

Dynamic Energy Dispatch Strategy for Integrated Energy System Based on Constrained Reinforcement Learning

Qingkai Sun, *Student Member, IEEE, Member, CSEE*, Xiaojun Wang, *Member, IEEE, Member, CSEE*, Zhao Liu, *Member, IEEE, Member, CSEE*, Weishan Lin, *Student Member, IEEE, Member, CSEE*, Jinghan He, *Fellow, IEEE, Fellow, CSEE*, and Wei Pei, *Member, IEEE, Member, CSEE*

Abstract—An integrated energy system includes multiple subsystems of electricity, heating, and gas. It is difficult to achieve dynamic dispatch of multi-energy flows due to challenges in establishing detailed mathematical models and the impacts of uncertain factors. Reinforcement learning has the ability to fit non-convex and non-linear problems. It avoids simplifying detailed mathematical models and complex, uncertain factors in the solution process, like traditional methods, which provides a new idea for achieving dynamic energy dispatch. However, existing research generally relies on soft constraints, such as reward function penalties, to obtain the dispatch strategy, which may violate the system's operational safety. This paper ensures the security of dynamic energy dispatch strategies by adding two constraint mechanisms on reinforcement learning. First, boundary conditions of the action security domain are established through continuous interaction and feedback between the agent and the environment. It can effectively restrict infeasible actions. Second, the strategy is constrained by the truncation function in the trust domain, and strategy update is always kept within a controllable range to improve convergence and stability of the IES dynamic energy dispatch model. Finally, simulation results indicate that the proposed method can effectively constrain agent action selections and strategy updates during the dynamic energy dispatch process. This is of great significance to ensure the safe and stable operation of the system.

Index Terms—Constrained reinforcement learning, dynamic energy dispatch, integrated energy system, security issue.

I. INTRODUCTION

THE conflict between energy demand and environmental protection is becoming increasingly severe worldwide. Therefore, establishing an integrated energy system (IES) with interconnected heterogeneous energy networks and complementary use of multiple resources is crucial to achieving low-carbon and sustainable development [1]–[3].

IES includes multiple subsystems of electricity, heating, and gas. Compared with the electric network, the infrastructures of the heating network and the natural gas network are easily prone to aging. At the same time, intermittent energy sources such as wind power and photovoltaics are widely used in the electric network. Therefore, IES's complexity and uncertainty have become more apparent. How to perform real-time, accurate, and safe, dynamic energy dispatch in this situation has become a new research direction [4], [5]. Now, the problem has been investigated using scenario analysis, robust optimization, and model predictive control. In [6], considering the uncertainty of intermittent energy sources and load in a combined heat and power microgrid, the system optimization dispatch problem based on robust optimization is studied. In [7], a two-layer robust model is constructed using an uncertainty approach to obtain a system dispatch scheme for the worst-case scenario. The above references require establishing accurate mathematical models and are mostly limited to a fixed dispatch plan. It cannot dynamically respond to the impacts of system structure changes, network infrastructures' aging, and random fluctuations in source and load. Reference [8] proposes a model-based predictive control method for dynamic dispatch of cold-heat-electricity cogeneration type microgrids, which can optimize the device output based on rolling optimization. In [9], a model predictive control method is used to optimize the output of each device for grid-connected building energy systems. Although the above research can effectively deal with uncertain factors and optimize equipment output in real-time, it still relies on fixed system parameter settings and accurate prediction of intermittent energy output. It also cannot be effectively applied to the entire life cycle of system operation.

In recent years, under widespread application of artificial intelligence technology, reinforcement learning has provided a new idea for solving the problems mentioned above [10]–[12]. Reinforcement learning is a data-driven method. It has better adaptive learning ability and decision-making ability for non-convex, non-linear problem optimization. Therefore, it avoids simplifying the non-convex and non-linear parts of the model to solve the problem, unlike the traditional method. At the same time, reinforcement learning focuses on how the agent takes actions in the environment to obtain maximum cumulative reward, which is consistent with the dynamic dispatch of the integrated energy system. It focuses on how the

Manuscript received September 29, 2021; revised December 25, 2021; accepted January 4, 2022. Date of online publication October 12, 2022; date of current version July 18, 2025. This work was supported by National Natural Science Foundation of China (51977005).

Q. K. Sun, X. J. Wang (corresponding author, email: xjwang1@bjtu.edu.cn), Z. Liu, W. S. Lin, and J. H. He are with the School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China.

W. Pei is with the Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing 100190, China.

DOI: 10.17775/CSEEJPES.2021.08770

system makes decisions to obtain optimal operating cost in a particular dispatch period [13], [14]. At present, reinforcement learning has been preliminarily applied to the economic dispatch problem of the IES. In [15], the Q learning algorithm is combined with game theory to optimize the dispatch problem of the smart grid. In [16], the deep Q learning network was used to solve the real-time dispatch strategy of the microgrid. However, all the above methods require discretization of action and suffer from the obvious challenge of dimensional disaster. In [17], the uncertain economic dispatch of the IES based on the DDPG algorithm can effectively deal with continuous state and action spaces. However, the algorithm is sensitive to parameters, and it is a daunting task to adjust them. In [18], the CPO algorithm is used to study optimal dispatch of microgrids. Policy updates are constrained by Kullback-Leibler (KL) divergence but suffer from disadvantages such as excessive computational effort and long training time. Therefore, it is not suitable for practical engineering. Additionally, a common problem with existing research is the over-reliance on reward functions for unconstrained exploration, which can easily lead to dispatch decisions that violate the system's operational safety [19].

Analyzing the shortcomings of existing methods, it was found that there is still a gap between the front-end of academic research and practical applications based on data-driven methods. As a black-box model, the reinforcement learning method of data-driven theory still needs further study, algorithm convergence, and result security verification. This paper formulates the IES multi-energy flow dynamic dispatch problem within a data-driven framework. It proposes an optimization solution strategy for IES dynamic energy dispatch based on a constrained reinforcement learning algorithm. The specific performance is as follows:

1) The proposed method is based on data-driven theory, which can avoid simplifying non-convex and non-linear parts of the model to solve the problem. It can use interactive data with the environment to dynamically respond to system changes, such as system structure changes, infrastructure aging, and impact of uncertain factors.

2) Aiming at the security issues caused by traditional reinforcement learning relying on the reward function for unconstrained exploration, boundary conditions of the action security domain are used to restrict the selection of actions, which effectively avoids over-limit actions of the IES units.

3) Aiming at the instability of decision-making results caused by the influence of uncertain external factors. In the trust domain, the strategy update is constrained by the truncation function, and the strategy update is always kept within a controllable range to improve the convergence and stability of the IES optimal dispatch decision-making model.

II. IES DYNAMIC ENERGY DISPATCH MATHEMATICAL MODEL

The goal of IES dynamic energy dispatch is to pursue economic optimization under the premise of ensuring safety. Thus, IES effectively arranges the output of various devices in each period under the premise of meeting user load requirements and system operation safety [20]. This paper takes

the IES shown in Fig. 1 as an example. It is a complex system composed of an electric network, a heating network, and a natural gas network. The system also includes a variety of devices, such as combined heat and power units (CHP), photovoltaics (PV), battery energy storage (BES), gas boilers (GB), electric boilers (EB), and electricity/heat loads.

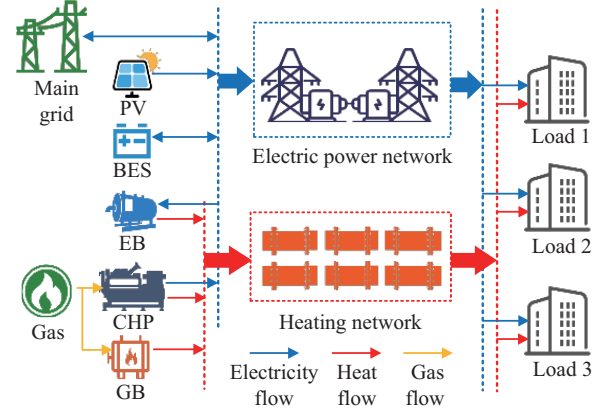


Fig. 1. Schematic diagram of the integrated energy system.

A. Objective Function

Since IES dynamic energy dispatch needs to take into account the mutual influence between different dispatch periods, minimizing the expected value of the total cost of all dispatch periods is considered as the objective function [21].

$$\min E \left(\sum_{t=1}^T F(t) \right) = E \left(\sum_{t=1}^T (F_e(t) + F_{BES}(t)) \right) \quad (1)$$

where T is the length of the dispatch period; $F_e(t)$ is the cost of purchasing energy at time step t ; and, $F_{BES}(t)$ is the charge or discharge depreciation cost of BES at time step t .

1) Cost of Purchasing Energy

The cost of purchasing energy at time step t can be expressed as:

$$F_e(t) = \varepsilon_e(t) p_{\text{grid}}(t) \Delta t + \varepsilon_{\text{gas}}(t) \left(\frac{p_{\text{CHP}}(t) \Delta t}{\eta_{\text{CHP-P}}} + \frac{h_{\text{GB}}(t) \Delta t}{\eta_{\text{GB}}} \right) \quad (2)$$

where $p_{\text{grid}}(t)$ is the exchanged power between the IES and the main grid at time step t . If the value is positive, it means that the system purchases electricity from the main grid, and if it is negative, it means that the surplus power of the system is sold to the grid. $p_{\text{CHP}}(t)$ is the electric power output of CHP; $h_{\text{GB}}(t)$ is the heat output of GB; $\eta_{\text{CHP-P}}$ is the electrical efficiency of the CHP; η_{GB} is the efficiency of GB; $\varepsilon_e(t)$ and $\varepsilon_{\text{gas}}(t)$ are the purchase price of electricity and natural gas at time step t , respectively; and Δt is time interval.

2) Cost of Battery Storage Depreciation

Charge and discharge depreciation cost of BES can be expressed as:

$$F_{\text{BES}}(t) = \rho_{\text{BES}} |p_{\text{BES}}(t)| \quad (3)$$

where $p_{\text{BES}}(t)$ is the charging/discharging power of BES at time step t . The positive and negative values of $p_{\text{BES}}(t)$

respectively, denote discharging and charging. ρ_{BES} is depreciation cost factor for BES.

B. Constraints

IES dynamic energy dispatch includes a series of constraints such as power balance, external energy supply, and equipment operation.

1) Power Balance Constraints

Neglecting the system's network loss, the balance between generation and load constraints at time step t can be expressed as:

$$p_{\text{load}}(t) = p_{\text{grid}}(t) + p_{\text{PV}}(t) + p_{\text{BES}}(t) + p_{\text{CHP}}(t) - p_{\text{EB}}(t) \quad (4)$$

$$h_{\text{load}}(t) = h_{\text{CHP}}(t) + h_{\text{GB}}(t) + h_{\text{EB}}(t) \quad (5)$$

where $p_{\text{PV}}(t)$ is the output of PV; $p_{\text{EB}}(t)$ is the input electrical power of EB; $h_{\text{EB}}(t)$ is the heat power output of EB; $h_{\text{CHP}}(t)$ is the heat power output of CHP; and, $p_{\text{load}}(t)$ and $h_{\text{load}}(t)$ are the electrical load and heat load, respectively.

2) IES and Main Grid Interactive Power Constraints

Considering operational stability on the grid side, there are upper and lower constraints on power interaction between the main grid and the IES.

$$P_{\text{grid}}^{\min} \leq p_{\text{grid}}(t) \leq P_{\text{grid}}^{\max} \quad (6)$$

where $P_{\text{grid}}^{\min}$ and $P_{\text{grid}}^{\max}$ are the upper and lower limits of the interaction power between IES and the main grid, respectively.

3) Device Operation Constraints

Taking the IES selected in this paper as a case, the upper and lower limits of the electrical power of CHP, charging/discharging power of BES, the heat power of GB, and the heat power of EB can be expressed as:

$$P_{\text{CHP}}^{\min} \leq p_{\text{CHP}}(t) \leq P_{\text{CHP}}^{\max} \quad (7)$$

$$P_{\text{BES}}^{\min} \leq p_{\text{BES}}(t) \leq P_{\text{BES}}^{\max} \quad (8)$$

$$H_{\text{GB}}^{\min} \leq h_{\text{GB}}(t) \leq H_{\text{GB}}^{\max} \quad (9)$$

$$H_{\text{EB}}^{\min} \leq h_{\text{EB}}(t) \leq H_{\text{EB}}^{\max} \quad (10)$$

where $P_{\text{CHP}}^{\min}$ and $P_{\text{CHP}}^{\max}$ are the upper and lower limits of the electric power of CHP, respectively; $P_{\text{BES}}^{\min}$ and $P_{\text{BES}}^{\max}$ are the upper and lower limits of the charging/discharging power of BES, respectively; $H_{\text{GB}}^{\min}$ and $H_{\text{GB}}^{\max}$ are the upper and lower limits of the heat power of GB, respectively; $H_{\text{EB}}^{\min}$ and $H_{\text{EB}}^{\max}$ are the upper and lower limits of the heat power of EB, respectively.

4) Device Output Ramp Rate Constraints

The output of each device in IES needs to meet upward and downward ramp rate constraints.

$$-R_{\text{CHP}}^{\text{down}} \Delta t \leq p_{\text{CHP}}(t) - p_{\text{CHP}}(t-1) \leq R_{\text{CHP}}^{\text{up}} \Delta t \quad (11)$$

$$-R_{\text{GB}}^{\text{down}} \Delta t \leq h_{\text{GB}}(t) - h_{\text{GB}}(t-1) \leq R_{\text{GB}}^{\text{up}} \Delta t \quad (12)$$

$$-R_{\text{EB}}^{\text{down}} \Delta t \leq h_{\text{EB}}(t) - h_{\text{EB}}(t-1) \leq R_{\text{EB}}^{\text{up}} \Delta t \quad (13)$$

where $R_{\text{CHP}}^{\text{up}}$, $R_{\text{CHP}}^{\text{down}}$, $R_{\text{GB}}^{\text{up}}$, $R_{\text{GB}}^{\text{down}}$, $R_{\text{EB}}^{\text{up}}$ and $R_{\text{EB}}^{\text{down}}$ are the upward and downward ramping rates of CHP, GB and EB, respectively.

5) BES Constraints

BES should consider the impact of excessive charging/discharging during its operation. As a result, state of charge (SOC) of BES needs to be limited within a certain range.

$$C_{\text{SOC}}^{\min} \leq c_{\text{SOC}}(t) \leq C_{\text{SOC}}^{\max} \quad (14)$$

where $C_{\text{SOC}}^{\min}$ and $C_{\text{SOC}}^{\max}$ are the upper and lower limits of the BES state of charge, respectively; $c_{\text{SOC}}(t)$ is the state of charge of BES at time step t . $c_{\text{SOC}}(t)$ can be determined using the following equation as:

$$c_{\text{SOC}}(t) = c_{\text{SOC}}(t-1) - \frac{\eta_{\text{BES}} p_{\text{BES}}(t) \Delta t}{Q_{\text{BES}}} \quad (15)$$

$$c_{\text{SOC}}(0) = C_{\text{SOC}}^{\text{ini}} \quad (16)$$

where Q_{BES} is BES capacity; $C_{\text{SOC}}^{\text{ini}}$ is the initial state of charge of BES; η_{BES} is the charging/discharging coefficient for BES, which can be expressed as:

$$\eta_{\text{BES}} = \begin{cases} \eta_{\text{ch}} & p_{\text{BES}} < 0 \\ \frac{1}{\eta_{\text{dis}}} & p_{\text{BES}} \geq 0 \end{cases} \quad (17)$$

where η_{ch} and η_{dis} are the charging and discharging efficiency of BES, respectively.

6) Coupling Link of Electric Network and Heating Network

This paper considers two coupling devices, including EB and CHP. EB can be regarded as the electric network's load and the heating network's source, consuming electricity to generate heat. CHP can be regarded as the load of the natural gas network, the source of the electric network, and the heating network. It mainly generates electricity by consuming natural gas, and waste heat can also provide heat energy for users.

For EB,

$$p_{\text{EB}}(t) = \frac{h_{\text{EB}}(t)}{\eta_{\text{EB}}} \quad (18)$$

where η_{EB} is the conversion efficiency of electricity to heat.

For CHP, the coupling relationship between its output of electric power and heat power is called an electric-heat characteristic. CHP in this paper uses the method of determining the heat-to-electricity ratio.

$$p_{\text{CHP}}(t) = \frac{h_{\text{CHP}}(t)}{\eta_{\text{CHP}}} \quad (19)$$

where η_{CHP} is heat-to-electricity ratio of CHP.

7) Uncertainty Distribution of Power and Load Output

Due to privacy protection and copyright issues, we are not allowed to use actual confidential data in this paper. Therefore, we generalized the data, mainly load data, by combining actual data and simulated data to form experimental data in this work, in which simulated data is obtained by simulations with a distribution function. Experimental data still hold the general characteristics of load in an integrated energy system and qualify for verification of the proposed method [22]. This function can be expressed as:

$$f(P_{\text{sto}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(P_{\text{sto}} - \mu)^2}{2\sigma^2} \right] \quad (20)$$

where P_{sto} is uniformly expressed as electrical load, heat load, and photovoltaic output; μ and σ are the expected and standard deviation of the distribution, respectively.

At the same time, considering that the constrained reinforcement learning method proposed in this paper is not limited to a single distribution function, in order to further verify its applicability to photovoltaic and load fluctuations, the following two measures are added based on the above simulation data: 1) randomness of the distribution interval is superimposed based on historical data and simulated data 2) random sampling is performed within the constraint interval of the initial state. Together, they form the training data for the reinforcement learning model.

III. CONSTRAINED REINFORCEMENT LEARNING FRAMEWORK FOR IES DYNAMIC DISPATCH PROBLEM

The IES dynamic energy dispatch mathematical model in Section II was transformed into a constrained reinforcement learning framework following a Markovian Decision Process [23]. This part is mainly used to explain critical parameter settings of the reinforcement learning method, which is to transform the specific physical problem studied in this paper into a detailed reinforcement learning practical application process.

A. Constrained Environment Model

The constrained environment model refers to the IES that interacts with the agent. Every control action given by the agent through reinforcement learning can obtain corresponding feedback rewards in this constrained environment. The environmental model in this paper mainly includes the power system, natural gas system, and related constrained links, which are expressed by (1) through (20).

B. State Space

The state space is a concrete representation of the environment in which the agent is currently located. States of the IES dynamic dispatch include $p_{load}(t)$, $h_{load}(t)$, $p_{PV}(t)$, and $c_{soc}(t-1)$. Therefore, the state space can be defined as:

$$s_t = [p_{load}(t), h_{load}(t), p_{PV}(t), c_{soc}(t-1)] \quad (21)$$

C. Action Space

The action space consists of decisions made by the agent through reinforcement learning, and the output of each device can represent an action. $h_{CHP}(t)$, $h_{EB}(t)$, $p_{EB}(t)$ and $p_{grid}(t)$ can be derived from $p_{CHP}(t)$, $p_{BES}(t)$ and $h_{GB}(t)$ [24], so to simplify analysis, action variables are reduced to 3. Action space can be expressed as:

$$a_t = [p_{CHP}(t), p_{BES}(t), h_{GB}(t)] \quad (22)$$

D. State Transition Probability

State transition probability is the agent will move from current state s_t to new state s_{t+1} after performing the selected action a_t . For the IES dynamic dispatch problem, $c_{soc}(t)$ in the state at the next moment after the agent makes an action can be obtained from (15). The remaining uncertainty state

values $p_{load}(t)$, $h_{load}(t)$, and $p_{PV}(t)$ can only be obtained after the agent has interacted with the environment. Together, these variables make up the state transition probability.

E. Reward Function and Discount Factor

First, for the IES dynamic dispatch problem, the agent reward function designed in this paper consists of system operating costs, BES depreciation costs, and action violation constraint penalties. Second, minimization of the IES dispatch cost is transformed into maximization of the reward function. Meanwhile, to prevent the model from converging to a failure, the value needs to be scaled if the reward function is too large. Therefore, the reward function obtained by the agent at time step t can be expressed as:

$$r_t(s_t, a_t) = -\frac{1}{1000}(F_e(s_t, a_t) + F_{BES}(s_t, a_t) + F_{pen}(s_t, a_t)) \quad (23)$$

$$F_{pen}(s_t, a_t) = \sum_i^N \max(\max(0, a_t^i - a_t^{i,max}), a_t^{i,min} - a_t^i) + \max(\max(0, S_{oc,t} - S_{oc}^{max}), S_{oc}^{min} - S_{oc,t}) + \text{abs}(p_t^{\text{supply}}/p_t^{\text{load}} - 1) + \text{abs}(h_t^{\text{supply}}/h_t^{\text{load}} - 1) \quad (24)$$

where $F_{pen}(s_t, a_t)$ is the penalty for violating constraints; a_t^i is the action of the i -th unit at time step t ; p_t^{supply} is the sum of the electric output provided all units at time step t ; h_t^{supply} is the sum of heat output provided by all units at time step t .

IES dynamic dispatch aims to find a policy π^* that maximizes cumulative discount benefit over the whole cycle.

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^T \gamma^t r_t \right] \quad (25)$$

$$\pi^* = \arg \max_{\pi \in \Pi_\pi} J(\pi) \quad (26)$$

where Π_π represents the set of policy π ; discount factor γ indicates the importance of the reward at the future moment to the current reward. For dynamic energy dispatch, decisions made in the current dispatch period have an essential impact on subsequent ones, so it should be taken as a more significant value. In this paper, this value of γ is set 0.8.

IV. CONSTRAINED REINFORCEMENT LEARNING ALGORITHM FOR SOLVING IES DYNAMIC DISPATCH PROBLEM

Traditional reinforcement learning algorithms generally obtain optimal strategies by maximizing rewards. This process is unconstrained exploration. However, in practical applications, some explorations will hurt the safe and stable operation of the system, which is not allowed [25], [26]. Therefore, this paper adds constraint measures based on traditional methods and then proposes the Constrained Proximal Policy Optimization (CPPO) algorithm to solve this problem.

First, the CPPO algorithm is based on the AC framework. AC framework combines policy learning and value learning to

make the model easier to converge. It approximates policy π_θ through the actor neural network, estimates advantage function A_{π_θ} through the critic neural network, and then evaluates and improves current policy π_θ [27]. Second, security domain boundary conditions are added to enforce constraints in the action selection process. Third, the importance sampling technique is used in the policy update process to convert the policy update method from on-policy to off-policy. Ratio of old and new policies is truncated in the trust domain. These compulsory restraints always keep strategy updates within a controllable range, effectively alleviating the impact of uncertain factors on the security of dispatch strategies [28].

A. Action Security Domain Boundary

In order to ensure the decision results of the reinforcement learning model meet security requirements, boundary conditions of the security domain are added to restrict action. Boundary conditions of the security domain can be gradually improved as the model is trained offline and judgments are made before actual action is executed. If the selected action in the current state is not in the security domain, the action will be forced to skip and wait for a new round of action decision results.

1) Action Constraint

If action a_{t+1} selected by the agent from state s_t to state s_{t+1} causes a security issue, the action will be restricted, and the action will be stored in constraint set C_{s_t} of state s_t .

2) Action Permissibility

If action a_{t+1} selected by the agent from state s_t to state s_{t+1} is not in constraint set C_{s_t} of state s_t , it is considered to be allowed to be executed.

3) Skip Action

Suppose the action obtained from the model decision is outside the security domain boundary in an actual-world application. In that case, the action is forced to skip and wait for the outcome of the next decision.

When the agent selects action a_t^i , $a_t^i \in a_t$ of the i -th unit under state s_t , it appears that the sum of the power provided by the units exceed the load demand. Then the constraint interval of the corresponding action in state s_t is $X_{s_t} = [a_t^{i,\min}, a_t^i]$. As training proceeds, the action constraint interval is updated to $X_{s_t} = [a_t^{i,\min}, a_t^{i'}]$ if action $a_t^{i'}$ still occurs in state s_t with the unit providing more power than the load demand and $a_t^{i'} < a_t^i$. Similarly, when the power provided by units is less than the load demand, the constraint action interval is set to $X_{s_t} = [a_t^i, a_t^{i,\max}]$. Then action constraint set X_{s_t} , state s_t , state s_{t+1} , action a_t , and reward function r are combined to form a set $(s_t, a_t, r_t, s_{t+1}, X_{s_t})$ and stored in the experience replay buffer. Through repeated training in a simulated environment, the model can gradually refine action constraint set X_{s_t} corresponding to state s_t . Thus, we can set X_{s_t} as the boundary of the action security domain. In the actual environment, it is judged whether the selected action meets the requirements of the security domain boundary. If it is not satisfied, the action will be forced to empty and wait for a new round of action decisions; if it is satisfied, each unit will execute the corresponding actions according to the decision result.

B. Constraint Update of Policy

In the traditional reinforcement learning parameter update process, the gradient itself only provides local first-order information under current parameter θ and ignores the curvature of the surface of the reward function defined according to the actual problem. If a larger step size is selected in a highly curved area, then the algorithm is prone to a sudden drop in performance; on the contrary, if the step size is too small, it can fail to converge for a long time. Even more, the gradient needs to be estimated from samples collected based on the current policy π_θ . Fluctuations in policy performance can, in turn, affect the quality of the collected samples, which further increases the degree of fluctuation during policy updates and raises the issue of system operational safety [29].

The purpose of introducing Theorems 1 and 2 in this paper is to illustrate how to transform the optimal scheduling problem of an integrated energy system into a constrained reinforcement learning process [30]. Theorem 1 is used to find a better strategy π'_θ than original strategy π_θ in scheduling strategy optimization to ensure the stable increase of strategy performance. Theorem 2 shows that when $D_{\text{KL}}^{\max}(\pi_\theta \parallel \pi'_\theta)$ is small, the optimization objective can be reasonably transformed into $L_{\pi_\theta}(\pi'_\theta)$ to simplify the iterative optimization process of the strategy.

1) Theorem 1

$$J(\theta') = J(\theta) + \mathbb{E}_{\tau \sim \pi'_\theta} \left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \quad (27)$$

where $J(\theta)$ represents the total discount cumulative income that the agent can obtain under current policy π_θ ; $A_{\pi_\theta}(s_t, a_t)$ is the advantage function, defined as the following equation $A_{\pi_\theta}(s_t, a_t) = Q_{\pi_\theta}(s_t, a_t) - V_{\pi_\theta}(s_t)$; τ is the same state action trajectory generated by π'_θ ; θ is the parameter of the actor network; φ is the parameter of the critic network.

It is known that obtaining optimal policy π_θ is equivalent to maximizing goal $\mathbb{E}_{\tau \sim \pi'_\theta} [\sum_{t=0}^{\infty} \gamma^t A_{\pi_\theta}(s_t, a_t)]$. However, the expectation in the equation is based on π'_θ instead of π_θ , which is difficult to optimize directly. Therefore, important sampling technology needs to be used to approximate the solution. The approximation process is as follows:

$$\begin{aligned} & \mathbb{E}_{\tau \sim \pi'_\theta} \left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \\ &= \mathbb{E}_{s \sim \rho_{\pi_\theta'}(s)} [\mathbb{E}_{a \sim \pi'_\theta(a|s)} [A_{\pi_\theta}(s, a)|s]] \\ &\approx \mathbb{E}_{s \sim \rho_{\pi_\theta}(s)} [\mathbb{E}_{a \sim \pi'_\theta(a|s)} [A_{\pi_\theta}(s, a)|s]] \\ &= \mathbb{E}_{s \sim \rho_{\pi_\theta}(s)} \left[\mathbb{E}_{a \sim \pi_\theta(a|s)} \left[\frac{\pi'_\theta(a|s)}{\pi_\theta(a|s)} A_{\pi_\theta}(s, a)|s \right] \right] \\ &= \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi'_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} A_{\pi_\theta}(s_t, a_t) \right] \end{aligned} \quad (28)$$

Also set

$$L_{\pi_\theta}(\pi'_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi'_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} A_{\pi_\theta}(s_t, a_t) \right] \quad (29)$$

2) Theorem 2

Set $D_{\text{KL}}^{\max}(\pi_\theta \parallel \pi'_\theta) = \max_s D_{\text{KL}}(\pi_\theta(\cdot|s) \parallel \pi'_\theta(\cdot|s))$, then

$|J(\theta') - J(\theta) - L_{\pi_{\theta}}(\pi'_{\theta})| \leq CD_{KL}^{\max}(\pi_{\theta} \parallel \pi'_{\theta})$.
where C is a constant independent of π'_{θ} .

Therefore if $D_{KL}^{\max}(\pi_{\theta} \parallel \pi'_{\theta})$ is small, then $L_{\pi_{\theta}}(\pi'_{\theta})$ can reasonably be used as the optimization objective. The constraint update of the policy can be translated into the following equation.

$$\begin{aligned} \max_{\pi'_{\theta}} & L_{\pi_{\theta}}(\pi'_{\theta}) \\ \text{s.t. } & \mathbb{E}_{s \sim \rho_{\pi_{\theta}}} [D_{KL}(\pi_{\theta} \parallel \pi'_{\theta})] \leq \delta \end{aligned} \quad (30)$$

C. Actor Neural Network

Based on the above theorem, the importance sampling technique is adopted. Two actor networks with the same structure and different parameters are used to complete the approximate policy task jointly [31]. The sampling actor network interacts with the environment to obtain sampled data continuously, and the sampled data is reused to train the updated policy actor network.

1) Sampling Actor Network

Given that traditional reinforcement learning uses an on-policy approach, actor network parameters will change after training with sampled data once. As a result, original data can no longer be used, and new data must be collected by interacting with the environment again. This results in much interaction with the environment during training, which reduces training efficiency and lengthens training time.

The CPPO algorithm uses important sampling techniques to convert the on-policy method to an off-policy method during policy updates, allowing the training of an actor network based on parameter θ using data sampled from an actor network on parameter θ' . Parameter θ' is fixed in this process. Data from parameter network can be reused to train parameter θ network several times, increasing data utilization and improving training speed.

The sampling actor network has the same structure as the policy actor network. According to the following equation, the sampling actor network parameters are updated at regular intervals.

$$\theta' \leftarrow \alpha\theta + (1 - \alpha)\theta' \quad (31)$$

where α is a weighting factor.

2) Policy Actor Network

Given the constrained policy update equation (30) has a KL divergence constraint term, the gradient descent algorithm is difficult to solve directly. Therefore, it is considered from two aspects: ensuring the stability of policy updates and simplifying calculations. First, the KL divergence constraint is replaced by the truncation function method, and the objective function is truncated. Subsequently, the smaller of the truncated and non-truncated objective functions is taken to reduce the fluctuation of the policy, so the direction of the policy update shows stable improvement. Specific performance is shown in (32) and Fig. 2.

$$\begin{aligned} L(\pi'_{\theta}) = \mathbb{E}_{\pi_{\theta}} & \left[\min \left(\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)} A_{\pi_{\theta}}(s_t, a_t), \right. \right. \\ & \left. \left. \text{clip} \left(\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)}, 1 - \varepsilon, 1 + \varepsilon \right) A_{\pi_{\theta}}(s_t, a_t) \right) \right] \end{aligned} \quad (32)$$

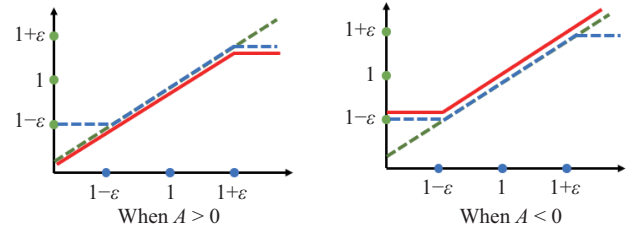


Fig. 2. The constrained value of the policy.

where ε is the truncation factor.

The green line in Fig. 2 is $\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)}$; the blue line is $\text{clip} \left(\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)}, 1 - \varepsilon, 1 + \varepsilon \right)$; the red line represents the result of the policy after constraint processing.

(1) Advantage function $A > 0$ means that the state-action pair is a positive promotion, and the probability of this state-action pair should be increased, and the larger $\pi_{\theta'}(a_t | s_t)$ is, the better. However, the policy is prone to instability when $\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)}$ is larger than $1 + \varepsilon$. Therefore, it is necessary to add a constraint to force it to stop when it reaches $1 + \varepsilon$.

(2) Advantage function $A < 0$ means the state-action pair is a negative promotion, and constraints should also be added to it to force the model to stop when it is trained to $1 - \varepsilon$.

After adding constraints for the policy update, the policy actor network uses the Adam stochastic gradient ascent algorithm to maximize the objective function to update the policy.

$$\begin{aligned} \theta_{k+1} = \arg \max_{\theta} & \frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{t=0}^T \\ & \min \left(\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)} A_{\pi_{\theta}}(s_t, a_t), \right. \\ & \left. \text{clip} \left(\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)}, 1 - \varepsilon, 1 + \varepsilon \right) \right. \\ & \left. * A_{\pi_{\theta}}(s_t, a_t) \right) \end{aligned} \quad (33)$$

where D_k is the set of saved trajectories, defined as $D_k = \{\tau_i\}$.

D. Critic Neural Network

The objective of the critic network is to accurately evaluate the state-action value function, which can be learned by minimizing mean square error using a gradient descent algorithm. Critic network parameter φ is updated as follows:

$$\varphi_{k+1} = \arg \min_{\varphi} \frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{t=0}^T (V_{\varphi}(s_t) - R_t)^2 \quad (34)$$

E. The Interaction Process of Actor Network and Critic Network

The interactive architecture of the actor network and the critic network designed in this paper is shown in Fig. 3. Both network inputs are system state values $\{p_{\text{load}}(t), h_{\text{load}}(t), p_{\text{PV}}(t), c_{\text{soc}}(t-1)\}$. These input data have a certain degree of randomness. The actor network extracts features from state information, and through the neural network output mean μ_{θ} and logarithmic variance $\log(\sigma_{\theta}^2)$ together form a normal

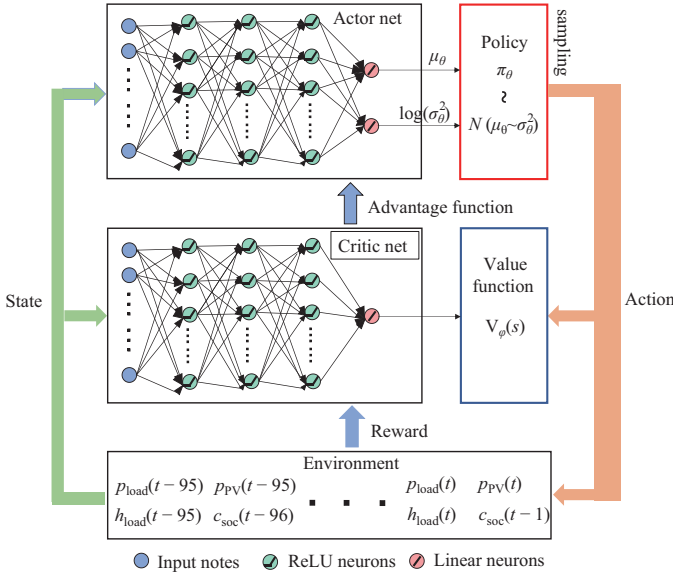


Fig. 3. The process of interaction between the actor network and the critic network.

distribution, which is sampled by the normal distribution to obtain policy π_θ , which in turn generates action a_t for IES dynamic dispatch. State value $V_\phi(s)$ output from the critic network is used to estimate advantage function A_{π_θ} to evaluate and improve current actor network policy π_θ . Therefore, optimal dynamic energy dispatch strategy can be obtained by repeatedly learning the random environment through the actor network and critic network.

F. Optimal Solution Process of Constrained Reinforcement Learning Model

Interaction process of agent based on the CPPO algorithm in the simulated environment and actual environment is shown in Fig. 4. Training and optimization solutions flow for the agent in the simulated environment shown in Fig. 4 is depicted in Fig. 5.

Stochastic quantities of intermittent energy output and load demand are superimposed on historical operating data of the system, which is used as input data for the establishment of the simulation environment and training of the agent model. In the simulated environment, the agent adapts itself to intermittent energy and load uncertainty changes using continuous interaction with the environment while gradually refining the action security domain boundary. After training, the trained agent model and learned action security domain boundary in the simulated environment are transferred to the actual environment. In an actual environment, optimal dispatch actions for the units are obtained by passing real-time data into the agent model. If the action does not meet the requirements of the action security domain, the action is forced to skip, and the result of the next round of agent decisions is awaited. Each unit will execute the corresponding action according to the decision result if the action meets the requirements. In the meantime, experience learned in the actual environment can be updated regularly to improve the agent's real-time decision-making ability.

Agent's optimal solution process based on the CPPO algorithm is as follows:

1) State s_t is fed into the policy actor network to obtain mean μ_θ and logarithmic variance $\log(\sigma_\theta^2)$. Since the effect of directly obtaining $\sigma_\theta^2(s)$ by neural network fitting is not good, the neural network is often used to fit $\rho = \log(\sigma_\theta^2)$ in practical applications. The variance is obtained indirectly by the $\sigma_\theta^2(s) = \exp(\rho)$ formula. Then, they are used to construct a normal distribution function representing action distribution. Then action is sampled and interacts with the environment to obtain reward r , next state s_{t+1} , and security domain boundary X_{s_t} . Finally, it is stored in experience replay buffer in the form of the set $(s_t, a_t, r_t, s_{t+1}, X_{s_t})$. Afterwards, s_{t+1} is fed into the policy actor network, and step 1 is cycled until the preset storage quantity requirement is reached.

2) State s_T obtained in the last step of step 1 is fed into critic network to obtain value function V_T . Value function V_T is used to calculate the discounted reward $R(t) = r(t) + \gamma r(t+1) + \dots + \gamma^{T-t+1} r(T-1) + \gamma^{T-t} V_T$, which are combined into set $R = [R(0), R(1), \dots, R(t), \dots, R(T)]$.

3) The combination of all stored states is fed into critic network to obtain value function V corresponding to each state and calculate advantage function A .

4) The loss function of the critic network is solved, and the backpropagation algorithm and (32) are used to update the critic network parameters.

5) Combining all stored states and inputting them into the sampling actor network and policy actor network to obtain normal distribution function 1 and function 2, respectively. Combining all stored actions and inputting them to normal distribution functions 1 and 2 to obtain probabilities 1 and 2 for each action. Finally, importance sampling weights are obtained by dividing probability 2 by probability 1.

6) Policy actor network parameters can be updated jointly through the back-gradient propagation algorithm and equation (31) above.

7) After a certain number of steps in the loop of steps 5 and 6, policy actor network parameters are updated by the weighted assignment method to sample actor network parameters.

8) Repeat steps 1–7.

V. CASE STUDY AND SIMULATION RESULTS

To validate the effectiveness of the proposed method, several simulations have been performed using the IES shown in Fig. 1. The system's electrical, heat, and PV data are sourced from the European energy open-source website Elia [32]. IES operating parameters are shown in Table AI. Other parameters are shown in Table AII. Time-of-use electricity prices and gas prices are shown in Table AIII.

A. Algorithm Parameter Setting

Actor network and critic network structures are set up as shown in Fig. 3 above. Considering the dimensions of state space and action space are 4 and 3, hidden layers are determined to be three after iterative adjustment and optimization. The number of hidden layer neurons in the actor network is

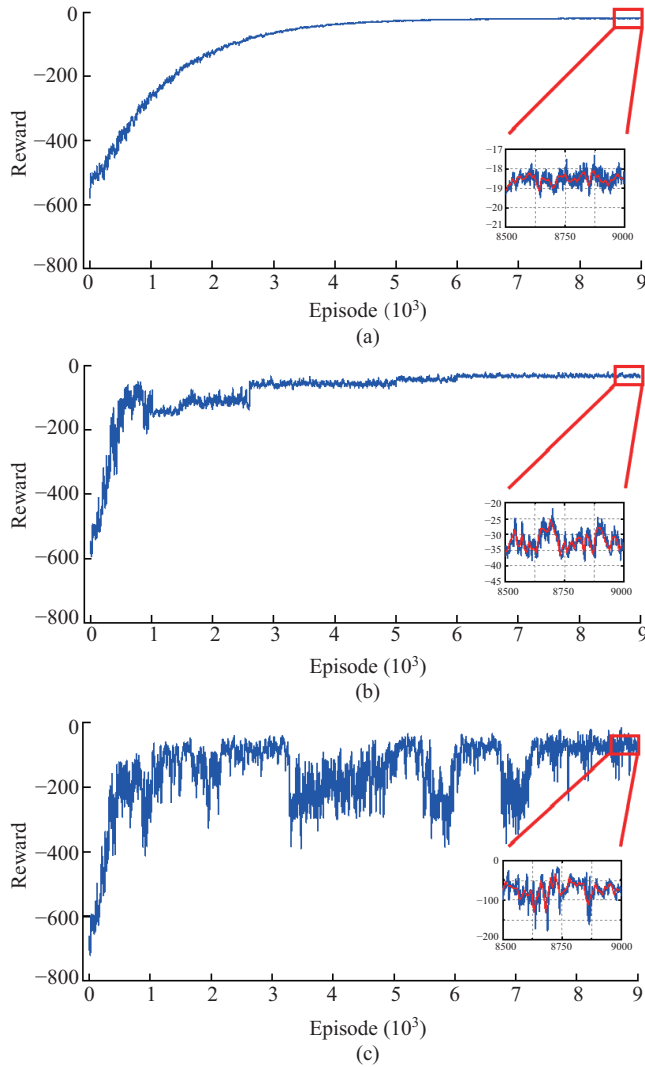


Fig. 6. Convergence analysis of different algorithms. (a) CPPO algorithm. (b) DDPG algorithm. (c) DQN algorithm.

met by the cooperation of CHP and GB. When the electricity price is at a peak period, the system first calls on CHP to supply electricity, while the heat generated by CHP becomes the primary source for meeting the heat load. When the heat supply is insufficient, the system chooses to use GB for heating instead of EB to achieve an optimum operating economy.

It can be seen that the agent based on the CPPO algorithm can adapt to intermittent energy and load changes. Furthermore, it can dynamically adjust the output of various types of units with economic optimization under the guidance of external electricity prices.

2) Security Analysis of Dispatch Results

a) Analysis of Over-limit of Unit Output and Action Security Domain Boundary: Due to space limitations, this article only takes the CHP's electric output over-limit situation and its action security domain boundary as a case for analysis. Over-limit situation of CHP's electric output during training is shown in Fig. 8; the boundary of the action security domain is shown in Fig. 9.

Combined with the analysis in Figs. 8 and 9, it can be

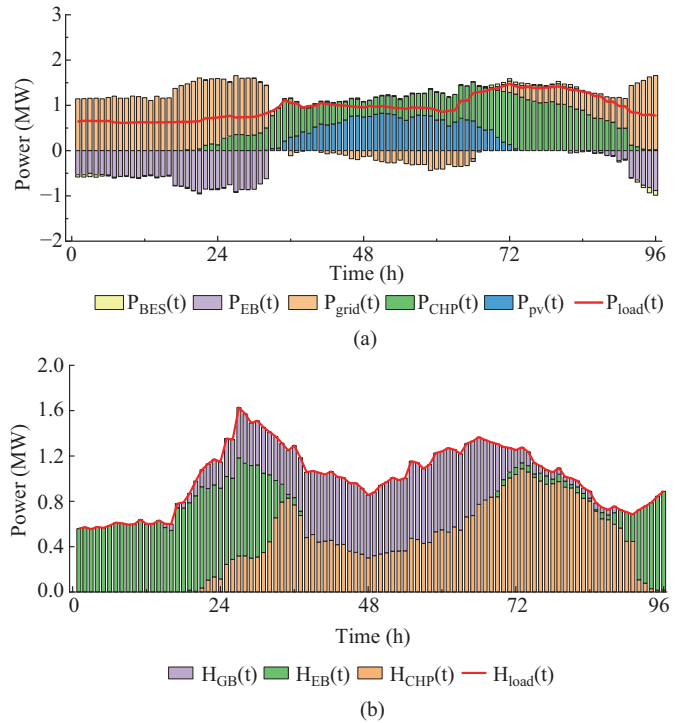


Fig. 7. Optimal dispatch results of units output. (a) Dispatch results for electric power. (b) Dispatch results for heat power.

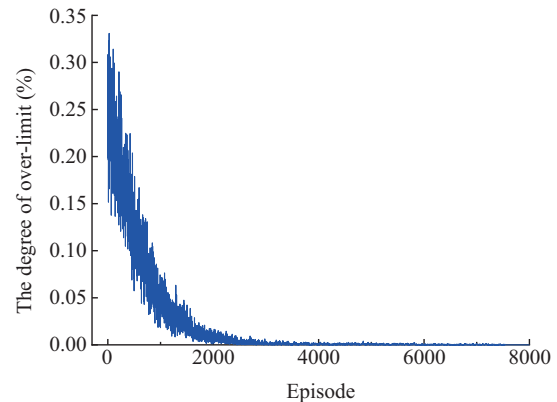


Fig. 8. Analysis of CHP's over-limit situation of electricity output.

seen that at the beginning of the training period, the agent has not interacted deeply with the environment, and the action security domain boundary has not yet been established. It leads to the apparent overrun of the unit's output. A significant mismatch between the power provided by the unit and the load demand in the dispatch decision process can easily lead to system operational safety problems. However, with continuous interaction between the agent and the environment, the agent can adapt to intermittent energy and load changes. Meanwhile, the action security domain boundary is gradually established and improved. The results of optimal dispatch decisions made by the agent after 7000 generations do not appear to exceed the output limit of the unit, which can effectively avoid the situation where the power provided by the unit does not match the load demand.

According to Figs. 9 and 7(a), it can be seen that in Fig. 7(a),

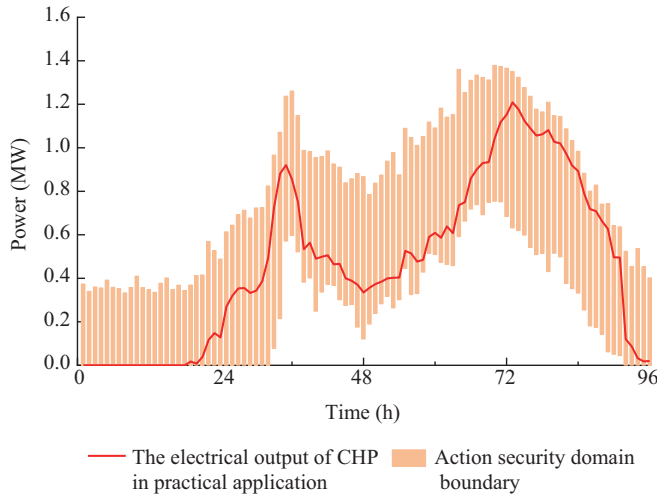


Fig. 9. Action security domain boundary.

the agent's electrical output dispatch result of CHP in actual environment is shown as the red line in Fig. 9. The action curve always meets requirements of the action security domain boundary in Fig. 9. This is mainly because agent needs to judge the security domain boundary of the selected action before executing the action in the actual environment. If it is not within the security domain boundary, it will force the action to be empty and wait for the next round of action decision results. Therefore, in practice, all unit's actions meet constraints of the action security domain. The sum of the power provided by all units in the system can match load demand in real-time within a specific reasonable range, thus effectively avoiding the problem of system operational safety.

b) Applicability Analysis of Uncertainty Factors:

1. Applicability analysis for random fluctuation of external intermittent energy and load

In order to verify the applicability of the proposed method to random fluctuations of external intermittent energy and load, and to compare and analyze with other algorithms, three scenarios were set up with standard deviations of the random variables of power supply and load being 5%, 10%, and 15% of the expected value. Samples were sampled using the Monte

TABLE I
COMPARISON OF ALGORITHMS UNDER DIFFERENT
FLUCTUATION LEVELS

Fluctuation level	Mismatched amount of electrical and heat power throughout the day (MW)		
	CPPO algorithm	DDPG algorithm	DQN algorithm
5%	0.079/0.016	0.243/0.087	1.401/0.432
10%	0.081/0.022	0.452/0.129	2.241/0.745
15%	0.203/0.062	0.721/0.417	3.855/1.104

Carlo method, and results are shown in Fig. 10 and Table I.

In Fig. 10, only the electrical output of the CHP unit is taken as an example for analysis due to space limitations. It can be seen that when the random fluctuation of intermittent energy and load increases, the unit output exceeds the limit signs in the early stage of algorithm optimization iteration, and the system shows apparent instability. However, with continuous interaction between the algorithm and the integrated energy system, it can effectively learn the output limit of the unit and gradually adjust the output scheduling plan of the unit. Finally, the purpose of effectively restraining the unit's output in the face of varying degrees of external intermittent energy and load fluctuations is achieved.

Referring to Table I, as uncertainty increases, total power mismatch between electricity and heat in the whole cycle increases. In the results, the agent based on the DQN algorithm has the most significant mismatch between electricity and heat power throughout the day. When the fluctuation level is 15%, the electric and heat power mismatch reaches 3.855 MW and 1.104 MW, respectively. It is straightforward to raise or lower the system voltage due to a power mismatch. In contrast, the calculation results of the algorithm proposed in this paper are more adaptable and stable, and can better deal with the uncertainty of intermittent energy and load. The reason is that the CPPO algorithm does not need to discretize the continuous action and state space. Compared with the DQN algorithm, it effectively retains a large amount of valuable information. In addition, compared to the DDPG algorithm, it adds mandatory constraints. Even when the system has large fluctuations, it will not be affected and will diverge or fall into a locally optimal solution.

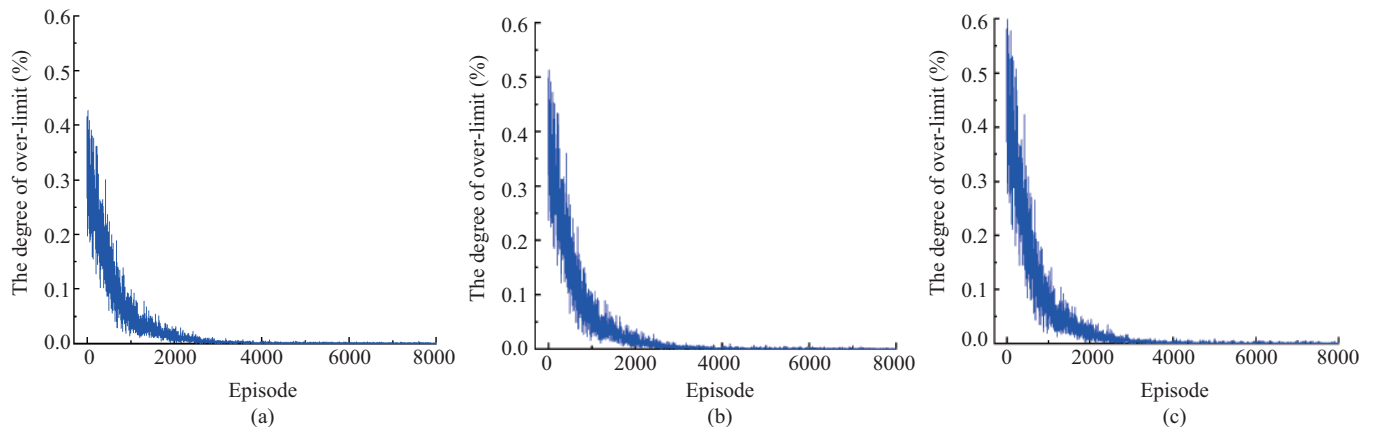


Fig. 10. Analysis of CHP's over-limit situation of electricity output under different fluctuation levels. (a) 5% fluctuation level. (b) 10% fluctuation level. (c) 15% fluctuation level.

2. Analysis of inaccurate applicability of long-term operating parameters of the system

In order to verify that the proposed method can adapt to the inaccuracy of internal parameters caused by long-term operation of the system, CHP unit efficiency parameters are set to 0.35, 0.30, and 0.25 to simulate the change in the internal parameters of the system with time. The result is shown in Fig. 11.

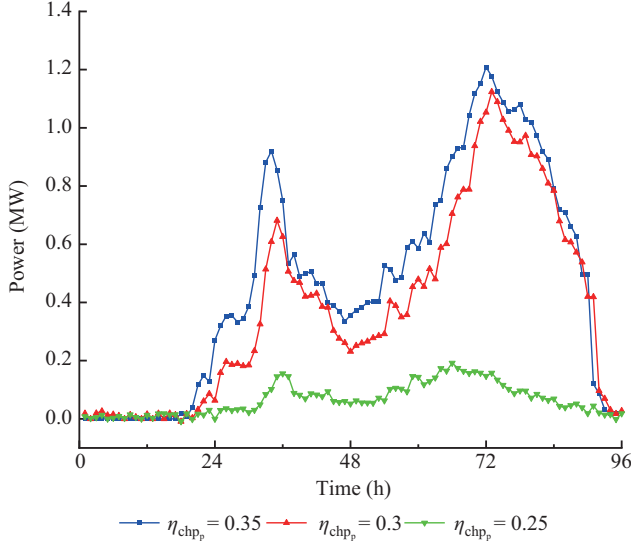


Fig. 11. Analysis of the electrical output of CHP units under the influence of different parameters.

From the analysis in Fig. 11, it can be seen that the original system will dispatch a large number of CHP units to meet demand during the peak load period. However, with long-term operation of the system, the electrical efficiency of CHP units gradually decreases, and system dispatching is no longer mainly based on CHP units and seeks a better dispatching scheme. The reason is that the method proposed in this paper can perform real-time learning based on actual operating data and learn information about CHP unit efficiency decline implicit in the data through the data-driven method. It has a specific adaptive adjustment capability for a series of problems caused by the system running for a long time.

VI. CONCLUSION

Aiming at the problem of IES dynamic energy dispatch, this paper proposes an IES dynamic energy dispatch strategy based on a constrained reinforcement learning algorithm. The conclusions are as follows:

1) The proposed method is based on data-driven theory. It can use interactive data to modify the model and dynamically respond to system changes.

2) The proposed method adds constraints based on traditional reinforcement learning. It can effectively constrain unit output and mitigate strategy update fluctuations, thereby avoiding system operation safety issues.

As an adaptive learning algorithm, reinforcement learning can dynamically respond to system changes. How to combine reinforcement learning with digital twin technology and apply

it to the full life cycle of an integrated energy system is a research direction for the future.

APPENDIX

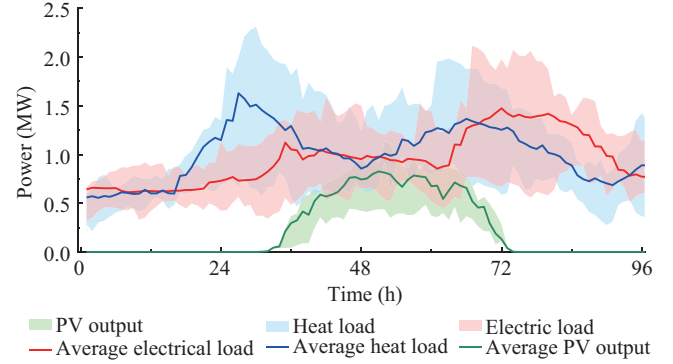


Fig. A1. A typical day load data from training data set.

TABLE AI
OPERATION PARAMETERS OF DEVICES

Device	$P^{\min}/H^{\min}$ MW	$P^{\max}/H^{\max}$ MW	R^{up} MW/min	R^{down} MW/min
CHP	0	1.6	0.11	0.1
BES	-0.2	0.2	-	-
GB	0	1.6	0.1	0.08
EB	0	1.5	0.1	0.1

TABLE AII
OTHER SYSTEM PARAMETERS

Parameter	Value	Parameter	Value
η_{CHP_p}	0.35	η_{dis}	0.95
b	1.20	ρ_{BES}	0.05
η_{GB}	0.80	$C_{\text{SOC}}^{\text{ini}}$	0.30
η_{EB}	0.95	$C_{\text{SOC}}^{\text{min}}$	0.20
η_{ch}	0.95	$C_{\text{SOC}}^{\text{max}}$	0.80

TABLE AIII
TIME-OF-USE ELECTRICITY PRICE

Time	Electricity purchase price (¥·(kW/h) ⁻¹)	Electricity selling price (¥·(kW/h) ⁻¹)
Peak hour	0.98	0.5
Normal hour	0.49	0.2
Valley hour	0.17	0

TABLE AIV
REINFORCEMENT LEARNING RELATED PARAMETERS

Parameter	Value
Dimension of state space	4
Dimension of action space	3
Hidden layer neurons of actor network	64/100/80
Hidden layer neurons of critic network	64/80/80
Activation function	Relu
Learning rate of actor network	0.001
Learning rate of critic network	0.0001
Experience replay buffer	12000
Mini-batch	128
Weighting factor	0.8
Truncation factor	0.2

REFERENCES

- [1] E. A. M. Ceseña, E. Loukarakis, N. Good, and P. Mancarella, "Integrated electricity-heat-gas systems: techno-economic modeling, optimization, and application to multienergy districts," *Proceedings of the IEEE*, vol. 108, no. 9, pp. 1392–1410, Sep. 2020.
- [2] C. Q. Pan, Z. H. Bie, G. F. Li, C. Wang, and C. Yan, "Reliability evaluation of integrated energy systems based on exergy," *CSEE Journal of Power and Energy Systems*, vol. 10, no. 6, pp. 2507–2516, Nov. 2024.
- [3] W. Wang, S. H. Huang, G. M. Zhang, J. Z. Liu, and Z. Chen, "Optimal operation of an integrated electricity-heat energy system considering flexible resources dispatch for renewable integration," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 699–710, Jul. 2021.
- [4] Y. Gao and Q. Ai, "Novel optimal dispatch method for multiple energy sources in regional integrated energy systems considering wind curtailment," *CSEE Journal of Power and Energy Systems*, vol. 10, no. 5, pp. 2166–2173, Sep. 2024.
- [5] J. Yu, X. W. Shen, and H. B. Sun, "Economic dispatch for regional integrated energy system with district heating network under stochastic demand," *IEEE Access*, vol. 7, pp. 46659–46667, Mar. 2019.
- [6] C. H. Peng, P. Xie, L. Pan, and R. Yu, "Flexible robust optimization dispatch for hybrid wind/photovoltaic/hydro/thermal power system," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 751–762, Mar. 2016.
- [7] S. Sharma, A. Verma, Y. Xu, and B. K. Panigrahi, "Robustly coordinated bi-level energy management of a multi-energy building under multiple uncertainties," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 3–13, Jan. 2021.
- [8] X. Dou, J. Wang, Z. Wang, L. J. Li, L. Q. Bai, S. H. Ren, and M. Gao, "A dispatching method for integrated energy system based on dynamic time-interval of model predictive control," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 5, pp. 841–852, Sep. 2020.
- [9] Y. Zhao, Y. Lu, C. Yan, and S. W. Wang, "MPC-based optimal scheduling of grid-connected low energy buildings with thermal energy storages," *Energy and Buildings*, vol. 86, pp. 415–426, Jan. 2015.
- [10] W. R. Liu, P. Zhuang, H. Liang, J. Peng, and Z. W. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, Jun. 2018.
- [11] E. O. Arwa and K. A. Folly, "Reinforcement learning techniques for optimal power control in grid-connected microgrids: a comprehensive review," *IEEE Access*, vol. 8, pp. 208992–209007, Nov. 2020.
- [12] P. C. Dai, W. W. Yu, G. H. Wen, and S. Baldi, "Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2258–2267, Apr. 2020.
- [13] R. S. Sutton and A. G. Barto, *Reinforcement Learning- An Introduction*, 2nd ed., Cambridge: MIT Press, 2018.
- [14] Y. T. Zhao, Z. Wu, W. Gu, J. X. Wang, F. J. Wang, Z. J. Ma, and M. Q. Hu, "Optimal dispatch and pricing of industrial parks considering CHP mode switching and demand response," *CSEE Journal of Power and Energy Systems*, vol. 10, no. 5, pp. 2174–2185, Sep. 2024.
- [15] Z. Ni and S. Paul, "A multistage game in smart grid security: a reinforcement learning solution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2684–2695, Sep. 2019.
- [16] T. Yang, L. Y. Zhao, W. Li, and A. Y. Zomaya, "Reinforcement learning in sustainable energy and electric systems: a survey," *Annual Reviews in Control*, vol. 49, pp. 145–163, Apr. 2020.
- [17] T. Yang, L. Y. Zhao, W. Li, and A. Y. Zomaya, "Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning," *Energy*, vol. 235, pp. 121377, Nov. 2021.
- [18] Y. Ji and J. H. Wang, "Online optimal scheduling of a microgrid based on deep reinforcement learning," *Control and Decision*, vol. 37, no. 7, pp. 1675–1684, 2022.
- [19] B. Feng, J. Y. Zhao, G. Huang, Y. J. Hu, H. T. Xu and C. X. Guo, "Safe deep reinforcement learning for real-time AC optimal power flow: A near-optimal solution," *CSEE Journal of Power and Energy Systems*, doi: 10.17775/CSEEJPES.2023.02070.
- [20] Y. J. Qin, L. L. Wu, J. H. Zheng, M. S. Li, Z. X. Jing, Q. H. Wu, X. X. Zhou, and F. Wei, "Optimal operation of integrated energy systems subject to coupled demand constraints of electricity and natural gas," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 2, pp. 444–457, Jun. 2020.
- [21] S. Moslehi and T. A. Reddy, "A new quantitative life cycle sustainability assessment framework: application to integrated energy systems," *Applied Energy*, vol. 239, pp. 482–493, Apr. 2019.
- [22] M. Fan, V. Vittal, G. T. Heydt, and R. Ayyanar, "Preprocessing uncertain photovoltaic data," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 1, pp. 351–352, Jan. 2014.
- [23] X. Wu, B. Liu, S. Yuan, B. Cao, Z. Zhang and Y. Hu, "Pricing strategy of regional integrated energy system considering privacy protection based on deep reinforcement learning," *CSEE Journal of Power and Energy Systems*, doi: 10.17775/CSEEJPES.2021.08300.
- [24] S. Yao, W. Gu, S. Lu, S. Y. Zhou, Z. Wu, G. S. Pan, and D. He, "Dynamic optimal energy flow in the heat and electricity integrated energy system," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 179–190, Jan. 2021.
- [25] X. W. Pu, H. Xiao, W. Pei, Y. H. Yang, L. Ma, and T. F. Ma, "Optimal energy management of networked multi-energy microgrids based on improved multi-agent federated reinforcement learning," *CSEE Journal of Power and Energy Systems*, doi: 10.17775/CSEEJPES.2023.05920.
- [26] Z. H. Zhou, O. S. Oguz, M. Leibold, and M. Buss, "Learning a low-dimensional representation of a safe region for safe reinforcement learning on dynamical systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 5, pp. 2513–2527, May 2023.
- [27] B. Huang and J. H. Wang, "Deep-reinforcement-learning-based capacity scheduling for PV-battery storage system," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2272–2283, May 2021.
- [28] N. D. H. Khoi, C. P. Van, H. V. Tran, and C. Dung Truong, "Multi-objective exploration for proximal policy optimization," in *Proceedings of 2020 Applying New Technology in Green Buildings*, 2020, pp. 105–109.
- [29] C. M. Liu, X. Xu, and D. W. Hu, "Multiobjective reinforcement learning: a comprehensive overview," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 3, pp. 385–398, Mar. 2015.
- [30] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 1889–1897.
- [31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. (2017, Aug. 28). Proximal policy optimization algorithms. [Online]. Available: <https://arxiv.org/pdf/1707.06347.pdf>.
- [32] The European energy open-source website Elia. [Online]. Available: <https://www.elia.be/en/grid-data/open-data>.



Qingkai Sun received the Ph.D. degree in Electrical Engineering at Beijing Jiaotong University, Beijing, China. He is currently employed by State Grid Energy Research Institute Co., Ltd., Beijing, China. His research interests include the application of artificial intelligence technology in power systems and integrated energy systems.



Xiaojun Wang received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from North China Electric Power University, Beijing, China, in 2001, 2004, and 2008, respectively. In 2014, he was a Bond Scheme Scholar with ALTOM Grid, Stafford, U.K. He is currently an Associate Professor at Beijing Jiaotong University. His research interests mainly include planning and operation of integrated energy systems, the application of machine learning in power system operations, and fault location.



Zhao Liu received the Ph.D. degree in Electrical Engineering from State University of New York at Binghamton, New York, the USA, in 2020. He is currently an Associate Professor with School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include power system operation and stability, and applications of machine learning technology in power systems.



Weishan Lin received the B.S. degree in Electrical Engineering from Beijing Jiaotong University, Beijing, China. He is currently employed by State Grid Gansu Co., Ltd., Lanzhou, China.



Wei Pei received the B.S. and M.S. degrees in Electrical Engineering from Tianjin University, Tianjin, China, in 2002 and 2005, respectively, and the Ph.D. degree from the Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing, China, in 2008. He is working in Institute of Electrical Engineering Chinese Academy of Sciences, Beijing, China. His research interests include distributed energy and microgrids, and AC/DC hybrid distribution networks.



Jinghan He received the M.Sc. degree in Electrical Engineering from Tianjin University, Tianjin, China, in 1994, and the Ph.D. degree in Electrical Engineering from Beijing Jiaotong University, Beijing, China, in 2007. Since 2000, she has been a Professor with the School of Electrical Engineering, Beijing Jiaotong University. Her research interests include ac-dc power system protection and control, monitoring, and protection of the railway traction power supply systems. Dr. He is currently a Chairman of IEEE Power and Energy Society (PES) China Membership

Committee and a member of DC System Protection and Control Working Group under the IEEE PES China Subcommittee.