Electrical Power and Energy Systems 120 (2020) 106016

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Combined heat and power system intelligent economic dispatch: A deep reinforcement learning approach

Suyang Zhou[a,b], Zijian Hu[a], Wei Gu[a,*], Meng Jiang[c], Meng Chen[d], Qiteng Hong[e], Campbell Booth[e]

[a] School of Electrical Engineering, Southeast University, 2 Sipailou Xuanwu Qu, Nanjing, China
[b] Jiangsu Key Laboratory of Smart Grid Technology and Equipment, Nanjing, China
[c] Department of Computer Science and Engineering, University of Notre Dame, USA
[d] Yangzhou Power Supply Company, Jiangsu Electric Power Company of State Grid Corporation of China, China
[e] Department of Electronic and Electrical Engineering, University of Strathclyde, UK

ARTICLE INFO

Keywords:
Combined heat and power economic dispatch
Deep reinforcement learning
Proximal policy optimization

ABSTRACT

This paper proposed a Deep Reinforcement learning (DRL) approach for Combined Heat and Power (CHP) system economic dispatch which obtain adaptability for different operating scenarios and significantly decrease the computational complexity without affecting accuracy. In the respect of problem description, a vast of Combined Heat and Power (CHP) economic dispatch problems are modeled as a high-dimensional and non-smooth objective function with a large number of non-linear constraints for which powerful optimization algorithms and considerable time are required to solve it. In order to reduce the solution time, most engineering applications choose to linearize the optimization target and devices model. To avoid complicated linearization process, this paper models CHP economic dispatch problems as Markov Decision Process (MDP) that making the model highly encapsulated to preserve the input and output characteristics of various devices. Furthermore, we improve an advanced deep reinforcement learning algorithm: distributed proximal policy optimization (DPPO), to make it applicable to CHP economic dispatch problem. Based on this algorithm, the agent will be trained to explore optimal dispatch strategies for different operation scenarios and respond to system emergencies efficiently. In the utility phase, the trained agent will generate optimal control strategy in real time based on current system state. Compared with existing optimization methods, advantages of DRL methods are mainly reflected in the following three aspects: 1) Adaptability: under the premise of the same network topology, the trained agent can handle the economic scheduling problem in various operating scenarios without recalculation. 2) High encapsulation: The user only needs to input the operating state to get the control strategy, while the optimization algorithm needs to re-write the constraints and other formulas for different situations. 3) Time scale flexibility: It can be applied to both the day-ahead optimized scheduling and the real-time control. The proposed method is applied to two test system with different characteristics. The results demonstrate that the DRL method could handle with varieties of operating situations while get better optimization performance than most of other algorithms.

1. Introduction

Co-generation units plays an increasingly important role in the latest power system for their high energy efficiency, excellent environmentally friendly performance and high flexibility. Considering the mutual conversion between various energies, there is a plenty headroom for us to optimize the current conventional CHP system, despite some widespread concerns over the way to improve the economy of the CHP operation [1,2].

The combined heat and power economic dispatch (CHPED) is a significant brunch in CHP researches, which aims at minimizing the total production cost or maximizing the operating income while keeping all constraints satisfied. CHPED problem is generally descripted as an optimization problem with one or more optimizing objectives and a set of highly nonlinear and non-smooth constraints including energy supply–demand balance, capacity limits and other constraints.

The researches on CHPED mainly concentrate on two aspect: models and solutions. Several works have already been done in the CHP economic dispatch models domain. A thermal-electrolytic coupling method was proposed in [3], in which the CHP economic dispatch problem was

* Corresponding author.
E-mail address: wgu@seu.edu.cn (W. Gu).

https://doi.org/10.1016/j.ijepes.2020.106016
Received 22 October 2019; Received in revised form 11 March 2020; Accepted 12 March 2020
Available online 19 March 2020

Nomenclature

$\mathscr{S}$	State of CHP system
$\mathscr{A}$	Action for devices.
I	Indicator function
o	Equipment operating status vector
d	Power mismatch vector
v	Random variables.
R_π	Reward function
$V_\pi(s_t)$	Value function
A_π:	Advantage function.
$\eta(\pi)$	Expected return
t	The t-th time slot.
p_{wind}	Wind power
p_{grid}	Trading electricity with grid
p_l,	Electricity load
h_l,	Thermal load
α	Thermoelectric conversion efficiency of the GT
$c_e/c_m/c_k$	Natural gas cost /The grid interaction cost
ρ_{gas} /ρ_{tou}	unitprice of natural gas/time of use price respectively
ϑ	Reasonable operating capacity
θ	Network parameter
$z_t(\theta)$	Probability ratio of updating parameter
ϵ	*Clip* hyperparamert
γ	Discounting factor
P_e^{max}	Maximum power output of power only units
P_e^{min}	Minimum power output of power only units
P_m^{max}	Maximum power output of CHP units
P_m^{min}	Minimum power output of CHP units
h_m^{min}	Maximum heat output of CHP units
h_m^{max}	Minimum heat output of CHP units
h_k^{min}	Maximum heat output of heat only units
h_k^{max}	Minimum storage capacity of heat only units
h_{tst}^{min}	Maximum storage capacity of TST
h_{tst}^{max}	Minimum storage capacity of TST
$\pi_\theta(a\|s)$	Parameterized policies
r	Reward
$\mathbb{E}_{a_t,s_{t+1},\ldots}[\bullet]$	actions are sampled $a_t\ \pi(\bullet\|s_t)$

decomposed into two heat and electricity sub-problems. Paper [4] and [5] proposed the CHP dispatch models which considered the detailed heat transfer process of the heat storage device and the cogeneration unit respectively. [6] established a two-stage dispatch model based on quantity adjustment and presented an iterative solution algorithm. An integrated response method for electro-thermal demand was proposed in [7] to improve the economy of CHP systems. An operational and structural Model based on efficiency matrices was proposed in [8,9], which was used for the dispatch of multi-energy system [10]. All economic dispatch problems are ultimately mathematically transformed into optimization problems, and the operation region of CHP systems can be modelled either convex or non-convex. Convex operation region is modelled by convex combination of electricity and heat extreme points [11,12] while non-convex operation region is usually modelled as mixed-integer model [13,14].

Some classical numerical methods have been successfully applied to CHPED including two-layer Lagrangian relaxation technique [3], Efficient Branch and Bound algorithm [15], dual and quadratic programming [16], etc. However, these methods have been criticized for their inability to cope with complex optimization problems which have highly nonlinear objective function and constraints. On the contrary, genetic algorithms, simulated annealing and evolutional algorithms could solve non-linear, non-smooth and non-convex optimization problem efficiently. Evolution programming-based algorithm was adopted in [17], in which the mutation search range could be controlled and the neighborhood of the best individual in a population could be searched. [18] presented multi-player harmony search algorithm for large-scale CHPED problem and obtained better convergence performance. Cuckoo optimization algorithm was powered by penalty function in [19]. This algorithm could yield better evolution and constraints handling methods. [20] improved basic genetic algorithm from avoiding excessive losses, excavating the information of parents and improving crossed offspring's quality three aspects. [21] applies newly proposed exchanged market algorithm (EMA) on CHPED problem and the result shows that the algorithm's efficiency and reliability. In addition, there are many excellent optimization methods in economic dispatch domain [22,23].

In addition to models and algorithms, [24,25] try to solve the uncertainty problem by updating optimization time horizons dynamically. In these works, data prediction errors are modeled and model predictive control (MPC) is introduced to reduce the impact of prediction errors on optimization results.

These pioneering researches laid the foundation for the optimal dispatch of CHP system. However, it is worth noting that the solutions proposed by the existing research depend upon strict description of the CHP system. When the operating state changes, the strategy generated according to the original optimization problem is no longer the optimal strategy. In addition, Optimization methods do not achieve good encapsulation in engineering applications because the user needs to adjust the optimization target and constraint equation according to the operating state of the system.

We aim to address both of the two challenges by modeling CHP system as MDP problem and solving it by deep reinforcement learning (DRL) method. MDP is a discrete time stochastic control process and provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker [26]. At each time step, the process is in some CHP operating state, and the decision maker may choose any control action that is available for the current state. The process will return a corresponding reward to evaluate the quality of this question. By solving MDP, the decision maker could learn to choose optimal action for the current state to achieve maximum cumulative reward. By this model, the user only needs to consider the input of the system and the corresponding output, without having to consider the complex mathematical description of the system while retaining strict constraints.

DRL methods have so far attracted great attentions to apply them to power system optimization. [27] combined the artificial neural network and the Q-learning algorithms to achieve the optimal management of operation and maintenance of power grids. [28] applied the fuzzy reinforcement learning to energy trading process to improve the users' economy. [29] presented two variants of RL algorithms to solve economic problem and tested their performance on the IEEE 30 bus system. In this paper, a variant of Distributed Proximal Policy Optimization (DPPO) algorithm [30] for CHP economic dispatch problem has been introduced to our research. This algorithm is capable of handling different operation conditions without sacrificing stability or accuracy. When the system parameters change, the dispatch strategy can be directly given without long-term calculation by the chosen optimization methods. The Asynchronous Advantage Actor-Critic (A3C) [31] based agents and the Clipped Surrogate Objective [32] are adopted to improve the learning efficiency and stability. A comparison has been performed between the performance of this algorithm and two other common benchmark algorithms in CHP dispatch problem. Furthermore, the algorithm has been applied to day-ahead dispatch and real-time dispatch in our research, and the result has been compared with that from the mathematical optimization method. The contribution of this paper could be summarized as the following:

1) The CHP economic dispatch problem is modeled as Markov Decision Process (MDP). we have strict treatment towards constraints and objective functions which ensures that the results obtained by MDP are still the optimal solution in the feasible domain.
2) A variant DPPO algorithm for economic dispatch problem has been developed to improve algorithm exploring ability, which ensures the performance of optimality and convergence. The paper gives detailed proof of the convergence of this algorithm.
3) The proposed method can be reused under a variety of operating scenarios without re-calculation, which improves the adaptability and provides more convenience for users. When the operating states change or emergency happen, users only need to input the current status to get an optimal control strategy instantly, instead of re-writing the constraint equation.
4) The proposed method has time scale flexibility. It can be applied to both day-ahead economic dispatch and real-time control.

This paper begins in Section 3 by describing the CHP system and the MDP model. Section 4 details the completion of the DPPO algorithm and the proof of its stability. Case studies are presented in Section 5 and Section 6 gives the conclusions.

2. Problem statements

In this section, the CHP economic dispatch environments and learning scenes are described.

In a CHP system, the electricity and heat networks are linked through the coupling components (e.g., CHP units, heat pump, electric boilers and circulation pumps). These coupling components allow the flows of energy between the two networks. These coupling components increase the flexibility of the electricity and heat supply system [33].

CHPED aims to minimize total fuel cost of all CHP units while meeting heat power and electric power demand and other constraints. Mathematically, the problem is to minimize the following objective function:

$$\sum_{e=1}^{n_e} c_e(p_e) + \sum_{m=1}^{n_m} c_m(p_m, h_m) + \sum_{k=1}^{n_k} c_k(h_k), \tag{1}$$

subject to:

$$\sum_{e=1}^{n_e} p_e + \sum_{m=1}^{n_m} p_m + \sum_{w=1}^{n_w} p_w = p_d, \tag{2}$$

$$\sum_{k=1}^{n_k} h_k + \sum_{m=1}^{n_m} h_m + \sum_{tst=1}^{n_{tst}} h_{tst} = h_d, \tag{3}$$

$$p_e^{min} \le p_e \le p_e^{max},\ e = 1, 2, \cdots, n_e \tag{4}$$

$$p_m^{min}(h_m) \le p_m \le p_m^{max}(h_m),\ m = 1, 2, \cdots, n_m \tag{5}$$

$$h_m^{min}(p_m) \le h_m \le h_m^{max}(p_m),\ m = 1, 2, \cdots, n_m \tag{6}$$

$$h_k^{min} \le h_k \le h_k^{max},\ k = 1, 2, \cdots, n_k \tag{7}$$

$$h_{tst}^{min} \le h_{tst} \le h_{tst}^{max},\ k = 1, 2, \cdots, n_k \tag{8}$$

where

c is the unit production cost;
p is the unit electrical power generation;
h is the unit heat generation;
h_d and p_d are the system heat and power demands;
e, m, k, w and tst are indices of conventional power units, CHP units, heat-only units, renewable energy source and thermal storage tank (TST) respectively;
n_e, n_m, n_k, $n_w$and n_{tst} are the numbers of the kind of units mentioned above;
$p^{max}$and $p^{min}$are the unit electricity power capacity limits;
$h^{max}$and $h^{min}$are the unit heat capacity limits, in particular, h_{tst}^{max} and h_{tst}^{min} are the charge/discharge rate limits of TST;

Usually, the production cost of different unit types defined as:

$$c_e(p_e) = \alpha_e(p_e)^2 + \beta_e p_e + \gamma_e(\$/h), \tag{9}$$

$$c_m(p_m, h_m) = a_m(p_m)^2 + b_m p_m + c_m + d_m(h_m)^2 + e_m h_m + f_m p_m h_m(\$/h), \tag{10}$$

$$c_k(h_k) = \alpha_k(h_k)^2 + b_k h_k + c_k(\$/h), \tag{11}$$

where α_e, β_e and γ_e represents the cost coefficients of the conventional power units; a_m, b_m, c_m, d_m, e_m and f_m are cost coefficients of co-generation unit; α_k, b_k, and c_k are heat-only units' coefficients.

For simplicity, the operating cost of the units could be calculated by gas price and time of use price directly:

$$c_e(p_e) = \rho_{tou}\left(\frac{p_e^t}{\eta_e}\right)\Delta t \tag{12}$$

$$c_m(p_m, h_m) = \rho_{tou}\left(\frac{p_m^t}{\eta_m}\right)\Delta t_d + \rho_{tou}\left(\frac{h_m^t}{\eta_m}\right)\Delta t, \tag{13}$$

$$c_k(h_k) = \rho_{gas}\left(\frac{h_k^t}{\eta_k}\right)\Delta t, \tag{14}$$

where ρ_{gas}andρ_{tou} are unit price of natural gas and time of use electricity price respectively, $t$denotes one timestep in economic dispatch.

Fig. 1 shows the schema diagram of the CHP network, two units are chosen as heat slack node and electricity slack node to guarantee the heat and electricity demand [33].

A. Problem modelling

The CHP economic dispatch problem is to determine the minimized unit cost of generating heat and power on the foundation that the heat and power loads along with other constraints are all met. To achieve the first-rate control strategy, optimal methods are publicly applied in CHP economic dispatch area, where the problem is described as a series of constraints and one or more objective functions [34]. Varieties of optimization algorithms [1] could be used to find optimal solution in feasible operation region.

MDP model is chosen in this research for its simplicity and computational efficiency, in which an agent interacts with environment over several time steps. Fig. 2 gives an illustration of MDP process: At each time step t, the agent receives a state $\mathscr{S}$ and selects an action $\mathscr{A}$ according to its policy π. After performing the selected action $\mathscr{A}$, the agent, in return, receives the next state and receive a scalar reward r. The agent then repeats the above process until the set conditions are met. The goal of the agent is to maximize the expected return from each state. In this paper, we modeled CHP system operation as an infinite-horizon discounted Markov Decision Process (MDP), defined by the tuple $(\mathscr{S}, \mathscr{A}, \mathscr{P}, r, \rho_0, \gamma)$, where $\mathscr{S}$ is an array of states, $\mathscr{A}$ is the array of actions, $P: \mathscr{S} \times \mathscr{A} \times \mathscr{S} \to \mathbb{R}$ is the transition probability

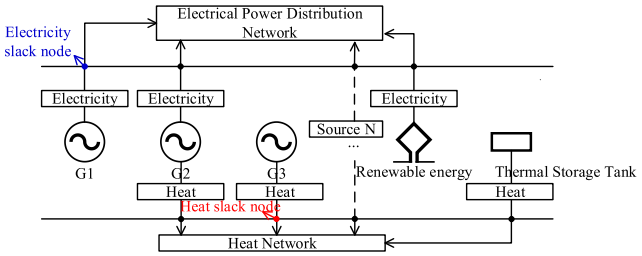

Fig. 1. Schematic diagram of the combined electricity and district heating networks.

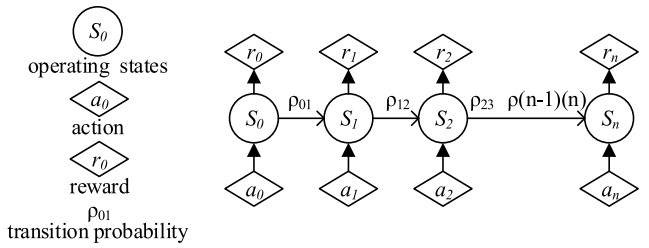

Fig. 2. Illustration of MDP process.

distribution, r: $S \to \mathbb{R}$ is the reward function, ρ_0: $\mathscr{S} \to \mathbb{R}$ is the distribution of the initial state s_0, and $\gamma \in (0, 1)$ is the discount factor. The detailed relationship can be described as the following:

$$\mathscr{S} = (I, o, d, v), \tag{15}$$

$$\mathscr{A} = (\Delta p_e, \Delta p_m, \Delta h_k, \Delta h_{tst}), \tag{16}$$

$$r = -\sum_{e=1}^{n_e} c_e(p_e) - \sum_{m=1}^{n_m} c_m(p_m, h_m) - \sum_{k=1}^{n_k} c_k(h_k) - 0.5 * ||d||^2 + 5I[||d||^2 < \varepsilon] - 0.1 * (s_{tst} - \vartheta)^2 \tag{17}$$

1. **states**: The system operating status contains four parameter vectors, including power mismatch, equipment information and random variables.

IB. is an indicator function which represents the degree of imbalance between supply and demand. In a training episode, I equals to 1 if power mismatch is lower than the limit ε for more than N consecutive time steps, otherwise I equals to 0. The stability of the strategy is improved by I.

$o = [p_e, p_e, h_e, p_{tst}, p_{grid}, p_{wind}]$C. is the normalized equipment operating status vector, indicates the output of all devices

$d = [(p_d - p_s), (h_d - h_s), p_d, h_d)]$D. is the power mismatch vector and indicates the difference between the energy production and the load demand, where p_d is the electricity load, p_s is the electricity supplied, and d_d is the heat load and h_s is the heat supplied.

$v = [tst_i, rtp]$E. denotes the value of random variables, where tst_i is the initial state of the TST and rtp is time-of-use price.

2. **action:** $\mathscr{A}$ suggests an action set for the decision variables which denotes the change amount of the normalized decision variables in every single time step.
3. **reward**: r represents the reward agent gets from system, where d is the power mismatch vector, $I[\bullet]$ is the indicator function, ε is the maximum cumulative power mismatch, s_{tst} is the current heat storage capacity of the heat storage tank and ϑ is the reasonable operating capacity. In a MDP, all objective function mentioned in optimization problems can be described as maximizing the expected cumulative reward signal [35]. Reasonable rewards must be set in order to guide the algorithm to continuously learn from the target. In this research, the rewards for all operational status were kept simple and consistent in different environments (i.e. different output of wind turbine, different electricity load, different heat load and different time-of-use price). The reward consists of 3 sub-targets in "(17)": 1) minus total operating costs $-\sum_{e=1}^{n_e} c_e(p_e) - \sum_{m=1}^{n_m} c_m(p_m, h_m) - \sum_{k=1}^{n_k} c_k(h_k)$: encouraging the agent to reduce the operating cost; 2) power mismatch $(-0.5 * ||d||^2 + 5 * I[||d||^2 < \varepsilon])$: besides the penalty of power mismatch, additional rewards were added when the system reached a power balance, encouraging the agent to minimize the power mismatch; 3) storage tank status $((s_{tst} - \vartheta)^2)$: the penalty for heat storage was added in order to guarantee the stored heat is in a safer range, i.e. there should be enough storage to deal with unexpected situations but not too much storage.

B. Constraints

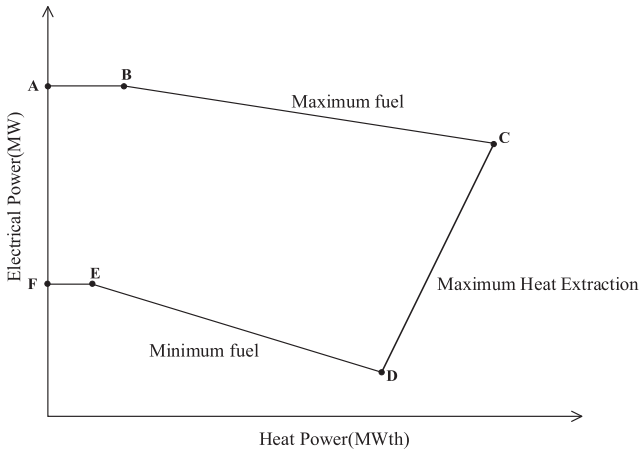

Fig. 3. Heat-Power Feasible Region for a CHP unit.

Constraints are essential part in the mathematical optimization problem. The premise of the optimal solutions is to set the allowable range for the constraints. To simulate the real operation of the CHP system, the strict constraints are set in state transition of MDP. In this section, we demonstrate how we handle constrain in MDP model. For example, if the GT output has reached the maximum in current state S and the action choose by decision maker is still increasing the output of GT, the output of the GT in next state is still maximum to meet equipment operation limit.

Power demands: Electric and thermal power need to reach a supply and demand balance (Eqs. (2) and (3)).

As we mentioned in Fig. 1, usually, two units are chosen as heat slack node and electricity slack node respectively to guarantee the demand. Besides, we convert this part constrains to part of reward in equation (17).

Equipment operation limit: all units must meet their upper and lower limits of output (Eqs. (4)–(7)). In particular, the CHP unit needs to be within the feasible range. Fig. 3 shows the heat-electricity feasible operation region of a coupling unit. The feasible operation region is enclosed by the boundary curve ABCDEF.

In a practical generation unit, steam value admission effects lead to the ripple in the production cost. In order to model this effect, a sinusoidal term is added to the quadratic cost function (9) [36]:

$$c_e(p_e) = \alpha_e(p_e)^2 + \beta_e p_e + \gamma_e + |\varphi_e \sin(\rho_e(p_e^{min} - p_e))|, \tag{18}$$

where φ_e and ρ_e are cost coefficients for modeling valve-point effects.

In MDP, if devices output in next state is beyond restriction, the probability of moving from the current state to next state P: $s_{t-1} \times a \times s_t$ is 0 which means that the agent would not take action that will cause the device to exceed the limit.

Energy storage device constraint: Energy storage device operating constraints are in (8).

The constraints on charge and discharge rate are reflected in the action $\mathscr{A}[\Delta h_{tst}]$ in the MDP model: $h_{dis/char}^{min} < \mathscr{A}[\Delta q_{tst}] < h_{dis/char}^{max}$. The treat to heat storage capacity limits is same as Equipment operation limit.

C. Proof of Optimality

Let π denote a stochastic policy where agent collects action from, the value of $\pi(a|s)$ is the probability distribution of action a at state s. The following are standard relationship between the return function R_π, the value function V_π, and the advantage function Q_π. The return function $R_\pi$is the total discounted reward from time t. Value function V_π denotes the expected return of the agent that acting in accordance with policy π from state s_t.

$$R_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \cdots}\left[\sum_t^{t+l} \gamma^l r(s_t, a_t)\right], \tag{19}$$

$$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \cdots}\left[\sum_t^{t+l} \gamma^l r(s_t)\right], \tag{20}$$

where the notation $\mathbb{E}_{a_t, s_{t+1}, \cdots}[\bullet]$ indicates that actions are sampled $a_t \ \pi(\bullet|s_t)$. γ is the discount factor.

Bellman expectation equation are adopted to describe the recursive relationship of value function and return function:

$$V_\pi(s_t) = \sum_{a_t \in \mathscr{A}} \pi(a_t|s_t) R_\pi(s_t, a_t), \tag{21}$$

$$R_\pi(s_t, a_t) = \mathrm{r}(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathscr{S}} P_{s_t \to s_{t+1}}^{a_t} V_\pi(s_{t+1}), \tag{22}$$

Recursive relationship of value function itself: By inserting Eq. (22) into (21) (For simplicity of the following formula, the subscript t will be omitted):

$$V_\pi(\mathrm{s}) = \sum_{\mathrm{a} \in \mathscr{A}} \pi(a|\mathrm{s}) \left[\mathrm{r}_s^a + \gamma \sum_{s' \in \mathscr{S}} P_{ss'}^a V_\pi(\mathrm{s}')\right], \tag{23}$$

Recursive relationship of return function itself: By inserting Eq. (21) into (22):

$$R_\pi(s, \mathrm{a}) = \mathrm{r}_s^a + \gamma \sum_{s' \in \mathscr{S}} P_{ss'}^a \sum_{a' \in \mathscr{A}} \pi(a'|s') R_\pi(s', a'), \tag{24}$$

Then is the definition of the optimal value function and the optimal return function. The optimal value function is the maximum value function over all policies, the optimal return function is the maximum return function over all policies:

$$V_*(\mathrm{s}) = \max_\pi V_\pi(\mathrm{s}), \tag{25}$$

$$R_*(\mathrm{s}, \mathrm{a}) = \max_\pi R_\pi(s, \mathrm{a}), \tag{26}$$

According to Eqs. (14) and (15):

$$\begin{aligned} V_*(\mathrm{s}) &= \max_\pi V_\pi(\mathrm{s}) \\ &= \max_\pi \sum_{a \in \mathscr{A}} \pi(a|\mathrm{s}) R_\pi(\mathrm{s}, \mathrm{a}) = \max_a R_\pi(\mathrm{s}, \mathrm{a}) \\ &= \max_\pi \left(\mathrm{r}_s^a + \gamma \sum_{s' \in \mathscr{S}} P_{ss'}^a V_\pi(\mathrm{s}')\right) \\ &\max_a \mathrm{r}_s^a + \gamma \sum_{s' \in \mathscr{S}} P_{ss'}^a V_*(\mathrm{s}'), \end{aligned} \tag{27}$$

Similarly,

$$R_*(\mathrm{s}, \mathrm{a}) = \mathrm{r}_s^a + \gamma \sum_{s' \in \mathscr{S}} P_{ss'}^a \max_a R_*(\mathrm{s}', \mathrm{a}'), \tag{28}$$

Define a partial order relationship between strategies: $\pi > \pi' if V_\pi(\mathrm{s}) \geq V_{\pi'}(\mathrm{s})$, then for all MDP:

- There is an optimal policy π_*,which satisfies $\pi_* \geq \pi, \forall \pi$
- The value function of all optimal strategies is equal to the optimal value function $V_{\pi_*}(\mathrm{s}) = V_*(\mathrm{s})$
- The return functions of all optimal strategies are equal to the optimal return function $R_{\pi_*}(\mathrm{s}, \mathrm{a}) = R_*(\mathrm{s}, \mathrm{a})$

Therefore, if there is a strategy π which satisfies $V_\pi(\mathrm{s}) \geq V_\pi(\mathrm{s}), \forall \pi$, this strategy is one of the optimal solutions for MDP.

3. DPPO for economic dispatch problem

Purpose of this algorithm is for economic dispatch problem in rich simulated CHP environments with continuous/sequential state and action spaces. It is required that the algorithms are robust across a wide range of state variation and are effective for CHP systems with high uncertainty. Finally, the strategy learnt by the algorithm should satisfy all the constraints and ultimately achieve the optimal function. It is described in this section how to derive a practical algorithm for the CHP system. A DPPO setup has been considered to learn the parameterized policies $\pi_\theta(a|s)$ with the neural network parameter vector θ, and a baseline function V_π. The architecture consists of a set of agents, the repeatedly generating trajectories of experience, and one chief learner that uses the experiences sent from agents to learn π off-policy. Fig. 4. Illustrates how to train the intelligent agent and how to use it in different scenarios. In training process, the generated random variables were passed into action network, and the action network will generate actions accordingly. Then the value network will evaluate the action strategy. Besides, the simulation CHP environment will proceed to the next state according to the current action and return the reward value. In order to simulate as many situations as possible, each state will be executed 500 times. Finally, update the action network and value network parameters with the goal of maximizing the product of the reward and the evaluation value. In other words, the action network can be understood as an experience pool and accumulate experience about CHP economic dispatch during the training process. When the training is completed, the user can use the trained action network in real-time scheduling or day-ahead optimization scheduling.

The detailed design of each part will be introduced next.

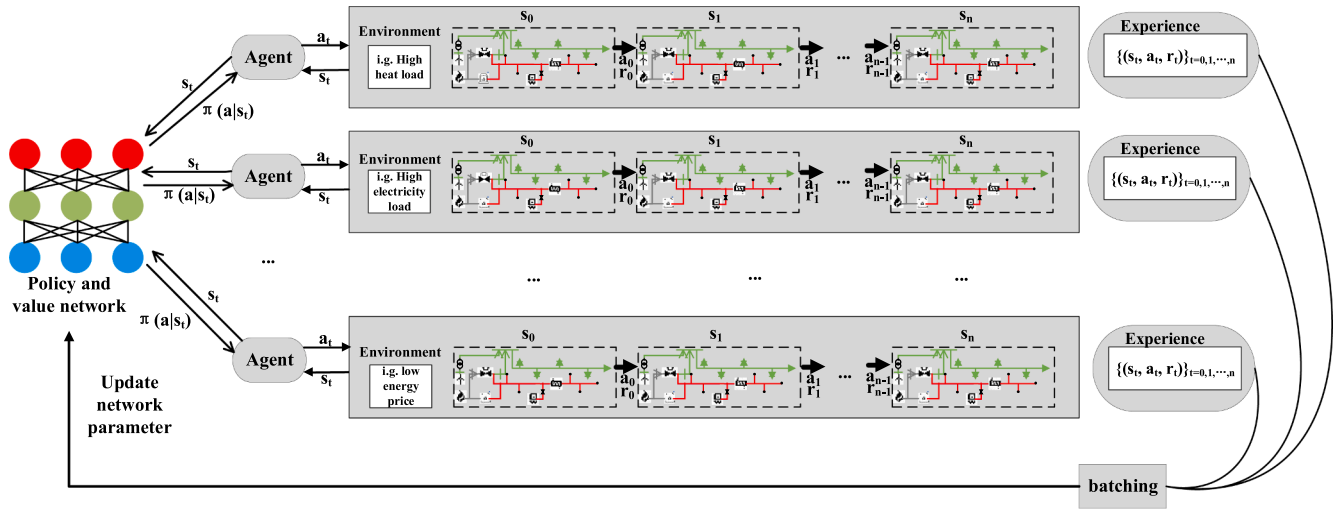

Fig. 4. DPPO Training flowchart.

A. Proximal Policy Optimization with Clipped Surrogate Objective

Advantage function $A_\pi(s, a)$ expresses how good the selection action $a$is in state s. If action a is better than average, then the advantage function is positive; otherwise, it is negative.

$$A_\pi(s, a) = R_\pi(s, a) - V_\pi(s), \tag{29}$$

The following useful identity expresses the expected return of another new policy π in terms of the advantage over π, accumulated over time steps:

$$\eta(\pi) = \eta(\pi) + \mathbb{E}_{a_t, s_{t+1}, \cdots}\left[\sum_{t=0}^{\infty} A_\pi(s, a)\right], \tag{30}$$

Eq. (30) implies that any policy update $\pi \leftarrow \pi$ that has a non-negative expected advantage at every state s ($\sum_t \pi(a|s)A_\pi(s, a) \geq 0$ is guaranteed to increase the policy performance η, or leave it constant in the case that the expected advantage is zero everywhere). This implies the classic result that the update performed by exact policy iteration, which uses the deterministic policy $\pi(s) = argmax V_\pi(s, a)$, improves the policy if there is at least one state-action that pairs with a positive advantage value and nonzero state visitation probability, otherwise the algorithm has converged to the optimal policy.

The accurate way to update the algorithm is mentioned in the previous section. Unfortunately, this proposed update is not possible in continuous dispatch problem since the computation of $\pi(s) = argmax V_\pi(s, a)$ is excessively time consuming. Hence, the accurate update way is inaccessible in the approximate setting for the estimation and approximation error. As a result, there will be some states s for which the expected advantages are negative, i.e. $\sum_a \pi(a|s)A_\pi(s, a) < 0$. The complex dependency of $\rho_\pi(s)$ on π makes it difficult to converge to the optimal policy.

Instead, PPO algorithms is introduced to make some changes on target value function $\eta(\pi)$. It is implied by Eq. (6) that our approach is guaranteed to improve the true objective η by performing the following maximization:

$$J = \max_{\theta} \mathbb{E}_{a_t, s_{t+1}, \cdots}\left[\sum_{t=0}^{\infty} A_\pi(s, a)\right]. \tag{31}$$

In practice, Clipped Surrogate Objective function has been chosen to replace this maximization for the higher robustness than updating policies directly as J. (Clipped Surrogate Objective and Kullback-Leibler divergence (KL) penalty are two widely accepted methods for policy update in PPO algorithms. John Schuman found that KL penalty performed worse than clipped surrogate objective in [32]. Besides, several modifications were added to the core algorithm in both [30] and [32], which includes the normalization of inputs and the accumulation of rewards through timestep with a window of length n as well as bootstrap from the value function after n-steps. Similar augmentations were adopted in this paper.)

Given a parameterized policy π_θ, where $\pi_\theta(a_t|s_t)$ is a differentiable function of network parameter vector θ, our research supposed that trajectory $(s_t, a_t, r_t)_{t=k}^{t=k+n}$ was generated by the agents with the policy π_θ. Let $z_t(\theta)$ denote the probability ratio $\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$, so $z_t(\theta_{old}) = 1$ means the strategy π_θ has not changed. Hence, the n-steps clipped target value function J could be re-written as:

$$J(\theta) = \mathbb{E}\left[\min(z_t(\theta)\widehat{A}_t, clip(z_t(\theta), 1 - \epsilon, 1 + \epsilon)\widehat{A}_t)\right] \tag{32}$$

where ϵ is a hyperparameters, the second *clip* term modifies the surrogate objective function $J(\theta)$ by clipping the probability ratio, by which the range of action changes were clipped in a reasonable scope. Finally, the minimum objective is taken to ensure the objective function is a lower bound on the unclipped objective. In that case, the change in probability ratio was ignored only when it improves the objective, otherwise it was included when it deteriorates the objective. The key idea of this target value function is that the probability ratio ($z_t(\theta)$ was clipped at $1 - \epsilon$ or $1 + \epsilon$ depending on whether the advantages is positive or negative. This assures that the policy change would not be too intense when the advantage is positive, and the update direction is correct when the advantage is negative. As aforementioned, given $A_\pi(s, a)$ was estimated in continuous problems, $\widehat{A}_t$ represents an advantage estimating value for n timesteps as:

$$\widehat{A}_t = \sum_{t=k}^{k+n-1} \gamma^{t-k}(r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)), \tag{33}$$

B. Distributed settings

To achieve good performance in various randomly generated scenes, agents must be guaranteed to explore in as many different environments as possible. Therefore, distributed setup has been introduced to the PPO algorithm. Data was collected in different environments by multiple threads simultaneously and all parallel threads share a global learner. The chief learner learns and develops through the experience collected by different threads. The chief learner setting is similar to A3C in [31]. The difference exists where in our setting that each thread does not compute nor push the gradient of its own policy update to the global PPO net, which promotes the efficiency of the multi-threaded data collection.

A Distributed Proximal Policy Optimization algorithm that uses clipped surrogate objective and distributed architecture is shown in Algorithm I. In each episode, each of the N (parallel) workers (agents) runs policy π_θ for K timesteps, collecting data $\{s_t, a_t, r_t\}$ and estimating the reward function $R_\pi(s_t, a_t)$, the value function $V_\pi(s_t)$ and the advantage function A_π. Besides, workers are required to push data to the chief net. Then the surrogate loss is constructed on NK timesteps of data, and optimized with Adam optimization [38]. Pseudocode are provided in Algorithm II. U is the number of sub-iterations with policy update when a batch of data was collected. Detailed hyperparameter in algorithm was show in Table 6 in Appendix.

Algorithm I.. *DPPO-chief*

```
for iteration = 1, 2...M do
    for actor = 1, 2... N do
        Run policy π_θ for K timesteps, collecting{s_t, a_t, r_t}
        Estimate R_π(s_t, a_t), V_π(s_t) andA_π
    end for
    push data to main PPO
    π_old ← π_θ
    Optimize surrogate loss and update global action π and critic network parameters
end for
```

Algorithm II.. *Agents*

```
for iteration = 1, 2... do
    for actor = 1, 2... N do
        Run policy π_θ for K timesteps, collecting{s_t, a_t, r_t}
        Estimate R_π(s_t, a_t), V_π(s_t) andA_π
    end for
    push data to main PPO
    π_old ← π_θ
    for m ∈ {1, 2, ···U} do
J_CLIP(θ) = E_{ρ^{θ_old}}(τ)[min(z_t(θ)Â_t, clip(z_t(θ), 1 − ϵ, 1 + ϵ)Â_t)]
        Send collect data to chief
        Wait until all agents end this episode
Chief compute main∇_θ J
Update chief-policy parametersθ
    end for
end for
```

C. Observations and random setting

To simulate all possible CHP system operating status, the agent is

trained on different types of courses. We collected the real CHP operational data of CHP system from [40] and determine the upper and lower limits of four variables. Fig. 5. Illuminate the process of operating scenarios generation. In every episode a new course is generated within the upper and lower limits. Load, time-of-use price, wind turbine output and initial state of TST are considered as random variables, which could include all possible operating scenarios, including the following typical types: a) morning: high heat load with very low electricity load; b) midday: higher electricity load and lower heat load; c) evening: high electricity load and high heat load. There are examples in Section 6 which consist of a sequence of random instantiations of the above environment types within user-specified parameter ranges. Both the time-of-use price and the wind turbine output are generated randomly within the pre-defined range, which means that the algorithm can not only cope with load changes, but also cope with different energy prices and different renewable energy output.

When applied in economic dispatch in CHP system, the agent receives two sets of observation: 1) **decision variables**: A set of states information, containing the operating states of the heat only units, power only units, the CHP units, TST, and the Grid. The agents collect this data set in every timestep and then push it to the main PPO net. 2) **random variables**: A set of uncertain information, including the output of the wind turbine, the energy price and the load, initial state of TST. Hence, these data sets are generated stochastically in each iteration, due to the high randomness of the wind power, the energy price and the load. Then the action network and the value network compute the action set and $V_\pi(s_t)$, respectively, with the input of observations.

4. Case study

Two different test system were adopted in this paper to test the performance of proposed method. Test system I (Fig. 6) is presented in this paper for the first time. This system is a grid-connected CHP system with four decision variables (Gas Turbine(GT), Gas Boiler(GB), Power Grid and Thermal Storage Tank(TST)) and four random variables (Wind Turbine, Energy Price, Heat Load and Electricity Load) which was adopted to test whether our method could cope with variable operating states without recalculation. To simplify the problem, the units cost in this case was calculated by equation (12) ~ (14). Test system II [39] is considered to show the optimization quality and computation time of proposed method. This experiment aims to prove that the proposed algorithm is applicable to the optimization problems with stochastic environments, therefore, we do not use a simplified formula. The performance of the DPPO was compared with other state-of-art optimization methods. For the modelling of the MDP and DPPO algorithm, Python is selected as the programming language and pytorch is used as the deep-learning framework and all our code are publicly available.[1]

A. Test Case I

Random variables in test system I are presented in Table 6 in appendix. This new small test case with the effect of time-of-use price, random renewable energy and variable load was proposed to evaluate the performance of DRL algorithm for different operating states.

1. Details in one episode

First, we will show how agent work in one episode. The specific parameters are set as follows and Fig. 6 demonstrate the detailed adjustment process. To meet the device operating constraints, action range is [−0.02, 0.02]. By comparing TABLE 1 and TABLE 8 in the appendix, it can be found that the current situation has a lower electrical load level and a higher thermal load level, wind power and time-of-use electricity price are relatively low. As usual, user should increase the output of GB to meet the heat load without excessive electrical load. At the same time, due to the lower energy price, the heat storage tank should reserve some heat.

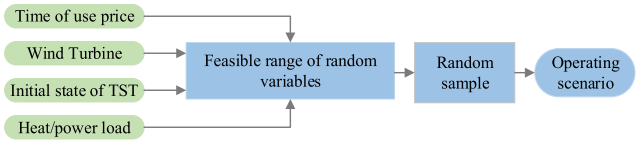

Fig. 5. Random scenarios generation process.

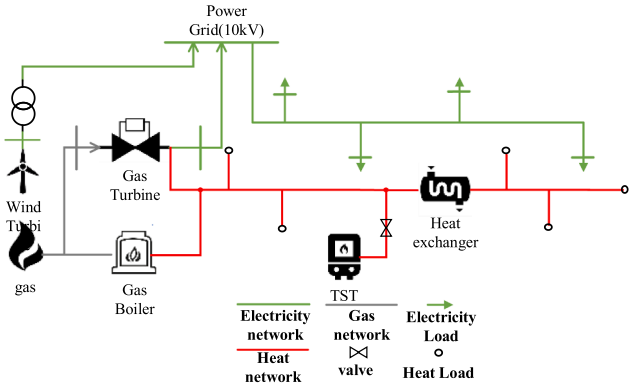

Fig. 6. Test CHP system 1.

Table 1
Random variables.

electricity load (kW)	wind (kW)	heat load (kW)	TOU price ($/kWh)
6000	700	9000	0.0627

Fig. 7 shows the actual adjustment process of DRL agent. It increases the output of GT and GB to meet the user load within the feasible domain and sells excess power to the grid to reduce operating costs. Furthermore, it finds that the energy price is lower currently, which is suitable for charging TST.

The strategy generated by DRL agent is in line with theoretical analysis and take the economy into account. In actual operation, the user only needs to input the detailed information of the current load, electricity price, etc., to get the control strategy which increases the flexibility and ease of use.

2. Hour-ahead Economic Dispatch in Different Operating States

It is essential to investigate whether the DPPO could deal with different emergency. To evaluate the DPPO on different tasks, the trained network was subjected to an hour-ahead CHP economic dispatch problem both in normal and extreme environment, i.e. wind turbine failure. The comparison of the dispatch strategies in the two operating status is shown in Fig. 8 (In the heat subplots, If the bottom of the histogram is less than 0, that part is used to charge the TST. In the electricity subplots, if the bottom of the histogram is lower than 0, that part means selling the electricity to the grid.) When there was no wind turbine output, the DPPO algorithm acquired a robust dispatch strategy compared with the normal strategy: 1) In the morning setting with the low electricity load and the high heat load, the DPPO managed to increase the output of the GT appropriately in order to slightly reduce the heat output of the GB, to use the stored heat in the TST and to sell the same amount of power to the grid. The strategy can be rationalized by the fact that the GT has the best economic efficiency in the system. By

[1] https://github.com/BeardHealth/Combined-Heat-and-Power-System-Economic-Dispatch.

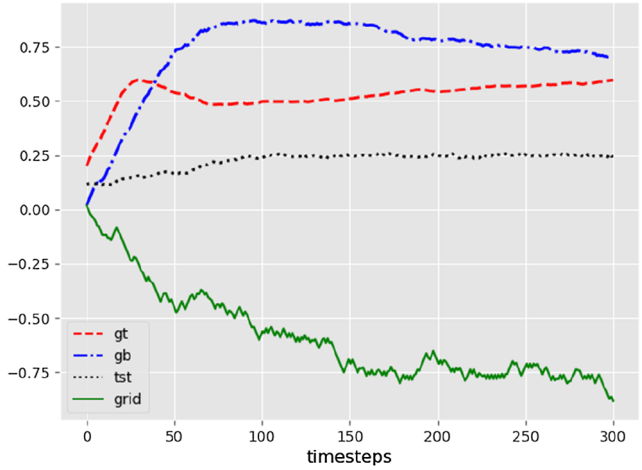

Fig. 7. Decision variables in one episode.

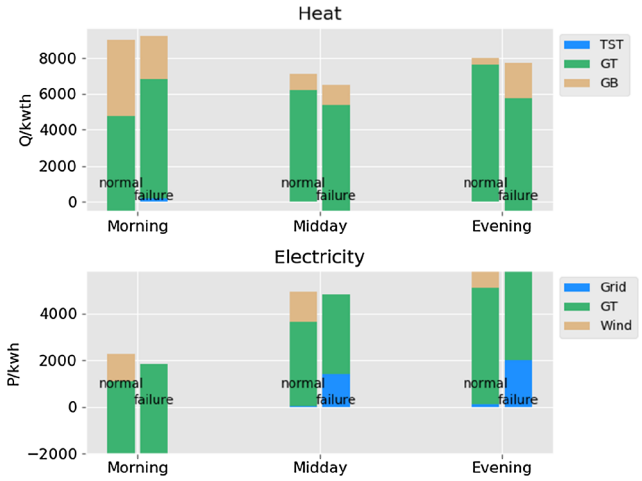

Fig. 8. Comparison of results under different operating scenarios.

Table 2
Detailed Results.

	Morning		Midday		Evening	
Condition	Normal	Failure	Normal	Failure	Normal	Failure
Cost/($)	594.04	685.02	588.28	691.50	810.9	897.65
Heat error	0.04	0.03	0.0036	0.04	0.007	0.0027
Electric error	0.009	0.03	0.012	0.03	0.064	0.037

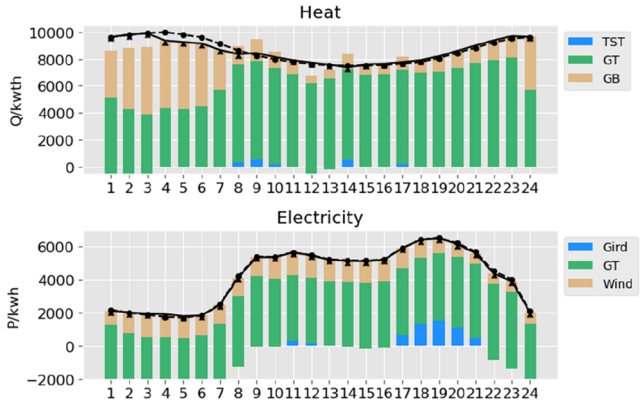

Fig. 9. Day-ahead dispatch strategies.

Table 3
Heat Load Mismatch and Cost.

	heat load error	Cost/$	
	DPPO	optimization methods	DPPO
0:00	0.0049	674.77	580.74
1:00	0.00035	592.65	564.15
2:00	0.002	595.18	557.27
3:00	0.025	528.56	554.05
4:00	0.008	529.33	544.08
5:00	0.009	567.91	546.31
6:00	0.009	650.84	526.36
7:00	0.033	702.68	616.83
8:00	0.03	800.41	737.72
9:00	0.03	747.36	694.55
10:00	0.025	709.5	708.3
11:00	0.026	706.25	651.15
12:00	0.032	690.84	646.26
13:00	0.034	762.13	657.53
14:00	0.02	868.03	726
15:00	0.02	691.24	657.45
16:00	0.03	753.85	731.47
17:00	0.02	770.38	807.27
18:00	0.03	799.15	847.59
19:00	0.0292	794.57	840.43
20:00	0.032	921.02	810.2
21:00	0.031	712.18	718.76
22:00	0.031	815.96	728.76
23:00	0.031	538.56	613.26
total	0.007	16924.029	16874.28

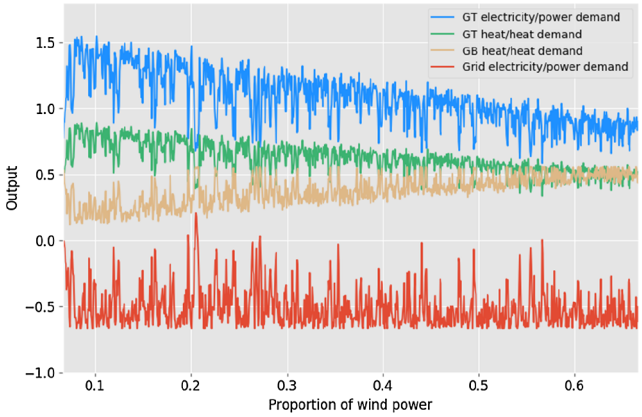

Fig. 10. Output under different wind power.

turning up the output of the GT, the gap in the electricity supply caused by the absence of the fan was accurately met and, simultaneously, excessive heat was generated to relieve the burden of the GB.

2) In the midday and the evening settings with the high enough electricity load and the declined heat load, the DPPO decided to reduce the output of the GT and to increase the power purchase from the grid, for optimal economic efficiency since it is cheaper to buy electricity from the grid rather than to generate. Meanwhile, the DPPO adjusted the output of the GB accordingly to meet the rest of the heat load and stored the excessive heat in the TST for future use, which further promotes the economic performance. In Table 2, the detailed result changes are demonstrated.

3. Day-ahead Economic Dispatch Problem

At last, the DPPO was applied to day-ahead generation dispatch problem, whose result was subsequently compared with that of the optimization method. Operating parameter settings are provided in Table 8 in appendix. Fig. 9 (In the heat subplots, the black dashed curve marked with point means heat load, the black solid curve marked with caret shows the heat generated by CHP system. In the electricity

Table 4
Cost function parameters of large test system IV.

Unit	α	β	γ	λ	ρ	P^{min}	P^{max}
Power only units							
1	0.00028	8.1	550	300	0.035	0	680
2	0.00056	8.1	309	200	0.042	0	360
3	0.00056	8.1	309	200	0.042	0	360
4	0.00324	7.74	240	150	0.063	60	180
5	0.00324	7.74	240	150	0.063	60	180
6	0.00324	7.74	240	150	0.063	60	180
7	0.00324	7.74	240	150	0.063	60	180
8	0.00324	7.74	240	150	0.063	60	180
9	0.00324	7.74	240	150	0.063	60	180
10	0.00284	8.6	126	100	0.084	40	120
11	0.00284	8.6	126	100	0.084	40	120
12	0.00284	8.6	126	100	0.084	55	120
13	0.00284	8.6	126	100	0.084	55	120

	a	b	c	d	e	f	Feasible region coordinates$[P^c, H^c]$
CHP units							
14	0.0345	14.5	2650	0.03	4.2	0.031	[98.8, 0], [81, 104.8], [215, 180], [247, 0]
15	0.0435	36	1250	0.027	0.6	0.011	[44, 0], [44, 15.9], [40, 75], [110.2, 135.5], [125.8, 32.4], [125.8, 0]
16	0.0345	14.5	2650	0.03	4.2	0.031	[98.8, 0], [81, 104.8], [215, 180], [247, 0]
17	0.0435	36	1250	0.027	0.6	0.011	[44, 0], [44, 15.9], [40, 75], [110.2, 135.5], [125.8, 32.4], [125.8, 0]
18	0.1035	34.5	2650	0.025	2.203	0.051	[20, 0], [10,40], [45, 55], [60, 0]
19	0.072	20	1565	0.02	2.34	0.04	[35, 0], [35,20], [90, 45], [90, 25], [105, 0]

	a	b	c	H^{min}	H^{max}
Heat only units					
20	0.038	2.0109	950	0	2695.2
21	0.038	2.0109	950	0	60
22	0.038	2.0109	950	0	60
23	0.052	3.0651	480	0	120
24	0.052	3.0651	480	0	120

subplots, the black dashed curve marked with point means ideal electricity load, the black solid curve marked with caret shows the electricity generated by CHP. Beside the illustration, if the bottom of the histogram is less than 0, that part is used to charge the TST. In the electricity subplots, if the bottom of the histogram is lower than 0, that part means selling the electricity to the grid.) shows the day-ahead economic dispatch strategies generated by the DPPO algorithm. The result demonstrates the following characteristics: 1) The GB output was time period dependent. For example, the output of GB was relatively higher when the heat load is higher from 0 am to 5 am and from 19 pm to 24 pm. 2) The GT undertook most of the electricity and heat loads. 3) The TST was used less frequently, for only 7 time periods, than other devices in the system. load demand reliably, with only acceptable variations in heat load across the time. The strategy learnt through DRL are like that through the optimization method, despite a slight numerical gap at every time step.

Besides, results imply that the DPPO succeeded in discovering the economical approaches on handling the load changes by choosing the GT as the main load bearer for its more economical performance and adjusting other decision variables based on the environment.

In addition to the qualitative analysis, the heat load error and the cost are listed in Table 3. The heat load error in DPPO, as shown in the second column in Table 3, was successfully kept at a very low level, indicating the user's comfort zone was well preserved, which approves the accuracy of the DPPO algorithm. The economic performance, i.e. the cost, of the two methods was also compared. The DPPO operated at lower costs for the majority time periods as the lower costs are highlighted in green in Table 3. Judging from the total cost of the day, the DPPO has the tiny advantage by making a 0.03% saving as costing of the optimization method.

Table 5
Optimal dispatch results for test system II using proposed method and other methods.

Output	CPSO [39]	TVAC-PSO [39]	EMA [21]	Ours
P1	680	538.5587	628.3171	630
P2	0	224.4608	299.1859	283.50
P3	0	224.4608	299.1859	283.50
P4	180	109.8666	109.8605	94.10
P5	180	109.8666	109.8605	94.10
P6	180	109.8666	109.8605	94.10
P7	180	109.8666	60	94.10
P8	180	109.8666	109.8605	94.10
P9	180	109.8666	109.8605	94.10
P10	50.4304	77.5210	40	72.56
P11	50.4304	77.5210	77.0195	72.56
P12	55	120	55	72.56
P13	55	120	55	72.56
P14	117.4854	88.3514	81	81
P15	45.9281	40.5611	40	44.09
P16	117.4854	88.3514	81	81
P17	45.9281	40.5611	40	44.09
P18	10.0013	10.0245	10	14.53
P19	42.1109	40.4288	35	35
H14	125.2754	108.9256	104.8002	104.8
H15	80.1174	75.4844	75	78.03
H16	125.2754	108.9256	104.8002	104.8
H17	80.1174	75.484	75	78.03
H18	40.0005	40.0104	40	41.94
H19	23.2322	22.4676	20	20
H20	415.9515	458.702	470.3996	453.50
H21	60	60	60	60
H22	60	60	60	60
H23	120	120	120	120
H24	120	120	120	120
Min cost ($)	59736.2635	58122.746	57829.4792	57990.15
Time(s)	13.34	7.84	–	≈ 0(after trained)

4. Variability and uncertainty of renewables

This subsection was designed to verify whether the proposed method can deal with renewable uncertainty. In order to eliminate the impact of the remaining variables, we assume that the electrical load and thermal load are constant (3000kWh and 9000kWth respectively), change the ratio of wind power to the electrical demand and observe the output of each device. All the results are shown in Fig. 10.

It can be found that with the increase of the proportion of wind power, the gas turbine continues to reduce its electricity power output, and at the same time, which will lead to insufficient heat supply and the gas boiler will increase its own heat output to make up for this shortfall. The results prove that the DRL method could cope with the instability of renewable energy.

Unlike model predictive control, DRL method is trained on a large amount of data to cover multiple operating scenarios. In the case of data prediction errors, users need to input real-time data into the trained model to get new results.

B. Test Case II

Test system II was a large test case widely used. This test system consists of 24 units where units 1–13 are power only units, units 14–19 are CHP units and 20–24 are heat only units. The total power demand of this case is 2350 MW and thermal demand is 1250 MWth. The total number of decision variables is 60. Test system data are presented in Tables 4 and 5 presents the optimal power and heat dispatches using the proposed method and other state-of the-art methods.

As it can be observed from Table 5, the proposed DRL method reaches a better solution comparing to CPSO and TVAC-PSO, while the result is not as good as EMA algorithm. In addition to the optimality of the solution, the proposed algorithm emphasizes adaptability, and the trained agent can handle multiple CHP operating states without re-calculation.

Using the trained neural network weights to calculate the dispatch strategy takes almost no time, which can help the CHP system generate control instructions under constant power and handle emergence failures.

5. Conclusion

We proposed and analyzed the DPPO algorithms for optimizing the stochastic CHP economic dispatch problem. We modeled the CHP economic dispatch problem as infinite-horizon discounted Markov decision process and set constrains to simulate the real environment. A form of reward signal was designed to lead the algorithm to the goal. We introduced proximal policy optimization methods that use multiple epochs of stochastic gradient ascent to perform each policy update and proved the convergence of the algorithm. Besides, we also used asynchronous advantage actor-critic to improve the convergence rate of the distributed framework, which subsequently improved the data collection speed, making it applicable to CHP settings where samples are expensive.

In the domain of the CHP economic dispatch, we successfully taught the agents to schedule the devices in the CHP system when chasing the economic optimum while satisfying load demand. Our analysis shows the DPPO algorithm could optimize the certain objective to a constraint.

In case study, the proposed algorithm was tested on two different cases with different characteristics. The obtained result demonstrates that the proposed algorithm can cope with more situations, have better time scale flexibility, and is easier to use on the basis of the same economic performance as the optimization method.

However, there are still shortcomings in solving economic dispatch problems with DRL methods. For examples, all optimization goals are reflected in the reward formula, which is not conducive to achieving multi-objective optimization, and optimization goals closer to the user's needs.

As future work, CHPED problem can be extended by considering more practical constraints and the DRL method could be enhanced for more complex optimization problems.

Funding

This work was supported in part by the Jiangsu Key Laboratory of Smart Grid Technology and Equipment, in part by the Science and Technology Project of Jiangsu Electric Power Company “Key Technologies of Smart Energy Service for City Energy Internet”, and in part by the National Natural Science Foundation of China under Grant 51807024.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Appendix A. Integration and ensemble models for *k*-alternative lineups

(See. Tables 6–8)

Table 6
DPPO Hyperparameters.

Hyperparameter	Value
Discountγ	0.9
Adam update rate for actor	0.0001
Adam update rate for critic	0.0005
Update step	10
Minibatch size	24
Clipping parametersϵ	0.2

Table 7
Device Operating Parameters.

name	parameters	value
Gas	Gas price	0.052$/kWh
GT	Upper limit	5000 kWh
	Lower limit	1000 kWh
	Effectiveness	0.3
	Thermoelectric ratio	2.3
GB	Upper limit	5000 kWh
	Lower limit	1000 kWh
	Effectiveness	0.8
TST	Capacity	5000 kWh
	Maximum charging power	1000 kWh
	Maximum discharging power	500 kWh
HE	Upper limit	12,000 kWh
	Lower limit	0 kWh
	Effectiveness	0.75
Grid	Capacity	5000 kWh
	Maximum purchase power	2000 kWh
	Maximum sell power	2000 kWh

Table 8
Day-ahead Environment Variables.

time interval	electricity load (kW)	wind (kW)	heat load (kW)	TOU price ($/kWh)
00:00–01:00	2178	875	9600	0.065
01:00–02:00	2009	1234.00	9792	0.065
02:00–03:00	1873	1390.00	9907.2	0.065
03:00–04:00	1755	1392.00	9984	0.065
04:00–05:00	1704	1336.00	9792	0.065
05:00–06:00	1839	1223.00	9600	0.065
06:00–07:00	2517	1173.00	9120	0.08
07:00–08:00	4211	1136.00	8640	0.08
08:00–09:00	5397	1158.00	8256	0.095
09:00–10:00	5735	1312.00	7968	0.095
10:00–11:00	5651	1369.00	7776	0.095
11:00–12:00	5481	1376.00	7603.2	0.08
12:00–13:00	5227	1315.00	7516.8	0.08
13:00–14:00	5176	1301.00	7488	0.08
14:00–15:00	5143	1343.00	7497.6	0.08
15:00–16:00	5227	1310.00	7545.6	0.08
16:00–17:00	5909	1208.00	7641.6	0.08
17:00–18:00	6417	1055.00	7776	0.095
18:00–19:00	6545	896	8064	0.095
19:00–20:00	6206	773	8448	0.095
20:00–21:00	5698	672	8832	0.095
21:00–22:00	4510	626	9216	0.095
22:00–23:00	3025	624	9504	0.065
23:00–24:00	2093	703	9600	0.065

References

[1] Nazari-Heris M, Mohammadi-Ivatloo B, Gharehpetian GB. A comprehensive review of heuristic optimization algorithms for optimal combined heat and power dispatch from economic and environmental perspectives. Renew Sustain Energy Rev 2018;81:2128–43. https://doi.org/10.1016/j.rser.2017.06.024.

[2] Gu W, Wang Z, Wu Z, Luo Z, Tang Y, Wang J. An online optimal dispatch schedule for CCHP microgrids based on model predictive control. IEEE Trans Smart Grid 2017;8(5):2332–42. https://doi.org/10.1109/TSG.2016.2523504.

[3] Tao Guo, M. I. Henwood, and M. van Ooijen, "An algorithm for combined heat and power economic dispatch," IEEE Trans. Power Syst., vol. 11, no. 4, pp. 1778–1784, Nov. 1996, 10.1109/59.544642.

[4] Dai Y, et al. A general model for thermal energy storage in combined heat and power dispatch considering heat transfer constraints. IEEE Trans. Sustain. Energy 2018;9(4):1518–28. https://doi.org/10.1109/TSTE.2018.2793360.

[5] Dai Y, et al. Dispatch model of combined heat and power plant considering heat transfer process. IEEE Trans. Sustain. Energy 2017;8(3):1225–36. https://doi.org/10.1109/TSTE.2017.2671744.

[6] Lu S, Gu W, Zhou J, Zhang X, Wu C. Coordinated dispatch of multi-energy system with district heating network: modeling and solution strategy. Energy 2018;152:358–70. https://doi.org/10.1016/j.energy.2018.03.088.

[7] Bahrami S, Sheikhi A. From demand response in smart grid toward integrated demand response in smart energy hub. IEEE Trans Smart Grid 2015:1. https://doi.org/10.1109/TSG.2015.2464374.

[8] M. Geidl and G. Andersson, "A modeling and optimization approach for multiple energy carrier power flow," in 2005 IEEE Russia Power Tech, St. Petersburg, Russia, 2005, pp. 1–7, 10.1109/PTC.2005.4524640.

[9] Chicco G, Mancarella P. Matrix modelling of small-scale trigeneration systems and application to operational optimization. Energy 2009;34(3):261–73. https://doi.org/10.1016/j.energy.2008.09.011.

[10] Mancarella P. MES (multi-energy systems): an overview of concepts and evaluation models. Energy 2014;65:1–17. https://doi.org/10.1016/j.energy.2013.10.041.

[11] Sondergren C, Ravn HF. A method to perform probabilistic production simulation involving combined heat and power units. IEEE Trans Power Syst 1996;11(2):1031–6. https://doi.org/10.1109/59.496191.

[12] Lahdelma R, Hakonen H. An efficient linear programming algorithm for combined heat and power production. Eur J Oper Res 2003;148(1):141–51. https://doi.org/10.1016/S0377-2217(02)00460-5.

[13] Makkonen S, Lahdelma R. Non-convex power plant modelling in energy optimisation. Eur J Oper Res 2006;171(3):1113–26. https://doi.org/10.1016/j.ejor.2005.01.020.

[14] Qiu H, Zhao B, Gu W, Bo R. Bi-Level two-stage robust optimal scheduling for AC/DC hybrid multi-microgrids. IEEE Trans Smart Grid 2018;9(5):5455–66. https://doi.org/10.1109/TSG.2018.2806973.

[15] Rong A, Lahdelma R. An efficient envelope-based branch and bound algorithm for non-convex combined heat and power production planning. Eur J Oper Res 2007;183(1):412–31. https://doi.org/10.1016/j.ejor.2006.09.072.

[16] Rooijers FJ, van Amerongen RAM. Static economic dispatch for co-generation systems. IEEE Trans Power Syst 1994;9(3):1392–8. https://doi.org/10.1109/59.336125.

[17] Wong KP, Algie C. Evolutionary programming approach for combined heat and power dispatch. Electr Power Syst Res 2002;61(3):227–32. https://doi.org/10.1016/S0378-7796(02)00028-7.

[18] Nazari-Heris M, Mohammadi-Ivatloo B, Asadi S, Geem ZW. Large-scale combined heat and power economic dispatch using a novel multi-player harmony search method. Appl Therm Eng 2019;154:493–504. https://doi.org/10.1016/j.applthermaleng.2019.03.095.

[19] Mellal MA, Williams EJ. Cuckoo optimization algorithm with penalty function for combined heat and power economic dispatch problem. Energy 2015;93:1711–8. https://doi.org/10.1016/j.energy.2015.10.006.

[20] Zou D, Li S, Kong X, Ouyang H, Li Z. Solving the combined heat and power economic dispatch problems by an improved genetic algorithm and a new constraint handling strategy. Appl Energy 2019;237:646–70. https://doi.org/10.1016/j.apenergy.2019.01.056.

[21] Ghorbani N. Combined heat and power economic dispatch using exchange market algorithm. Int J Electr Power Energy Syst 2016;82:58–66. https://doi.org/10.1016/j.ijepes.2016.03.004.

[22] M. Rahmani-andebili and G. K. Venayagamoorthy, "Combined emission and economic dispatch incorporating demand side resources," in 2015 Clemson University Power Systems Conference (PSC), Clemson, SC, USA, 2015, pp. 1–6, 10.1109/PSC.2015.7101676.

[23] Zhou S, Zou F, Wu Z, Gu W, Hong Q, Booth C. A smart community energy management scheme considering user dominated demand side response and P2P trading. Int J Electr Power Energy Syst 2020;114:105378https://doi.org/10.1016/j.ijepes.2019.105378.

[24] M. Rahmani-Andebili and G. K. Venayagamoorthy, "Stochastic Optimization for Combined Economic and Emission Dispatch with Renewables," in 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 2015, pp. 1252–1258, 10.1109/SSCI.2015.179.

[25] Rahmani-Andebili M. Dynamic and adaptive reconfiguration of electrical distribution system including renewables applying stochastic model predictive control. IET Gener Transm Distrib 2017;11(16):3912–21. https://doi.org/10.1049/iet-gtd.2016.1549.

[26] D. P. Bertsekas, "Dynamic Programming and Optimal Control 3rd Edition, Volume II," p. 233.

[27] Rocchetta R, Bellani L, Compare M, Zio E, Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids. Appl Energy 2019;241:291–301. https://doi.org/10.1016/j.apenergy.2019.03.027.

[28] S. Zhou, Z. Hu, and W. Gu, "Artificial intelligence based smart energy community management: A reinforcement learning approach," CSEE J. Power Energy Syst., 2019, 10.17775/CSEEJPES.2018.00840.

[29] Jasmin EA, Imthias Ahamed TP, Jagathy Raj VP. Reinforcement learning approaches to economic dispatch problem. Int J Electr Power Energy Syst 2011;33(4):836–45. https://doi.org/10.1016/j.ijepes.2010.12.008.

[30] N. Heess et al., "Emergence of Locomotion Behaviours in Rich Environments," ArXiv170702286 Cs, Jul. 2017.

[31] V. Mnih et al., "Asynchronous Methods for Deep Reinforcement Learning," ArXiv160201783 Cs, Feb. 2016.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy

Optimization Algorithms," ArXiv170706347 Cs, Jul. 2017.
[33] Liu X, Wu J, Jenkins N, Bagdanavicius A. Combined analysis of electricity and heat networks. Appl Energy 2016;162:1238–50. https://doi.org/10.1016/j.apenergy.2015.01.102.
[34] Gu W, et al. Residential CCHP microgrid with load aggregator: Operation mode, pricing strategy, and optimal dispatch. Appl Energy 2017;205:173–86. https://doi.org/10.1016/j.apenergy.2017.07.045.
[35] Silver D, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 2018;362(6419):1140–4. https://doi.org/10.1126/science.aar6404.
[36] Basu M. Combined heat and power economic dispatch by using differential evolution. Electr Power Compon Syst 2010;38(8):996–1004. https://doi.org/10.1080/15325000903571574.
[38] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ArXiv14126980 Cs, Dec. 2014.
[39] Mohammadi-Ivatloo B, Moradi-Dalvand M, Rabiee A. Combined heat and power economic dispatch problem solution using particle swarm optimization with time varying acceleration coefficients. Electr Power Syst Res 2013;95:9–18. https://doi.org/10.1016/j.epsr.2012.08.005.
[40] Open Power System data. 2019. Data Package Time Series. Version 2018-03-13. URL https://data.open-power-system-data.org.