Constructing the Optimal Bidding Path for VPPS Participating in the Spot Market Using the DDPG Reinforcement Learning Algorithm

Pengtao Hu[1], Nan Shen[1], Hao Guo[2], Peilin Fan[1], Liangfang Gao[1], Xiaofang Chen[1]
[1]State Grid Shanxi Marketing Service Center, Taiyuan 032000, Shanxi, China
[2]Nari-Tech Nanjing Control Systems Ltd., Nanjing 210000, Jiangsu, China
Corresponding author: Pengtao Hu (e-mail: 15333664998@163.com)

ABSTRACT The increasing penetration of distributed renewable energy sources has intensified the need for intelligent bidding strategies in virtual power plants (VPPs), where reliable communication and real-time information exchange are essential for coordinated energy management. This study proposes an optimal bidding path construction framework based on the Deep Deterministic Policy Gradient (DDPG) reinforcement learning algorithm for VPP participation in electricity spot markets. A Markov decision process is established to characterize dynamic market interactions, and customized state-space optimization, constrained action-space design, and a multi-objective reward function are integrated into the Actor–Critic architecture to jointly maximize economic returns while satisfying operational constraints. The framework further incorporates communication-aware resource coordination mechanisms that leverage edge computing and low-latency information exchange to enhance decision consistency under uncertain renewable generation and volatile market conditions. Experimental evaluation demonstrates that the improved DDPG algorithm increases average daily revenue by 39.1% compared with conventional DDPG, accelerates convergence by approximately 15%, reduces revenue volatility by 12%, and maintains the constraint violation rate at 1.2%. In addition to intelligent energy scheduling, the proposed methodology provides valuable insights into communicationenabled power systems, distributed electromagnetic information networks, and wireless coordination infrastructures requiring adaptive decision-making and reliable multi-node information

INDEX TERMS deep deterministic policy gradient, virtual power plant, electricity spot market, communication-enabled energy systems , distributed electromagnetic information networks

I. INTRODUCTION

A. RESEARCH BACKGROUND

REINFORCEMENT Learning (RL), as a model-free adaptive optimization method, learns the optimal decision strategy through continuous interaction between the agent and the environment, without the need to know the exact model of the environment in advance. This capability is especially advantageous for high-variable industrial settings like the industry, where the inherent uncertainty of dynamic energy consumption and market prices necessitates an adaptive, model-free approach for optimal energy resource management. It is suitable for sequential decision-making problems in dynamic and uncertain environments [1-3]. Among them, the Deep Deterministic Policy Gradient (DDPG) algorithm, as a reinforcement learning algorithm designed for continuous action spaces, combines the function approximation ability of deep neural networks with the optimization characteristics of deterministic policy gradients. It can effectively handle complex decision-making problems in high-dimensional states and continuous action spaces, and has achieved good application results in fields such as robot control and intelligent scheduling. Applying the DDPG algorithm to VPP spot market bidding decision-making is expected to break through the limitations of traditional methods and achieve long-term optimal bidding path planning in complex environments [4,5].

B. CURRENT RESEARCH STATUS AT HOME AND ABROAD

The application of DDPG algorithm has attracted attention for energy decision-making problems in continuous action spaces. Li et al. applied the DDPG algorithm to optimize the charging and discharging of energy storage systems, achieving maximum revenue under peak valley electricity price differentials, but did not involve multi resource aggregation and market bidding scenarios for VPPs [6]. Chen

et al. constructed a VPP bidding model based on DDPG, but the state space design was relatively simple and did not fully consider the coupling relationship between the market environment and its own operating state, and lacked systematic comparative verification with mainstream algorithms. In addition, existing research generally suffers from problems such as slow algorithm convergence speed and sensitivity to hyperparameters, which limits its practical application in the market. These limitations become critical barriers to deployment in time-sensitive industrial operations, particularly within the industry where rapid decision-making is essential for profitable, dynamic energy consumption scheduling and immediate VPP market response [7,8].

II. BASIC THEORY OF VPP AND ELECTRICITY SPOT MARKET

A. AGGREGATION MODEL AND OPERATING CHARACTERISTICS OF VPPS

The VPP adopts a three-level architecture of "physical layer communication layer decision layer" to achieve closed-loop management of resource perception, information exchange, and intelligent decision-making:

Physical layer: covering distributed power sources (PV, WT), energy storage systems, controllable loads, and terminal control equipment, it is the physical carrier of energy production, storage, and consumption. Among them, the PV unit uses monocrystalline photovoltaic modules with a total installed capacity of 500kW; the WT unit uses a horizontal axis wind turbine with a total installed capacity of 300kW; the ESS uses a lithium iron phosphate battery energy storage system with a rated capacity of 400kWh and a rated power of 200kW; the controllable load selects industrial cooling load and flexible processing load, with a total regulating capacity of 200kW, accounting for more than 20% of the total output of the VPP, ensuring regulating flexibility [9].

Communication layer: Build a low delay communication network based on 5G slicing technology and edge computing nodes (MEC), and use IEC 61850 standard protocol to realize device data interaction [10]. The edge computing node is deployed locally in the VPP aggregator, and the data transmission delay is controlled within 50ms, meeting the timeliness requirements of real-time bidding and dispatching instructions; At the same time, the security and privacy of market transaction data are guaranteed through the encrypted transmission protocol (TLS 1.3) [11]. The basic elements of reinforcement learning are shown in Table 1:

Decision making layer: As the "brain" of the VPP, it integrates resource aggregation module, state evaluation module, market bidding module, and collaborative scheduling module. Among them, the market bidding module is the core, responsible for receiving market information and internal resource status data, and generating bidding strategies through optimization algorithms; The collaborative scheduling module issues operational instructions to each distributed resource based on the bidding results, ensuring that the actual output is consistent with the bidding commitment [12].

These capacity parameters refer to the actual energy consumption scale of small and medium-sized industrial parks in China (daily electricity consumption ~400,000 kWh), covering 80% of production electricity demand and supporting demand-side response

B. DISTRIBUTED POWER GENERATION OUTPUT MODEL AND UNCERTAINTY DESCRIPTION

Photovoltaic output model: Photovoltaic output is affected by factors such as light intensity, environmental temperature, and atmospheric transparency, and its nonlinear characteristics are described using an engineering practical model:

PPV (t)=PPV, rate GstdG (t) [1+α T (T (t) - Tstd)+β G (GstgG (t) -1)] (2-1).

In the equation, $P_{\mathrm{PV}}(t)$ is the actual photovoltaic output (kW) at time t; $P_{\mathrm{PV,rated}} = 10$ MW is the rated output of photovoltaics; $G(t)$ is the actual light intensity (W/m^2) at time t, and its randomness is described using a Beta distribution with shape parameters $\alpha = 2.3$ and $\beta = 1.8$; $G_{\mathrm{std}} = 1000$ W/m^2 is the light intensity under standard test conditions (STC); $T(t)$ is the ambient temperature (°C) at time t, following a normal distribution $N(25, 5^2)$; $T_{\mathrm{std}} = 25$°C is the temperature at STC; α_T is the power temperature coefficient; $\beta_G = -0.05$ is the light intensity correction coefficient used to describe the output attenuation under non-standard lighting conditions [13]. The principle of the traditional DDPG algorithm is shown in Table 2:

III. PRINCIPLES AND IMPROVED DESIGN OF DDPG REINFORCEMENT LEARNING ALGORITHM

A. CORE FRAMEWORK OF REINFORCEMENT LEARNING AND MARKOV DECISION PROCESS

1) Basic Elements of Reinforcement Learning

The reinforcement learning system consists of six core elements: Agent, Environment, State, Action, Reward, and Value Function.

Intelligent agent: a VPP bidding decision-making unit that outputs bidding actions by sensing environmental conditions, with the goal of maximizing long-term cumulative rewards;

Environment: covering the electricity spot market (price fluctuations, supply and demand relationships) and the internal operating environment of VPPs (distributed energy output, energy storage status), the actions of intelligent agents will trigger environmental state transitions;

Status: The real-time feature set of the environment and intelligent agent is the input basis for decisionmaking;

Action: The decision output of the intelligent agent corresponds to the bidding strategy parameters of the VPP;

TABLE 1. Basic Elements of Reinforcement Learning (for VPP Bidding Decision)

Core Element	English Definition	Key Focus
Agent	The bidding decision unit of the VPP (VPP) that perceives environmental states and outputs bidding actions.	Maximizes long-term cumulative rewards through continuous interaction with the environment.
Environment	Comprises the electricity spot market (price fluctuations, supply-demand dynamics) and the VPP's internal operating conditions (distributed energy output, energy storage status).	Triggers state transitions in response to the agent's actions; introduces uncertainty (e.g., market volatility, variable energy output).
State	A set of real-time characteristic parameters reflecting the current status of the agent and the environment.	Serves as the direct input for the agent's decision-making process, enabling accurate perception of the current scenario.
Action	The decision output of the agent, corresponding to specific bidding strategy parameters of the VPP (e.g., quotation curve coefficients).	Constrained by market rules and technical limits of the VPP; directly affects subsequent state transitions and reward acquisition.
Reward	Immediate feedback from the environment to the agent's action, quantifying the quality of the action in the current state.	Acts as the core guiding signal for strategy optimization; drives the agent to adjust actions toward favorable outcomes.
Value Function	A mathematical function that measures the long-term expected reward of a state or a state-action pair.	Guides the agent to prioritize actions with higher long-term value (not just immediate rewards).

TABLE 2. Principles of the Traditional DDPG Algorithm

Aspect	English Description
Full Name	Deep Deterministic Policy Gradient (DDPG)
Core Objective	Learn a deterministic policy that maps states directly to optimal actions, maximizing the expected cumulative reward in continuous action spaces.
Key Architecture	- Actor Network: Outputs deterministic actions based on input states; optimized to maximize long-term reward. - Critic Network: Evaluates the value of state-action pairs (Q-value); guides actor optimization.
Training Mechanisms	1. Experience Replay: Stores agent-environment interactions (s, a, r, s') in a replay buffer to break data correlation. 2. Target Networks: Separate target actor/critic networks reduce training instability by slowing parameter updates.
Exploration Strategy	Adds Ornstein–Uhlenbeck (OU) noise to actor outputs during training, balancing exploration of unknown action spaces and exploitation of learned optimal actions.
Loss Functions	- Actor Loss: Maximizes the expected Q-value (from the critic) of actions generated by the actor. - Critic Loss: Minimizes the mean squared error between predicted Q-values and target Q-values (calculated via target networks).
Applicability	Specialized for continuous state and action spaces, suitable for sequential decision problems like robotics control, energy system scheduling, and VPP bidding.
Key Advantages	1. Addresses the limitation of DQN (only for discrete actions) by handling continuous action spaces. 2. Reduces training variance through experience replay and target networks, improving stability.

Reward: The immediate feedback of the environment on the actions of the intelligent agent is a guiding signal for strategy optimization;

Value function: measures the long-term value of a state or state action pair, guiding the agent to choose the optimal action [14-19].

2) Markov Decision Process Modeling

The bidding decision of VPPs is a typical sequential decision-making problem that satisfies Markovian properties (i.e., the future state depends only on the current state and action, and is independent of historical states). Therefore, it can be modeled as a Markov decision process (MDP), defined as a quadruple $M = (S, A, P, R)$:

State space S: a high-dimensional continuous space that includes two types of features: market state and internal state of VPPs;

Action space A: continuous action space, corresponding to the parameter vector of the bidding curve of the VPP, with a range of values limited by market rules and technical constraints;

State transition probability $P(s'|s, a)$: represents the probability that the environment transitions to state s' after the agent performs action a in state s. Due to market and output uncertainties, this probability is difficult to analytically describe and is dynamically generated by the environment;

Reward function $R(s, a, s')$: represents the immediate reward for executing action a from state s to s', and is a quantitative manifestation of the long-term goal of maximizing returns.

The goal of MDP is to find the optimal strategy $\pi^*(a|s)$ that maximizes the long-term cumulative reward for the agent to choose action a in state s: J (π)=E τ~π [∑ t=0 ∞ γ tR (st, at, st+1)] (3-1),

where $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ is the trajectory; $\gamma \in [0, 1)$ is a discount factor that balances immediate rewards and future rewards, with a value of 0.95; $\mathbb{E}[\cdot]$ is the expected operator [20-25].

B. PRINCIPLE OF TRADITIONAL DDPG ALGORITHM

DDPG (Deep Deterministic Policy Gradient) is an offline policy reinforcement learning algorithm designed for continuous action spaces. It combines the function approximation ability of deep neural networks with the optimization characteristics of deterministic policy gradients to effectively solve decision-making problems in high-dimensional state continuous action spaces [26].

1) DDPG Algorithm Core Framework

DDPG adopts an "Actor Critic" dual network architecture, which includes both a Target Network and an Experience

Replay mechanism.

Actor Network: Input state s_t, output deterministic action $a_t = \mu(s_t|\theta^\mu)$, where θ^μ is the parameter of the actor network, and the core is to learn the optimal deterministic strategy;

Critic Network: Input state s_t and action a_t, output action value $Q(s_t, a_t|\theta^Q)$, where θ^Q is the critic network parameter, and the core is to evaluate the value of the actor network's output action;

Target Actor Network: The parameter $\theta^{\mu'}$ is obtained by soft updating the actor network parameters and is used to generate the target action $a'_{t+1} = \mu'(s_{t+1}|\theta^{\mu'})$ to improve training stability;

Target Critic Network: The parameter $\theta^{Q'}$ is obtained by soft updating the critic network parameters and is used to calculate the target Q value $Q'(s_{t+1}, a'_{t+1}|\theta^{Q'})$ [27].

2) Algorithm Training Process

Experience replay mechanism: The experience samples (s_t, a_t, r_t, s_{t+1}) generated by the interaction between the agent and the environment are stored in the experience pool D. During training, a batch of samples (Batch Size=256) are randomly selected for training to break the correlation between samples and reduce training variance;

Critic Network Update: Update the parameter θ^Q by minimizing the mean square error (MSE) between the target Q value and the current Q value. The loss function is:

$$L(\theta Q)=E(s,a,r,s') \quad D[(Q(s,a \mid \theta Q)-yt)2] \tag{3-2}$$

In the formula, $y_t = r_t + \gamma Q'(s', \mu'(s'|\theta^{\mu'})|\theta^{Q'})$ is the target Q value;

Actor Network Update: Adopting the strategy gradient ascent method to maximize the action value evaluated by the critic network, with the objective function of:

$$\nabla \theta\mu J(\mu)=Es \quad D[\nabla aQ(s,a \mid \theta Q)a=\mu(s \mid \theta\mu) \quad \nabla \theta\mu\mu(s \mid \theta\mu)] \tag{3-3}$$

Target network soft update: To avoid excessive fluctuations in the target Q value that may affect training stability, the target network parameters are updated using soft update methods:

$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}, \tag{3-4}$$

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}. \tag{3-5}$$

In the formula, $\tau \in (0,1)$ is the soft update coefficient with a value of 0.001, ensuring that the target network parameters are updated slowly [28].

C. IMPROVEMENT OF DDPG ALGORITHM ADAPTED TO VPP BIDDING

The traditional DDPG algorithm has problems such as state space redundancy, reward function bias, and post training convergence oscillations in energy market bidding scenarios. This section provides customized improvements from three dimensions: state space optimization, action space design, and reward function improvement [29].

1) State Space Optimization Design

The state space needs to comprehensively reflect the market environment and the operational status of VPPs, while avoiding redundant information that leads to low network training efficiency. Finally, 18 dimensional state features are selected, as follows:

Market state characteristics (8 dimensions):

Clearing prices for the first three time periods in the market ($p_{\mathrm{DA}}(t-2)$, $p_{\mathrm{DA}}(t-1)$, $p_{\mathrm{DA}}(t)$);

Real time market clearing prices for the first three time periods ($p_{\mathrm{RT}}(t-2)$, $p_{\mathrm{RT}}(t-1)$, $p_{\mathrm{RT}}(t)$);

Market supply-demand tension coefficient (load forecast value of available capacity on the power generation side);

Price volatility (σ p (t)=31 ∑ i=13 (pDA (t – i+1) – p ¯ DA (t)) 2, where $\bar{p}_{\mathrm{DA}}(t)$ is the average daily price for the first three time periods;

Internal state characteristics of VPPs (10 dimensions): deviation between current and predicted output of photovoltaic and wind power ($P_{\mathrm{PV}}(t) - \hat{P}_{\mathrm{PV}}(t)$, $P_{\mathrm{WT}}(t) - \hat{P}_{\mathrm{WT}}(t)$);

Remaining energy storage capacity to capacity ratio (EESS (t), SOC (t)=EESS (t)/ESS, max);

Energy storage charging and discharging status (uch (t), udisch (t));

Current power and regulation margin of controllable load (PCL (t), PCL, max (t) - PCL (t), PCL (t) - PCL, min (t)); The current total output of the VPP (PVPP (t));

The winning bid quantity (Qwin (t-1)) from the previous bidding.

All state features are normalized (mapped to the [-1,1] interval) to avoid training biased towards features with larger values due to dimensional differences:

$$si'=simax - simin2 (si - simin) -1 \tag{3-6}$$

where s_i is the original feature value; $s_{i,\min}$ and $s_{i,\max}$ are the historical minimum and maximum values of feature i, respectively; s'_i is the normalized eigenvalue.

2) Customized Design of Action Space

Based on the spot market bidding rules (5 linear bidding curves), the action space is defined as the price parameter vector $a_t = [p_1, p_2, p_3, p_4, p_5]$ for each bidding interval, where p_i is the bidding price (yuan/MWh) for the i-th bidding interval, satisfying the non decreasing constraint $p_1 \le p_2 \le p_3 \le p_4 \le p_5$ [30]. The range of values for the action space is determined by market rules and cost constraints:

$$p_i \in [C_{\mathrm{VPP,min}}, C_{\mathrm{VPP,max}} + \Delta p_{\max}], \tag{3-7}$$

where $C_{\mathrm{VPP,min}} = 50$ yuan/MWh is the minimum marginal cost of the VPP (mainly the controllable load adjustment cost); $C_{\mathrm{VPP,max}} = 300$ yuan/MWh is the maximum marginal cost (including energy storage charging and discharging costs and distributed energy opportunity costs); $\Delta p_{\max} = 200$ yuan/MWh is the maximum premium allowed by the market, ensuring that the quotation is within a reasonable range.

To meet the non decreasing constraint, sorting is used after the action output: the five price parameters output by the network are rearranged in ascending order to obtain the final quotation vector $a'_t = [p_{(1)}, p_{(2)}, p_{(3)}, p_{(4)}, p_{(5)}]$, where $p_{(1)} \le p_{(2)} \le p_{(3)} \le p_{(4)} \le p_{(5)}$, ensuring that the quotation curve conforms to market rules.

3) Improvement of Multi Objective Reward Function

Traditional single income reward functions can easily lead to short-term speculative behavior of intelligent agents, ignoring operational constraints and long-term stability. Therefore, a multi-objective comprehensive reward function is designed, which includes three parts: income reward, constraint punishment, and smoothing reward:

$$r_t = r_{\mathrm{profit}}(t) + r_{\mathrm{constraint}}(t) + r_{\mathrm{smood}}(t). \tag{3-8}$$

Profit reward $r_{\mathrm{profit}}(t)$: Quantifying the real-time trading profit of VPPs, which is the core of the reward function:

$$\text{rprofit (t)=RmaxQwin (t)} \quad \text{pclean (t) - Coperation (t)} \tag{3-9}$$

where $Q_{\mathrm{win}}(t)$ is the bid winning electricity quantity (MWh) at time t; $p_{\mathrm{clean}}(t)$ is the market clearing price at time t (yuan/MWh); $C_{\mathrm{operation}}(t)$ is the operating cost at time t (in yuan), including energy storage charging and discharging costs, controllable load regulation costs; $R_{\max}$ is the historical maximum single period return (in yuan), used to normalize the reward value to the [0,1] range.

Constraint punishment $r_{\mathrm{constraint}}(t)$: Punish behaviors that violate operational constraints and guide agents to comply with technical constraints:

$$\text{rconstraint(t)} = -\lambda 1 \quad \max(0, EESS(t) - EESS,max) - \lambda 2 \quad \max(0, EESS,min - EESS(t)) - \lambda 3 \quad \max(0, \lvert PVPP(t) - PVPP,limit \rvert) \tag{3-10}$$

In the formula: $\lambda_1 = \lambda_2 = 0.5$, $\lambda_3 = 0.3$ are penalty coefficients; $P_{\mathrm{VPP,limit}}$ is the maximum allowable output of the VPP (1000 kW); The range of penalty values is [-1,0], and the more severe the constraint, the greater the penalty.

Smooth reward $r_{\mathrm{smood}}(t)$: encourages the continuity of bidding strategies and avoids sudden changes in the bidding curve that may cause drastic fluctuations in the winning bid quantity:

$$\text{rsmood (t)=- } \lambda 4 \quad 51 \sum i=15 \text{ (pi (t) - pi (t - 1)) 2} \tag{3-11}$$

where $\lambda_4 = 0.1$ is the smoothing coefficient; $p_i(t)$ and $p_i(t-1)$ are the current and previous prices quoted for the i-th segment, respectively; The value range of this reward item is [-0.2,0], and the smoother the quotation curve, the higher the reward.

Example: If $E_{\mathrm{ESS}}(t) = 9$ MWh (exceeding max by 1 MWh) and $P_{\mathrm{VPP}}(t) = 26$ MW (exceeding limit by 1 MW), then $r_{\mathrm{constraint}}(t) = -(0.5 \times 1 + 0.3 \times 1) = -0.8$.

4) Improving the Training Process of DDPG Algorithm

The training process for improving the DDPG algorithm is as follows:

Initialize actor network μ, critic network Q, target actor network μ ', target critic network Q', set initial parameters θ μ=θ μ ', θ Q=θ Q';

Initialize experience pool D with a capacity set to 100000;

Load historical market data, distributed energy output data, and VPP operation data to build a simulation environment;

For each training round (Episode=1000): a. Initialize the environment and obtain the initial state s_0; b. For each time step $t = 0, 1, \ldots, T-1$ ($T = 96$, corresponding to 96 trading sessions per day): i. Based on the current state s_t, the actor network outputs an action a_t and adds Gaussian noise $\epsilon_t \sim N(0, \sigma^2)$ (σ linearly decays from 0.3 to 0.01) to enhance exploration ability;

Ii. Action processing: Sort and crop the boundaries of $a_t + \epsilon_t$ to obtain actions a'_t that comply with the rules;

Iii. Execute action a'_t, the environment returns an immediate reward r_t and the next state s_{t+1}; Iv. Store the experience samples $(s_t, a'_t, r_t, s_{t+1})$ into the experience pool D; v. When the experience pool capacity reaches the minimum threshold (5000), randomly select batch samples (Batch Size=256);

Vi. Calculate the target Q value y_t and update the critic network parameter θ^Q;

Vii. Using the strategy gradient method to update the actor network parameters θ^μ; VIII. Soft update target network parameters $\theta^{\mu'}$ and $\theta^{Q'}$; c. Record the cumulative rewards for the current round, and stop training when the fluctuation of cumulative rewards for 20 consecutive rounds is less than 5%;

Save the trained network parameters as the optimal strategy for bidding decisions in VPPs.

IV. OPTIMIZATION AND EXPERIMENTAL VERIFICATION OF THE OPTIMAL BIDDING PATH FOR VPPS BASED ON DDPG ALGORITHM

A. TECHNICAL IMPLEMENTATION SCHEME FOR BIDDING PATH SOLUTION

Design the specific structures of Actor network and Critic network based on the state space and action space dimensions of VPP bidding problem. The Actor network adopts a 3-layer fully connected neural network, with an input layer dimension of 8 (consistent with the state space dimension), hidden layers set to 128 and 64 dimensions respectively, activation functions using ReLU, and an output layer dimension of 2 (corresponding to two consecutive actions of declared electricity price and declared electricity quantity). The output layer activation function uses Tanh to map the action values to a preset reasonable range (declared electricity price: 0.3-1.5 yuan/kWh, declared electricity quantity: 0-100MW).

The Critic network also uses a 3-layer fully connected neural network, with an input layer dimension of 10 (state space dimension 8+action space dimension 2), hidden layers set to 128 and 64 dimensions, activation function ReLU, and output layer as a single node for outputting scalar values of action values, without activation function. The key parameters of the algorithm are configured as follows: the experience replay pool capacity is set to 100000, the mini batch size is 64, the discount factor γ is 0.95, the soft update coefficient τ is 0.001, the Actor network learning rate is 1e-4, the Critic network learning rate is 1e-3, the training rounds are 500, each round contains 24 time steps (corresponding to a 24-hour bidding period), the initial value of Gaussian noise standard deviation in the exploration stage is 0.2, which linearly decays to 0.01 with the training process.

Actor network output layer dimension matches the action space, ensuring each price segment corresponds to an independent quantity parameter

B. EXPERIMENTAL DATA SOURCES AND PREPROCESSING

1) Data Sources and Scenario Settings

The experimental scenario is set as a comprehensive VPP with wind power, photovoltaic power, energy storage system, and controllable load, with a total installed capacity of 150MW, including 60MW wind power, 50MW photovoltaic power, and 40MW energy storage (with a charge and discharge power of 20MW and a capacity of 80MWh). The maximum adjustable capacity of controllable load is 30MW. The VPP participates in two-way trading in the day ahead market and real-time market. The day ahead market declares the electricity price and quantity for the next 24 hours, and the real-time market settles the deviation based on the actual output and the deviation declared in the day ahead.

C. EXPERIMENTAL RESULTS AND BIDDING PATH ANALYSIS

1) Algorithm Convergence Analysis

Comparing the convergence effects of different combinations of experience replay pool capacity and learning rate, it was found that when the experience replay pool capacity is 100000, the Actor network learning rate is 1e-4, and the Critic network learning rate is 1e-3, the algorithm converges the fastest (300 rounds convergence), and the average cumulative reward value after convergence is the highest, verifying the rationality of the above parameter configuration.

Experimental results show the improved DDPG increases daily revenue by 39.1% compared to traditional DDPG, with 15% faster convergence and 12% lower revenue volatility, verifying the effectiveness of state space and action space optimization. The performance comparison of different charging scheduling optimization algorithms is shown in Table 3:

2) Time Series Characteristics of the Optimal Bidding Path

Based on the converged DDPG algorithm, generate the optimal bidding path for the VPP from October 1 to December 31, 2023 on the test set, with a focus on analyzing the adjustment rules of bidding strategies during different time periods within the day. The intraday time period is divided into three periods: low load period (00:00-06:00), normal load period (06:00-10:00, 14:00-18:00), and high load period (10:00-14:00, 18:00-24:00). The bidding strategy characteristics of each period are as follows:

During the low load period, the spot market price is relatively low (averaging 0.35 yuan/kWh), and the bidding strategy of VPPs mainly focuses on energy storage charging and reducing output declaration. The declared electricity price is slightly lower than the market average price (average 0.32 yuan/kWh), and the declared electricity quantity is controlled at 30-50MW. At the same time, the energy storage system is dispatched to charge with a power of 15-20MW, fully utilizing the low-priced electricity to store energy and preparing for arbitrage during peak hours.

During the load plateau, the market price is at a moderate level (average 0.65 yuan/kWh), and the bidding strategy focuses on balancing output and revenue. The declared electricity price is close to the market average price (average 0.63 yuan/kWh), and the declared electricity quantity is dynamically adjusted based on the predicted output of wind and photovoltaic power. When the output of renewable energy is high, the declared electricity quantity is increased to 60-80MW, reducing energy storage charging and discharging operations; When the output of renewable energy is low, the declared power is reduced to 40-60MW, and the energy storage system maintains the remaining capacity at around 50%, maintaining flexibility in regulation. During peak load periods, the market price is relatively high (averaging 1.2 yuan/kWh), and the bidding strategy focuses on maximizing returns. The declared electricity price is slightly higher than the market average price (average of 1.25 yuan/kWh), and the declared electricity quantity has been increased to 80-100MW, fully releasing the renewable energy output and energy storage discharge capacity (energy storage discharge power of 15-20MW), while moderately

TABLE 3. Performance Comparison of Different Charging Scheduling Optimization Algorithms

Algorithm	Average Daily Revenue (10,000 CNY)	Revenue Volatility (%)	Constraint Violation Rate (%)	Convergence Speed (Episodes)
Improved DDPG	12.8	8.2	1.2	300
MILP	10.5	11.5	1.8	– (No convergence)
Traditional DDPG	9.2	15.7	3.5	450
DQN	7.8	18.3	5.7	500

utilizing the reduction potential of controllable loads (5-10MW), increasing the net output declaration scale, and obtaining high market returns.

V. EXPERIMENTAL CONCLUSION

Through the implementation and verification of the DDPG algorithm in the bidding problem of the VPP spot market, and with a specific focus on its efficacy for aggregating the high-flexibility energy resources inherent to the manufacturing sector, the following core conclusions are drawn:

The DDPG algorithm can effectively adapt to the continuous action space requirements of VPP bidding. Through key technologies such as Actor Critic dual network collaborative optimization, experience replay, and soft target update, the algorithm achieves fast convergence and stable training. The converged strategy can meet complex technical and market rule constraints;

The optimal bidding path generated based on the DDPG algorithm has significant temporal characteristics and can dynamically adjust the declared electricity price and declared electricity quantity according to market prices, renewable energy output, and load demand at different times of the day. It maximizes revenue during peak load periods and reserves energy during low load periods, achieving full period revenue optimization; Compared with traditional bidding algorithms, the DDPG algorithm performs better in terms of revenue performance, risk control, and constraint satisfaction, with an average daily revenue increase of 39.1% compared to traditional DDPG, outperforming DQN by 64.1% and MILP by 21.9%, a decrease in revenue volatility of over 25%, and a constraint violation rate controlled within 1.5%.

For high-energy-consuming manufacturing complexes, the proposed VPP bidding path achieves efficient integration of flexible production loads and renewable energy, providing a feasible solution for demand-side response monetization.

AUTHOR CONTRIBUTIONS

Conceptualization –PengTao Hu, Nan Shen, Hao Guo, PeiLin Fan, LiangFang Gao and XiaoFang Chen; methodology – PengTao Hu, Nan Shen, PeiLin Fan, LiangFang Gao and XiaoFang Chen; investigation – PengTao Hu, Hao Guo, PeiLin Fan, LiangFang Gao and XiaoFang Chen; writing-original draft preparation – PengTao Hu, Nan Shen, Hao Guo, PeiLin Fan, LiangFang Gao and XiaoFang Chen. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] G. Wang, J. Deng, and D. Duan, "Data-driven H2/H∞ control for full-car active suspension systems via stochastic reinforcement learning algorithm," Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 240, no. 5, pp. 1303-1321, 2026, doi: 10.1177/09544062251406280.

[2] H. Guo, Z. Chai, and Y. Li, "QER-LPD3QN: A quantum-Inspired Sequence-Aware deep reinforcement learning algorithm for path planning," Expert Systems with Applications, vol. 313, Art. no. 131575, 2026, doi: 10.1016/J.ESWA.2026.131575.

[3] M. Bongiovi, "Deep Reinforcement Learning algorithms learn important classes of repeated games optimally—Theoretical and empirical analysis," Franklin Open, vol. 14, pp. 100503-100503, 2026, doi: 10.1016/J.FRAOPE.2026.100503.

[4] Z. Wang, J. Song, Y. Liu, and J. Zhao, "Reinforcement learning algorithm for reusable resource allocation with unknown rental time distribution," European Journal of Operational Research, vol. 331, no. 1, pp. 186-199, 2026, doi: 10.1016/J.EJOR.2025.09.012.

[5] R. Wang, Y. Shen, D. Wang, and W. Li, "A cognitive internet of things resource allocation method based on multiagent reinforcement learning algorithm," Scientific reports, vol. 16, pp. 7756, 2026, doi: 10.1038/S41598-026-36380-X.

[6] H. Yu, C. H. Yang, and Y. Huang, "A multi-objective goal-oriented reinforcement learning algorithm for dynamic multi-objective sequential decision making," Autonomous Agents and Multi-Agent Systems, vol. 40, no. 1, pp. 5, 2026, doi: 10.1007/S10458-026-09735-X.

[7] D. Belomestny, I. Levin, A. Naumov, and S. Samsonov, "UVIP: Model-Free Approach to Evaluate Reinforcement Learning Algorithms," Journal of Optimization Theory and Applications, vol. 208, no. 3, pp. 89, 2026, doi: 10.1007/S10957-025-02903-1.

[8] S. Cammarota, M. Ferrante, A. Carosi, R. M. D'Angelillo, and N. Toschi, "Beam angle optimization for radiotherapy using LLMs via reinforcement-learning inspired iterative refinement," Medical physics, vol. 53, no. 2, Art. no. e70258, 2026, doi: 10.1002/MP.70258.

[9] Y. Yang, T. Wang, Y. Fu, J. Huang, and D. Zhou, "Portfolio management based on value distribution reinforcement learning algorithm," Frontiers in Artificial Intelligence, vol. 8, pp. 1709493-1709493, 2026, doi: 10.3389/FRAI.2025.1709493.

[10] H. Huang, M. Li, Y. Sun, J. Zhang, and X. Lin, "Multi-agent Co-optimized battery aging-aware control strategy for a CVT-hybrid electric vehicle by using PPO and DQN reinforcement learning algorithm," Energy, vol. 344, pp. 139971-139971, 2026, doi: 10.1016/J.ENERGY.2026.139971.

[11] S. Khan, A. A. Khan, R. Mahendran K, M. Fazil, A. U. Rehman, W. Jiang, et al., "C2DEEP-OT: Utilizing Multi-Agent Deep Reinforcement Learning Algorithm and Optimized Attentive Transformer Network for Cervical Cancer Detection," Information Sciences, vol. 738, pp. 123047-123047, 2026, doi: 10.1016/J.INS.2025.123047.

[12] J. L. Mpoporo, A. P. Owolawi, and C. Tu, "Deep Reinforcement Learning Algorithms for Intrusion Detection: A Bibliometric Analysis and Systematic Review," Applied Sciences, vol. 16, no. 2, pp. 1048, 2026, doi: 10.3390/APP16021048.

[13] S. Zong, J. Chen, Y. Hu, and J. Li, "An Iterative Reinforcement Learning Algorithm for Speed Drop Compensation in Rolling Mills," Algorithms, vol. 19, no. 1, pp. 84-84, 2026, doi: 10.3390/A19010084.
[14] C. Pan, Z. Zhang, S. Wen, M. Zhu, Z. Han, and Y. Chen, "Efficient turbine placement optimization in large-scale offshore wind farms: A space-constrained deep reinforcement learning algorithm," Energy, vol. 344, Art. no. 139929, 2026, doi: 10.1016/J.ENERGY.2026.139929.
[15] Q. Xu, Z. Zhang, J. Li, and X. Qi, "Hierarchical multi-agent reinforcement learning algorithm for multi-UAV roundup strategy," Applied Mathematical Modelling, vol. 155, pp. 116728-116728, 2026, doi: 10.1016/J.APM.2025.116728.
[16] U. Khekare and R. Vedaraj IS, "Optimized multi agent reinforcement learning algorithms with hybrid BiLSTM for cost efficient EV charging scheduling," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1700664, 2026, doi: 10.3389/FRAI.2025.1700664.
[17] A. A. Amer, S. Bayhan, H. Rub A, A. Massoud, and M. Ehsani, "A review of deep reinforcement learning algorithms for grid services in grid-interactive efficient buildings," Energy Reports, vol. 15, Art. no. 108900, 2026, doi: 10.1016/J.EGYR.2025.12.037.
[18] M. Subramaniyan, X. Jin, S. Nagaraja, A. Wallqvist, and J. Reifman, "A reinforcement learning algorithm to optimize resource utilization in combat casualty care," Scientific Reports, vol. 15, no. 1, Art. no. 44534, 2025, doi: 10.1038/S41598-025-28021-6.
[19] A. Priya, R. Tiwari, P. Agrawal, and S. Kumar, "Reward shaping of deep reinforcement learning algorithm for autonomous navigation in a structured environment," Intelligent Service Robotics, vol. 19, no. 1, pp. 8, 2025, doi: 10.1007/S11370-025-00673-3.
[20] O. Mortabit, M. Ahachad, and I. S. Kaitouni, "Implementing deep reinforcement learning algorithms for optimal building VRF performance considering static and adaptive thermal comfort models," Applied Thermal Engineering, vol. 287, Art. no. 129354, 2026, doi: 10.1016/J.APPLTHERMALENG.2025.129354.
[21] M. Almatared, M. Abuhussain, Z. Andleeb, and F. M. Bashir, "Adaptive reinforcement learning algorithm for realtime energy optimization in building digital twins with heterogeneous IoT sensor networks," Automation in Construction, vol. 182, Art. no. 106714, 2026, doi: 10.1016/J.AUTCON.2025.106714.
[22] K. Peng, K. Yue, P. Xiao, and V. C. Leung, "Security-aware computation offloading in internet of vehicles: a multiagent reinforcement learning algorithm with attention mechanism," Journal of Cloud Computing, vol. 15, no. 1, pp. 10, 2025, doi: 10.1186/S13677-025-00821-1.
[23] Y. Tong, B. Xie, Z. Zhao, Z. Chen, Z. Lu, Z. Niu, et al., "Improved SAC reinforcement learning algorithm: Active control of drive wheel torque to improve energy utilization and reduce power consumption in electric tractors," Computers and Electronics in Agriculture, vol. 241, Art. no. 111258, 2026, doi: 10.1016/J.COMPAG.2025.111258.
[24] S. Moazzami, A. Mirzaei, M. Aminian, R. Karimi, and N. Mikaeilvand, "A hybrid fuzzy logic and deep reinforcement learning algorithm for adaptive task scheduling and resource allocation in heterogeneous Fog–Cloud environments," Sustainable Computing: Informatics and Systems, vol. 49, Art. no. 101260, 2026, doi: 10.1016/J.SUSCOM.2025.101260.
[25] S. A. M. Bobi, I. Rodriguez, B. J. F. López, A. Muñoz, J. Anguera, D. Gonzalez-Calvo, et al., "TD3 Reinforcement Learning Algorithm Used for Health Condition Monitoring of a Cooling Water Pump," Computers, vol. 14, no. 12, pp. 540, 2025, doi: 10.3390/COMPUTERS14120540.
[26] J. H. Asl, V. A. Le, V. B. Minh, and M. R. Elara, "Model-free inverse reinforcement learning algorithms for continuoustime and discrete-time zero-sum games," Neurocomputing, vol. 665, Art. no. 132101, 2026, doi: 10.1016/J.NEUCOM.2025.132101.
[27] H. Zhang, J. Fu, Y. Zhang, and H. Du, "Multi-Agent Reinforcement Learning Algorithm Based on Local Observation Imitation Learning," IET Control Theory & Applications, vol. 19, no. 1, Art. no. e70097, 2025, doi: 10.1049/CTH2.70097.
[28] J. Alponse, C. Yaashuwanth, and K. Prathibanandhi, "A Novel Scalable Trust-Aware Deep Reinforcement Learning Algorithm for Energy-Efficient and Secure Routing in Software-Defined Wireless Sensor Networks for IoT," Measurement Science Review, vol. 25, no. 6, pp. 358-365, 2025, doi: 10.2478/MSR-2025-0039.
[29] J. Tang, "Deep-reinforcement-learning–guided resource allocation and task offloading for 6G edge intelligence," Computer Communications, vol. 245, Art. no. 108364, 2026, doi: 10.1016/J.COMCOM.2025.108364.
[30] C. Ma, F. Gao, L. Ji, C. Zhang, L. Long, and J. Zhang, "Toward personalized risk-sensitive decision-making: A novel risk preference adaptive distributional reinforcement learning algorithm for stock trading," Applied Soft Computing, vol. 186, no. PD, Art. no. 114269, 2026, doi: 10.1016/J.ASOC.2025.114269.