+2,n}) + \cdots + \gamma^{T-t-1} r(e_{T,n}|E_{T,n},\lambda_{T,n})] = r(e_{t,n}|E_{t,n},\lambda_{t,n}) + \gamma\cdot R_{t+1} \qquad (18)$$

We denote the policy that maps states to actions by ν: $\lambda_{t,n} = \nu(E_{t,n})$. The goal of the dynamic pricing problem is to find an optimal policy ν that always chooses an action (retail price) to maximize the expected discounted reward.

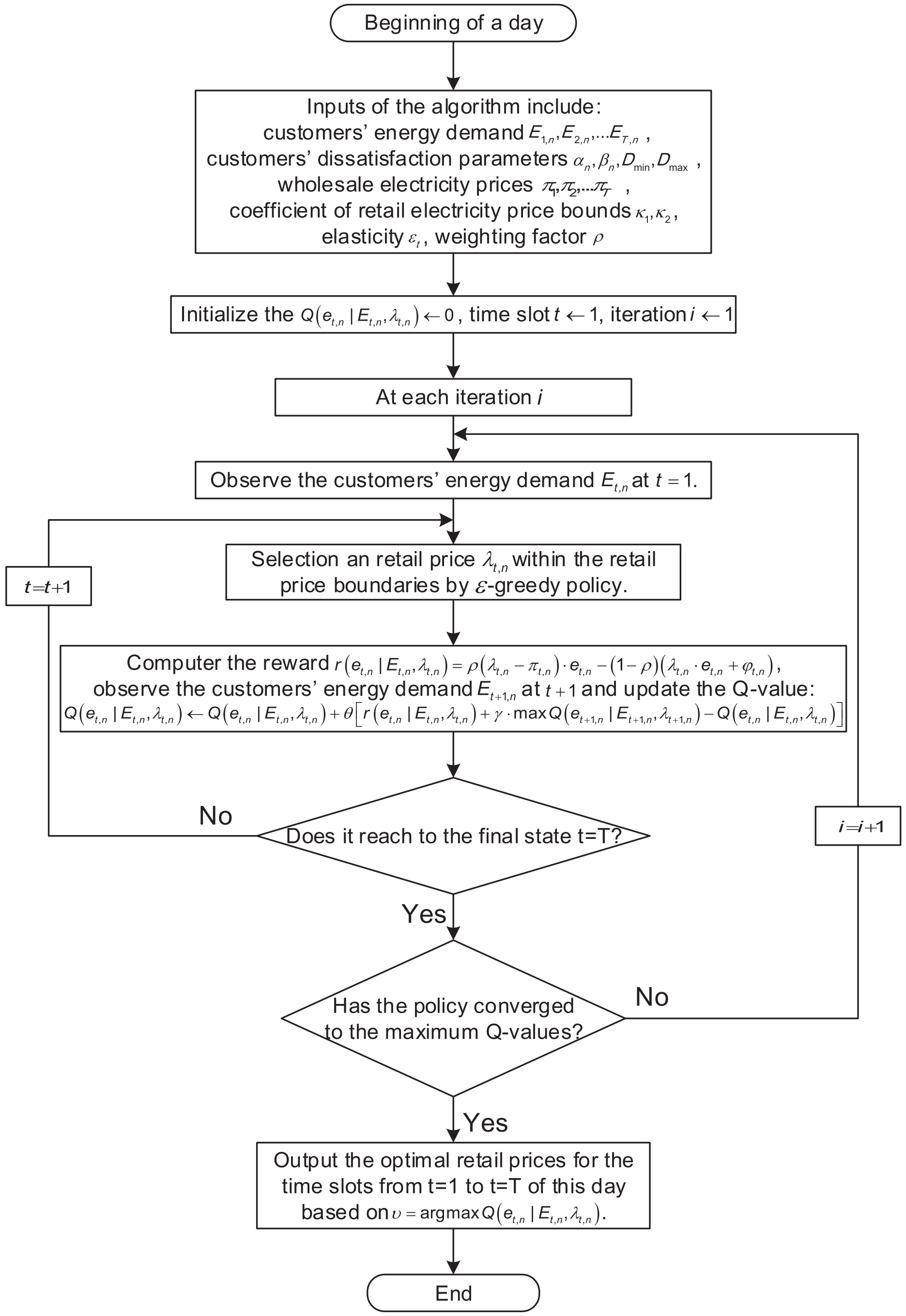

Beginning of a day
Inputs of the algorithm include: customers' energy demand $E_{1,n},E_{2,n},\ldots E_{T,n}$, customers' dissatisfaction parameters $\alpha_n,\beta_n,D_{\min},D_{\max}$, wholesale electricity prices $\pi_1,\pi_2,\ldots\pi_T$, coefficient of retail electricity price bounds κ_1,κ_2, elasticity ε_t, weighting factor ρ
Initialize the $Q(e_{t,n}|E_{t,n},\lambda_{t,n}) \leftarrow 0$, time slot $t \leftarrow 1$, iteration $i \leftarrow 1$
At each iteration i
Observe the customers' energy demand $E_{t,n}$ at $t=1$.
Selection an retail price $\lambda_{t,n}$ within the retail price boundaries by ε-greedy policy.
t=t+1
Computer the reward $r(e_{t,n}|E_{t,n},\lambda_{t,n}) = \rho(\lambda_{t,n}-\pi_{t,n})\cdot e_{t,n} - (1-\rho)(\lambda_{t,n}\cdot e_{t,n}+\varphi_{t,n})$, observe the customers' energy demand $E_{t+1,n}$ at $t+1$ and update the Q-value: $Q(e_{t,n}|E_{t,n},\lambda_{t,n}) \leftarrow Q(e_{t,n}|E_{t,n},\lambda_{t,n}) + \theta[r(e_{t,n}|E_{t,n},\lambda_{t,n}) + \gamma\cdot\max Q(e_{t+1,n}|E_{t+1,n},\lambda_{t+1,n}) - Q(e_{t,n}|E_{t,n},\lambda_{t,n})]$
No
Does it reach to the final state t=T?
i=i+1
Yes
Has the policy converged to the maximum Q-values?
No
Yes
Output the optimal retail prices for the time slots from t=1 to t=T of this day based on $\upsilon = \arg\max Q(e_{t,n}|E_{t,n},\lambda_{t,n})$.
End

Fig. 4. Flowchart for implementing the Q-learning mechanism presented in Table 1.

3.2. Adopting Q-learning to dynamic pricing problem

RL is an approach to sequential decision making in an unknown environment. It can change the policy in real time based on-line learning from past experience.

Q-learning [50], which is a type of model-free RL technique, is used to acquire the optimal policy (a sequence of retail prices in this study). The basic principle behind Q-learning is to assign a Q-value $Q(e_{t,n}|E_{t,n},\lambda_{t,n})$ to each state-action pair at time slot t, and update it at each iteration, in a way that reinforces good behavior. The optimal Q-value $Q^*(e_{t,n}|E_{t,n},\lambda_{t,n})$ denotes the maximum discounted future reward when taking action $\lambda_{t,n}$ at starting state $S(E_{t,n})$, and continue optimal following policy, which satisfies the Bellman equation [51] based on Eq. (18) as below:

$$Q^*(e_{t,n}|E_{t,n},\lambda_{t,n}) = r(e_{t,n}|E_{t,n},\lambda_{t,n}) + \gamma\cdot\max Q(e_{t+1,n}|E_{t+1,n},\lambda_{t+1,n}) \tag{19}$$

The updating mechanism to obtain the maximum Q-value using the Bellman equation is concise as shown in Table 1 [52]. $\theta\in[0,1]$ in the updating mechanism is a learning rate that indicates to what extent the newly acquired Q-value will override the old Q-value. A factor of 0 will make the agent (SP) learns nothing, while a factor of 1 would make the agent considers only the most recent information.

In the Q-learning algorithm, the agent (SP) interacts with the environment (CUs) through executing a set of actions ($\lambda_{t,n}$). Then the environment is changed and the agent receives a new state and a reward signal. Learning is taking place though trial and error during this process. Over the course of the learning process, Q-values $Q(e_{t,n}|E_{t,n},\lambda_{t,n})$ are stored and updated. After updating over a sufficient number of iterations, the Q-value will converge to a maximum value. The detailed proof can be referred to the existing studies [53–55]. Since $Q(e_{t,n}|E_{t,n},\lambda_{t,n})$ is the maximum expected system profit with action $\lambda_{t,n}$ at state $S(E_{t,n})$, we can obtain the optimal policy

$$\nu = \mathrm{argmax}Q(e_{t,n}|E_{t,n},\lambda_{t,n}) \tag{20}$$

Then the optimal retail prices are acquired.

The flowchart in Fig. 4 shows how the Q-learning algorithm presented in Table 1 is implemented to obtain the maximum Q-value (optimal retail price), in which the inputs and outputs of the algorithm are also specified.

As shown in Fig. 4, the Q-learning algorithm runs at the beginning of a day. Inputs to the algorithm include the CUs' energy demand and the wholesale electricity prices for the following T time slots, the coefficients of the retail price bounds, and other related parameters as defined in Fig. 4. Upon receiving these parameters, the SP then initializes the Q-value $Q(e_{t,n}|E_{t,n},\lambda_{t,n})$ to 0, the time slot t and the iteration i to 1. Afterwards, the SP will compute the optimal retail prices in an iterative way, i.e., at each iteration i, the SP observes the CUs' energy demand information at each time slot t, then it selects an retail price using an epsilon greedy (ε-greedy) policy [31] within the retail price boundaries. RL requires clever exploration mechanisms. Randomly selecting actions usually produces poor performance without reference to an estimated probability distribution [56]. The most common method is to use an ε-greedy policy, it is a way of selecting an action with uniform distribution from a set of available actions. Using this policy, we can either select a random action with probability ε (where ε is a fraction between 0 and 1) or we can select an action with probability $1-\varepsilon$ from the Q-values at given state in each iteration, where the random selection represents the agent selecting a retail price randomly within the price boundaries at the given state, and the selection from the Q-values indicates that the agent will scan the stored Q-values at the given state to find the "maximum" Q-value and then choose its corresponding retail price. Here, it should be noted that the "maximum" Q-value is not fixed, and can be replaced with further iterations. This gives the system some randomness, but prevents it from being completely random, to promote exploration of the action space. After choosing the retail price, the SP can gain the immediate reward arising from (12), meanwhile, the SP will also observe the CUs' energy demand at time slot $t+1$ and update the Q-value using the Q-learning mechanism in Table 1, then repeat this process until it reaches the final time slot T. After that, the SP will compare the current Q-value to the previous Q-value to verify whether it has converged to the maximum Q-value, and if not, the system will move to the next iteration $i+1$ and do this iteration again.

The termination condition of iteration is given as $|Q^i - Q^{i-1}| \leqslant \delta$: when the gap between the current Q-value and the previous Q-value is less than δ, the Q-value converges to the maximum value, the value of δ depends on the system design [53]. Finally, the SP will obtain the optimal retail prices for the following T time slots and inform these prices to its enrolled CUs.

4. Numerical simulation results

This section presents numerical simulation results to assess the performance of the proposed dynamic pricing DR program. For ease of illustration, simulations are conducted based on one SP and three CUs. The entire time cycle is divided into 24 time slots representing the 24 h of a day, thus the value of T defined in Sections 2 and 3 is 24 that indicates the final time slot of a day. Example of three CUs' load demand profiles at each time slot were obtained on the date of June 22, 2017 from SDG&E [57] as shown in Fig. 5, and were used as input $E_{1,n},E_{2,n},\ldots,E_{T,n}$ in the flowchart given in Fig. 4. Table 2 lists these three CUs' dissatisfaction related parameters [44]. The elasticity ξ_t values are shown in Table 3, derived from [48], with a differentiated response in off-/mid-/on-peak hours [47,45].

For the example of wholesale electricity prices shown in Fig. 6, we accessed the on-line data provided by ComEd [58] on the date of June 22, 2017. The retail electricity price bounds are formulated using a certain coefficient of the wholesale electricity price [25], i.e., ($\varkappa_1\pi_{t,\min} \leqslant \lambda_{t,n} \leqslant \varkappa_2\pi_{t,\max}$). In this simulation, $\varkappa_1$ and $\varkappa_2$ are set to 1.5 [25] which is an acceptable value that considers both SP's profit and CUs' costs, thus the retail price bounds are [2.4, 8.2]. The weighting factor ρ in this study is assumed to 0.9 to indicate that the SP's profit carries more relative importance than the CUs' costs. However, the impact of varying ρ from 0 to 1 is also investigated in Section 4.3. These parameters are also used as inputs to the flowchart shown in Fig. 4.

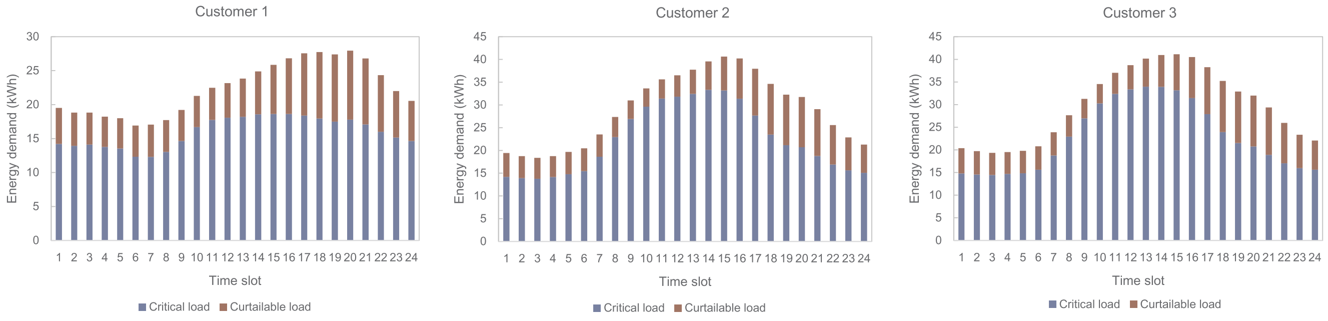


Fig. 5. Example of three customers' energy demand profiles on June 22, 2017.

Table 2
Customers' dissatisfaction related parameters.

	α_n	β_n	$D_{\min}$	$D_{\max}$
Customer 1	0.8	0.1	$0.1E_{t,n}^{curt}$	$0.5E_{t,n}^{curt}$
Customer 2	0.5			
Customer 3	0.3			

Table 3
Elasticity.

	Off-peak (1–12 am)	Mid-peak (13–16 pm, 22–24 pm)	On-peak (17–21 pm)
ξ_t	−0.3	−0.5	−0.7

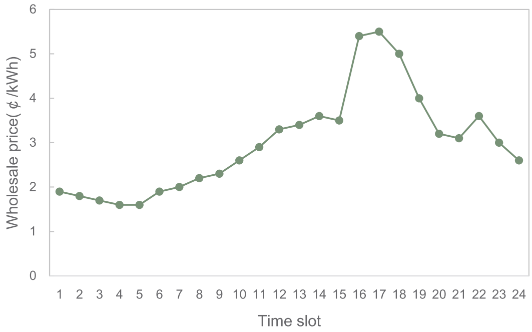


Fig. 6. Example of wholesale electricity prices on June 22, 2017.

It should be pointed out that all parameter values in the simulation scenario are specific, and may change according to the electricity market design, and the characteristics of the SP and the CUs. However, this will not distort the analysis of the simulation results.

Based on the scenario defined above, the simulations are run with iterations from which the optimal retail prices are computed, i.e., at the beginning of a day, the SP receives the wholesale electricity price from GO, the electricity load demand from CUs and other parameters defined in the scenario; then SP calculates the Q-value (retail price) by using the learning algorithm in Fig. 4, and finally obtains the maximum Q-value (optimal retail price) for the following 24 h of that day. Fig. 7 shows the convergence of the Q-value for the scenario on June 22, 2017. From this figure, it can be observed that at beginning, the agent (SP) does not know how to select an action to result a high Q-value, however, as iteration goes by, the Q-value increases as SP learns from the environment by trial and error, finally converging to the maximum value. The computation time and convergence iterations are addressed in Section 4.4.2.

Once the optimal electricity prices are obtained, each CU's optimal energy consumption is determined according to (1) and (2). Next, the performance of this dynamic pricing DR algorithm are examined from various aspects in detail.

4.1. Optimal retail prices

After running the simulation, the main output is the optimal retail electricity prices for each CU. Fig. 8 shows the optimal retail prices along with the wholesale price signals, and the curtailable load energy demand and energy consumption for the three CUs. Since the critical load demand will not change in response to the retail prices, only the curtailable load energy information is shown.

When looking into Fig. 8, we can see that the trend of retail price is similar to that of the wholesale price, reflecting the cost of buying energy from the GO; however, the price bounds are not exceeded (11). The retail price for each CU increases from time slot 6 to time slot 12 to get more profit for the SP, but a sudden decrease is observed at time slot 13. This is because at time slot 13, the elasticity changes from −0.3 to −0.5 to reflect the mid-peak period, and a continuing increase in retail price will lead to a greater reduction in energy during this period. However, this value should not surpass the reduced energy bounds (8), thus the retail price decreases at time slot 13. Comparing on-peak hours (from 17:00 to 21:00) to off-peak hours (from 7:00 to 12:00) for each CU, the price gap (retail price – wholesale price) in on-peak hours is smaller than that in off-peak hours, but the energy reduction gap (energy demand - energy consumption) is larger. This is because during on-peak hours, the electricity demand is more elastic that resulting in a greater energy reduction based on Eq. (2).

It also can be observed that the average retail price for CU 3 is larger than the other two CUs; this is because CU 3 has a smaller dissatisfaction factor α_n and prefers to reduce more energy demand, leading to a greater average retail price.

4.2. Total energy reduction

Fig. 9 shows the total energy reduction of each CU when they join the proposed dynamic pricing DR algorithm, where the yellow bar represents energy demand and the green bar represents energy consumption. As shown in Fig. 9, energy consumption is reduced from CU 1, 2, and 3 by 43.802, 51.733, and 58.947 kWh, respectively. CU 1 ($\alpha_n = 0.8$) reduces a much smaller amount of energy compared with other two CUs. This phenomenon coincides with the physical meaning of α_n, i.e., a CU with a larger α_n, prefers a smaller energy reduction to experience less dissatisfaction, in contrast, a CU with a smaller α_n, will choose a greater energy reduction during the DR process. In this situation, DR provides an opportunity to balance the energy supply and demand in the electricity market, which can effectively remove system overloads and improve the reliability of the electric power system.

4.3. Impact of weighting factor ρ

To investigate the impact of the weighting factor ρ, simulations were conducted with ρ varying from 0 to 1. Figs. 10 and 11 show the average retail price and the average profit of the SP and CUs coupled with ρ, separately. From these two figures, we can observe that an increase in ρ from 0 to 1 leads to an increase in the average retail price and the SP's average profit; however, the CUs' average profit decreases. The reason is obvious: as ρ increases, the SP's profit becomes more important compared to the CUs' costs. In particular, in the case where $\rho = 1$, the SP aims at maximizing its own profit that does not consider the CUs' costs, hence, the SP chooses relatively high retail prices. On the contrary, when $\rho = 0$, the system tends to minimize the CUs' costs; thus, the SP chooses relatively low retail prices to the CUs.

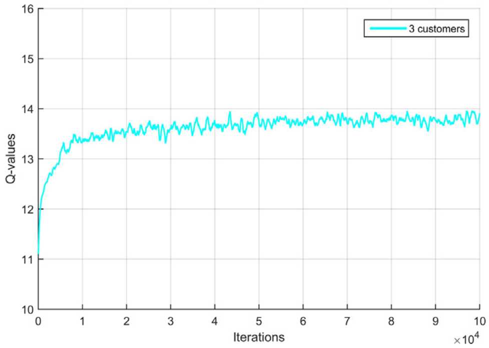
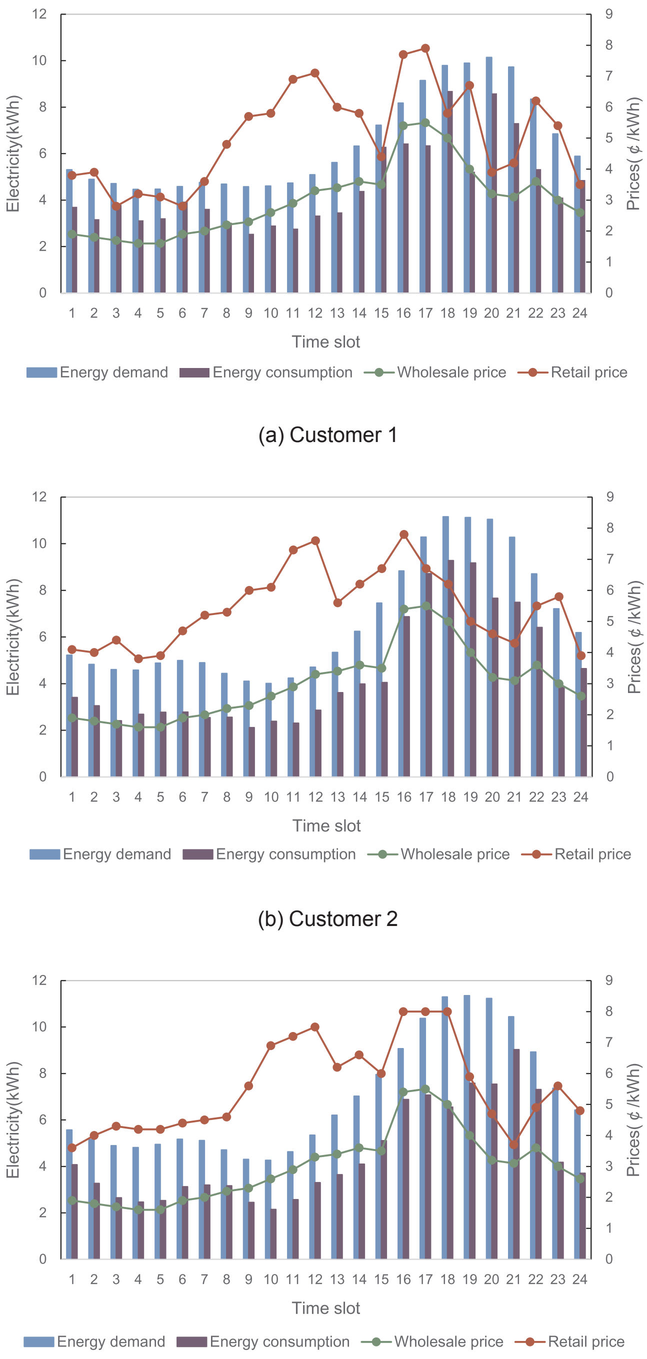


Fig. 7. Convergence of Q-value for three customers on June 22, 2017.



(a) Customer 1

(b) Customer 2

(c) Customer 3

Fig. 8. Optimal retail price and energy consumption at each time slot.

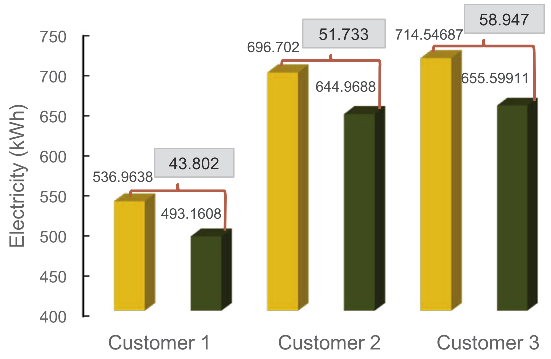


Fig. 9. Three customers' energy reduction.

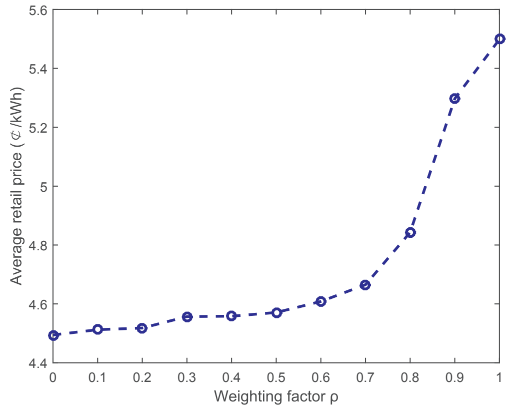


Fig. 10. Impact of the weighting factor on average retail price.

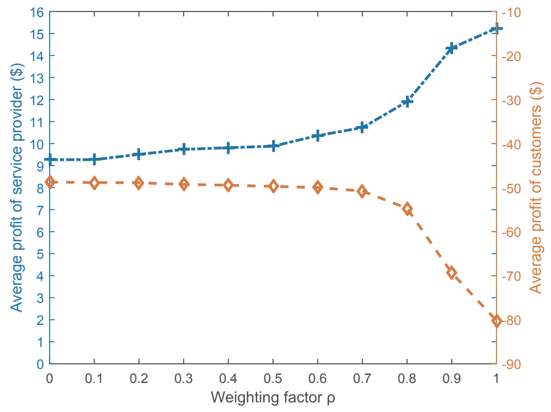


Fig. 11. Impact of the weighting factor on the SP and CUs' average profit.

4.4. System performance

To evaluate the system performance, we conducted the simulation from a single day to different days, and also extended the simulation to consider more customers.

4.4.1. Simulation with different days

We did the simulation from a single day to three different days, wherein the load data of the CUs are retrieved from SDG&E [57] and the wholesale electricity prices are obtained from ComEd [58] on the date from June 26 to June 28, 2017. Fig. 12 shows the optimal retail prices and curtailable load energy consumption for one customer within these three days. As shown in Fig. 12, similar trend of retail prices and energy consumption profiles with the previous single day were repeated on each of the three days that verifies time-differentiated retail rates can be optimized with SP to reflect the cost of buying energy from the GO and the profit of selling energy to the CUs, and the load curtailment with CUs in each day can ensure their bill savings; which further enhances the simulation analysis before, indicating that this proposed DR algorithm with RL methodology can handle the energy management between the SP and CUs.

4.4.2. Simulation with more customers

The number of CUs was increased from 3 to 10, and the simulation run with 10^5 iterations. The simulations were conducted using software with java-programmed source code in the Eclipse tool, and a 3.30 GHz, 4-core i5-6600 CPU, 8 GB RAM Window PC hardware. Fig. 13 shows the learning speed with different numbers of CUs, and Table 4 lists the

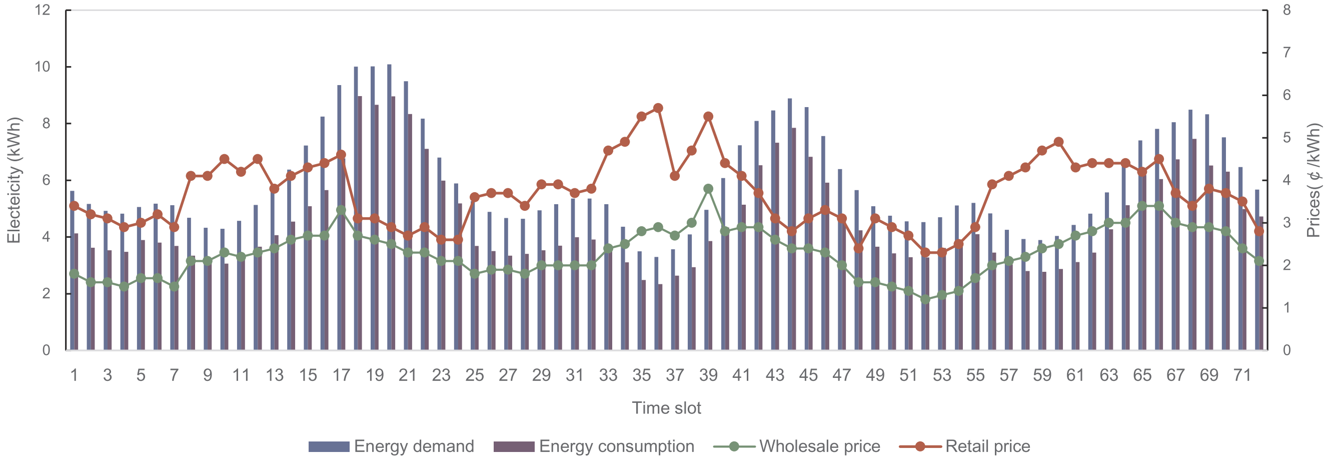


Fig. 12. Optimal retail prices and energy consumption for one customer from June 26 to June 28, 2017

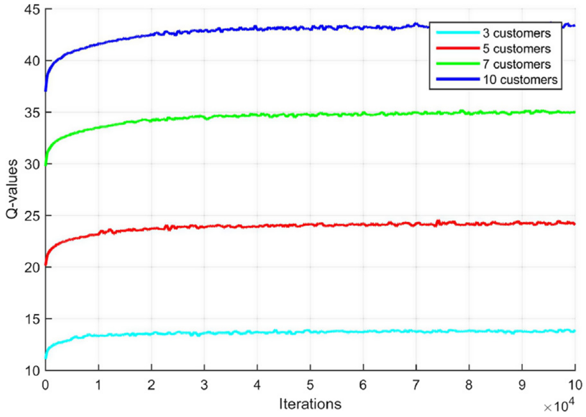


Fig. 13. Learning speed with different numbers of customers.

Table 4
Computation time and number of convergence iterations.

	Computation time (10^5 iterations)	Convergence iterations
3 Customers	34 min	2×10^4
5 Customers	55 min	3×10^4
7 Customers	71 min	4×10^4
10 Customers	98 min	5×10^4

corresponding computation time and number of convergence iterations. It can be seen that as the number of CUs increases, the computation time and number of convergence iterations also increase, respectively. Specifically, when the number of CUs is set to 10, the simulation takes 98 min with 10^5 iterations, and converges to the optimal value at close to 5×10^4 iterations. The reason for this long time is that for each CU, it has 24 time slots and at each time slot, it has a maximum of 59 actions that can be chosen, where this value of the number of actions (59) is calculated by $(\kappa_2\pi_{t,\max}-\kappa_1\pi_{t,\min})/0.1+1$. In this study, the minimum interval of the retail price is set to 0.1 to maintain the same minimum interval as wholesale electricity price, giving a value of $(8.2-2.4)/0.1+1$. Thus, for only one CU, with the ε-greedy policy, there are M^{24} permutations, where $M \in [1,...,59]$. However, modern advanced technologies, such as cloud computing, mean that this is unlikely to be a significant issue.

5. Guideline for real system implementation

The smart grid is expected to be comprehensively equipped with smart metering infrastructure and integrate information and communication technology, which enables two-way communication between different DR entities named GO, SP and CUs in this study. As shown in the electricity market model in Fig. 2, the SP buys the electricity from the GO at wholesale market prices and sells the electricity to CUs at retail market prices. The proposed dynamic pricing DR algorithm using RL methodology in this work is installed at the SP side and realized between SP and CUs to promote SP profitability and reduce CUs costs. The communication between SP and CUs can be achieved by Open Automated Demand Response (OpenADR) [59], which was created to standardize, automate and simplify DR to enable utilities to cost-effectively meet growing energy demand, and CUs to control their energy future using a common language (i.e., XML) over any existing IP-based communication network, such as Internet.

To better explore the communication between different entities, a sequence diagram is also drew to illustrate the detailed information exchanged during the DR process. As shown in Fig. 14, the GO calculates and announces the wholesale electricity prices to the SP though its own algorithms within taking into account the power generation capacity and procurement cost. Upon the receipt of wholesale electricity prices from GO, the SP will launch the DR program to its enrolled CUs. In specific, the SP will firstly collect the energy demand and private parameters from its CUs at the precondition of taking actions on behalf of its CUs, then SP will calculate the optimal retail electricity prices for CUs by maximizing Eq. 12 via RL methodology. And the detailed steps

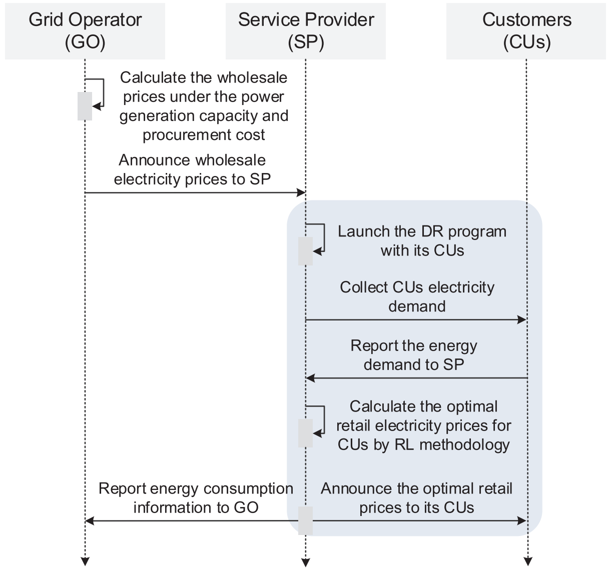

Grid Operator (GO)
Service Provider (SP)
Customers (CUs)
Calculate the wholesale prices under the power generation capacity and procurement cost
Announce wholesale electricity prices to SP
Launch the DR program with its CUs
Collect CUs electricity demand
Report the energy demand to SP
Calculate the optimal retail electricity prices for CUs by RL methodology
Report energy consumption information to GO
Announce the optimal retail prices to its CUs

Fig. 14. The diagram of information exchange between different entities.

of adopting RL to obtain the optimal retail prices is discussed in Section 3.2. Once getting the optimal retail prices, the SP will announce these prices to its CUs and report the energy consumption information to GO.

6. Conclusions and future work

In this paper, a dynamic pricing DR algorithm between an SP and CUs in a hierarchical electricity market is investigated, wherein the SP can adaptively decide the retail electricity price using the RL methodology according to the CUs' load demand profiles and dissatisfaction levels, and wholesale electricity prices.

We first formulate the dynamic pricing problem to a finite discrete MDP, and then Q-learning is employed to solve this decision-making problem. Though the use of the RL methodology, the SP does not require a pre-specified model of the CUs on which a retail price action needs to be selected; instead, the relationship between states, actions, and rewards is learned through dynamic on-line interaction with the CUs. It is capable of responding to a dynamic changing environment though ongoing learning and adaption that considers the uncertainty in CUs' load demand profiles and the flexibility of wholesale electricity prices. The numerical simulation results show that this proposed dynamic pricing DR algorithm, can promote SP profitability, reduce energy costs for CUs, balance energy supply and demand in the electricity market, and improve the reliability of the electric power system, which can be regarded as a win-win strategy for both SP and CUs.

In the future, the presented model will perform a further analysis towards the weighting factor ρ and investigate how to determine the optimal ρ between the SP's profit and CUs' costs.

Acknowledgements

This work was partially supported by National Research Foundation of Korea (NRF-2016K2A9A2A11938310), and in part by Ansan-Si hidden champion fostering and supporting project (Development of Industrial Communication-IoT Gateway & Industrial Wireless IoT Sensor) funded by Ansan city, and partially supported by the Human Resources Program in Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning (KETEP).

References

[1] Vardakas JS, Zorba N, Verikoukis CV. A survey on demand response programs in smart grids: pricing methods and optimization algorithms. IEEE Commun Surv Tut 2015;17(1):152–78.
[2] Nolan S, O'Malley M. Challenges and barriers to demand response deployment and evaluation. Appl Energy 2015;152:1–10.
[3] Qdr Q. Benefits of demand response in electricity markets and recommendations for achieving them. US Dept. Energy, Washington, DC, USA, Tech. Rep, 2006.
[4] Shen B, Ghatikar G, Lei Z, Li J, Wikler G, Martin P, et al. The role of regulatory reforms, market changes, and technology development to make demand response a viable resource in meeting energy challenges. Appl Energy 2014;130:814–23.
[5] Siano P. Demand response and smart grids–a survey. Renew Sustain Energy Rev 2014;30:461–78.
[6] Faria P, Vale Z. Demand response in electrical energy supply: an optimal real time pricing approach. Energy 2011;36(8):5374–84.
[7] Yi P, Dong X, Iwayemi A, Zhou C, Li S. Real-time opportunistic scheduling for residential demand response. IEEE Trans Smart Grid 2013;4(1):227–34.
[8] McKenna K, Keane A. Residential load modeling of price-based demand response for network impact studies. IEEE Trans Smart Grid 2016;7(5):2285–94.
[9] Wang Z, Paranjape R. Optimal residential demand response for multiple heterogeneous homes with real-time price prediction in a multiagent framework. IEEE Trans Smart Grid 2017;8(3):1173–84.
[10] Lee J-W, Lee D-H. Residential electricity load scheduling for multi-class appliances with time-of-use pricing. In: GLOBECOM workshops (GC Wkshps), 2011 IEEE. IEEE; 2011. p. 1194–8.
[11] Torriti J. Price-based demand side management: assessing the impacts of time-of-use tariffs on residential electricity demand and peak shifting in northern Italy. Energy 2012;44(1):576–83.
[12] Yang P, Tang G, Nehorai A. A game-theoretic approach for optimal time-of-use electricity pricing. IEEE Trans Power Syst 2013;28(2):884–92.
[13] Jessoe K, Rapson D. Commercial and industrial demand response under mandatory time-of-use electricity pricing. J Indust Econ 2015;63(3):397–421.
[14] Jang D, Eom J, Kim MG, Rho JJ. Demand responses of Korean commercial and industrial businesses to critical peak pricing of electricity. J Clean Product 2015;90:275–90.
[15] Zhou Z, Zhao F, Wang J. Agent-based electricity market simulation with demand response from commercial buildings. IEEE Trans Smart Grid 2011;2(4):580–8.
[16] Li XH, Hong SH. User-expected price-based demand response algorithm for a home-to-grid system. Energy 2014;64:437–49.
[17] Gao D-c, Sun Y, Lu Y. A robust demand response control of commercial buildings for smart grid under load prediction uncertainty. Energy 2015;93:275–83.
[18] Ding YM, Hong SH, Li XH. A demand response energy management scheme for industrial facilities in smart grid. IEEE Trans Indust Inf 2014;10(4):2257–69.
[19] Luo Z, Hong S-H, Kim J-B. A price-based demand response scheme for discrete manufacturing in smart grids. Energies 2016;9(8):650.
[20] Vanthournout K, Dupont B, Foubert W, Stuckens C, Claessens S. An automated residential demand response pilot experiment, based on day-ahead dynamic pricing. Appl Energy 2015;155:195–203.
[21] Li Y-C, Hong SH. Real-time demand bidding for energy management in discrete manufacturing facilities. IEEE Trans Indust Electron 2017;64(1):739–49.
[22] 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.
[23] Huang X, Hong SH, Li Y. Hour-ahead price based energy management scheme for industrial facilities. IEEE Trans Indust Inf 2017;13(6):2886–98.
[24] Rana R, Oliveira FS. Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning. Omega 2014;47:116–26.
[25] Jin M, Feng W, Marnay C, et al. Microgrid to enable optimal distributed energy retail and end-user demand response. Appl Energy 2018;210:1321–35.
[26] Meng F-L, Zeng X-J. A stackelberg game-theoretic approach to optimal real-time pricing for the smart grid. Soft Comput 2013;17(12):2365–80.
[27] Jia L, Tong L. Dynamic pricing and distributed energy management for demand response. IEEE Trans Smart Grid 2016;7(2):1128–36.
[28] Dagoumas AS, Polemis ML. An integrated model for assessing electricity retailer's profitability with demand response. Appl Energy 2017;198:49–64.
[29] Nojavan S, Zare K, Mohammadi-Ivatloo B. Optimal stochastic energy management of retailer based on selling price determination under smart grid environment in the presence of demand response program. Appl Energy 2017;187:449–64.
[30] Behboodi S, Chassin DP, Djilali N, Crawford C. Transactive control of fast-acting demand response based on thermostatic loads in real-time retail electricity markets. Appl Energy 2018;210:1310–20.
[31] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602.
[32] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[33] Li F-D, Wu M, He Y, Chen X. Optimal control in microgrid using multi-agent reinforcement learning. ISA Trans 2012;51(6):743–51.
[34] Kuznetsova E, Li Y-F, Ruiz C, Zio E, Ault G, Bell K. Reinforcement learning for microgrid energy management. Energy 2013;59:133–46.
[35] Jiang B, Fei Y. Smart home in smart microgrid: a cost-effective energy ecosystem with intelligent hierarchical agents. IEEE Trans Smart Grid 2015;6(1):3–13.
[36] Vandael S, Claessens B, Ernst D, Holvoet T, Deconinck G. Reinforcement learning of heuristic ev fleet charging in a day-ahead electricity market. IEEE Trans Smart Grid 2015;6(4):1795–805.
[37] Chiş A, Lundén J, Koivunen V. Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Trans Veh Technol 2017;66(5):3674–84.
[38] Kofinas P, Vouros G, Dounis AI. Energy management in solar microgrid via reinforcement learning using fuzzy reward. Adv Build Energy Res 2017:1–19.
[39] Wen Z, O'Neill D, Maei H. Optimal demand response using device-based reinforcement learning. IEEE Trans Smart Grid 2015;6(5):2312–24.
[40] Ruelens F, Claessens BJ, Vandael S, Iacovella S, Vingerhoets P, Belmans R. Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning. Power systems computation conference (PSCC), 2014. IEEE; 2014. p. 1–7.
[41] Ruelens F, Claessens BJ, Quaiyum S, et al. Reinforcement learning applied to an electric water heater: from theory to practice. IEEE Trans Smart Grid 2016;99:1–9.
[42] Ruelens F, Claessens BJ, Vandael S, et al. Residential demand response of thermostatically controlled loads using batch reinforcement learning. IEEE Trans Smart Grid 2017;8(5):2149–59.
[43] Wang H, Huang T, Liao X, Abu-Rub H, Chen G. Reinforcement learning in energy trading game among smart microgrids. IEEE Trans Indust Electron 2016;63(8):5109–19.
[44] Yu M, Hong SH. Incentive-based demand response considering hierarchical electricity market: a Stackelberg game approach. Appl Energy 2017;203:267–79.
[45] Gyamfi S, Krumdieck S, Urmee T. Residential peak electricity demand response—highlights of some behavioural issues. Renew Sustain Energy Rev 2013;25:71–7.
[46] Filippini M. Short-and long-run time-of-use price elasticities in swiss residential electricity demand. Energy Pol 2011;39(10):5811–7.
[47] Thimmapuram PR, Kim J. Consumers' price elasticity of demand modeling with economic effects on electricity markets using an agent-based model. IEEE Trans Smart Grid 2013;4(1):390–7.
[48] Miller M, Alberini A. Sensitivity of price elasticity of demand to aggregation, unobserved heterogeneity, price trends, and price endogeneity: evidence from US data. Energy Policy 2016;97:235–49.
[49] Yu M, Hong SH. Supply–demand balancing for power management in smart grid: a Stackelberg game approach. Appl Energy 2016;164:702–10.
[50] Watkins CJ, Dayan P. Q-learning. Mach Learn 1992;8(3–4):279–92.
[51] Wikipedia. Bellman equation < https://en.wikipedia.org/wiki/Bellman_equation > .

[52] Hwang K-S, Lin C-J, Wu C-J, Lo C-Y. Cooperation between multiple agents based on partially sharing policy. In: Advanced intelligent computing theories and applications. With aspects of theoretical and methodological issues; 2007. p. 422–32.

[53] Melo FS. Convergence of q-learning: a simple proof. Institute of Systems and Robotics, Tech. Rep; 2001. p. 1–4.

[54] Tsitsiklis JN. Asynchronous stochastic approximation and q-learning. Mach Learn 1994;16(3):185–202.

[55] Jaakkola T, Jordan MI, Singh SP. Convergence of stochastic iterative dynamic programming algorithms. In: Advances in neural information processing systems; 1994. p. 703–10.

[56] Wikipedia. Reinforcement learning < https://en.wikipedia.org/wiki/Reinforcement_learning > .

[57] Sdge.com. Home — san diego gas & electric < https://www.sdge.com > .

[58] Comed.com. Powering lives — comed - an exelon company < https://www.comed.com/Pages/default.aspx > .

[59] OpenADR.org. Home — openadr < http://www.openadr.org/ > .