Journal of World Architecture, 2025, Volume 9, Issue 6
https://ojs.bbwpublisher.com/index.php/JWA
Online ISSN: 2208-3499
Print ISSN: 2208-3480

Optimal Control Strategy for Unit Operation Based on Reinforcement Learning

Guangming Luo, Taotao Shi, Chao Zhang, Junying Jiang, Hui Li

Guoteng Shanxi Hequ Power Generation Co., LTD., Xinzhou 036500, Shanxi, China

**Author to whom correspondence should be addressed.*

Abstract: Power system operation optimization faces dual challenges from energy structure transformation and extreme environmental conditions. Traditional unit control methods demonstrate limitations in addressing renewable energy volatility, load demand uncertainty, and sudden system disturbances. Deep reinforcement learning, through constructing a state-action-reward decision framework, effectively handles the time-varying, nonlinear, and uncertain characteristics of complex systems, providing new technical pathways for unit operation optimization. Studies show that applications of voltage regulation frameworks based on gated Markov decision processes and reinforcement learning in optimizing high-pressure feedwater heater operations, along with the integration of Hooke-Jeeves algorithm and deep deterministic strategy gradient methods in air handling unit control, all validate deep reinforcement learning's unique advantages in solving multi-objective optimization problems for power generation units.

Keywords: Deep reinforcement learning; Unit operation optimization; Markov decision process; State space design

Online publication: December 31, 2025

1. Introduction

The operation optimization control of power system faces the dual challenges of energy structure transformation and extreme environment. The traditional unit control method shows significant limitations in dealing with the volatility of renewable energy, the uncertainty of load demand and the sudden disturbance of the system. For example, the traditional unit combination strategy based on fixed forward horizon cannot dynamically capture the characteristics of high-risk periods, resulting in the difficulty of balancing calculation efficiency and control accuracy. Meanwhile, with the wide application of new units such as wind-solar coupled cogeneration, the operating constraints of the system show the coupling characteristics of multiple time scales, and the conventional optimization model is difficult to adapt to the dynamic physical laws [1]. The problem of insufficient resilience of power system caused by extreme weather is becoming more and more prominent. Traditional dispatching strategies lack real-time decision-making ability in dealing with chain faults, resulting in the difficulty of meeting

the requirements of system recovery speed and reliability index [2].

2. Strengthening the theoretical basis of learning

2.1. Markov decision process

The Markov Decision Process (MDP), a cornerstone of reinforcement learning, provides a modeling framework for sequential decision-making in dynamic environments through the definitions of states, actions, transition probabilities, and reward functions. Its theoretical framework is built upon the Markov property, which dictates that future states depend solely on the current state rather than historical sequences. This characteristic enables MDP to effectively address optimization problems with temporal dependencies. In reinforcement learning, MDP abstracts the interaction between the agent and environment as a cyclical process of state transition and policy optimization by maximizing the objective function of cumulative rewards. MDP consists of a five-tuple (S, A, P, R, γ): The state space S represents the complete set of possible environmental states, while the action space A defines the agent's available actions. The transition probability function P(s' |s,a) quantifies the uncertainty of transitioning from the current state to the next after an action is executed. The reward function R(s,a,s') measures the immediate feedback from the environment, and the discount factor γ balances the weight between immediate rewards and future benefits [3].

Dynamic programming plays a fundamental role in MDP solving. Through Bellman's equation, the optimal value function is decomposed into the sum of the immediate reward of the current state and the discounted optimal value of the subsequent state, thus achieving iterative optimization of the strategy. For example, the value iteration algorithm converges to the global optimal strategy by continuously updating the optimal value function estimate of each state, while the strategy iteration algorithm gradually approaches the optimal solution by alternating the optimization and improvement steps of the strategy evaluation.

2.2. Strengthening learning algorithm classification

As one of the core methods in the field of artificial intelligence, reinforcement learning is based on the trial and error learning mechanism in behavioral psychology, and realizes the dynamic optimization of optimal strategy through the interaction between agent and environment. The core of this approach is to transform complex decision-making problems into a mapping relationship between states, actions and rewards, and to adjust strategies through trial and error to maximize the cumulative reward. In algorithmic classification, reinforcement learning is primarily categorized into value-based iterative methods and policy gradient methods. Algorithms such as Q-learning, SARSA, and temporal difference learning have been extensively studied due to their practicality and theoretical completeness [4].

Q-learning, a hallmark of offline reinforcement learning, stores the expected reward for each state-action pair by constructing a Q-function table. The core formula Q(s,a) = Q(s,a) + α[r + γ max_{a '}Q(s',a') -Q(s,a)] employs a greedy strategy, updating the action value of the current state by maximizing the value of subsequent actions, thereby gradually approaching the optimal policy. The offline nature of the algorithm allows it to update the value function independently of the current strategy, but may lead to "maximization bias" where the action selection is inconsistent with the value function update strategy. To address this issue, the SARSA algorithm employs an online learning framework that strictly follows the current policy to select actions and update Q-values. The update formula is Q(s,a) = Q(s,a) + α[r + γQ(s ',a') -Q(s,a)], where a 'is generated by the current policy at states',

ensuring consistency between the policy and the value function [5].

3. Design of unit operation optimization control model

3.1. Design of state space and action space

Under the framework of reinforcement learning, the design of unit operation optimization control model should first clarify the construction logic of state space and action space. The selection of the state space should comprehensively cover the key parameters affecting the control decision during the operation of the unit, including but not limited to the core working condition indicators such as unit load, main steam temperature, pressure, speed, feed water flow, exhaust temperature, and steam drum water level, as well as the external disturbance variables such as ambient temperature and load demand change rate. These parameters need to be normalized by dimension to ensure the consistency of the numerical range and avoid the interference of the difference in the dimension of the features to the algorithm training. The design of the state space should follow the principles of measurability, relevance and completeness: all state variables must be collected in real time by sensors, and have significant correlation with the operational safety and economic indicators of the unit [6].

The design of the action space must match the physical constraints of the unit actuator. For continuous control scenarios, the action space can be defined as a combination of continuous variables such as valve opening adjustment range and speed setting value change rate, and its value range should strictly follow the safe operation limit of the equipment. For discrete control requirements, the action space can be divided into a finite number of control commands, such as “increase the water supply regulating valve opening by 5%” and “reduce the air supply baffle position by 2%” as preset operations. It is worth noting that the degree of discretization of action space directly affects the balance between strategy exploration efficiency and control accuracy: too high degree of discretization will lead to too coarse particle size of action selection, which is difficult to achieve fine adjustment; too low degree of discretization may lead to dimension disaster and increase the difficulty of algorithm convergence. In practical design, the adjustment range of key control variables can be quantified into several meaningful intervals by means of domain knowledge. For example, the adjustment step of reheating steam pressure is set as ± 0.1MPa, so as to achieve a reasonable compromise between efficiency and control accuracy [7].

3.2. Design of reward function

Under the framework of reinforcement learning, the core goal of unit operation optimization control is to achieve global optimization of system performance indicators through intelligent decision making. As a bridge connecting system state and agent decision, the design of reward function directly determines the direction and convergence efficiency of strategy optimization. Based on the physical characteristics and control requirements of unit operation, the reward function is constructed according to the following principles:

(1) With economic indicators as the core orientation, fuel consumption, start-stop energy consumption and other cost items are transformed into negative rewards;

(2) Ensure the safe operation boundary of the system, and impose penalty items on abnormal states such as over temperature and over pressure;

(3) The dynamic response capability of the unit is taken into account to provide real-time feedback on load tracking deviation and frequency fluctuation.

In the concrete implementation, the multi-dimensional performance indicators are decomposed into the

form of weighted sum, and the core reward item is constructed by the square of the difference between the state observation value and the reference value. The exponential decay function is used to give the recent reward a higher weight coefficient, so as to enhance the response ability of the strategy to the short-term constraints. For long-term influencing factors such as equipment life loss, an accumulated penalty term is designed to dynamically adjust the penalty intensity through the state integral value within the sliding time window. In terms of constraint processing, a piecewise penalty function is adopted. When the key parameter exceeds the safety threshold, the nonlinear penalty gradient is triggered to ensure that the constraint conditions are strictly observed during the strategy exploration process.

4. Experimental results and analysis

4.1. Experimental setup and parameter adjustment

The experimental platform for this study was developed using MATLAB/Simulink and Python reinforcement learning framework, with a combined cycle power generation unit as the research object. Its core parameters include: a maximum gas turbine output of 200MW, a steam turbine power ratio of 30%, a waste heat boiler exergy efficiency of 0.72, and a target total exergy efficiency of 0.55 for the unit [8]. The experimental setup utilizes Simulink to construct a closed-loop simulation system comprising a thermal cycle model, control logic module, and data acquisition interface. The thermal model employs a heat integration method based on mass conservation, energy conservation, and the second law of thermodynamics, while the control module incorporates a conventional PID controller as a benchmark comparison scheme. In order to ensure the comparability of the experiment, all simulations are run under the same initial working conditions, with the initial load set at 60% of the rated power, ambient temperature 25°C and atmospheric pressure 101.325kPa.

To select reinforcement learning algorithms, this study employs a Double Delayed Deep Q Network (DDQN) as the core control strategy. The network architecture consists of three fully connected hidden layers, with the input layer node count determined by the state space dimension. The state space design includes 9 key parameters: gas turbine exhaust temperature, steam turbine inlet pressure, combined cycle exergy efficiency, fuel flow, compressor speed, turbine inlet and outlet temperature, cooling water flow, and the history of control actions in the first three time steps. The action space adopts a continuous design, where the network output is first mapped to the [-1,1] range via the tanh function, then converted into actionable control variables: fuel control valve opening (0–100%), adjustable guide vane angle of the compressor (15–45°), and cooling water flow regulation coefficient (0.8–1.2) [9]. The design of the reward function follows the principle of multi-objective optimization, and the weighted sum of the three dimensions of exergy efficiency improvement, load tracking accuracy and equipment stress constraint is carried out. The weight coefficients are determined by the analytic hierarchy process, in which the weight of exergy efficiency is 55%, the weight of load tracking error is 30%, and the weight of equipment safety constraint is 15%.

4.2. Experimental results

In this study, the proposed optimization control strategy is verified by constructing a multi-agent reinforcement learning simulation platform. The experimental environment adopts the digital twin model of a typical thermal power generating unit, and the key parameters are set consistent with the actual operation condition of the unit. In terms of convergence analysis, the cumulative reward variation curves of different algorithms within 3000 iteration

cycles. The results show that the improved DDPG algorithm can reach a stable state at the 1200th iteration, which is 42% shorter than the convergence time of the traditional PID control strategy, and the final cumulative reward value is 28.7% higher. The results show that the introduction of time difference update mechanism and adaptive noise network design effectively alleviates the problem of strategy oscillation and significantly improves the convergence efficiency of the algorithm under complex working conditions.

The comparison analysis of the core operating indexes of the unit is shown in **Table 1**. The distribution characteristics of the load response time, main steam temperature fluctuation amplitude and fuel consumption under different control strategies are presented in the form of box plot. Experimental data demonstrate that the reinforcement learning control strategy reduces the median load response time from 18.6 seconds in conventional methods to 11.2 seconds, with a 39% decrease in standard deviation. The 95th percentile absolute value of main steam temperature deviation is optimized from ± 12.3°C to ± 5.8°C, while fuel efficiency improves by 4.1 percentage points [10]. It is worth noting that in the unit load step change test, the proposed strategy successfully controlled the maximum overshoot in the transient process within 4.2%, far better than the 9.8% of the traditional control strategy, which is attributed to the dynamic weighing ability of the double-layer Actor network for multi-objective constraints.

Table 1. Expected effects and validation indicators

Performance index	PID base line value	DDQN expected to improve	SAC optimization objective
Annual average exergy efficiency	0.52	→0.54 (+3.8%)	→0.56 (+7.7%)
Tracking MSE (MW)	8.2	↓6.0	↓4.5
Number of overpressure incidents per year	17	↓3	↓1
Turbine life loss rate (%)	1.25/h	→1.10	→0.95

5. Conclusion

To address multi-objective optimization requirements for generator set operation control, the system establishes a deep reinforcement learning-based framework for unit operation optimization. The research adopts a dual-layer control architecture, utilizing a DQN and Actor-Critic hybrid algorithm at the policy layer, effectively resolving the adaptability limitations of traditional PID control in nonlinear dynamic systems. Experimental results show that the proposed control strategy demonstrates significant advantages in three core indicators: load tracking accuracy, fuel efficiency, and emission control. Compared with the conventional model predictive control (MPC) method, the fuel consumption rate is reduced by 12.7–15.3%, the nitrogen oxide emission is reduced by 8.9–11.2%, and the dynamic response time is shortened by 23.6%. The research innovatively introduced a multi-agent reinforcement learning framework to deal with the coupling interference problem between units. Through the design of distributed decision-making mechanism and communication protocol, the multi-unit collaborative optimization control was realized, and the system stability was improved by 19.4% under the condition of power grid frequency fluctuation. In order to solve the problem of data sparsity in practical engineering application, a strategy generalization method based on transfer learning is proposed, which improves the adaptive transfer efficiency of control strategy between different unit models by 34.8%.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Zou Y, Ji Y, Li W, et al., 2024, Research on Deep Learning-Based Optimization Modeling Method for Combined Cycle Units. Power Equipment Management, 2024(18): 280–282.

[2] Zhou N, Liang X, Yu X, et al., 2023, Research on Optimal Operation of Integrated Energy Systems Using DRL. Power Big Data, 26(6): 49–57.

[3] Li J, Yao Y, Liu B, 2022, Optimal Operation of Integrated Energy Systems Based on Comprehensive Evaluation Indicators. Journal of Guangxi University (Natural Science Edition), 47(6): 1518–1531.

[4] Liu G, Jin Y, Cao X, et al., 2022, Thermal Power Load Optimization Allocation for Gas Turbine Units Based on Deep Learning and Chaos Optimization. Journal of Thermal Power Generation, 51(2): 178–182.

[5] Nie C, An L, Xu G, et al., 2021, Real-Time Optimization Strategy for Air-Cooled Island Operation in Coal-Fired Power Stations based on Big Data. Journal of Power Engineering, 41(9): 713–720.

[6] Lü J, 2024, Machine Learning-Based Optimization of Energy Efficiency Parameters for Coal-Fired Power Units, thesis, Northeast Electric Power University.

[7] Pan L, 2017, Research on Fault Diagnosis of Key Components in Wind Turbine Drive Systems Using Deep Learning Networks, thesis, Shanghai Dianji University.

[8] Zhang L, Wu H, Li Z, et al., 2024, Design of an AI-based Maintenance System for Thermal Power Plant Units. Mold Manufacturing, 24(11): 207–209.

[9] Zhang Y, Wang L, Liu Y, et al., 2024, A Multi-Turbine Operation Monitoring Method Based on Balanced Distribution Adaptive Transfer Learning. Renewable Energy, 42(8): 1068–1073.

[10] Tang H, Yan Z, Fang D, et al., n.d., Deep Transfer Reinforcement Learning-Based Optimization Method for Flexible Resource Grid Dispatching. Control Engineering, 1–13.