Dynamic Pricing Strategies for the Electricity Retail Market Using Reinforcement Learning Algorithms

Xu Weiting[1,3], Liu Jiajia[2], Liu Chang[1,3], Zhang Fengxi[2], Xu Ke[1,3], and Yuan Cheng[4,*]
[1]Economic Research Institute of State Grid Sichuan Electric Power Company, Chengdu 610041, China
[2]Sichuan Electric Power Trading Center Co., Ltd., Chengdu 610041, China
[3]Sichuan New Electric Power System Research Institute, Chengdu 610041, China
[4]Beijing Tsintergy Technology Co., Ltd., Beijing 100080, China

Corresponding author: Cheng Yuan (e-mail: scjj_conne@163.com)

ABSTRACT The dynamic pricing mechanism in the electricity retail market plays a pivotal role in achieving optimal resource allocation on both supply and demand sides. However, the complex heterogeneity of consumer behavior at the retail level coupled with frequent price fluctuations in the wholesale market poses significant challenges to traditional optimization approaches. This paper models the dynamic pricing problem of power retailers as a Markov decision process within a continuous action space and proposes a solution framework based on proximal policy optimization algorithms. Centered on maximizing cumulative profits for power retailers, the framework integrates user price response characteristics, procurement cost constraints in wholesale markets, and regulatory boundaries for retail pricing into a unified modeling system, enabling adaptive generation of retail electricity prices across time periods through continuous interaction between agents and their environment. To address common issues with standard PPO algorithms–including premature policy degradation and insufficient exploration efficiency–the study introduces an adaptive exploration coefficient adjustment scheme based on policy entropy, along with a generalized advantage estimation method to reduce gradient estimation variance. Simulation experiments using actual operational data from a provincial electricity market in China during 2023 demonstrated that the proposed algorithm outperforms benchmark algorithms such as the Deep Deterministic Policy Gradient and Dual Delayed Deep Deterministic Policy Gradient in key metrics including average daily retailer profits, peak load reduction rates, and price series stability, validating the feasibility and practical effectiveness of reinforcement learning approaches for dynamic pricing in electricity retail markets.

INDEX TERMS Electricity retail market; Dynamic pricing; Reinforcement learning; Near-field policy optimization; Demand response

I. INTRODUCTION

The market-oriented reform of the power industry is fundamentally transforming the operational model of traditional power systems. On the retail side, electricity retailers serve as pivotal intermediaries linking upstream wholesale markets with downstream end-users, with their core mission being to maximize operational profits while meeting diverse user demand. The determination of retail electricity prices not only directly influences consumer behavior patterns and grid load characteristics but also significantly determines retailers' competitiveness in the market. Consequently, developing rational and effective retail pricing strategies within a complex and uncertain market environment has become a fundamental research focus in power retail market studies.

Dynamic pricing enables electricity retailers to dynamically adjust retail electricity prices across different time periods based on multidimensional factors such as wholesale market price signals, real-time system load levels, and user demand elasticity. Compared to traditional fixed tariffs or simple time-of-use pricing schemes, dynamic pricing more accurately reflects the marginal costs and market value of electricity under varying spatiotemporal conditions, theoretically offering potential to enhance economic efficiency and optimize resource allocation on the demand side. Numerous studies have explored this topic from multiple perspectives: some have developed pricing optimization models for retailers using microeconomic theories to derive pricing strategies through solving constrained nonlinear programming problems; others focus on modeling and analyzing user demand response behaviors, examining load shifting patterns and reduction characteristics across different user groups under various pricing signals. However, most approaches rely either on precise functional models of user response behavior or require strong prior assumptions about the probability distributions of market stochastic parameters. When systems face significant uncertainties arising from high renewable energy integration, the applicability and robustness of these methods become substantially challenged.

Reinforcement learning provides an effective paradigm for addressing sequential decision-making problems in complex and uncertain environments. Its fundamental concept involves agents learning decision strategies through continuous interaction with their environment, without requiring explicit mathematical modeling of environmental dynamics. In recent years, deep reinforcement learning has combined the feature representation capabilities of deep neural networks with the strategy search advantages of reinforcement learning, demonstrating promising applications across various domains such as power system dispatching, generation-side market bidding, and demand-side response management. Specifically in electricity pricing, researchers have applied reinforcement learning methods to real-time tariff generation and retail tariff package design. Literature [1] proposes a dynamic price optimization algorithm based on reinforcement learning that integrates power providers' profit objectives with users' electricity costs into a unified framework. Literature [2] investigates data-driven transaction decision support techniques from the perspective of agency-based power purchasing. Literature [3] models the pricing problem of multi-grid systems (power, gas, heat) as a Starkleberg game and solves it using reinforcement learning. Literature [4–5] examines bidding and pricing strategies for power generators from a multi-agent perspective. Literature [6] develops a deep reinforcement learning-based decision model for quantity-price combination bidding in power markets. Literature [7] further applies two-stage deep reinforcement learning to analyze multi-agent coordinated bidding scenarios. Furthermore, literature [8] examines profit strategies for energy storage systems in electricity markets; literature [9] applies the PPO algorithm to optimize interactive operations of park-scale wind-solar-storage systems; literature [10] investigates deep reinforcement learning modeling methods for incentivized demand response; literature [11–12] explores approaches from the perspectives of regional integrated energy pricing and system scheduling under uncertain source-load conditions; literature [13] employs reinforcement learning for demand response control in regional cooling systems under real-time pricing mechanisms; while literature [14–16] provides foundational research on retail electricity price mechanism design, market trading strategies for thermal power companies, and economic dispatching with security constraints in renewable energy systems.

Nevertheless, existing research still exhibits several notable shortcomings. Most studies focus on bidding strategies for power generators on the wholesale side or multi-energy complementary pricing within microgrids, while paying insufficient attention to the formulation of dynamic electricity prices for retail electricity suppliers serving end-users. In constructing pricing models, the description of user demand response behaviors remains often crude, failing to adequately capture the distinct price sensitivity and load characteristics among different user types. From an algorithmic perspective, current approaches predominantly employ discrete-action-space methods such as Q-learning or Deep Q-Networks (DQN), which suffer from precision degradation when handling continuous electricity price decision variables due to action discretization, and their algorithmic stability and convergence speed require further improvement.

To address the aforementioned shortcomings, this paper proposes a dynamic pricing strategy for electricity retail markets based on the Proximal Policy Optimization (PPO) algorithm. The main contributions include: systematically modeling the retail dynamic pricing problem as a Markov decision process in a continuous action space, precisely defining the state space, action space, and reward function structure tailored to practical pricing scenarios; designing an adaptive exploration coefficient adjustment mechanism based on policy entropy at the algorithmic level to achieve a dynamic balance between exploration and exploitation during training; constructing a simulation environment using actual operational data from a provincial-level electricity market in China, and validating the effectiveness and superiority of the proposed algorithm through multiple comparative experiments and ablation analyses. The paper is structured as follows: Section 2 presents the system model and problem formulation; Section 3 details the PPO-based dynamic pricing algorithm; Section 4 reports the design, results, and analysis of simulation experiments; and Section 5 summarizes the findings and discusses future research directions.

II. SYSTEM MODEL AND PROBLEM DESCRIPTION

A. THE BASIC STRUCTURE OF THE ELECTRICITY RETAIL MARKET

Consider a retail electricity market served by a single power retailer serving multiple end-users. The retailer purchases electricity from upstream wholesale markets and resells it to end-users at retail prices. On the wholesale side, a nodal marginal pricing mechanism is employed to generate price sequences across different time periods, denoted as P_t^w, $t = 1, \ldots, T$. Here, T represents the total number of time periods within a scheduling cycle (in this paper, $T = 24$, corresponding to all 24 hours). On the retail side, electricity retailers announce retail prices to users during each period t. P_t^r Users adjust their actual electricity consumption based on the price signals they receive.

User demand response behavior is modeled using the price elasticity model. Let the base electricity consumption of User Type i at period t be $D_{i,t}^0$, That is, the actual electricity consumption when not influenced by price signals.

The relationship with the retail price can be expressed as

$$D_{i,t} = D_{i,t}^0 \left(\frac{P_t^r}{P_{i,t}^{\mathrm{ref}}} \right)^{\epsilon_i}. \tag{1}$$

Among $P_{i,t}^{\mathrm{ref}}$ As the reference price benchmark, $\epsilon_i < 0$ This represents the price elasticity coefficient for users of category i. The economic implication of this functional form is that when the retail price exceeds the reference price, users' actual electricity consumption decreases correspondingly; when the retail price falls below the reference price, consumption increases. It should be noted that price elasticity coefficients vary significantly among different user types, with residential users exhibiting particularly strong price elasticity due to their highly inelastic electricity demand. ϵ_i They are typically small; industrial and commercial users respond more sensitively to price signals. ϵ_i The scale is generally large.

However, the aforementioned elastic model only captures users' immediate price responses within a single time period and fails to reflect their cross-period substitution behavior over time. In practical scenarios, when electricity prices rise during a specific period, users may not only reduce their consumption during that period but also shift part of their demand to adjacent peak or off-peak periods. To incorporate this effect into the model while avoiding causal mismatch issues arising from future information, this paper employs a load-shifting description based solely on historical period data. Specifically, retail price decisions are formulated for period t. P_t^r At this stage, the electricity retailer can only estimate the load transfer effect based on price data from the past period $k < t$. Let the actual electricity consumption of User Type i during period t consist of two components: one component is the consumption directly influenced by the period's pricing, $D_{i,t}^{\mathrm{self}}$. The other component represents the load transfer amount induced by price differences that have occurred over past periods, $\Delta D_{i,t}^{\mathrm{shift}}$. That is

$$D_{i,t} = D_{i,t}^{\mathrm{self}} + \Delta D_{i,t}^{\mathrm{shift}}, \qquad D_{i,t}^{\mathrm{self}} = D_{i,t}^0 \left(\frac{P_t^r}{P_{i,t}^{\mathrm{ref}}} \right)^{\epsilon_i}. \tag{2}$$

The load transfer amount is defined as

$$\Delta D_{i,t}^{\mathrm{shift}} = \sum_{k=1}^{t-1} \eta_{i,k\to t} \cdot D_{i,k} \cdot \frac{P_k^r - P_t^r}{P_{i,t}^{\mathrm{ref}}}. \tag{3}$$

Several key design choices require clarification. First, the summation range is limited to $k = 1$ through $t-1$, accounting solely for load transfers from past periods to the current period and thereby avoiding causal errors arising from future information. Second, the transfer base is set as $D_{i,k}$ (Actual electricity consumption during the past period) rather than $D_{i,k}^0$ (Basic electricity consumption): This is because the actual load capacity that users can adjust should be based on their real electricity usage after being affected by pricing policies, rather than their original, unaffected electricity demand. Third, $\eta_{i,k\to t} \geq 0$ The load transfer coefficient for User Class i from period k to period t reflects the intensity of the impact of historical electricity price differences on current period electricity consumption. Fourth, in the formula: $P_k^r - P_t^r$ The symbol determines the direction of transfer: when the price in the past period k is higher than that in the current period t, $\Delta D_{i,t}^{\mathrm{shift}}$ The positive value indicates that partial load demand originally scheduled for use during the high-price period k has been postponed to the current low-price period t, reflecting users' tendency to shift electricity consumption to later periods when prices are lower–a typical response pattern observed in practice during peak-hour price increases. This coefficient satisfies the required conditions, $\sum_{t>k} \eta_{i,k\to t} \leq 1$, To ensure that the total load transferred out from any historical period does not exceed the actual electricity consumption of that period.

It should be noted that the aforementioned load transfer model involves a theoretical simplification: when a load is transferred from period k to period t $(t > k)$ Subsequently, the electricity consumption during the subsequent period $tD_{i,t}$ This transfer amount is already included; however, if the price at period t exceeds that of the subsequent period $t+1$, this load may be incorporated again into the transfer from t to $t+1$, creating a chain reaction. In the 24-hour short-term scheduling scenario presented herein, the actual impact of such cross-period chain reactions is limited (price fluctuations typically occur within hours, and continuous one-way price gradients across multiple periods are uncommon). Nevertheless, this limitation does exist in the theoretical rigor of the model. To minimize this effect as much as possible, the simulation applies a 50% discount attenuation to transfer amounts with chain lengths exceeding two.

The profit function of the electricity retailer during period tt can be expressed as

$$R_t = P_t^r \sum_i D_{i,t} - P_t^w \sum_i D_{i,t} - C_t^{\mathrm{op}}. \tag{4}$$

This C_t^{op} represents the operating costs for period tt, including allocation of transmission and distribution line losses, system reserve capacity fees, and the power retailer's own management and maintenance expenses. Retail price P_t^r Must comply with the price upper and lower limits set by regulatory authorities. $P_{\min} \leq P_t^r \leq P_{\max}$.

B. MODELING OF THE MARKOV DECISION PROCESS FOR DYNAMIC PRICING PROBLEMS

The dynamic pricing problem of electricity retailers can be formulated as a Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ over a finite time horizon, defined by a five-element tuple.

State space $\mathcal{S}$, s_t. The state vector at time t should encompass all critical pricing-related information observable by the electricity retailer during decision-making. This paper defines it as

$$s_t = \left(P_t^w, D_{t-1}, D_{t-1}^{\mathrm{avg}}, L_{t-1}, H_t, \Delta P_t^w \right). \tag{5}$$

Among P_t^w The wholesale market price for the current period; $D_{t-1} = \{D_{1,t-1}, D_{2,t-1}, \ldots, D_{I,t-1}\}$ This data represents the actual electricity consumption of various user types during a specific period and can be directly collected from smart meters. D_{t-1}^{avg} The moving average of actual electricity consumption for each user type during the same period over the past week, used to assist agents in identifying short-term cyclical patterns in power usage behavior. L_{t-1} The total system load level for the previous period; H_t Identifies the time period type (peak, off-peak, or low-demand period); $\Delta P_t^w = P_t^w - P_{t-1}^w$ The change in the wholesale market price. It should be noted that, theoretically, ΔP_t^w can be determined completely by P_t^w and P_{t-1}^w. These variables are redundant. The paper deliberately retains them in the state space as an engineering feature engineering technique, presenting price trend changes in explicit difference form to the neural network, thereby accelerating convergence during the initial training phase. This approach is widely adopted in practice and does not compromise the algorithm's theoretical correctness.

Additionally, the paper intentionally omits the baseline electricity consumption for the current period. The prediction D_t^0 values are included in the state space. Since this predicted value is not a directly observable state parameter in real-world scenarios but rather an estimate derived from external forecasting models, using it as state input implicitly assumes that power retailers possess precise day-ahead forecasts when making decisions–a assumption that does not hold in practice. The proposed approach employs historical electricity consumption data as state input, enabling the agent to autonomously extract consumption patterns and trends from the data through training. The total dimensionality of the state space is $1 + I + I + 1 + 1 + 1 = 2I + 4$, where I represents the number of user types.

Action Space $\mathcal{A}$, P_t^r. The electricity sales company's action at each time period t is the retail price, a continuous variable with a defined range $[P_{\min}, P_{\max}]$. The advantage of using a continuous action space is its ability to more precisely capture the marginal effects of electricity price adjustments, avoiding the information loss and reduced strategic precision associated with discretizing continuous prices.

State transition probability $\mathcal{P}$, $P(s_{t+1} \mid s_t, a_t)$. State transition probability Describes performing $a_t = P_t^r$ actions in a given state s_t. Then move to the next state s_{t+1}. The transition probability is unknown in real-world scenarios, as wholesale market prices are influenced by various external factors and end-users' electricity consumption is subject to random fluctuations caused by weather conditions and social activities. This precisely motivates the adoption of a model-free reinforcement learning approach in this study: agents do not require explicit modeling of environmental dynamics but learn strategies through direct interaction with the environment.

Reward Function $\mathcal{R}$. Immediate Reward $r_t = R(s_t, a_t)$ Defined as the profit obtained by the electricity sales company during period t:

$$r_t = P_t^r \sum_i \left[D_{i,t}^0 \left(\frac{P_t^r}{P_{i,t}^{\mathrm{pref}}} \right)^{\epsilon_i} + \sum_{k=1}^{t-1} \eta_{i,k\to t} D_{i,k} \frac{P_k^r - P_t^r}{P_{i,t}^{\mathrm{pref}}} \right] - P_t^w \sum_i D_{i,t} - C_t^{\mathrm{op}}. \tag{6}$$

It should be noted that the calculation $r_t D_{i,t} r_t$ depends on the cumulative electricity consumption over the entire period t, so this reward can only be claimed after period tt ends and is therefore a delayed reward. In MDP implementation, this means the reward is awarded upon state transition. $(s_t, a_t) \to s_{t+1}$ It is assigned when the event occurs and aligns with the temporal conventions of a standard MDP.

The discount factor $\gamma \in (0,1)$ balances the relative importance between current rewards and future rewards. Within this framework, $\pi^* : \mathcal{S} \to \mathcal{A}$ is the optimization objective for electricity retailers to identify an optimal pricing strategy that maximizes the expected cumulative discount reward throughout the entire dispatch cycle:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} \gamma^{t-1} r_t. \tag{7}$$

III. DYNAMIC PRICING ALGORITHM BASED ON PROXIMAL POLICY OPTIMIZATION

A. THE BASIC FRAMEWORK OF THE STRATEGIC GRADIENT METHOD

The policy gradient method directly parameterizes the policy $\pi_\theta(a \mid s)$ and then updates the policy parameters θ in the direction that maximizes the expected cumulative reward. The gradient estimation formula for the standard policy gradient algorithm is given in (8):

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(a \mid s) \cdot \hat{A}(s,a) \right]. \tag{8}$$

This represents $\hat{A}(s,a)$ the estimate of the loss function, used to measure how well action a performed compared to the average performance under state s. Although the policy gradient method theoretically exhibits good convergence properties, its standard implementation faces challenges such as difficulty in step size selection, low sampling efficiency, and sensitivity to hyperparameters. The proximal policy optimization algorithm introduces confidence interval constraints and importance sampling techniques, significantly improving sample utilization efficiency while maintaining algorithmic stability, making it one of the most widely adopted policy gradient algorithms for continuous control problems today.

B. PROXIMAL POLICY OPTIMIZATION ALGORITHM

The core principle of the PPO algorithm is to limit the deviation between the new strategy and the old one during each update, thereby preventing performance degradation caused by excessive single-step updates. The algorithm employs a objective function based on the truncation importance sampling ratio:

$$L^{\mathrm{CLIP}}(\theta)=\mathbb{E}_t\big[\min\big(\rho_t(\theta)\hat{A}_t,\mathrm{clip}(\rho_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t\big)\big],$$
$$\rho_t(\theta)=\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t\mid s_t)}. \tag{9}$$

Here, $\rho_t(\theta)$ represents the importance sampling ratio, reflecting the probability ratio between the old and new strategies in sampling actions; ϵ is the truncation hyperparameter (typically set to 0.1–0.2); $\hat{A}_t$ is the estimate of the advantage function, as defined in Section 3.3 below. $\mathbb{E}_t$ denotes the empirical mean over all time steps along the sampling trajectory. The truncation operation aims to limit the growth of the objective function when deviations from 1 become excessively large, thereby preventing overly aggressive policy updates.

C. ADAPTIVE EXPLORATION COEFFICIENT ADJUSTMENT MECHANISM

During the actual training of the PPO algorithm, policy entropy typically exhibits a gradual decline: as training progresses, the policy distribution becomes increasingly concentrated, and exploratory behavior diminishes accordingly. In the specific context of electricity pricing, excessive determinism may lead to the agent prematurely settling on a particular price level, thereby preventing the discovery of potentially superior pricing strategies. To address this issue, this paper introduces a policy entropy regularization term into the PPO objective function, representing an adaptive improvement to the PPO algorithm:

$$L(\theta)=L^{\mathrm{CLIP}}(\theta)+\beta_t\cdot\mathcal{H}_{\pi_\theta},$$
$$\mathcal{H}_{\pi_\theta}=-\sum_{a\sim\pi_\theta}\pi_\theta(a\mid s)\log\pi_\theta(a\mid s). \tag{10}$$

Here, $\mathcal{H}_{\pi_\theta}$ represents the entropy of the policy, and β_t is the time-varying regularization coefficient. A higher policy entropy indicates greater randomness and richer exploration behavior, while a lower entropy suggests a more deterministic policy. Policy entropy regularization is not a novel approach; it has been extensively implemented in the Soft Actor-Critic algorithm. This paper introduces this concept into the power pricing problem within the PPO framework and designs an adaptive adjustment mechanism tailored to the specific characteristics of pricing scenarios. β_t The update follows the following rule:

$$\beta_{t+1}=\max\left(0,\beta_t+\alpha_\beta\cdot(\mathcal{H}_\alpha-\mathcal{H}_t)\right). \tag{11}$$

Here, $\mathcal{H}_\alpha$ represents the preset target entropy value, $\mathcal{H}_t$ denotes the measured entropy value of the current policy, and α_β indicates the adjustment step size. Adopting the maximum value ensures non-negativity; when the entropy exceeds the target value, the weight of the entropy regularization term should be reduced or even set to zero rather than becoming negative, which would produce opposite effects. The fundamental logic of this mechanism is: when policy entropy falls below the target value, exploration bias is increased; when it exceeds the target value, exploration bias is decreased, thereby achieving a dynamic balance between exploration and utilization. The selection of the initial value significantly impacts early-stage training behavior; in this study, we determined the optimal value to be 0.01 through grid search within the range $\{0.005,0.01,0.02,0.05\}$.

The estimation of the advantage function employs the generalized advantage estimation method:

$$\hat{A}_t^{\mathrm{GAE}(\lambda)}=\sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l},\qquad \delta_t=r_t+\gamma V(s_{t+1})-V(s_t). \tag{12}$$

This represents $V(s)$ the state value function, approximated by a value network operating in parallel with the policy network; $\lambda\in[0,1]$ is a hyperparameter used to balance bias and variance. By introducing the λ parameter, GAE performs smooth interpolation between one-step difference estimation and Monte Carlo estimation, typically yielding stable and reliable advantage estimates in practical applications.

D. ALGORITHM FLOW AND NETWORK STRUCTURE

The algorithm employs an Actor-Critic architecture. Both the policy network (Actor) and value network (Critic) are three-layer fully connected neural networks with 256 and 128 hidden layers respectively, using ReLU as the activation function. The policy network outputs the mean and standard deviation of actions sampled from a normal distribution, $a\sim\mathcal{N}(\mu_\theta(s),\sigma_\theta(s))$, while the value network provides estimates of state values, $V_\phi(s)$.

The training process proceeds as follows: At the start of each episode, the environment state is reset; at each time step, the agent samples and executes an action according to its current policy, receives an immediate reward and the next state from the environment, and stores the transition tuple (s_t,a_t,r_t,s_{t+1}) in the experience buffer; after collecting data from K consecutive time steps, the advantage function estimate and objective function are computed using the buffer data, followed by multiple rounds of small-batch gradient updates for both the policy network and value network; upon completion of updates, the buffer is cleared and the process proceeds to the next round of data collection. The entire training cycle spans M episodes until the policy converges or the preset maximum number of training rounds is reached.

IV. SIMULATION EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL DESIGN AND DATA SOURCES

The data foundation for the simulation experiment originates from the full-year operational records of a provincial electricity market in China during 2023. This provincial electricity market employs a nodal marginal pricing mechanism for day-ahead market clearing, with wholesale price data sourced from the publicly released day-ahead market clearing price series by the provincial power trading center–

totaling 365 complete daily price datasets, each containing clearing prices for 24 time periods. This study uniformly adopts the clearing prices at the system reference node (i.e., the benchmark node excluding transmission congestion effects) to avoid confusion caused by price variations across nodes. User load data were extracted from typical daily load collection records in the provincial grid company's marketing system. End-users were categorized into three types based on usage patterns: residential users, general industrial/commercial users, and large industrial users. The daily electricity consumption ratios, peak load factors, and price elasticity coefficients for each category are presented in Table 1. The price elasticity coefficient values were derived from relevant research findings on demand price elasticity in China's retail electricity market [17]. The load transfer coefficient values were determined as follows: for adjacent time periods ($|k-t|=1$), the transfer coefficients for residential users, general industrial and commercial users, and large industrial users are set at 0.05, 0.12, and 0.10, respectively; for intervals exceeding two periods. (When $|k-t| \geq 2$), the transfer coefficient decreases by 50% for each additional time interval. Operating cost C_t^{op} In the simulation, the amount is estimated at 3% of electricity sales revenue.

TABLE 1. Load Characteristics and Price Elasticity Parameters for Different User Categories

User Type	Proportion of daily electricity consumption (%)	Peak load rate	Price elasticity coefficient
Residential Users	32	0.28	-0.15
General Industry and Commerce	41	0.35	-0.32
Large Industrial Users	27	0.42	-0.28

Note: Peak load rate is defined as the ratio of the average load during peak hours to the daily average load.

The upper and lower limits for retail prices are set as $P_{\min}$=¥0.2/kWh. The $P_{\max}$ rate is ¥1.2/kWh, reference price $P_{i,t}^{\mathrm{ref}}$ Calculate the weighted average electricity purchase cost for various users.

Discount factor. The selection of $\gamma = 0.95$ was based on the following preliminary experiments: A grid search was conducted for γ within the range $\{0.9, 0.95, 0.99\}$, revealing that $\gamma = 0.95$ strikes an optimal balance between convergence speed and final cumulative reward: $\gamma = 0.9$ leads to rapid convergence but suboptimal performance, while $\gamma = 0.99$ results in slow convergence and significant training fluctuations; thus, 0.95 was ultimately selected. The PPO algorithm employs a truncation parameter $\epsilon = 0.2$, a generalized advantage estimation parameter $\lambda = 0.95$, and learning rates set uniformly for both the policy network and value network, 3×10^{-4}. Each round collects $K = 2048$ steps, with a total training episodes count of $M = 2000$. The target entropy value $\mathcal{H}_\alpha$ is determined as follows: Before formal training, run 100 episodes using a random policy (an untrained network), calculate the average policy entropy, use this value as the initial reference, and then fine-tune it via a small grid search $\{-0.5, -1.0, -1.5\}$. Final value $\mathcal{H}_\alpha$ determined = -1.0. Step size $\alpha_\beta = 0.001$. The comparison algorithms include the Deep Deterministic Policy Gradient (DDPG) and the Double Delayed Deep Deterministic Policy Gradient (TD3). Both algorithms employ an Actor-Critic architecture and support continuous action spaces, serving as representative benchmark methods for continuous control problems. All algorithms were trained and tested under identical simulation environments and hyperparameter configurations. To minimize the impact of randomness on training outcomes in reinforcement learning, each experiment was repeated five times with different random seeds, and results are presented as the mean ± standard deviation across all runs.

B. ANALYSIS OF THE TRAINING PROCESS

Figure 1 illustrates the average reward curves of the three algorithms during training (a moving average over every 50 episodes was applied to facilitate trend analysis). Overall, the PPO algorithm exhibits the fastest initial ascent in its learning curve: within approximately the first 200 episodes, the average reward rapidly rises from an initial negative range to positive values and continues to improve. The DDPG algorithm follows closely behind in convergence speed, but demonstrates significant reward fluctuations during training–reflecting its inherent instability when tackling high-dimensional continuous control problems such as pricing. The TD3 algorithm mitigates the overestimation issue in DDPG by introducing a dual Q-network and a delayed policy update mechanism; it achieves better convergence stability than DDPG but converges more slowly than PPO.

It is noteworthy that during the later stages of training (after approximately 1,500 episodes), the reward curves of the PPO algorithm continue to exhibit a gradual upward trend, whereas those of DDPG and TD3 have approached saturation. This phenomenon may be attributed β_t to the adaptive exploration coefficient adjustment mechanism in the PPO algorithm: as the policy converges, the exploration coefficient decreases adaptively but does not drop to zero entirely, enabling the agent to perform fine-grained search and optimization near already explored optimal policies. In contrast, the Ornstein-Uhlenbeck noise or Gaussian noise employed by DDPG and TD3 contribute minimally to policy optimization in the late training phase.

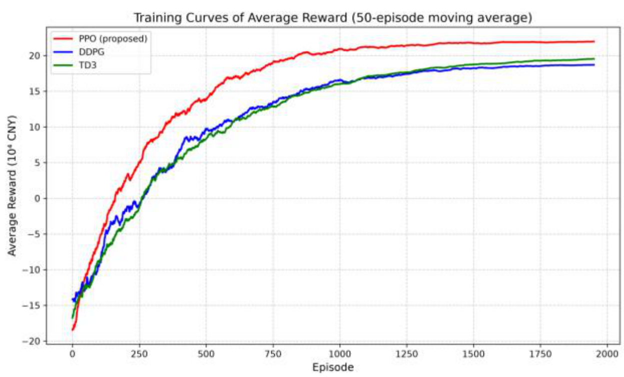

FIGURE 1. Training curves of average reward (50-episode moving average)

C. COMPARATIVE ANALYSIS OF PRICING STRATEGIES

To evaluate the effectiveness of the trained pricing strategy in real-world scenarios, a systematic comparative analysis was conducted on the daily pricing sequences generated by each algorithm across the test set. The test set comprises data from 30 typical days, covering various types including weekdays, weekends, and holidays. Evaluation metrics included: retailer's average daily profit; peak-hour load reduction rate (relative to the baseline scenario without dynamic pricing); electricity price volatility coefficient (defined as the ratio of intraday price standard deviation to mean value); and the change rate of users' average electricity expenditure. Statistical results for all metrics are summarized in Table 2.

TABLE 2. Comparison of pricing strategy performance across different algorithms

Evaluation Metrics	Fixed Electricity Price	DDPG	TD3	PPO (this text)
Average Daily Profit per 10,000 Yuan	12.4±0.0	18.7±1.2	19.3±0.9	21.6±0.7
Peak load reduction rate (%)	–	11.3±1.8	12.8±1.5	16.5±1.2
Electricity Price Fluctuation Coefficient	0.18±0.00	0.15±0.03	0.15±0.02	0.13±0.02
Average Electricity Bill Change Rate per User (%)	0	+2.1±0.4	+1.6±0.3	+0.8±0.2

Note: In the fixed electricity price scenario, the user's electricity cost change rate is set to a baseline value of 0; the electricity consumption is set as the base consumption $D_{i,t}^{0}$.

As shown in Table 2, the PPO algorithm demonstrates superior performance across all four evaluation metrics. Its average daily profit reached ¥216,000, representing a 15.5% increase compared to DDPG and an 11.9% improvement over TD3. The peak-hour load reduction rate achieved 16.5%, outperforming both comparison algorithms, indicating that the pricing strategy generated by PPO exhibits stronger practical effectiveness in guiding users to balance peak and off-peak demand. With a price fluctuation coefficient of 0.13–the lowest among the three dynamic pricing algorithms–it demonstrates PPO's ability to optimize profit maximization while maintaining temporal stability in electricity pricing. This is particularly crucial in real-world retail power markets, where excessive price volatility may lead to customer dissatisfaction or even user attrition. User average electricity costs increased only 0.8% compared to fixed-price scenarios, significantly lower than DDPG's 2.1% and TD3's 1.6%, highlighting PPO's balanced approach between maximizing utility profits and protecting consumer interests.

To verify the statistical significance of the aforementioned differences, an independent samples t-test was conducted on the disparity between PPO and TD3 in terms of average daily profit. $p = 0.038 < 0.05$, indicating a statistically significant difference between the two groups. The difference between PPO and DDPG was also significant ($p = 0.021 < 0.05$). Regarding the peak load reduction rate, the t-test results between PPO and TD3 showed: $p = 0.042 < 0.05$; the difference between PPO and DDPG was $p = 0.019 < 0.05$, both reaching statistical significance.

It is necessary to conduct a further analysis of the phenomenon where the PPO algorithm achieves superior performance in both "average daily profit" and "peak-hour load reduction rate." Intuitively, a higher peak-hour load reduction rate implies reduced electricity consumption during peak hours, which should directly decrease power retailers' revenue during these periods. However, the results from the PPO algorithm demonstrate that this assumption is not entirely valid–the underlying mechanism may be as follows: the DDPG and TD3 algorithms set excessively high prices during peak hours, leading to excessive consumption cuts that ultimately harm retailers' overall profits; in contrast, the PPO algorithm employs a more refined peak-valley pricing structure, effectively guiding load shifting while avoiding excessive suppression of consumption due to high tariffs, thereby achieving a win-win outcome for both profitability and peak shaving. This explanation is further supported by the PPO algorithm's lowest tariff fluctuation coefficient–excessively high peak-hour tariffs not only severely suppress demand but also cause significant volatility in electricity price trends.

Figure 2 illustrates the retail electricity price sequence generated by the PPO algorithm over a typical working day, alongside corresponding wholesale market prices. The analysis reveals that retail prices generally follow wholesale price trends, yet exhibit significant premiums during peak hours (10:00–12:00 and 18:00–20:00) while remaining lower during off-peak periods (0:00–6:00 and 23:00–24:00). This pricing mechanism that widens the peak-off-peak price gap effectively incentivizes users to shift their electricity consumption from peak to off-peak hours. Notably, retail prices do not simply reflect linear adjustments of wholesale prices–during certain periods (e.g., 14:00–16:00), even when wholesale prices decline, retail prices do not decrease proportionally. This demonstrates the algorithm's nonlinear pricing adjustments driven by profit maximization objectives, highlighting the advantages of data-driven approaches over rule-based methods.

FIGURE 2. Retail vs. wholesale prices on a typical weekday

D. ABLATION EXPERIMENT AND SENSITIVITY ANALYSIS

To evaluate the practical contribution of the adaptive exploration coefficient adjustment mechanism, this study designed a series of ablation experiments by replacing the adaptive entropy adjustment in the PPO algorithm with a fixed entropy coefficient ($\beta = 0.01$), denoted as PPO-Fixed, was trained and tested under identical conditions. Additionally, a standard PPO model without any entropy regularization was included as an additional baseline (denoted as

PPO-NoEnt). Experimental results showed that PPO-Fixed achieved a final average daily profit of 201,000 ± 8,000 yuan, 6.9% lower than the full PPO algorithm; PPO-NoEnt achieved 194,000 ± 1,100 yuan, 10.2% lower. Regarding peak-hour load reduction rates, PPO-Fixed recorded 14.2 ± 1.3%, while PPO-NoEnt achieved 13.5 ± 1.6%, both below the full PPO algorithm's 16.5 ± 1.2%. These findings demonstrate that entropy regularization itself positively enhances algorithm performance (PPO-Fixed outperforms PPO-NoEnt), with the adaptive adjustment mechanism further improving results (full PPO surpasses PPO-Fixed). This indicates that the adaptive exploration coefficient adjustment mechanism dynamically maintains appropriate exploration capacity, enabling the algorithm to escape local optima and identify superior pricing strategies during training.

A sensitivity analysis was conducted on the key parameter of price elasticity coefficient. The elasticity coefficients for various user groups were adjusted by ±20% above and below the baseline values to examine changes in PPO algorithm performance. Results indicate that when the absolute value of the elasticity coefficient increases (i.e., users become more price-sensitive), retailers' optimal profits decline by approximately 4% to 7%, while peak load reduction rates rise significantly by about 18% to 25%. Conversely, when the elasticity coefficient decreases (indicating low price sensitivity), retailer profits increase slightly but peak shaving effectiveness diminishes markedly. These findings align with fundamental economic principles: higher price sensitivity expands the potential for power retailers to guide load shifting through pricing signals; however, excessive user sensitivity also severely limits price adjustment space, thereby compressing profit margins. Given the lack of direct $\eta_{i,k\to t}$ empirical evidence for determining the optimal value of the load transfer coefficient, this study conducts a sensitivity analysis specifically targeting this parameter. By adjusting the transfer coefficients for various user categories by ±30% relative to baseline values, we examined changes in PPO algorithm performance. Results indicate that a 30% increase in the transfer coefficient enhances peak-hour load reduction rates by approximately 8% while reducing average daily profits by about 3%; conversely, a 30% decrease lowers peak-hour load reduction rates by roughly 7% but increases average daily profits by around 2%. Overall, the transfer coefficient significantly impacts peak shaving effectiveness but has limited influence on profitability metrics, demonstrating that the core conclusion (the PPO algorithm outperforms competing algorithms) is not highly sensitive to this parameter value. This sensitivity analysis provides valuable insights for power retailers to refine pricing strategies and parameter settings under diverse user profile characteristics.

V. CONCLUSION

This paper applies reinforcement learning methods to the dynamic pricing problem in the electricity retail market and proposes a pricing framework based on the proximal policy optimization algorithm. Through the aforementioned theoretical analysis and simulation validation, the following key conclusions can be drawn.

First, modeling the dynamic pricing problem as a Markov decision process in a continuous action space is a rational and effective approach. This modeling strategy fully leverages the advantages of reinforcement learning in addressing complex sequential decision-making problems while avoiding the challenges associated with precise probabilistic modeling of user response behaviors and market stochastic parameters inherent in traditional methods. Simulation results demonstrate that the PPO algorithm outperforms benchmark algorithms such as DDPG and TD3 across multiple metrics, including retailer profitability, peak-hour load reduction, and electricity price stability.

Second, the adaptive exploration coefficient adjustment mechanism tailored to pricing scenarios effectively mitigates premature convergence during policy training. By dynamically adjusting the weight of the policy entropy regularization term, this mechanism achieves an optimal balance between exploration and exploitation, enabling the algorithm to continuously optimize the pricing strategy throughout the later stages of training.

Third, the pricing strategy proposed in this paper enhances the profits of electricity retailers while having a relatively limited impact on users' electricity bills (an increase of only 0.8% compared to a fixed-price scenario), demonstrating a good balance between benefits and costs. The peak-hour load reduction rate reached 16.5%, indicating that dynamic pricing, as a demand-side management tool, holds practical value in guiding rational electricity consumption and alleviating grid pressure during peak hours.

Of course, this study also has several limitations. First, the model assumes a monopolistic scenario with a single electricity retailer and does not account for game-theoretic interactions under competition among multiple retailers. Second, while the user demand response model incorporates load transfer descriptions based on historical periods and approximates chain-transfer effects (with 50% attenuation applied to transfers exceeding two steps), its characterization of user behavior remains overly simplistic; real-world decision-making processes likely involve more complex learning and adaptation mechanisms. Additionally, the load transfer coefficients in this study are primarily empirically determined without direct support from large-scale user behavior data, necessitating parameter calibration using actual measurement data in future work. The study employs an online reinforcement learning framework, which may face operational risks during the exploration phase in practical deployment. Future research could focus on: extending the single-agent framework to a multi-agent game-theoretic model to examine competitive market environments; introducing more refined user behavior models, such as behavioral economics frameworks based on prospect theory; and exploring offline reinforcement learning approaches in electricity pricing to mitigate operational risks associated with online exploration.

FOUNDATION ITEMS

This work was supported by Science and Technology Project of State Grid Sichuan Electric Power Company: 521996240003.

REFERENCES

[1] A. M. H. Shehzad, A. Ussama, F. Umar, et al., “Dynamic Price-Based Demand Response through Linear Regression for Microgrids with Renewable Energy Resources,” Energies, vol. 15, no. 4, pp. 1385–1385, 2022, doi: 10.3390/EN15041385.

[2] Li Bin, Guo Huifang, Xue Li, et al., “Research on Key Technologies for the Development of Power Procurement Agency Services and Auxiliary Decision Support,” Inner Mongolia Electric Power Technology, vol. 41, no. 3, pp. 51–56, 2023, doi: 10.19929/j.cnki.nmgdljs.2023.0039.

[3] Li Yuan, Chi Kun, Wang Zhou, et al., “Pricing strategies for multi-electricity-gas-heating microgrid systems based on reinforcement learning,” Southern Power Grid Technology, vol. 18, no. 1, pp. 94–101, 2024, doi: 10.13648/j.cnki.issn1674-0629.2024.01.010.

[4] Q. Liu, D. Feng, Y. Zhou, et al., “Bounded rational bidding strategy of GenCo in electricity spot market based on prospect theory and distributional reinforcement learning,” Sustainable Energy, Grids and Networks, vol. 44, pp. 102036–102036, 2025, doi: 10.1016/J.SEGAN.2025.102036.

[5] W. Huo, Y. Zhang, H. Zhao, et al., “Price signal forecasting for day-ahead offering strategy of price-maker renewable energy producers considering different risk preferences,” Applied Energy, vol. 401, no. PC, pp. 126819–126819, 2025, doi: 10.1016/J.APENERGY.2025.126819.

[6] Xu Dan, Hu Xiaojing, Hu Fei, et al., “Electricity market quantity-price combined bidding strategy based on deep reinforcement learning,” Grid Technology, vol. 48, no. 8, pp. 3278–3286, 2024, doi: 10.13335/j.1000-3673.pst.2023.2100.

[7] Liu Feiyu, Wang Jiwen, Wang Zhengfeng, et al., “Research on multi-agent free-cooperative bidding mechanisms based on a two-stage deep reinforcement learning algorithm,” China Journal of Electrical Engineering, vol. 44, no. 12, pp. 4626–4638, 2024, doi: 10.13334/j.0258-8013.pcsee.222935.

[8] L. Ting, Z. Weige, and D. Xiaowei, “Operation Strategy of Electricity Retailers Based on Energy Storage System to Improve Comprehensive Profitability in China’s Electricity Spot Market,” Energies, vol. 14, no. 19, pp. 6424–6424, 2021, doi: 10.3390/EN14196424.

[9] S. Yan, W. Yin, W. Yan, et al., “Optimizing intelligent startup strategy of power system using PPO algorithm,” Intelligent Decision Technologies, vol. 18, no. 4, pp. 3091–3104, 2024, doi: 10.3233/IDT-240122.

[10] W. Lulu, Z. Kaile, L. Jun, et al., “Modified deep learning and reinforcement learning for an incentive-based demand response model,” Energy, vol. 205, 2020, doi: 10.1016/j.energy.2020.118019.

[11] Z. Xu, S. Yuanzhang, Y. Jun, et al., “Day-ahead energy pricing and management method for regional integrated energy systems considering multi-energy demand responses,” Energy, vol. 251, 2022, doi: 10.1016/J.ENERGY.2022.123914.

[12] J. Zhao, Z. Gao, and Z. Chen, “Energy Management of Electric-Hydrogen Coupled Integrated Energy System Based on Improved Proximal Policy Optimization Algorithm,” Energies, vol. 18, no. 15, pp. 3925–3925, 2025, doi: 10.3390/EN18153925.

[13] Song Yonghua, Yu Peipei, and Zhang Hongcai, “Demand response operation control of regional cooling systems based on reinforcement learning with composite two-end sampling under real-time electricity pricing mechanisms,” China Science: Technical Science, vol. 53, no. 10, pp. 1699–1712, 2023.

[14] A. Natalia and V. Nikolai, “The Optimal Mechanism Design of Retail Prices in the Electricity Market for Several Types of Consumers,” Mathematics, vol. 9, no. 10, pp. 1147–1147, 2021, doi: 10.3390/MATH9101147.

[15] Li Chaoying and Tan Qinliang, “Market trading strategies for thermal power enterprises under the new power system based on agent modeling,” China Electric Power, vol. 57, no. 2, pp. 212–225, 2024, doi: 10.11930/j.issn.1004-9649.202303088.

[16] B. Mandal, K. P. Roy, and C. Paul, “Dynamic Economic Dispatch Problem in Hybrid Renewable Energy Sources Based Power Systems Using Chaotic Hippopotamus Optimization Algorithm,” Iranian Journal of Science and Technology, Transactions of Electrical Engineering, vol. 50, no. 2, pp. 1–28, 2025, doi: 10.1007/S40998-025-00893-4.

[17] “Research Analysis on Price Elasticity and Demand Response Behavior in the Electricity Retail Market,” Beijing: China Electricity Council, 2022, Report No.: CEC-2022-MR-087.