Postprint: Real-Time Pricing Strategy Using Reinforcement Learning Considering Load Uncertainty

Authors: Jingqi Wang, Gao Yan, Wu Zhiqiang, Li Renjie

Date: 2022-05-10T11:22:57+00:00

Abstract

In response to the current challenges of load uncertainty, renewable energy integration, and "dual carbon" objectives in power systems, this paper establishes a real-time pricing model that incorporates load uncertainty and carbon trading within the smart grid context, while fully considering the welfare of both supply and demand sides. Leveraging reinforcement learning' s capability to handle variable complexity and non-convex nonlinear problems, the Q-learning algorithm is employed to iteratively solve the model. First, the real-time interaction process between users and suppliers is transformed into a Markov Decision Process within the reinforcement learning framework. Second, the information exchange between users and suppliers is modeled through the repeated exploration of agents in a dynamic environment. Finally, the optimal value, i.e., the maximum social welfare value, is obtained through the Q-learning algorithm in reinforcement learning. Simulation results demonstrate that the proposed real-time pricing strategy can effectively enhance social welfare and reduce total carbon emissions, thereby validating the effectiveness of the proposed model and algorithm.

Full Text

Preamble

Vol. 39 No. 9
Application Research of Computers
ChinaXiv Partner Journal

Real-Time Pricing Strategy Based on Reinforcement Learning with Load Uncertainty

WANG Jingqi, GAO Yan†, WU Zhiqiang, LI Renjie
(Business School, University of Shanghai for Science & Technology, Shanghai 200093, China)

Abstract: In response to the current challenges of load uncertainty, renewable energy grid integration, and the "dual carbon" goals in power systems, this paper establishes a real-time pricing model for smart grids that considers both load uncertainty and carbon trading, with full consideration of welfare for both supply and demand sides. Leveraging the capability of reinforcement learning to handle variable complexity and non-convex, nonlinear problems, the Q-learning algorithm is employed to iteratively solve the model. First, the real-time interaction process between users and suppliers is transformed into a Markov decision process corresponding to the reinforcement learning framework. Second, the information exchange between users and suppliers is represented through the agent' s repeated exploration in a dynamic environment. Finally, the Q-learning algorithm seeks the optimal value—i.e., the maximum social welfare. Simulation results demonstrate that the proposed real-time pricing strategy can effectively enhance social welfare and reduce total carbon emissions, thereby validating the effectiveness of the model and algorithm.

Keywords: real-time pricing; reinforcement learning; Markov decision process; load uncertainty; "dual carbon" goal

0 Introduction

In smart grid systems, bidirectional flow of electricity and information enables the simultaneous achievement of economic, efficient, and environmentally friendly objectives. The deepening penetration of renewable energy generators introduces greater uncertainty into power generation systems. Demand-side management presents substantial industrial opportunities centered around generators, distributed renewable energy, carbon trading markets, and user demand.

With advances in information communication and smart terminals, intensified electricity price fluctuations in power markets increase ordinary users'willingness to participate in power system regulation. Demand-side management of power systems can effectively shave peak loads, optimize electricity consumption patterns, and enhance power system stability and security. Demand response (DR) represents one solution for demand-side management. Existing DR strategies [1-3] are typically categorized into incentive-based DR (IBDR) and price-based DR (PBDR). PBDR adjusts users' consumption patterns through electricity price modifications, while IBDR provides users with fixed or time-varying incentive payments. Many studies employ price-based DR by considering user

behavior, with real-time pricing representing a crucial research direction within PBDR. This strategy directly controls electricity prices to adjust user-side load demand, aiming to effectively flatten user electricity demand through real-time price signals.

Literature [4] first proposed a real-time pricing model maximizing social welfare, simultaneously considering supplier profits and user welfare, solved using distributed gradient descent methods. Numerical simulations verified that the model could achieve peak shaving while benefiting both users and suppliers. Building upon this foundation, real-time pricing models with social welfare maximization as the objective function have been widely applied. Literature [5] employed smoothing techniques to smooth the commonly used quadratic piecewise utility functions in real-time pricing and simulated user utility. Literature [6] established a real-time pricing optimization model minimizing peak-valley differences and proposed a simultaneous perturbation stochastic approximation algorithm dependent on online power fluctuations. Literature [7] integrated blockchain into real-time pricing models, effectively improving renewable energy utilization in microgrids. Users also participated as independent nodes in grid decision-making, with blockchain transactions enhancing user electricity precision and total social welfare. Literature [8] effectively combined social welfare maximization models with microgrids, establishing a two-level optimization model accounting for uncertainty and solving it using a PSO-BBA algorithm. Compared with deterministic functions, this approach better achieved peak shaving. Literature [9] discussed the role of minimum power supply constraints in social welfare maximization models, introduced effective cost functions, and proposed dual online algorithms for model improvement. Literature [10] formulated real-time pricing as a non-cooperative game problem and solved it using distributed online algorithms, providing a more precise description of user interaction processes.

From an optimization perspective, the aforementioned real-time pricing strategies fall into two categories: gradient-based optimization algorithms [4–7] and metaheuristic optimization algorithms [8–11]. The former, such as conjugate gradient and Newton methods, offer high computational efficiency but struggle with nonlinear, non-smooth functions, or chance constraints. Metaheuristic algorithms like genetic algorithms and particle swarm optimization, which possess strong global search capabilities, are largely model-independent and effectively address these challenges. However, existing pricing strategies often pre-determine model parameters and employ centralized algorithms, which neither adequately consider load uncertainty nor provide corresponding privacy protection measures. These approaches suffer from slow computation speeds and low reliability when facing large-scale batch data, making innovative real-time pricing mechanisms theoretically and practically significant.

From a temporal correlation perspective, most studies treat real-time pricing as multiple single-period problems [4–9], where each period exists independently without fully considering overall state transition characteristics. This limits

the accuracy of real-time pricing model interactions and ignores the temporal correlation between user consumption and supplier generation. Markov decision processes can describe load relationships across stages using state transition matrices, fully considering period correlations. Literature [11, 13] studied real-time pricing based on Markov processes, considering both known and unknown parameters, and validated model rationality and algorithm feasibility.

Most real-time pricing research relies on traditional algorithms based on analytical models and deterministic rules. In recent years, reinforcement learning has achieved new breakthroughs. Unlike traditional optimization algorithms, reinforcement learning can explore random actions in dynamic environments and learn from experience, providing crucial support for complex system decision-making. Reinforcement learning is concise and uses reward functions to evaluate decision-making behaviors, yielding effective solution strategies with convergent results. It has been applied in various fields such as game control and computer vision [12]. Research on reinforcement learning for power systems holds broad prospects, and its application in demand-side management will effectively expand new load-side consumption patterns [13].

Recent reinforcement learning applications in demand-side management fall into two categories. The first stands from the consumer perspective, designing effective response patterns to maximize consumer benefits when facing supplier pricing strategies [15]. The second stands from the utility company perspective, designing effective strategies to improve social welfare, thereby enhancing welfare encompassing both user and supply sides [14, 16]. Lu et al. [14] first applied reinforcement learning to demand-side management, proposing a real-time pricing algorithm for hierarchical electricity markets that represents supplier-user interactions as Markov decision processes to dynamically determine optimal electricity prices. Literature [15] used reinforcement learning to obtain energy scheduling for specific devices in demand response, maximizing user returns during scheduling periods. Literature [16] applied reinforcement learning frameworks to demand response strategies, considering industrial user-supplier interactions to maximize supplier long-term profits. Literature [17] treated microgrids as intelligent agents in a reinforcement learning approach, where microgrids could independently select energy trading strategies to maximize average returns. Literature [18] proposed a multi-microgrid energy management method based on neural networks and reinforcement learning algorithms, where operators predicted power exchanges among microgrids through deep neural networks and obtained retail pricing strategies via Monte Carlo methods, achieving profit maximization and minimizing peak-to-average ratios on the demand side while improving electricity reliability.

However, the aforementioned reinforcement learning-based demand-side management studies lack comprehensive consideration of social welfare, carbon trading, and load uncertainty [13–17]. Based on this analysis, it is necessary to extend real-time pricing models accordingly. Using reinforcement learning algorithms to solve real-time pricing models offers significant advantages. Con-

sidering carbon emission rights from power generation and associated costs or benefits from carbon trading, this paper promotes renewable energy consumption through carbon emission rights trading, thereby supporting "dual carbon" goal achievement.

The main contributions of this paper are:

a) Considering a supplier system comprising traditional and renewable energy suppliers and a user system comprising residential and large consumers, with full representation of supply-demand welfare and social welfare maximization as the objective.

b) Transforming the interaction process between users and suppliers into a Markov decision process through the reinforcement learning framework, where the agent (supplier) learns and obtains optimal real-time pricing strategies through iterative processes with all users.

c) Mapping real-time pricing model elements to reinforcement learning components while fully considering load uncertainty, enabling more refined modeling.

d) Effectively improving renewable energy consumption rates in power systems through carbon trading, which holds important practical significance for promoting sustainable green energy development.

---

1 System Model

Consider a smart grid system containing two types of suppliers and multiple terminal users (system framework shown in Figure 1, symbol descriptions in Table 1). Suppliers include traditional energy suppliers and renewable energy suppliers, with renewable generation comprising wind and photovoltaic power. Due to the intermittent and unstable nature of renewable generation, suppliers cannot control output per time period and must forecast each period' s output based on renewable unit characteristics and daily weather conditions. User electricity consumption is prioritized to be supplied by renewable energy to promote renewable consumption.

The user side considers residential and large consumers. Residential energy consumption is for daily life, while large industrial and commercial users consume energy for higher-level production activities. Users and suppliers directly interact through smart meters for bidirectional information exchange: suppliers can obtain user electricity consumption via smart meters, while users can receive price signals for the next period from suppliers. Suppliers maximize profits through real-time pricing strategies, while users dynamically adjust energy demand through demand response to reduce electricity costs, enabling dynamic price adjustment based on interactive load demand and generation costs.

Let $\mathcal{R}$ denote the residential user set, $\mathcal{L}$ the large user set, and $\mathcal{N} = \mathcal{R} \cup \mathcal{L}$ the complete user set. Supplier-user power interaction follows a daily cycle divided into T periods, with prices updated hourly ($T = 24$). Considering load uncertainty and carbon trading scenarios, this paper establishes a real-time pricing model under social welfare maximization.

Let $\mathcal{T} = \{1, 2, 3, \ldots, T\}$ be the set of all time periods. The model assumes:

Figure 1. System framework

1.1 User-Side Model

Generally, users' electricity quantity requirements and utility values from consuming the same amount differ. Based on load priority and demand characteristics, this paper assumes user load configuration divides into two categories: basic load and flexible load [19]. Loads with fixed demand within specific periods are called basic loads, while loads with flexible scheduling are called flexible loads. Users can achieve demand response by flexibly adjusting flexible loads such as air conditioners and water heaters. In demand response, suppliers guide users to change electricity demand in each period through dynamic price adjustments, thereby achieving supply-demand balance.

1.1.1 Load Function Assume basic loads must be strictly satisfied and cannot be regulated through demand response, such as essential living electricity. The relationship between user n' s basic load $X_{n,t}^{\text{basic}}$ and basic load demand $D_{n,t}^{\text{basic}}$ in period t is:

$$X_{n,t}^{\text{basic}} = D_{n,t}^{\text{basic}}, \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T}$$

Consider flexible loads that can be scheduled in time and power, defined as flexible loads. Flexible loads relate to current electricity prices and user price elasticity coefficients. The definition of user n' s flexible load in period t is [16]:

$$X_{n,t}^{\text{flex}} = D_{n,t}^{\text{flex}} \cdot \left(1 - c_n \cdot \frac{p_{n,t} - \pi_0}{\pi_0}\right), \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T}$$

where $p_{n,t}$ represents the electricity price user n pays in period t, c_n is user n' s price elasticity coefficient, and $D_{n,t}^{\text{flex}}$ is user n' s flexible load demand in period t. Price increases lead to actual loads smaller than expected demand. Electricity prices should remain within a fixed interval: π_0 is the benchmark price, while $c_n^{\min}$ and $c_n^{\max}$ represent lower and upper bounds of electricity price coefficients, which differ across user types. Electricity price constraints ensure reasonable transaction prices for both supply and demand sides [20].

Let $X_{n,t}$ denote user n' s total electricity load in period t, comprising basic and flexible loads:

$$X_{n,t} = X_{n,t}^{\text{basic}} + X_{n,t}^{\text{flex}}, \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T}$$

Due to real-world variations and user-side load randomness, power facilities typically face load fluctuations. Considering load uncertainty, user n' s total load $X_{n,t}$ in period t is:

$$X_{n,t} = X_{n,t}^{\text{basic}} + X_{n,t}^{\text{flex}} + \delta_{n,t}, \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T}$$

where $\delta_{n,t}$ is a random variable following a normal distribution $\mathcal{N}(0, \sigma_{n,t}^2)$, characterizing user-side load uncertainty [21].

Since flexible loads are price-sensitive, reasonable scheduling can effectively achieve peak shaving.

1.1.2 Utility Function In microeconomics, utility functions $U(x)$ characterize user satisfaction. Assume each user' s behavior toward different electricity prices is independent, with varying preferences for load demand. Elasticity coefficients β_n effectively reflect different user preferences. The utility function $U(x)$ must satisfy: $\frac{\partial U}{\partial x} > 0$, $\frac{\partial^2 U}{\partial x^2} < 0$, $U(0) = 0$, and $\beta_n > 0$.

In existing real-time pricing models, user utility functions are commonly represented by quadratic functions [22]. User n' s utility function in period t can be expressed as:

$$U_{n,t}(X_{n,t}) = \begin{cases} \alpha_n X_{n,t} - \frac{\beta_n}{2} X_{n,t}^2, & 0 \le X_{n,t} \le \frac{\alpha_n}{\beta_n} \\ \frac{\alpha_n^2}{2\beta_n}, & X_{n,t} > \frac{\alpha_n}{\beta_n} \end{cases}$$

where $X_{n,t}$ is user n' s total load in period t. Parameters α_n and β_n are user utility parameters [23,24] that should be estimated from historical data and user surveys in practical applications. Different user types' utility variations are characterized by parameters α_n and β_n.

Similar to residential users, large users' utility increases with electricity consumption within a certain range, remaining constant when reaching a predefined maximum load. However, user-side loads typically do not reach saturation.

In summary, user-side welfare can be expressed as the expectation of the user' s current-period utility function minus payment costs. Let C_π denote user-side welfare, expressed as:

$$C_\pi = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{N}} \left(U_{n,t}(X_{n,t}) - p_{n,t} X_{n,t}\right)\right]$$

1.2 Supplier Model

Suppliers provide electricity according to user demand, enabling power production and transmission. In recent years, renewable energy integration has significantly increased power system randomness. Let L_t^e and L_t^r represent traditional and renewable energy supplier generation in period t, respectively. Since total supply must cover all user demands, L_t^e and L_t^r must satisfy unit generation interval constraints:

$$\sum_{n\in\mathcal{N}} X_{n,t} \leq L_t^e + L_t^r, \quad \forall t \in \mathcal{T}$$

$$L_t^{e,\min} \leq L_t^e \leq L_t^{e,\max}, \quad \forall t \in \mathcal{T}$$

where $L_t^{e,\min}$ and $L_t^{e,\max}$ represent minimum and maximum generation of traditional energy suppliers in period t.

1.2.1 Traditional Energy Supplier Assume traditional energy supplier costs primarily derive from fossil fuel consumption and operation maintenance. The traditional energy generation cost function is monotonically increasing and strictly convex, commonly represented by quadratic functions [23]. The generation cost function $C_t^e(L_t^e)$ in period t is:

$$C_t^e(L_t^e) = a_t(L_t^e)^2 + b_t L_t^e + c_t, \quad \forall t \in \mathcal{T}$$

where L_t^e is the total electricity provided by traditional energy suppliers in period t, and a_t, b_t, c_t are preset parameters with $a_t > 0$, $b_t \geq 0$, $c_t \geq 0$.

1.2.2 Renewable Energy Supplier Due to intermittency in natural resources like solar irradiance and wind speed, renewable energy output exhibits significant uncertainty. Insufficient dispatchable capacity leads to wind/solar curtailment, severely compromising system stability. This paper assumes renewable energy lacks storage capability and has no coupling constraints between generation periods, with suppliers prioritizing renewable energy to improve consumption rates.

Photovoltaic output depends on solar irradiance, ambient temperature, and PV module characteristics. The actual PV output power in period t is [8]:

$$P_t^{PV} = P_{\text{rated}}^{PV} \cdot \frac{G_t^{PV}}{G_{\text{ref}}^{PV}} \cdot [1 + \eta^{PV}(T_t^{PV} - T_{\text{ref}}^{PV})] \cdot N^{PV}, \quad \forall t \in \mathcal{T}$$

where P_{rated}^{PV} is rated PV output power, G_t^{PV} is irradiance at the operating point, G_{ref}^{PV} is standard irradiance, η^{PV} is the power temperature coefficient, T_t^{PV} is

cell temperature at the operating point, T_{ref}^{PV} is reference temperature, and N^{PV} is the number of PV devices.

Wind power output relates to actual wind speed in each period. Generally, wind speed fluctuations follow a Rayleigh distribution. The actual wind turbine output power in period t is [8]:

$$P_t^{WT} = \begin{cases} 0, & v_t^{\text{rated}} < v_{\text{in}} \text{ or } v_t^{\text{rated}} > v_{\text{out}} \\ P_{\text{rated}}^{WT} \cdot \frac{v_t^{\text{rated}} - v_{\text{in}}}{v_{\text{rated}} - v_{\text{in}}}, & v_{\text{in}} \le v_t^{\text{rated}} < v_{\text{rated}} \\ P_{\text{rated}}^{WT}, & v_{\text{rated}} \le v_t^{\text{rated}} \le v_{\text{out}} \end{cases}, \quad \forall t \in \mathcal{T}$$

where v_t^{rated} is actual wind speed, v_{rated} is rated wind speed, v_{in} and v_{out} are cut-in and cut-out wind speeds, P_{rated}^{WT} is rated output power, and N^{WT} is the number of wind turbines.

Renewable energy supply comprises wind and PV outputs. Let L_t^r denote total renewable energy supply in period t:

$$L_t^r = P_t^{PV} + P_t^{WT}, \quad \forall t \in \mathcal{T}$$

Since renewable generation costs are negligible, assume renewable supplier costs derive from operation and maintenance expenses. This paper uses a quadratic cost function for renewable equipment maintenance loss costs in period t [25]:

$$C_t^r(L_t^r) = \delta_{\text{RE}}(L_t^r)^2 + \sigma_{\text{RE}} L_t^r, \quad \forall t \in \mathcal{T}$$

where δ_{RE} is the renewable equipment maintenance loss cost coefficient.

1.2.3 Carbon Trading Model Carbon emission rights trading promotes "dual carbon" goal achievement in power systems. Under carbon trading mechanisms, the state allocates carbon emission quotas based on suppliers' total generation. If actual emissions are less than allocated quotas, the surplus can be sold for profit; if actual emissions exceed quotas, suppliers must purchase excess emission rights, incurring carbon over-emission costs [26].

Suppliers obtain carbon emission rights through traditional and renewable generation. The carbon emission quota E_t^D allocated to generation units in period t is:

$$E_t^D = \delta_e L_t^e + \delta_r L_t^r, \quad \forall t \in \mathcal{T}$$

where δ_e and δ_r are unit carbon emission quota allocation rates for traditional and renewable generation, respectively.

Considering traditional energy generation as the carbon emission source, actual carbon emissions E_t from traditional generation units in period t are [27]:

$$E_t = \lambda_e L_t^e, \quad \forall t \in \mathcal{T}$$

where λ_e is the carbon emission coefficient per unit of electricity from traditional generators.

The carbon trading cost C_t^E in period t is calculated as:

$$C_t^E = p_e \cdot (E_t - E_t^D), \quad \forall t \in \mathcal{T}$$

where p_e is the market price per unit of carbon emission rights. $C_t^E \geq 0$ represents carbon trading costs from excess emissions, while $C_t^E < 0$ indicates carbon trading revenue.

1.3 Real-Time Pricing Model Under Load Uncertainty

Considering social welfare maximization objectives, the smart grid real-time pricing model accounting for load uncertainty is formulated as:

$$\max_{p_{n,t}, L_t^e, L_t^r} \quad \mu_1 \cdot \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\left(U_{n,t}(X_{n,t}) - p_{n,t}X_{n,t}\right)\right] + \mu_2 \cdot \mathbb{E}\left[\sum_{t\in\mathcal{T}}\left(\sum_{n\in\mathcal{N}} p_{n,t}X_{n,t} - C_t^e(L_t^e) - C_t^r(L_t^r) - C_t^E\right)\right]$$

subject to constraints (8), (9), and (13).

where μ_1 and μ_2 are weighting coefficients for user-side and supplier-side welfare, respectively, with $\mu_1, \mu_2 \in (0, 1)$ and $\mu_1 + \mu_2 = 1$. These values are determined jointly by supplier pricing strategies and user demand elasticity. Optimal social welfare occurs when total user load equals total supplier generation.

1.4 Objective Function Transformation

The objective function (19) can be separated into user and supplier components. Based on expectation properties:

$$\mathbb{E}[X_{n,t}] = X_{n,t}^{\text{basic}} + X_{n,t}^{\text{flex}} + \mathbb{E}[\delta_{n,t}] = X_{n,t}^{\text{basic}} + X_{n,t}^{\text{flex}} + \mu_\delta$$

where μ_δ represents the mean of random variable $\delta_{n,t}$.

The transformed deterministic model (21) becomes:

$$\max_{p_{n,t}, L_t^e, L_t^r} \quad \mu_1 \cdot \sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\left(\hat{U}_{n,t}(X_{n,t}) - p_{n,t}\mathbb{E}[X_{n,t}]\right) + \mu_2 \cdot \sum_{t\in\mathcal{T}}\left(\sum_{n\in\mathcal{N}} p_{n,t}\mathbb{E}[X_{n,t}] - C_t^e(L_t^e) - C_t^r(L_t^r) - C_t^E\right)$$

where $\hat{U}_{n,t}(X_{n,t})$ incorporates the variance of $\delta_{n,t}$.

---

2 Algorithm Design

This section transforms the real-time pricing model into a Markov decision process. Reinforcement learning based on Markov processes applies well to single-agent environments. This paper employs an efficient Q-learning algorithm adaptable to various environments.

Reinforcement learning (RL) is an optimal action decision-making technique that self-learns in different environments [28]. Its most important feature is that agents learn and record corresponding feedback, aiming to maximize long-term cumulative rewards. Agents spontaneously select actions with higher reward values through parameter adjustment, offering advantages in self-learning and self-updating. The interaction process is shown in Figure 2.

Temporal-difference (TD) learning is the core RL algorithm, with Q-learning being a common TD method. Its value function update formula is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right]$$

where $\alpha \in [0, 1]$ is the learning rate and $\gamma \in [0, 1]$ is the discount factor indicating the relative importance of future rewards.

Figure 2. The interaction process between agent and environment in reinforcement learning

TD learning combines Monte Carlo and dynamic programming (DP) methods. Similar to Monte Carlo, it learns directly from historical experience. Similar to DP, it updates current state value functions using successor state value functions.

In each time period, the agent aims to maximize cumulative discounted returns —the sum of current and future period returns:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

The Q-value function update for real-time pricing is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]$$

RL solving for optimal policies transforms into seeking optimal state-action value functions. Implementing policy π transfers state s to s' with transition

probability $\mathcal{P}^a_{ss'}$ and reward function $\mathcal{R}^a_{ss'}$, yielding the Bellman equation for action-value functions [14]:

$$Q^\pi(s,a) = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \sum_{a'} \pi(a'|s') Q^\pi(s',a') \right]$$

where $s \in \mathcal{S}$ represents the state set.

The optimal state-value function $V^*(s)$ under optimal policy π^* is:

$$V^*(s) = \max_{a \in \mathcal{A}} \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^*(s') \right]$$

where $V^*(s)$ is the state-value function under optimal policy and $\mathcal{A}$ represents all possible actions in state s.

When transition probabilities $\mathcal{P}^a_{ss'}$ and cumulative rewards $\mathcal{R}^a_{ss'}$ are known, the Bellman optimality equation is nonlinear. Optimal policy π^* is typically solved iteratively [29], with algorithms classified as value iteration or policy iteration based on whether value functions or state-action value functions are iterated.

The optimal policy is:

$$\pi^*(a|s) = \begin{cases} \arg\max_a Q^*(s,a), & \text{if } a = \arg\max_a Q^*(s,a) \\ 0, & \text{otherwise} \end{cases}$$

When applying Q-learning to solve real-time pricing, the electricity pricing problem can be formulated as a Markov decision process requiring RL model elements $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ [30]. Through continuous strategy selection by the agent against the environment and iterative feedback, the optimal strategy—i.e., the optimal real-time electricity price—is obtained. Suppliers set prices based on current-period user consumption (strategy), then users transition from previous to next states according to prices. This transition depends primarily on current actions and previous user states. The RL framework (Figure 3) represents energy trading strategies between suppliers and users to maximize overall social welfare.

a) State space $\mathcal{S}$: State space definition must comprehensively consider factors affecting decision-making. For real-time pricing, $\mathcal{S}$ comprises load demand, supply, and time periods. $p_{n,t}$ represents the electricity price suppliers offer user n in period t. $X_{n,t}$ represents user energy demand after receiving price signals, updated in real-time as user feedback to prices. The state space is:

$$\mathcal{S} = \{(p_{n,t}, X_{n,t}, L^e_t, L^r_t) \mid n \in \mathcal{N}, t \in \mathcal{T}\}$$

b) Action space $\mathcal{A}$: The agent outputs actions—i.e., supplier-offered prices $p_{n,t}$—as continuous variables without discretization. The action space is set as a continuous price interval:

$$\mathcal{A} = \{p_{n,t} \mid p_{n,t} \in [c_n^{\min}, c_n^{\max}]\}$$

c) State transition probability $\mathcal{P}$: Corresponding to equation (24), $\mathcal{P}_{ss'}^{a}$ represents the probability that the environment transitions to next state s' when the agent takes action a in state s.

d) Discount factor γ: $\gamma \in [0, 1]$ is the discount factor representing the proportion of future reward expectations under current decisions. Larger γ values increase future rewards' importance relative to current rewards, making current decisions more impactful on subsequent states. A discount rate of 0 (considering only current rewards) causes algorithmic "short-sighted" optimization.

e) Reward $\mathcal{R}$: In this section, the real-time pricing model considers social welfare maximization as its objective, mapping rewards to social welfare values. The single-stage reward is defined as:

$$r_t = \mu_1 \cdot \sum_{n \in \mathcal{N}} \left(U_{n,t}(X_{n,t}) - p_{n,t} X_{n,t}\right) + \mu_2 \cdot \left(\sum_{n \in \mathcal{N}} p_{n,t} X_{n,t} - C_t^e(L_t^e) - C_t^r(L_t^r) - C_t^E\right)$$

Figure 3. Real-time pricing mechanism based on reinforcement learning

At iteration start ($t = 0$), the model aims to maximize total benefits across all periods. After the first period, the objective converts to maximizing remaining periods' total rewards. Maximizing rewards for remaining periods at each period' s end fully considers temporal correlation. The Q-learning real-time pricing mechanism is as follows:

Algorithm 1: Q-Learning Real-Time Pricing Mechanism

Input: Preset parameters, initial load values $X_{n,0}$, generation L_0^e, L_0^r, and price $p_{n,0}$.

Output: Optimal action-value function Q^*, optimal generation L_t^{e*}, L_t^{r*}, and electricity price $p_{n,t}^*$.

1. Initialize data: Set $Q(s, a) = 0$ for all s, a, iteration counter $k = 0$.
2. Iterate:
   a) Repeat for each episode:
   b) If $|Q_k - Q_{k-1}| < \delta$, stop iteration and output Q^*. Otherwise proceed.
   c) Observe state s_t and select action a_t under initial policy.
   d) Agent observes reward r_t and next state s_{t+1}.

e) Update action-value function:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]$$

f) Check episode completion: If $t = T$, exit loop. Otherwise proceed.
g) Calculate real-time prices, generation, and loads using equations (2), (7), and (28).

A common RL optimization method is the ϵ-greedy policy [31], which selects random actions with given probability distribution. At day start, the agent (supplier) first randomly selects initial policy a_0 (initial price) within price bounds. After selecting the initial policy, the agent immediately receives a reward and observes the environment to update Q-values (social welfare values). Through repeated price adjustments, Q-values increase and converge to maximum values through agent-environment learning. When $|Q_k - Q_{k-1}| < \delta$, the termination condition is met, and the model converges to the optimal value—maximum social welfare—while obtaining optimal state-action pairs.

---

3 Case Study

3.1 Case Background

This section presents numerical simulation experiments to verify model rationality and algorithm effectiveness. Assume a region with suppliers and a community containing 20 residential users and 5 large users. Smart meters can aggregate same-type user information for unified scheduling while protecting user privacy. Typical daily PV and wind outputs are considered (see Figures A1-A3 in appendix). Users directly participating in electricity trading represent aggregated loads across different user types. Residential and large user load data from literature [33] are adjusted proportionally as data sources, with load demands shown in Figures A2 and A3.

Price elasticity coefficients are in Appendix Table 3. Carbon trading price p_e (price per carbon emission right in the carbon market) is set at 130 yuan/ton under the benchmark scenario [27]. Different utility parameters are set for different user types [32], with user utility parameters β_n following uniform distributions. User-side model parameters are detailed in Table 4. RL algorithm initial parameters and supply-side parameters are in Appendix Table 5, with weighting coefficient μ_1 adaptively selected by the algorithm. Shanghai' s time-of-use (TOU) pricing is compared with the proposed real-time pricing model, with TOU pricing shown in Table 6.

3.2 Results Analysis

User-side real-time prices and load reduction are shown in Figures 4 and 5. Figure 4 shows similar real-time price trends for both user types. Comparing peak periods (10:00-15:00, 18:00-21:00) with off-peak periods (21:00-7:00), peak-period price change rates and load reduction ratios are higher due to greater price elasticity coefficients, enabling better demand-side peak shaving with smaller price adjustments. Price interval constraints maintain reasonable price ranges. Figure 5 shows total load reduction: large users' reduction exceeds residential users due to higher prices and greater price volatility during peak periods.

Figure 4. Real-time electricity prices for the user side

Figure 5. Load reduction of the user side

Figure 6. Welfare values of the user side

Figure 7. Total power supply, the amount of power supplied by traditional and new energy suppliers

Figure 8. Welfare values and carbon trading costs of the supply side

User-side welfare values are shown in Figure 6. Large users exhibit higher welfare than residential users, with larger price change rates during peak periods, indicating higher willingness to participate in load regulation. Users adjust loads according to welfare maximization objectives when facing price changes.

Figures 7 and 8 reflect final supplier generation, supplier welfare, and carbon trading costs. At optimal social welfare, total user load equals total supplier generation. With carbon trading, suppliers prioritize wind and PV generation, reducing fossil fuel pressure and generation costs. Negative carbon emission costs in Figure 8 indicate carbon trading increases supplier welfare—renewable generation earns carbon emission quotas exceeding actual emissions, effectively improving supply-side welfare. The case validates model rationality and effectiveness under carbon trading, demonstrating that carbon trading promotes green energy development and renewable consumption at the societal level.

To further compare model rationality and effectiveness, the proposed real-time pricing scenario (Scenario 1) is compared with three alternatives: (2) real-time pricing under deterministic load, (3) "short-sighted" optimization under deterministic load, and (4) social welfare under TOU pricing. Table 2 shows model indicator values for four scenarios in a typical day.

Assuming identical base parameters across scenarios, simulation results show the proposed real-time pricing achieves similar social welfare values that are always superior to TOU pricing. Although welfare under uncertainty is slightly lower than under deterministic conditions, uncertain real-time pricing better matches actual user consumption patterns. The proposed strategy achieves favorable social welfare while ensuring model robustness, validating effectiveness

and rationality.

Table 2. Model indicator values in four types of scenarios

Scenario	Social Welfare Value	User-Side Welfare Value	Supplier-Side Welfare Value	Carbon Trading Cost (yuan)
Scenario 1 (Proposed)	[Value]	[Value]	[Value]	[Value]
Scenario 2 (Deterministic)	[Value]	[Value]	[Value]	[Value]
Scenario 3 (Shortsighted)	[Value]	[Value]	[Value]	[Value]
Scenario 4 (TOU)	[Value]	[Value]	[Value]	[Value]

4 Conclusion

This paper employs the Q-learning algorithm within a reinforcement learning framework to solve real-time pricing. Case simulations verify the proposed strategy' s effectiveness with the following advantages:

a) The RL framework transforms real-time pricing into a Markov decision process, where suppliers as agents learn and obtain optimal pricing strategies through iterative interactions with all users, enabling automatic price optimization.

b) User classification effectively improves system performance while matching actual consumption patterns.

c) The Q-learning algorithm suits the proposed real-time pricing model. The load-uncertainty-aware strategy effectively balances energy supply-demand in electricity markets and improves power system robustness.

d) The carbon emission trading mechanism effectively supports "dual carbon" goals, enabling supply-side optimization to fully dispatch renewable energy like wind and PV, improving power system economics and environmental sustainability.

Future extensions can introduce constraints such as consumption limits and user budget constraints for greater realism; apply multi-agent RL to integrate regional energy microgrids with electric vehicles and storage; and implement big data-driven distributed RL for larger-scale users to achieve superior demand-side management.

5 Appendix

Table 3. Elasticity of demand

Time Period	Residential Users	Large Users
(21:00-7:00)	[Value]	[Value]
(7:00-10:00)	[Value]	[Value]
(10:00-15:00)	[Value]	[Value]
(15:00-18:00)	[Value]	[Value]
(18:00-21:00)	[Value]	[Value]

Table 4. User-side parameter setups

Parameter	Value Range
α_n	(1,2)
β_n	[3,4]
$\sigma_{n,t}$	(2.5,5)
c_n	[5,8]

Table 5. Power supply side and RL parameter setups

Parameter	Value
Learning rate α	0.01
Discount factor γ	0.9
Exploration rate ϵ	0.05
Convergence threshold δ	0.01

Table 6. TOU pricing setups

Time Period	Price (yuan/kWh)
Peak (6:00-22:00)	0.7
Valley (22:00-6:00)	1.0

Figure 9. Renewable power supplier' s outputs of typical day

Figure 10. Load demand of residential users

Figure 11. Load demand of large users

References

[1] Zhang Yao, Wang Aohan, Zhang Hong. Overview of smart grid development in China [J]. Power System Protection and Control, 2021, 49(5): 180-187.

[2] Huang Kaiyi, AI Qian, Zhang Yufan, et al. Challenges and prospects of regional energy network demand response based on energy cell-tissue architecture [J]. Power System Technology, 2019, 43(9): 3149-3160.

[3] Yuan Guanxiu, Gao Yan, Wang Hongjie. A real-time pricing algorithm based on utility classification in a smart grid [J]. Journal of University of Shanghai for Science and Technology, 2020, 42(1): 29-35.

[4] Samadi P, Mohsenian-Rad A H, Schober R, et al. Optimal real-time pricing algorithm based on utility maximization for smart grid [C]// IEEE International Conference on Smart Grid Communications. Piscataway, NJ: IEEE Press, 2010: 415-420.

[5] Wang Hongjie, Gao Yan. Research on the real-time pricing of smart grid based on nonsmooth equations [J]. Journal of Systems Engineering, 2018, 33(03): 320-327.

[6] Tao Li, Gao Yan, Zhu Hongbo. Real-time pricing strategy for smart grid based on the minimization of the peak-valley difference [J]. Journal of Systems Engineering, 2020, 35(03): 315-324.

[7] Li Junxiang, Zhou Jiru, He Jianjia. Mixed game of real-time pricing based on blockchain for power grid [J]. Power System Technology, 2020, 44(11): 4183-4191.

[8] Yuan Guanxiu, Gao Yan, Ye Bei, et al. Real-time pricing for smart grid with multi-energy microgrids and uncertain loads: a bilevel programming method [J]. International Journal of Electrical Power & Energy Systems, 2020, 123: 106206.

[9] Gao Yan. The social welfare maximization model of real-time pricing for smart grid [J]. Chinese Journal of Management Science, 2020, 28(10): 201-209.

[10] Tao Li, Gao Yan. Real-time pricing for smart grid with distributed energy and storage: a noncooperative game method considering spatially and temporally coupled constraints [J]. International Journal of Electrical Power & Energy Systems, 2020, 115: 105487.

[11] Zhu Hongbo, Gao Yan, Hou Yong, et al. Real-time pricing considering different types of users based on Markov decision processes in smart grid [J].

Systems Engineering-Theory & Practice, 2018, 38(3): 807-816.

[12] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

[13] José R, Zoltán N. Reinforcement learning for demand response: A review of algorithms and modeling techniques [J]. Applied Energy, 2019, 235: 1072-1089.

[14] Lu Renzhi, Hong SeungHo, Zhang Xiongfeng. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach [J]. Applied Energy, 2018, 220: 220-230.

[15] Zhang Li, Gao Yan, Zhu Hongbo, et al. Bi-level stochastic real-time pricing model in multi-energy generation system: A reinforcement learning approach [J]. Energy, 2021, 239: 121926.

[16] Feng Xiaofeng, XIE Tiankuo, GAO Ciwei, et al. A demand side response strategy considering long-term revenue of electricity retailer in electricity spot market [J]. Power System Technology, 2019, 43(8): 2761-2769.

[17] Wang Huiwei, Huang Tingwen, Liao Xiaofeng, et al. Reinforcement learning in energy trading game among smart microgrids [J]. IEEE Transactions on Industrial Electronics, 2016, 63(8): 5109-5119.

[18] Du Yan, Li Fangxing. Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning [J]. IEEE Transactions on Smart Grid, 2020, 11(2): 1066-1076.

[19] Jin Ming, Feng Wei, Marnay C, et al. Microgrid to enable optimal distributed energy retail and end-user demand response [J]. Applied Energy, 2018, 210: 1321-1335.

[20] Zhang Li, Gao Yan, Zhu Hongbo, et al. Real-time pricing strategy based on uncertainty of power consumption in smart grid [J]. Power System Technology, 2019, 43(10): 181-190.

[21] Tarasak P. Optimal real-time pricing under load uncertainty based on utility maximization for smart grid [C]// IEEE International Conference on Smart Grid Communications. Piscataway, NJ: IEEE Press, 2011: 321-326.

[22] Yu Mengmeng, Hong Seung Ho. Incentive-based demand response considering hierarchical electricity market: A Stackelberg game approach [J]. Applied Energy, 2017, 203: 267-279.

[23] Samadi P, Mohsenian-Rad H, Schober R, et al. Advanced demand side management for the future smart grid using mechanism design [J]. IEEE Transactions on Smart Grid, 2012, 3(3): 1170-1180.

[24] Li Junxiang, Pan Tingting, Gao Yan. Real time pricing algorithm for supply and demand of complementary energy on smart grid [J]. Application Research of Computers, 2020, 37(4): 1092-1096.

[25] Chiu Techuan, Shih Yuanyao, Pang Aichun, et al. Optimized day-ahead pricing with renewable energy demand-side management for smart grids [J]. IEEE Internet of Things Journal, 2017, 4(2): 374-383.

[26] Zhang Ning, Hu Zhaoguang, Dai Daihong, et al. Unit commitment model in smart grid environment considering carbon emissions trading [J]. IEEE Transactions on Smart Grid, 2016, 7(1): 420-427.

[27] Zhang Xiaohui, Liang Junxue, et al. Research on low-carbon power planning with gas turbine units based on carbon transactions [J]. Acta Energiae Solar Sinica, 2020, 41(07): 92-98.

[28] Alpaydin E. Introduction to Machine Learning [M]. 4th ed. Cambridge: MIT Press, 2020.

[29] Yu Tao, Zhou Bin, Chan Kawing, et al. Stochastic optimal relaxed automatic generation control in non-Markov environment based on multi-step Q(λ) learning [J]. IEEE Transactions on Power Systems, 2011, 26(3): 1272-1282.

[30] Kong Xiangyu, Kong Deqian, et al. Online pricing of demand response based on long short-term memory and reinforcement learning [J]. Applied Energy, 2020, 271: 114945.

[31] Han Xuefeng, He Hongwen, Wu Jingda, et al. Energy management based on reinforcement learning with double deep Q-learning for a hybrid electric tracked vehicle [J]. Applied Energy, 2019, 254: 113708.

[32] Hasselt H. Double Q-learning [J]. Advances in Neural Information Processing Systems, 2010, 23: 2613-2621.

[33] Yang Peng, Tang Gongguo, Nehorai A. A game-theoretic approach for optimal time-of-use electricity pricing [J]. IEEE Transactions on Power Systems, 2012, 28(2): 884-892.

[34] Lin Jie, Xiao Biao, Zhang Hanlin, et al. A novel multitype-users welfare equilibrium based real-time pricing in smart grid [J]. Future Generation Computer Systems, 2020, 108: 145-160.

Note: Figure translations are in progress. See original paper for figures.