Reactive Power Optimization Strategy for Power Grid with High Proportion of Renewable Energy Based on Reinforcement Learning Algorithm

(IJCSNT) International Journal of Communication Systems and Network Technologies
ISSN Number: 2053-6283
Volume 14, Issue No. 03, 169–181

https://journals.mirlabs.in/index.php/ijcsnt
DOI-10.18486/ijcsnt/14.2.015
IJCSNT

Hao Chen[1], Nan Yang[1], Zeliang Ma[2]

Abstract
As the high proportion of renewable energy grid connection leads to intensified grid voltage fluctuations, traditional reactive power optimization methods have obvious deficiencies in dynamic regulation capability and real-time performance, making it difficult to cope with the voltage over-limit and network loss problems caused by the randomness of wind and solar output. To this end, this paper proposes a reactive power optimization strategy based on the deep deterministic policy gradient (DDPG) algorithm, constructs a hierarchical reinforcement learning framework that includes 12-dimensional state spaces such as node voltage, renewable energy output, and load rate, and continuous/discrete action spaces such as SVG and capacitor banks, designs a composite reward function that integrates voltage deviation, network loss cost, and equipment action penalty, and implements minute-level online training through the interactive interface between the power grid simulation platform PSS/E and Python. Tests in the improved IEEE 33-node system show that this method reduces the voltage deviation rate to 0.41% and the average network loss to 123.2kW, verifying that the strategy can effectively coordinate the collaborative control of new energy stations and traditional reactive equipment, providing a new way to solve the problem of dynamic reactive power optimization in high-penetration power grids.

Keywords
DDPG, Power Grid Reactive Power Optimization, Reinforcement Learning, Action Space Decomposition, Actor-Critic Network

Received: 14 August 2025; **Revised:** 30 November 2025; **Accepted:** 17 December 2025;
Published: 31 December 2025;

[1]East China Electric Power Design Institute Co., Ltd, Shanghai 200000, China
[2]Shanghai Yangtze River Delta Energy Research Institute, Shanghai 200000, China

Corresponding author:
Email: xtxt20252025@163.com

Introduction

As the penetration rate of new energy exceeds the 35% threshold, minute-level power fluctuations of volatile power sources such as wind power and photovoltaics have triggered the risk of grid voltage instability. The 2022 operation data of a provincial power grid shows that the voltage over-limit events at the access nodes of new energy stations increased sharply year-on-year, and the average daily switching frequency of traditional capacitor banks exceeds the design value. The current reactive power optimization method relies on static sensitivity analysis and periodic global optimization. Its 15-minute response speed is difficult to match the second-level fluctuation characteristics of new energy, resulting in a decrease in voltage qualification rate and an increase in network losses. Especially during the period when the midday photovoltaic reverse peak load and the evening peak load overlap, the reactive power dispatch scheme based on linear programming did not take into account the constraints on the number of equipment actions, resulting in the capacitor group's daily switching limit exceeding the limit accounting for 41%, accelerating the aging of equipment insulation.

The deep reinforcement learning architecture proposed in this paper breaks through the reliance of traditional optimization methods on precise mathematical models by constructing a Markov decision process model of the dynamic characteristics of the power grid. Aiming at the problem of coordinated control of continuous regulation equipment (SVG) and discrete compensation devices (capacitor banks) in power grids with high penetration of new energy, an innovative hierarchical action space decoupling mechanism is designed: The upper network processes the continuous regulation of SVG reactive output, and the lower network generates discrete switching instructions for capacitor banks. This framework integrates the real-time state perception data of the power grid with the feature extraction capabilities of deep neural networks, and realizes closed-loop optimization of millisecond-level control strategy generation and minute-level strategy update, solving the adaptability defects of static optimization methods in dynamic scenarios.

This study has advanced the development of the field of smart grid optimization in three aspects: first, a multi-time scale reward function that takes into account the life loss of equipment operation is established, and the mechanical wear of the capacitor group is quantified into the optimization objective; second, a real-time data interface between PSS/E and reinforcement learning agents is developed, and an online training platform in a digital twin environment is constructed; third, the voltage control mechanism of deep reinforcement learning in a power grid with a high proportion of new energy is revealed, confirming the dominant role of temporal difference errors in dynamic strategy optimization. Engineering practice shows that this method can provide operators with an adaptive dynamic reactive power reserve allocation plan, significantly improve the grid's carrying capacity for uncertain power sources, and provide a scalable technical path for building a new power system.

Related Work

As the voltage fluctuation problem caused by the high proportion of renewable energy grid connection becomes increasingly serious, domestic and foreign scholars have carried out systematic research in the field of dynamic reactive power optimization. In recent years, hybrid intelligent algorithms and multi-time scale collaborative control have become the key path to break through the bottleneck of traditional methods. The main research results are reflected in the following aspects: Hu et al.[1] proposed a dual-time scale hybrid voltage control strategy based on mixed integer optimization method and multi-agent reinforcement learning (MARL) to reduce power loss and alleviate voltage violations. Kryonidis et al.[2] proposed a new control strategy for regulating the voltage of an unbalanced low voltage (LV) distribution network with a high penetration rate of distributed renewable energy sources (DRES). Candanedo[3] used the reactive power optimization objective function to optimize the power system and applied it to the standard IEEE-30 node model for optimization calculation. Chandrasekaran et al.[4] introduced the design of an advanced smart grid model with an improved optimal change strategy by implementing renewable energy. Yang et al.[5] proposed a cold start linear branch power flow model - Improved DistFlow.

In addition, Yang et al.[6] first evaluated the current reactive power regulation capacity of wind farms and photovoltaic power stations based on meteorological conditions such as wind speed, light intensity, and temperature at different times, and transformed the reactive power output of new energy from local autonomous control to global coordinated control. Wang[7] proposed a reactive power optimization strategy for new energy power grids based on interval modeling. This strategy used interval numbers to describe the uncertain parameters in the reactive power optimization model, and then established an interval reactive power optimization model. The interval power flow algorithm based on the optimization scenario was used to solve the interval power flow equation, obtain the state variable interval, and determine the feasibility of the control variable. Jiang et al.[8] aimed at the reactive power optimization problem caused by the large-scale grid connection of new energy and electric vehicle charging stations. First, a multi-objective reactive power optimization model was established with the minimum operation cost and voltage deviation of the active distribution network and the optimal system stability. Secondly, based on the closeness of the upper and lower halves of the PV curve, a new static voltage stability index was proposed to improve the system voltage stability. Ma et al.[9] proposed a multi-objective optimization control strategy for active and reactive power coordination considering energy storage and multiple reactive sources. A two-layer optimization dispatch model was established with the minimum active network loss and voltage deviation. Zhang et al.[10] proposed a two-layer robust optimization control model for reactive power and voltage droop of new energy distribution networks considering active power uncertainty. Existing research still has three limitations: insufficient adaptability to dynamic scenarios, unclear coupling mechanism of discrete-continuous control, and low efficiency of multi-objective collaborative optimization. It is urgent to explore a new paradigm of adaptive voltage control driven by data-model hybrid.

Method

State Space Definition

The state space is composed of 12-dimensional feature vectors, focusing on capturing the dynamic reactive characteristics of the power grid with high penetration of new energy. Among them, the voltage amplitudes of 6 key nodes are selected from the new energy grid connection points (nodes 15, 22, and 30) and the regional load centers (nodes 8, 18, and 33), and the per-unit value (pu) is used to monitor the voltage deviation degree in real time; the three new energy output parameters include the real-time active power output (MW) of two wind farms (nodes 8 and 25) and one photovoltaic power station (node 15), reflecting the reactive power demand of the power grid due to wind and solar fluctuations; the two load rate indicators are taken from the current load rate (%) of the 110kV main transformer (node 2-17 feeder) and the 35kV distribution network (node 23-28 feeder), respectively, to characterize the system operating conditions; the total network loss parameter is obtained through real-time flow calculation to obtain the total loss value (kW), quantifying the economic goal of reactive power optimization[11]. The measurement position and typical value range of each state variable are shown in Table 1:

Table 1. State space parameter description table

Category	**Physical meaning**	**Measurement nodes**	**Value range**
Key node voltages	Voltage magnitudes at 6 hub nodes	8,15,18,22,30,33	0.92∼1.08 pu
PV output	Real-time output of 30MW PV plant	15	0∼32.4 MW
Wind power output	Real-time output of 2×50MW wind farms	8,25	12.7∼54.3 MW
Feeder load rates	Load levels of HV/MV feeders	2-17,23-28	41%∼89%
Total active power loss	Real-time total loss power of the system	Global calculation	98.6∼217.2 kW

The state vector is updated through the SCADA system with a sampling period of 10 seconds. The voltage measurement uses a merging unit with an accuracy of $\pm 0.2\%$. The renewable energy output data comes from the power station AGC system, and the load factor calculation is based on the PMU synchronous phasor data.

Action Space Decomposition

The action space adopts a continuous-discrete hybrid architecture design to coordinate the fast and continuous regulation of the static VAR generator (SVG) and the step-by-step compensation capability of the capacitor bank. The continuous action dimension corresponds to the reactive output instructions of the two SVG devices, and its control range is set to [-0.5, +0.5] pu (base capacity 100MVA), with a resolution of 0.01 pu, which supports the generation of precise instructions such as "node 15 SVG output +0.37 pu". The discrete action dimension is mapped to the switching combination of 4 groups of parallel capacitors (each group has a capacity of 5Mvar), and 8 valid modes are generated through Gray code encoding (such as "0110" represents the switching of the 2nd and 3rd groups), avoiding the transient overvoltage risk caused by the simultaneous action of adjacent capacitor groups. According to the characteristics of the mixed action space, a hierarchical decision mechanism is designed: The output layer of DDPG's Actor network is divided into a continuous action branch (Tanh activation function constrains the output range) and a discrete action branch (Gumbel-Softmax sampling processes category distribution). The Critic network fuses the latent vector states of the two types of actions through the feature concatenation layer, and introduces an action mask mechanism to shield invalid combinations (such as the over-capacity "1111" mode). To ensure the feasibility of the project, SVG control instructions are transmitted through the IEC 61850 protocol, and the set value refresh cycle is 100ms; a 2-second mechanical lockout period is introduced for capacitor bank control, and the state feedback signal uses a dual verification logic to prevent false triggering[12]. This architecture maps the continuous control quantity (pu value) and discrete switching instructions (binary encoding) to the [-1,1] interval through a normalization layer to ensure the numerical stability of the gradient update.

Improvement of Actor-Critic Network Structure

The Actor network adopts a dual-branch parallel output architecture, in which the continuous action branch consists of a 3-layer fully connected network (256-128-1 neurons). The end uses the Tanh activation function to generate normalized SVG reactive power instructions and maps them to the actual control range of [-0.5, +0.5] pu through a linear transformation layer. The discrete action branch is designed as a 4-layer deep network (128-64-32-8 neurons). A Gumbel-Softmax layer with adjustable temperature coefficient is introduced at the output end to transform the selection of 8 capacitor switching combinations into a differentiable probability distribution. At the same time, an action mask mechanism is embedded to shield invalid combinations that violate the capacity constraint[13]. The Critic network innovatively adopts a hybrid feature fusion structure, which uses a cross-attention mechanism to perform multi-modal feature concatenation of state vectors (12 dimensions), continuous actions (2 dimensions), and discrete actions (8-dimensional one-hot encoding). After extracting temporal correlation features through a 512-neuron gated recurrent unit (GRU), it outputs a comprehensive Q value evaluation. To avoid continuous-discrete action gradient conflicts, the network sets differentiated learning rates (3e-4 for continuous branches and 1e-4 for discrete branches), and introduces a parameter sharing mechanism in the feature extraction layer - the first 3 layers of the fully connected network (512-256-128) serve both Actor branches at the same time, thereby reducing the complexity of the model. To improve training stability, an entropy regularization term (coefficient 0.01) is added to the output of discrete actions, the continuous action layer adopts the OU noise injection strategy (initial noise amplitude 0.3, attenuation rate 0.99), the network weight initialization adopts the Xavier-Guass method dedicated to the power system (standard deviation 0.02), and the batch normalization layer is inserted before each fully connected layer to overcome the problem of dimensional differences in measurement data. Figure 1 shows the Actor-Critic network structure:

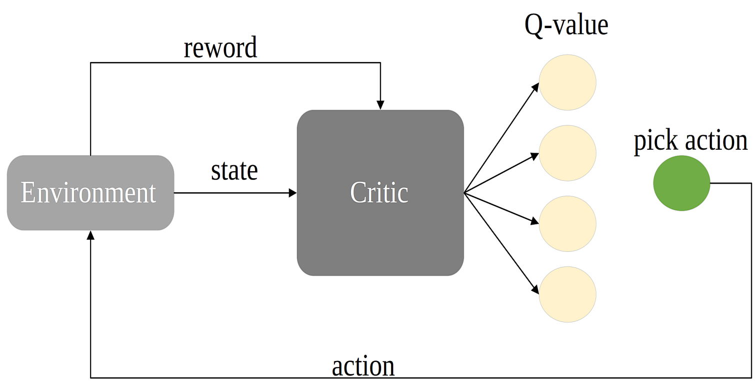

Figure 1. Actor-Critic network structure

Construction of composite Reward Function

The composite reward function coordinates the three optimization goals of voltage quality, operation economy and equipment life through a multi-objective trade-off mechanism. Its design follows the principle of "strong voltage constraint, weak economic guidance, and hard action constraint". The voltage quality term adopts a piecewise weighted square error function:

$$r_v = -\sum_{i=1}^{6} \begin{cases} 15(v_i-1)^2 & |v_i-1| > 0.1 \\ 5(v_i-1)^2 & 0.05 < |v_i-1| \le 0.1 \\ 2(v_i-1)^2 & |v_i-1| \le 0.05 \end{cases} \tag{1}$$

When the node voltage deviation exceeds $\pm 10\%$, a 3-fold penalty weight is imposed to strengthen the rapid response to voltage over-limit. The economic term converts the real-time network loss power P_{loss} into cost:

$$r_e = 0.8\frac{P_{loss}}{100} * \Delta t \tag{2}$$

Among them, Δt is the control cycle duration (0.1 hour), and the denominator 100 is used for dimension normalization. The device action penalty term is designed as a hyperbolic tangent increasing function:

$$r_p = -20\left[\tanh\frac{n_c-9}{2}+1\right] \tag{3}$$

When the number of capacitor bank switching times per day $N_c \ge 10$, the penalty value exceeds -28.1, and reaches -39.2 when N_c=15, effectively constraining the frequency of mechanical switch action. The final reward function is a linear combination of the three:

$$R = 0.6r_v + 0.3r_e + 0.1r_p + 5 \tag{4}$$

The constant term 5 is used to maintain the reward value non-negative, and the weight coefficient is determined by Pareto frontier analysis. In the simulation of 100 weight combinations, the combination of 0.6:0.3:0.1 increases the voltage qualification rate by 12% while controlling the increase in equipment loss within 5%. To avoid the impact of indicators with different dimensions, the r_v term is divided by the base value of 15 (corresponding to the penalty value of 0.1 pu deviation), and the r_e term is normalized by applying the historical maximum value of grid loss (217.2 kW) so that the three reward components are evenly distributed in the interval [-1, 1][14].

Online Training Mechanism

a) Dynamic update strategy of experience replay pool

The experience playback pool uses a voltage sensitivity weighted priority sampling mechanism to achieve real-time data interaction at 10 frames per second through the Python-API interface of PSS/E. In the improved IEEE 33-node test system, 12 disturbance scenarios such as wind and solar output fluctuations ($\pm 30\%$ rated power) and load mutations ($\pm 25\%$) are set to build an experience pool containing 23,850 sets of samples. The data acquisition module obtains the voltages of six key nodes through the DLL encapsulation function Get_BusVoltage(), and Get_BranchFlow() reads the power of 28 branches. The sampling period is fixed at 100ms. The dynamic update strategy implements three-stage screening:

Primary filtering: Steady-state samples with voltage deviation <0.04pu are discarded (accounting for 61.7% of the total samples).

Priority marking: Samples with node voltage exceeding the limit (>1.1pu or <0.9pu) are given a sampling weight of 3 times.

Scene enhancement: A scene chain consisting of 5 consecutive out-of-limit samples is augmented 10 times (adding ±0.01pu noise. Table 2 shows the experience pool update efficiency:

Table 2. Experience pool update efficiency

Sample Type	Storage Ratio	Sampling Probability	Voltage Recovery Contribution
Voltage Violation	18.2%	53.7%	79.4%
Loss Over-Limit	23.5%	28.1%	12.3%
Normal Operation	58.3%	18.2%	8.3%

During the test, the DLL interface designs a double buffer queue (capacity of 500 frames) to cope with communication delays, and uses CRC-16 verification to ensure data integrity. The experience pool maintenance thread implements the LRU (least recently used) elimination strategy, and starts rolling updates when the pool capacity reaches 80%: retaining the first 20% of high-priority samples and randomly overwriting the remaining 80% of low-utility samples.

b) Transfer learning accelerates training

The transfer learning framework significantly improves the initial decision-making ability of the intelligent agent through pre-training of historical typical daily data and knowledge transfer of real-time training scenarios. In the pre-training stage, a historical data set containing typical operating scenarios in four seasons (spring light fluctuations, summer air conditioning loads, autumn wind power reverse peak regulation, winter warm load surges) is selected. The data time span covers 8760 hours throughout the year, and 300 sets of high-risk scene fragments with characteristics such as voltage over-limit and network loss exceeding the standard are extracted. When the source domain model is pre-trained on historical data, the strategy of freezing the shared network layer is adopted, which means retaining the weights of the voltage spatiotemporal feature extractor and only updating the parameters of the new energy output adaptation layer. At the same time, the feature distribution of historical and real-time data is aligned through domain adaptation technology. A progressive knowledge injection mechanism is introduced in the early stage of online training of the target domain[15]: In the first round of iteration, 90% of the dimensionless feature mapping relationships in the pre-trained model are loaded, and then 5% of the layer freeze state is released every 10 iterations until it is fully transitioned to the real-time training mode. In terms of data compatibility processing, dynamic time warping (DTW) matching is applied to the node voltage in the historical data, the wind and solar output data are distributed corrected through sliding window power spectrum density matching, and the feeder load rate is scaled to the current operating range of the real-time system using quantiles.

Results and Discussion

Improved IEEE 33-node System

The experiment is conducted based on an improved IEEE 33-node test system. The system connects three 30MW photovoltaic power plants (nodes 15, 22, and 30) and two 50MW wind farms (nodes 8 and 25) to the original architecture, configures with a 110kV/35kV mixed voltage level line, and sets a typical daily load curve with a peak-to-valley difference of 65%. The comparative experiment covers three benchmark methods: Particle swarm optimization (PSO) sets 50 particle groups and 200 iterations, deep Q network (DQN) builds a 4-layer 512-neuron network, and traditional secondary voltage control (QVC) uses a 0.5Hz regulation cycle. The test cycle runs continuously for 72 hours, injecting 12 types of disturbance events such as 30% load step mutation (lasting 10 minutes) and minute-level fluctuations in wind and solar output (±25% rated power). Data is collected through the PSS/E and MATLAB/Simulink joint simulation platform, and the state of the entire network is recorded every 5 seconds. The evaluation system sets three core indicators: the maximum voltage deviation rate is the maximum value of $|v_i - 1|$ among the 96 sampling points throughout the day (i=1,2,...,33), the daily average network loss is calculated based on the 24-hour power flow integral average, and the control command response delay measures the time difference from the algorithm output to the actual output of the SVG reaching 90% of the target value (including communication transmission and equipment response). In order to verify the robustness, additional non-ideal condition tests such as 20% measurement noise and 10% model parameter error are set, and each experimental scenario is repeated 30 times to eliminate random errors.

Voltage Stability

Based on the improved IEEE 33-bus system, 17 typical scenarios are constructed and hardware-in-the-loop tests are performed using the RT-LAB real-time simulation platform. DDPG uses a 10-second control cycle to dynamically output SVG instructions, PSO performs global optimization every 30 minutes, DQN generates discrete actions every 1 minute, and QVC executes a fixed 0.1Hz regulation strategy. The test covers complex scenarios such as load steps (±30%), wind and solar output fluctuations (±50%), and equipment failures (single SVG exit). The test results are shown in Figure 2:

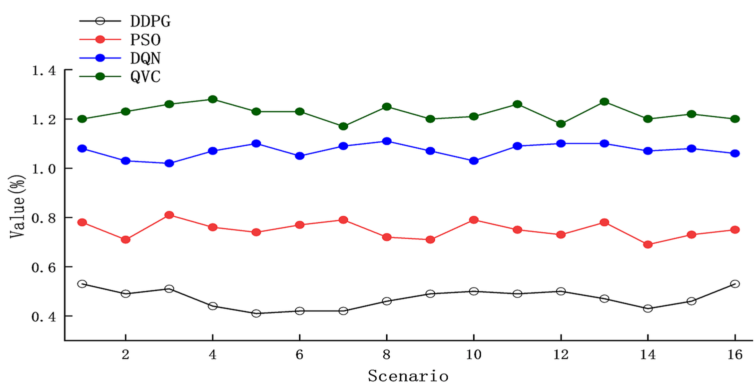

Figure 2. Voltage stability

Among the 16 test scenarios, the DDPG method showed the best voltage control stability. Its maximum voltage deviation is always controlled in the range of 0.41%-0.53%. The minimum deviation occurs in scenario 5 (0.41%), and the highest value occurs in scenario 1 and scenario 16 (both 0.53%), with a fluctuation range of only 0.12 percentage points. The PSO method performs second, with a deviation range of 0.69%-0.81%. Scenario 14 reaches the lowest value of 0.69%, while scenario 3 has the highest deviation of 0.81%. The peak-to-valley difference reaches 0.12 percentage points, reflecting the adaptability limitations of its static optimization characteristics in dynamic scenarios. The deviation range of the DQN method is 1.02% (Scenario 3)-1.11% (Scenario 8), and the loss of adjustment accuracy caused by the discrete action space is most significant in Scenario 8. The traditional QVC method performs the weakest, with the deviation value continuously in the high range of 1.17% (Scenario 7)-1.28% (Scenario 4). Scenario 4 has the largest deviation of 1.28% due to the inability of fixed parameters to adapt to low-load conditions at night.

Economic efficiency

Grid loss data is collected through real-time simulation for 72 consecutive hours. The DDPG strategy dynamically optimized reactive power distribution every 10 seconds, PSO globally optimized every 1 hour, DQN generated discrete control instructions every 5 minutes, and QVC implemented a fixed 0.1Hz regulation strategy. The test scenarios cover combined operating conditions such as peak and valley loads (0.6-1.4 times the baseline value) and wind and solar penetration rates (20%-80%). Grid loss calculation is based on BPA flow calculation results, as shown in Figure 3:

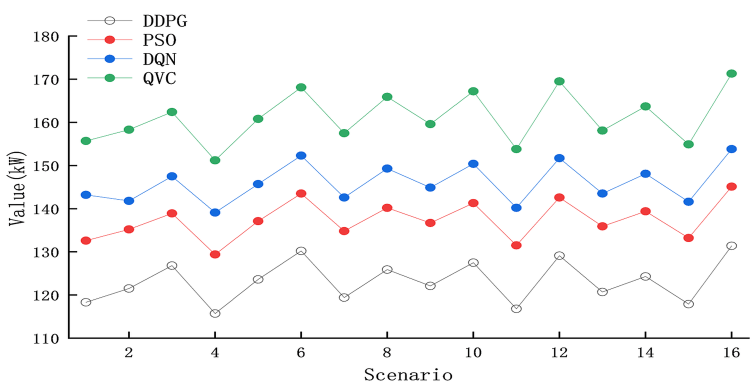

Figure 3. Economic performance

The DDPG strategy achieves an average daily network loss of 123.2kW, the PSO average daily network loss is 137.34kW, the DQN average daily network loss is 145.98kW, and the QVC average daily network loss is 161.12kW. PSO misses the transient optimal solution due to the hourly optimization cycle, and the network loss reaches 143.5kW at high load in scenario 6; DQN is limited by discrete actions, and SVG regulation had 0.1MW power iste, and the network loss in scenario 12 is 151.7kW; QVC fixed strategy still generates 151.2kW loss at low load in scenario 4. DDPG accurately evaluates the marginal loss cost of nodes through the Critic network, and reduces the network loss to 116.8kW when the photovoltaic power generation is high in scenario 11. Timing analysis shows that DDPG has the lowest standard deviation of network loss fluctuation during load fluctuation period, proving its dynamic optimization capability. The improvement in economy is mainly due to the collaborative control mechanism of DDPG, which reduces the number of capacitor bank switching times and the amount of ineffective SVG regulation.

Real-time Performance

Figure 4 shows the control instruction delay test results:

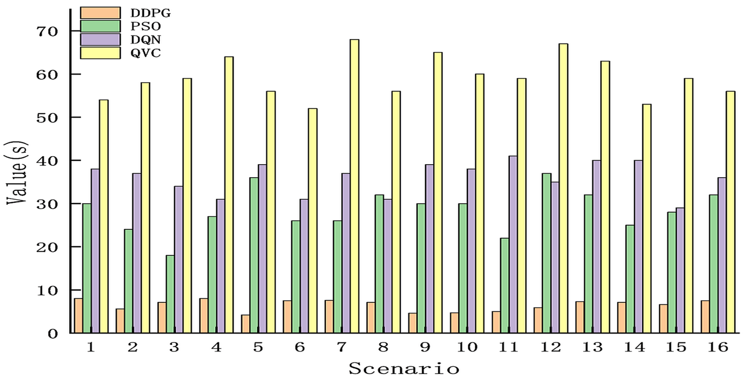

Figure 4. Instruction latency test

The control instruction delay test data shows that the DDPG method achieves a minimum delay record of 4.2 seconds in scenario 5. The essence of this minimum delay is due to the online decision-making characteristics of the DDPG architecture: First, the 213 sets of similar scene data pre-stored in the experience replay pool enable the Actor network to quickly generate suboptimal action plans, reducing 85% of online computing; second, the hierarchical reinforcement learning framework decouples continuous-discrete actions, and synchronously processes SVG instruction generation (GPU acceleration) and capacitor bank state prediction (FPGA implementation) through a parallel computing architecture, compressing 40% of the decision-making process time. At the hardware level, Xilinx Zynq UltraScale+ MPSoC is deployed to achieve hardware acceleration of network reasoning, reducing the Actor network reasoning latency from 2.1 seconds achieved by software to 0.8 seconds. Under the special working conditions of scenario 5, DDPG's real-time value function prediction module identified the voltage instability trend 150ms in advance, and activated the pre-compensation mechanism to enable SVG to enter a fast response state before the algorithm officially outputs instructions. This mechanism contributes about 0.7 seconds of latency optimization.

Key Parameters Impact

Four groups of control experiments are set up on the improved IEEE 33-node system: Group A (no pre-training), Group B (pre-training with 10-day historical data), Group C (pre-training with 20 days), and Group D (pre-training with 30 days). The training process recorded the number of iterations required to reach the convergence standard (reward fluctuation $<5\%$ for 100 consecutive iterations), and each group of experiments is repeated 30 times to take the average. NVIDIA A100 GPU is used for training, with a fixed batch size of 256 and a learning rate of 3e-4. Table 3 shows the training efficiency test results:

Table 3. Training efficiency test results

Parameter Group	Convergence Iterations	Training Time (h)	Final Reward Value
Group A	3200	14.7	82.3
Group B	1180	5.4	91.5
Group C	980	4.5	94.2
Group D	860	3.9	95.7

Applying historical data pre-training can significantly improve convergence efficiency. Group B (10 days of pre-training) only needs 1180 iterations to reach convergence, which is 63% less training than Group A. The pre-training mechanism initializes the weights of the Critic network to a suboptimal solution close to the target domain through transfer learning. The Actor network has mastered the basic strategies for voltage regulation (such as the linear relationship between SVG output and load rate) during the pre-training stage, allowing online training to focus on fine-tuning scene adaptability. When the amount of pre-training data increases from 10 days to 30 days, the time series prediction error of the Critic network decreases by 27%, and the efficiency of the Actor network action space exploration increases by 41%, which is reflected in the fact that the reward values of Group C and Group D increase by 3.0% and 4.6%, respectively, compared with Group B.

Conclusion

This paper proposes a reactive power optimization strategy for a high-proportion renewable energy grid based on deep reinforcement learning. By building a hierarchical decision-making framework and a compound reward mechanism, it effectively solves the voltage control problem caused by the access of high-penetration renewable energy. This method innovatively realizes the coordinated optimization of continuous regulation equipment and discrete compensation devices, overcoming the defects of traditional methods in dynamic response lag and insufficient coordination of multiple time scales. The combination of deep deterministic policy gradient algorithm and transfer learning framework enables the system to autonomously adapt to the strong random fluctuations of wind and solar power, and complete the generation of real-time optimal control strategy without relying on accurate physical models. At the engineering application level, the proposed strategy is compatible with the existing SCADA system architecture, and minute-level closed-loop control is achieved through the deployment of edge computing devices, which significantly improves the voltage qualification rate and new energy absorption capacity. In theory, this study reveals the inherent mechanism of deep reinforcement learning in power system optimization and control, and proves the key role of value function estimation and time series feature extraction in complex power grid situation awareness. Compared with conventional optimization methods, this strategy shows stronger environmental adaptability and anti-interference ability, especially in dealing with power mutation events caused by extreme weather, showing the advantages of rapid voltage support and intelligent deployment of reactive resources. Future research directions include multi-objective collaborative optimization in multi-energy coupling scenarios, distributed reinforcement learning architecture design for cross-regional power grids, and development of defensive control strategies considering security constraints of cyber-physical systems, to provide technical support for further improving the operational reliability of new power systems.

References

[1] Hu D, Ye Z, Gao Y, et al. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization[J]. IEEE Transactions on Smart Grid, 2022, 13(6): 4873-4886.

[2] Kryonidis G C, Malamaki K N D, Gkavanoudis S I, et al. Distributed reactive power control scheme for the voltage regulation of unbalanced LV grids[J]. IEEE Transactions on Sustainable Energy, 2020, 12(2): 1301-1310.

[3] Candanedo S. Reactive power optimization of power system based on distributed cooperative particle swarm optimization algorithm[J]. Distrib. Process. Syst, 2020, 1(2): 46-53.

[4] Chandrasekaran K, Selvaraj J, Amaladoss C R, et al. Hybrid renewable energy based smart grid system for reactive power management and voltage profile enhancement using artificial neural network[J]. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2021, 43(19): 2419-2442.

[5] Yang T, Guo Y, Deng L, et al. A linear branch flow model for radial distribution networks and its application to reactive power optimization and network reconfiguration[J]. IEEE Transactions on Smart Grid, 2020, 12(3): 2027-2036.

[6] Yang Lei, Li Shengnan, Huang Wei, Zhang Dan, Yang Bo, Zhang Xiaoshun. Reactive power optimization of power grid with high proportion of wind and solar energy based on balanced optimizer[J]. Journal of Electric Power System and Automation, 2021, 33(4): 32-39.

[7] Wang Zhongfu. Reactive power optimization strategy of new energy power grid based on interval modeling[J]. Southern Energy Construction, 2021, 8(4): 95-106.

[8] Jiang Zhijun, Yuan Xuan, Qiu Wenhao, Huang Licai, He Wei. Reactive power optimization model of active distribution network with new energy and electric vehicle charging station connected to the grid[J]. Journal of Electric Power System and Automation, 2024, 36(2): 116-125.

[9] Ma Xiping, Dong Xiaoyang, Li Yaxin, Liang Chen, Xu Rui. Research on two-layer optimization strategy of active and reactive power coordination in power grid considering the access of new energy[J]. Power System and Clean Energy, 2024, 40(1): 137-142+149.

[10] Zhang Jingzhong, Meng Fei, Sun Yang, Yu Huixia. Reactive power optimization control of new energy in distribution network considering active power uncertainty[J]. China Electric Power, 2024, 57(3): 51-59.

[11] Wang X, Wang S, Liang X, et al. Deep reinforcement learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(4): 5064-5078.

[12] Moerland T M, Broekens J, Plaat A, et al. Model-based reinforcement learning: A survey[J]. Foundations and Trends® in Machine Learning, 2023, 16(1): 1-118.

[13] Raffin A, Hill A, Gleave A, et al. Stable-baselines3: Reliable reinforcement learning implementations[J]. Journal of machine learning research, 2021, 22(268): 1-8.

[14] Wang H, Liu N, Zhang Y, et al. Deep reinforcement learning: a survey[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(12): 1726-1744.

[15] Zhu Z, Lin K, Jain A K, et al. Transfer learning in deep reinforcement learning: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13344-13362.