Optimal Scheduling Framework of Electricity-Gas-Heat Integrated Energy System Based on Asynchronous Advantage Actor-Critic Algorithm

JIAN DONG¹, HAIXIN WANG², (Member, IEEE), JUNYOU YANG¹, (Member, IEEE), XINYI LU², LIU GAO¹, AND XIRAN ZHOU¹

¹School of Electrical Engineering, Shenyang University of Technology, Shenyang 110870, China
²Electricity Intensive Control Department State Grid Liaoning Marketing Service Center, Shenyang 110000, China

Corresponding author: Junyou Yang (junyouyang@sut.edu.cn)

This work was supported in part by the Scientific Research Fund of Liaoning Provincial Education Department under Grant LQGD2019005, and in part by the Doctoral Start-up Foundation of Liaoning Province under Grant 2020-BS-141.

ABSTRACT The optimal scheduling of integrated energy system (IES) can improve energy efficiency and economic operation. However, the existing scheduling methods cannot accurately handle the dynamic changes of supply and demand sides in the electricity-gas-heat IES thanks to their power uncertainties. To tackle this problem, an optimal dispatch framework based on the asynchronous advantage actor-critic (A3C) method of IES is proposed. Firstly, we describe the dispatch problem of IES with multiple uncertainties as Markov decision process (MDP) according to the corresponding mathematical models and constraints. Then, the dispatch framework based on A3C is developed to optimize the control decision for supply and demand sides by asynchronous learning of agents and to reduce the relevance of parameters update of neural networks by the multi-agent utilization of the central processing unit (CPU) multi-threading function. Finally, our proposed methods are verified by the simulations. Compared with the previous optimization algorithms, the training time of the proposed methods is shortened by 37% and 30%, and the daily average operating cost is reduced by 3%, 5.2% and 8.7%.

INDEX TERMS Integrated energy system (IES), asynchronous advantage actor-critic (A3C), optimal scheduling, deep reinforcement learning (DRL), Markov decision process (MDP).

NOMENCLATURE

Acronyms

IES Integrated energy system.
A3C Asynchronous advantage actor-critic.
MDP Markov decision process.
CPU Central processing unit.
DRL Deep reinforcement learning.
WT Wind turbines.
PV Photovoltaic.
EB Electric boiler.
MG Microgrid.
RL Reinforcement learning.
DQN Deep Q network.
SOC State-of-charge.

Sets and Parameters

\( P_{CHP,t} \) The electric and heat power of CHP at time \( t \).
\( G_{CHP,t} \) The gas power consumed by CHP at time \( t \).
\( \eta_{CHP} \) The power generation efficiency of CHP.
\( \eta_{loss} \) The energy loss coefficient of CHP.

DDPG Deep deterministic policy gradient.
CHP Combined heat and power.
BS Battery storage.
HS Heat storage.
GS Gas storage.
PtG Power to gas.
PCE Policy cross entropy.
\( P_{\text{min}, \text{CHP}}, P_{\text{max}, \text{CHP}} \)

The lower and upper limits of electric power of CHP.

\( H_{\text{min}, \text{CHP}}, H_{\text{max}, \text{CHP}} \)

The lower and upper limits of heat power of CHP.

\( \Delta P_{\text{min}, \text{CHP}}, \Delta P_{\text{max}, \text{CHP}} \)

The lower and upper limits of electric power ramp of CHP.

\( \Delta H_{\text{min}, \text{CHP}}, \Delta H_{\text{max}, \text{CHP}} \)

The lower and upper limits of heat power ramp of CHP.

\( G_{\text{PtG}, t} \)

The natural gas output power of the PtG at time slot \( t \).

\( \eta_{\text{PtG}} \)

The conversion efficiency of the PtG.

\( P_j, t \)

The electrical power consumed by the PtG equipment.

\( P_{\text{min}, \text{PtG}}, P_{\text{max}, \text{PtG}} \)

The lower and upper limits of PtG.

\( \Delta P_{\text{min}, \text{PtG}} \) and \( \Delta P_{\text{max}, \text{PtG}} \)

The lower and upper limits of climbing power of PtG.

\( H_{\text{EB}, t} \)

The thermal power output by EB at time \( t \).

\( P_{\text{EB}, t} \)

The power consumption of EB at time \( t \).

\( \eta_{\text{EB}} \)

The conversion efficiency of the electric heating.

\( \mu_{\text{Loss}} \)

The heat loss rate of EB.

\( H_{\text{EB}, \text{min}}, H_{\text{EB}, \text{max}} \)

The lower and upper limits of the heat power output of EB.

\( \text{SOC}_{x, t+1}, \text{SOC}_{x, t} \)

The SOC of the energy storage device \( x \) during time \( t \) and \( t+1 \).

\( P_{x, \text{ch}, t}, P_{x, \text{disch}, t} \)

The energy storage device \( x \) in the charging and discharging power at time \( t \).

\( \text{SOC}_{x, 1}, \text{SOC}_{x, T} \)

The initial and final energy of the energy storage in a scheduling period.

\( Q_x \)

The capacity of the energy storage \( x \).

\( \eta_{x, \text{ch}, \text{EB}}, \eta_{x, \text{disch}} \)

The charging and discharging efficiency of the energy storage device \( x \).

\( a_{x, \text{ch}, t}, a_{x, \text{disch}} \)

The charging and discharging state parameters of energy.

\( P_{E, t} \)

The power purchased from the external power grid at time \( t \).

\( G_{G, t} \)

The power purchased from the gas grid at time \( t \).

\( P_{\text{Load}, t}, G_{\text{Load}, t}, H_{\text{Load}, t} \)

The electric, gas, and heat loads at time \( t \).

\( C_E \)

The electricity purchase cost of the system.

\( C_G \)

The gas purchase cost of the system.

\( C_O \)

The operation and maintenance costs of each unit.

\( P_i, t \)

The electric power produced by unit \( i \) at time \( t \).

\( P_j, t \)

The electric power consumed by unit \( j \) at time \( t \).

\( G_m, t \)

The gas power produced by unit \( m \) at time \( t \).

I. INTRODUCTION

As an important carrier for the development of the energy internet, integrated energy system (IES) maintains a significant effect in the coordinated operation of multiple energy sources and the improvement of energy utilization efficiency [1], [2]. However, with the couplings of multiple energy sources, the uncertainties of wind turbines (WT) and photovoltaic (PV) outputs and load demands challenge the optimal operation of IES [3], [4].

The main research methods about the optimal dispatch of IES are divided into model-based traditional methods and model-free DRL methods [5]. Traditional model-based methods need to establish a detailed system dynamics model and solve the problem through a solver [6]. Liu et al. [7] used electric boilers (EB), heat storage and batteries to establish a coordination model considering network transmission characteristics for the regional electricity-heat IES, and used a double \( \lambda \) iterative algorithm to solve the model. Ehsan and Yang [8] considered the constraints of multiple types of power generation and energy storage, established an operation optimization model for the comprehensive utilization and collaborative optimization of multi-energy demand. Wang et al. [9] considered the uncertainty of the load in the combined heat and power microgrid (MG) and optimize the economic dispatch of the time system. For the uncertainty of wind power forecasting, Qaddan et al. [10] used stochastic programming to establish an optimal scheduling model. However, the above methods are limited to the day-ahead scheduling plan, and cannot be adjusted according to the dynamic changes in load demands and renewable energy outputs.

In response to these problems, model predictive control (MPC) adopts a rolling optimization strategy, which utilizes online repeated optimization calculations and rolling...
implementations to obtain dynamic control performance and solve the above problems [11]–[14]. Li et al. [12] proposed a strategy that combines the advantages of combining two-stage stochastic planning and backward control. Petrollese et al. [15] proposed the MPC-based control strategy for the optimal management of energy storage systems in high-permeability renewable energy MGs. Wang et al. [16] proposed an IES scheduling method based on MPC dynamic time intervals. Although the above-mentioned methods have achieved remarkable results in the optimization and decision-making of multi-energy systems, the above-mentioned model-based methods require the establishment of detailed models of IES, and the selection of parameters depends to a large extent on expert knowledge.

With the rapid development of artificial intelligence technology, reinforcement learning (RL) has received more and more attention in the optimization control of power systems in recent years [17]–[19]. The RL model realizes the gradual accumulation of experience and improvement of strategies through continuous interaction with the environment. Especially the deep reinforcement learning (DRL) model combines with deep neural network and RL has better adaptive learning ability and optimization decision-making ability for non-convex and nonlinear problems [20]–[22].

Focusing on electricity-gas-heat IES optimal scheduling problems, Ji et al. [23] proposed real-time energy management for MGs based on DRL. MG energy management is modeled as a Markov decision process (MDP), which aims to minimize daily operating costs and is solved by a deep Q network (DQN) algorithm. Brandi et al. [24] use the deep Q learning method to control the indoor temperature of the building. Gorostiza and Gonzalez-Longatt [25] proposed a DRL-based state-of-charge (SOC) management controller for energy storage systems, use the deep deterministic policy gradient (DDPG) method for training. The above literature provides a research foundation for the application of DRL methods in IES. However, the above methods control the continuous actions in IES through discrete time, resulting in inaccuracy control. In addition, the above literature uses a centralized control method for optimization. When faced with the collaborative optimization of various forms of energy and energy storage in IES, the model training takes a long time and is easy to non-convergence.

In this paper, we handle an IES including the renewable energy, combined heat and power (CHP), power to gas (PtG), electric boiler (EB), battery storage (BS), heat storage (HS) and gas storage (GS) without considering system dynamics model and uncertain factor modeling. However, achieving the above goals is challenging for the following reasons. Firstly, the system contains a variety of energy sources, storage devices, and conventional methods used in optimization are inefficient and cannot meet the needs of real-time regulation. Secondly, the system contains a large number of determinants such as electricity price, wind power output, photovoltaic output and electricity-gas-heat load demand. Thirdly, the system exists in a variety of energy storage device-related coupled operation constraints, but the energy storage output at the current moment affects future decisions. In response to the above problems, this paper proposes an A3C-based optimal scheduling method for the electricity-gas-heat IES, which realizes the economic operation of the system by making decisions on the units and energy storage in each period. The main contributions of this paper are as follows:

1) We describe the integrated energy system mathematical models and optimization problem of the electricity-heat-gas network as an MDP including the action space, state space and reward mechanism of the agent.

2) An A3C-based optimal scheduling framework for the electricity-gas-heat IES is proposed, which uses the multi-threaded function of the computer CPU to control multiple agents in parallel and asynchronously.

3) Compared with the previous optimization algorithms, the training time of the proposed methods is shortened by 37% and 30%, and the daily average operating cost is reduced by 3% and 8.7%.

The rest of the paper is organized as follows: Section II describes the mathematical models of IES. Section III proposes the A3C-based optimal scheduling framework of IES. Section IV discusses simulation results. Section V draws conclusions.

II. MATHEMATICAL MODELS OF IES

The structure of IES is shown in Fig. 1. The electric network includes the external power grid, WT, PV, ES and power loads. The gas network includes natural gas stations, GS and gas loads. The heat network includes HS and heat loads, and energy conversion equipment mainly includes a CHP, PtG and an electric heating furnace. In the following section, the model of each unit in IES is provided.

A. ENERGY CONVERSION EQUIPMENT MODEL

1) CHP MODEL

The CHP unit can provide heat and electricity for the system [26]. The mathematical model of CHP can be described by (1) and (2).

\[ P_{CHP,1} = G_{CHP,1} \cdot \eta_{CHP}. \]  
\[ H_{CHP,1} = G_{CHP,1} \cdot (1 - \eta_{CHP} - \eta_{loss}). \]
The power and climbing constraints of CHP are shown in (3)-(6).

\[
P_{\text{min,CHP}} \leq P_{\text{CHP},t} \leq P_{\text{max,CHP}},
\]

\[
H_{\text{min,CHP}} \leq H_{\text{CHP},t} \leq H_{\text{max,CHP}},
\]

\[
\Delta P_{\text{min,CHP}} \leq P_{\text{CHP},t} - P_{\text{CHP},t-1} \leq \Delta P_{\text{max,CHP}},
\]

\[
\Delta H_{\text{min,CHP}} \leq H_{\text{CHP},t} - H_{\text{CHP},t-1} \leq \Delta H_{\text{max,CHP}}.
\]

2) PTG MODEL

The power-to-gas unit converts the electric power into gas power, which can transform renewable energy into natural gas for large-scale, long-term storage. The relationship between the natural gas output and the electric power consumption can be described by (7).

\[
G_{\text{PtG},t} = \eta_{\text{PtG}} \cdot P_{\text{PtG},t}.
\]

The power and constraints of PtG are shown in (8) and (9).

\[
P_{\text{min,PG}} \leq P_{\text{PG},t} \leq P_{\text{max,PG}},
\]

\[
\Delta P_{\text{min,PG}} \leq P_{\text{PG},t} - P_{\text{PG},t-1} \leq \Delta P_{\text{max,PG}}.
\]

3) ELECTRIC BOILER

The EB converts electric power into heat power, which is used to supplement the remaining heat load demand when CHP heat supply is insufficient. The relationship between the power consumption and the heat supply of EB can be described by (10).

\[
H_{\text{EB},t} = \eta_{\text{EB}} \cdot (1 - \mu_{\text{Load}}) \cdot P_{\text{EB},t}.
\]

The operating constraint of EB is shown in (11).

\[
H_{\text{min,EB}} \leq H_{\text{EB},t} \leq H_{\text{EB},t}.
\]

B. ENERGY STORAGE EQUIPMENT MODEL

Adding various energy storage equipment to the IES can alleviate the uncertainty of WT and PV output, stabilize load fluctuations, and increase the economic efficiency of the IES. The mathematical model of energy storage can be described as (12)-(14).

\[
SOC_{x,t+1} = SOC_{x,t} + a_{x,ch,t} \eta_{x,ch} \frac{P_{x,ch,t}}{Q_x} \Delta t + a_{x,disch,t} \frac{P_{x,disch,t}}{\eta_{x,disch}} \Delta t,
\]

where \( x \) is the energy category, which represents battery storage (BS), gas storage (GS), and heat storage (HS), \( a_{x,ch,t} = 1 \) indicates that the energy storage device operates in charging state at time \( t \), and \( a_{x,disch,t} = 1 \) indicates that the energy storage \( x \) operates in the discharge state.

The charge-discharge power and capacity constraints of energy storage units are shown in (15)-(18).

\[
SOC_{x,t}^\min \leq SOC_{x,t} \leq SOC_{x,t}^\max,
\]

\[
0 \leq P_{x,ch,t} \leq P_{x,ch,t}^\max,
\]

\[
0 \leq P_{x,disch,t} \leq P_{x,disch,t}^\max.
\]

C. SYSTEM BALANCE CONSTRAINT

To satisfy the electricity-gas-heat load demand in each period, the system balance constraints need to meet the constraints as shown in (19)-(21).

\[
P_{E,t} + P_{PV,t} + P_{\text{Wind},t} + P_{\text{CHP},t} + P_{\text{BS,disch},t} - P_{\text{BS,ch},t} - P_{\text{PG},t} - P_{\text{EB},t} = P_{\text{Load},t},
\]

\[
G_{G,t} + G_{\text{PtG},t} + G_{\text{GS,disch},t} - G_{\text{CHP},t} - P_{\text{GS,ch},t} = G_{\text{Load},t},
\]

\[
H_{\text{CHP},t} + H_{\text{EB},t} + P_{\text{HS,disch},t} - P_{\text{HS,ch},t} = H_{\text{Load},t}.
\]

D. OBJECTIVE FUNCTION

The operation optimization goal of IES is to improve the economic benefits while meeting the constraints of the safety operation of the system.

The optimization objective function \( F \) of IES is to save the operating cost of the system. According to the IES balance constraints (19)-(21), the electricity and gas purchase of the system is determined by the output of CHP, PtG, EB, BS, GS and HS. The output constraints of each unit have been described in Section II-A. The decision variables of the objective function are determined by electricity purchase power \( P_{E,t} \), gas purchase power \( G_{G,t} \), CHP output power \( P_{\text{CHP},t} \), PtG electrical power \( P_{\text{PtG},t} \), EB electrical power \( P_{\text{EB}} \), and energy storage equipment charge and discharge power \( P_{x,ch,t}, P_{x,disch,t} \). The optimization objective function \( F \) can be formulated as:

\[
F = \min (C_E + C_G + C_O),
\]

\[
s.t. \quad C_E = \sum_{t=1}^{T} \varepsilon_{E,t} P_{E,t} \Delta t,
\]

\[
C_G = \sum_{t=1}^{T} \varepsilon_{G,t} G_{G,t} \Delta t,
\]

\[
C_O = \sum_{t=1}^{T} \varepsilon_{O,t} P_{O,t},
\]

\[
P_{E,t} = P_{\text{Load},t} - \sum_{i\in \text{A}} P_{i,t} + \sum_{j\in \text{B}} P_{j,t},
\]

\[
G_{G,t} = G_{\text{Load},t} - \sum_{m\in \text{C}} P_{m,t} + \sum_{n\in \text{D}} P_{n,t},
\]

where \( \text{A} \) is the set of electric power generation units, \( \text{B} \) is the set of electric power consumption units, \( \text{C} \) is the set of gas power generation units, \( \text{D} \) is the set of gas power consumption units.

III. PROPOSED A3C-BASED OPTIMAL SCHEDULING FRAMEWORK OF IES

To address the limitations of previous approaches, this paper proposes the framework of A3C to solve the optimal
scheduling problem of IES with uncertain problems. In this section, the optimal scheduling problem of IES is transformed into an MDP, and then the principles of the A3C-based framework are explained.

A. MARKOV PROCESS CONVERSION

RL provides the control decision over a sequence of time steps, which can be described as MDP [20]. The five elements of MDP are state space \( S \), action space \( A \), transfer function \( T \), reward function \( R \) and discount coefficient \( \gamma \). Among them, the transfer function describes the probability of the agent changing from state \( s_t \) to \( s_{t+1} \) under a given action: \( T : S \times A \times S \rightarrow [0, 1] \). In this paper, a discrete-time finite MDP is used to describe an integrated energy system economic scheduling problem. Integrated energy system as an agent which learns control strategies through repeated interactions with the environment. The units of the integrated energy system are treated as the environment. Fig. 2 illustrates the interactions between the agent and the environment in the MDP.

In the optimization process, the agent obtains the state \( s_t \) at the beginning of each time slot. In each optimization time slot, the agent chooses the corresponding action from the possible action set \( a_t \) according to the current state \( s_t \) and the policy function \( \pi \). The agent receives the next state \( s_{t+1} \) information and the reward value \( R_t \), and then repeats the above steps until the end of optimization. The ultimate optimization goal of the agent is to maximize the accumulated reward value.

1) SYSTEM STATE

The state space \( S \) is referenced to make decisions for the agent at each period. \( S \) includes the controllable state space \( S_C \) and the uncontrollable state space \( S_{UC} \). Among them, the controllable state space contains environmental variables that can be directly/indirectly controlled by the agent, such as the SOC of BS \( S_{BS,t} \), the SOC of GS \( S_{GS,t} \), and the SOC of HS \( S_{HS,t} \). The uncontrollable state space contains variables that the agent cannot control, mainly including the wind power generation \( P_{Wind,t} \), the photovoltaic power generation \( P_{PV,t} \), and the electric load \( P_{Load,t} \), gas load \( G_{Load,t} \), heat load \( H_{Load,t} \). Therefore, the state space can be defined as:

\[
\begin{align*}
\text{State } s_t & \in S = S_C \times S_{UC}, \\
\text{State } s_t & = (S_{BS,t}, S_{GS,t}, S_{HS,t}, P_{Wind,t}, P_{Load,t}, G_{Load,t}, H_{Load,t}, t).
\end{align*}
\]

2) SYSTEM ACTION

The action space of the agent includes CHP output \( P_{CHP,t} \), the gas power output of PtG \( P_{PtG,t} \), the electric power output of BS \( P_{BS,t} \), the gas power output of GS \( P_{GS,t} \), the heat power output of HS \( P_{HS,t} \), the heat power output of EB \( H_{EB,t} \). The action space can be defined as:

\[
\begin{align*}
a_t & = (P_{CHP,t}, P_{PtG,t}, P_{BS,t}, P_{GS,t}, P_{HS,t}, H_{EB,t}).
\end{align*}
\]

3) REWARD FUNCTION

The reward function is set to guide the current action of the agent to obtain the accumulated maximum reward. The reward function includes the cost of purchasing electric power from the external power grid, the cost of purchasing gas power from the gas grid and the cost of operation and maintenance of each unit. Furthermore, to accelerate the convergence of the RL, the penalty cost of the agent action exceeding the limit of each unit is added to the reward function. The penalty function is designed to generate a negative reward value when the agent’s action exceeds the constraints at each time. Thus, the agent’s action value is constrained by penalty function to a specified range in the training process.

The penalty cost of the agent’s action is represented by \( \delta \), mainly includes two parts, agent action amplitude penalty and agent action change rate punishment. Among them, the action amplitude and change rate penalty of the agent are determined by the output constraints and ramp constraints of each equipment in the IES.

\[
\delta_t = \sum_{i=1}^{T} \sum_{a_t \in A} \left[ \varepsilon_{ur,t} \max (a_{i,t+1} - a_{i,t} - a_{i,ur,max}, 0) \\
+ \varepsilon_{dr,t} \max (a_{i,t} - a_{i,t+1} - a_{i,dr,min}, 0) \\
+ \varepsilon_{ur,t} \max (a_{i,t+1} - a_{i,t} - a_{i,ur,max}, 0) \\
+ \varepsilon_{da,t} \max (a_{i,t+1} - a_{i,t} - a_{i,da,min}, 0) \right].
\]

In the training progress, to reduce the penalty, the action change of the agent will be gradually constrained to within a certain range.

The reward, that the environment model feeds back to the agent, is a component of the above objective function at each time. Since the RL agent adopts the form of maximizing the cumulative return, the objective function in the reward takes a negative value as shown in (32).

\[
R_t = - (C_E + C_G + C_O + \delta_t) + r_0.
\]
CPU thread of a device [28]. Each agent obtains network parameters from the global network through the pull function before running. The agents on each thread interact independently of the environment and update their parameters. After the operation is completed, the parameters are synchronized to the global network through the push function [29]. The global network does not need to be trained and only stores Actor-critic structure parameters. The A3C algorithm can improve operating efficiency through multithreading. Different agent exploration strategies are independent of each other, which is conducive to the convergence of the system.

At each time $t$, the actor takes action $a_t$ according to the policy function $\pi(a_t|s_t; \theta)$ to make the environment state change from $s_t$ to $s_{t+1}$, and the obtained reward is $r_t(a_t, s_t)$. The optimized trajectory $\tau$ is expressed as $s_1, a_1, r_1, s_2, a_2, r_2, \cdots s_t, a_t, r_t, \cdots$. The obtained reward $R(\tau)$ of the trajectory $\tau$ is the sum of the obtained rewards at each stage and can be obtained by (33).

$$R(\tau)=r_1+y r_2+y^2 r_3+\cdots+y^{T-i} r_T = \sum_{i=1}^{T} y^{T-i} r_i(a_t, s_t),$$

where $\gamma$ is the discount factor, $0 \leq \gamma \leq 1$.

In the IES optimal scheduling, a trajectory in one day is divided into 24 h. The probability of trajectory $\tau$ is calculated by (32).

$$p_\pi(\tau) = p(s_1)p_\pi(a_1|s_1)p(s_2|s_1,a_1)p_\pi(a_2|s_2)\cdots p(s_T|s_{T-1},a_{T-1})p(s_{T+1}|s_1,a_T)$$

$$= p(s_1) \prod_{t=1}^{T} p_\pi(a_t|s_t) p(s_{t+1}|s_t,a_t).$$

The expectation reward of theactor under the policy $\pi$ can be calculated (33).

$$R_\theta = \sum_{\tau} R(\tau)p_\theta(\tau) = \mathbb{E}_{\tau \sim p_\theta(\tau)} [R(\tau)].$$

Therefore, the purpose of optimization is to guide the actor to take actions to maximize the expected reward value. The expectation reward has been given by (33), so the policy $\pi$ can be updated by solving the network parameter $\theta$ gradient.

In the gradient solution process, the log function derivation and the average value of $N$ sampling methods are used to approximate the expectation reward by (34).

$$\nabla R_0 = \sum_{\tau} R(\tau) \nabla p_\theta(\tau) = \sum_{\tau} R(\tau)p_\theta(\tau) \nabla \log p_\theta(\tau)$$

$$= \mathbb{E}_{\tau \sim p_\theta(\tau)} [R(\tau) \nabla \log p_\theta(\tau)]$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} R(\tau^n) \nabla \log p_\theta(\tau^n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{T} R(\tau^n) \nabla \log p_\theta(a^n_i|s^n_i).$$

Critic network mainly uses the value function to evaluate the performance of the policy, which is represented by $V(s_t; \omega)$, where $\omega$ is the critic network parameter. It can be obtained by (35).

$$V_\pi(s_t; \omega) = \mathbb{E} [R(\tau)|s_t; \omega].$$

The action value function $Q_\pi(a_t, s_t; \omega)$ represents the expected sum of rewards when starting in the state $s_t$, taking action $a_t$, and from then on following the policy $\pi$.

$$Q_\pi(a_t, s_t; \omega) = \mathbb{E} [R(\tau)|a_t, s_t; \omega].$$

The advantage function $A_\pi(a_t, s_t; \omega)$ in (39) can evaluate the effects of the actor taking an action $a_t$, compared to the average at the current state $s_t$. When the action value function $Q_\pi(a_t, s_t; \omega)$ is greater than the value function $V_\pi(s_t; \omega)$, the value of the advantage function is positive, indicating that it is advantageous to take the current action in this state, the policy improvement is positive, and the policy network parameters are updated in the direction of increasing returns.

$$A_\pi(a_t, s_t; \omega) = Q_\pi(a_t, s_t; \omega) - V_\pi(s_t; \omega).$$

In this paper, to speed up the learning process, the A3C algorithm adopts the N-step method, and its advantage function is calculated by (40).

$$A(a_t, s_t; \omega) = \sum_{i=0}^{k-1} \gamma^i r_i + \gamma^k V(s_{t+k}|\omega) - V(s_t|\omega).$$

Therefore, the critic network parameter $\omega$ can be trained as:

$$L(\omega) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1})} \left[ \gamma^t r_{t+i} + \gamma^i V(s_{t+i+1}|\omega) - V(s_t|\omega) \right]^2,$$

$$\omega_{t+1} = \omega_t + n_\omega \nabla \omega L(\omega).$$

To avoid local optimization, policy cross-entropy (PCE) is added to the A3C algorithm. Since the cross-entropy index can better describe the uncertainty of the probability distribution, the PCE can improve the ability of exploration diversity.
of the RL algorithm based on actor-critic. After adding the PCE, the gradient of the policy function is updated as:

$$\nabla R_{\theta} = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \nabla \theta \log \pi (a_t | s_t; \theta) A(a_t, s_t; \omega) + \beta \nabla \theta H (\pi (a_t | s_t; \theta)), \quad (43)$$

$$\theta_{t+1} = \theta_t + n_\theta \nabla \theta R_{\theta}. \quad (44)$$

The optimal scheduling framework of this paper is shown in Fig. 4. The global network parameters are randomly initialized at the beginning of training. The network parameters $\theta'$ and $\omega'$ of the actor and critic network in each thread are synchronously obtained from the parameters $\theta$ and $\omega$ of the global network through the pull function. Then the global pedometer $T$ and the local thread pedometer $t$ are initialized from an episode to the episode $T_{\text{max}}$. The output of each unit of IES is optimized in each episode. The parallel agent interacts with the environment and updates the parameters $\theta'$ and $\omega'$ through (42) and (44). In the training process of the A3C algorithm, agents on each thread interact independently with the environment. A3C algorithm adopts an asynchronous training mechanism. The global network preserves the global share parameters ($\theta$, $\omega$), which are updated asynchronously by the agents on each thread. Asynchronous update means that each agent updates the shared parameters after terminating a training episode. The global network saves the newly updated parameters, and the next agent will start the training with the new parameters. At the same time, the historical data is trained in the A3C model for the optimal scheduling of the electricity-gas-heat IES, which can guide the output of each unit to achieve the control purpose in actual operation.
parameters are shown in Table 3. Other parameters are shown in Table 4.

The A3C algorithm uses neural networks to fit actors and critic networks. The input of the actor network is the state set $S$, the output is the action set $A$, the number of hidden layers is 3, and the numbers of neurons are 400, 200, and 100. The input of the critic network is the state set $S$ and the action set $A$, the output is the action value $V$, the number of hidden layers is 3, and the numbers of neurons are 200, 100, and 100. All neural networks use the rectified linear unit (ReLU) activation function. The parameters of the A3C algorithm are shown in Table 5.

### A.3C Training Process

The data from the 1st November to the 31st January is used as the training set. The data of February is used as a test set to test the optimization effect. The training data is shown in Fig. 5.

The convergence characteristics of the proposed method are shown in Fig. 6. It can be seen in Fig. 6 (a) that the method obtains a low reward in the initial stage, which is a trial and error process. Then through further training, the reward is continuously improved. The agent can make more reasonable decisions through the training. The algorithm begins to converge when approaching 3300 episodes. At this time, the agent can obtain higher rewards steadily. The training process takes about 3.3 hours. The penalty value of the agent’s action exceeding the limit can also be seen. It can be seen that after the 2000 episodes of training, the agent’s actions can be controlled within the prescribed range. Fig. 6 (b) shows the value of the cumulative rewards of agents in different threads. Agents on different threads interact independently with the environment, resulting in different reward values at the same episode. The A3C algorithm can significantly accelerate the convergence speed of the algorithm through independent exploration of each thread agent through the asynchronous update.

Figure 7 shows the convergence of the A3C algorithm under different discount factors. Referring to the literature [27] and [29] on the A3C optimization algorithm, four discount factors of 0.85, 0.9, 0.95 and 0.99 are selected in the simulation. When the discount factor is 0.85, the algorithm cannot converge. The reward function fluctuates when the discount factor is 0.9 and 0.99, the algorithm can obtain a

### TABLE 3. Parameters of each energy storage.

| Energy storage | BS [30] [31] | GS [30] | HIES [30] [31] |
|----------------|--------------|---------|----------------|
| $S$            | 0.1          | 0.1     | 0.1            |
| $P_{in}/MW$    | 0.9          | 0.9     | 0.9            |
| $P_{out}/MW$   | 0.25         | 0.2     | 0.2            |
| $P_{in}/MW$    | 0.2          | 0.15    | 0.15           |
| $\eta$         | 0.92         | 0.98    | 0.95           |
| $\eta_{dout}$  | 0.85         | 0.98    | 0.95           |
| $S_{i}$        | 0.2          | 0.2     | 0.2            |

### TABLE 4. Other parameters.

| Parameter     | Value | Parameter | Value |
|----------------|-------|-----------|-------|
| $\eta_{act}$  | 0.045 | $\eta_{act}$ [32] | 0.6   |
| $\eta_{dout}$ | 0.1   | $\eta_{dout}$ [33] | 0.95  |
| $\eta_{in}$   | 0.15  | $\eta_{in}$ [30] [31] |       |

### TABLE 5. A3C algorithm parameters.

| Parameter | Value | Parameter | Value |
|-----------|-------|-----------|-------|
| $\gamma$  | 0.95  | $T_{max}$ | 10000 |
| $\rho$    | 0.0001| $\epsilon_{act}$ | 15    |
| $\rho_{d}$| 0.001 | $\epsilon_{dout}$ | 10    |
| $\beta$   | 0.05  | $\epsilon_{in}$ | 15    |
| $\eta_{act}$ | 96   | $\epsilon_{dout}$ | 10    |
| time slot | 15 mins |           |       |

![FIGURE 5. Historical sample data of IES.](image-url)
FIGURE 6. The cumulative rewards value of agents.

FIGURE 7. Convergence of A3C algorithm with different discount factors.

better convergence, when the discount factor is 0.95. Therefore, it is not conducive to the convergence of the algorithm when the discount is too large or too small.

B. OPTIMIZED CONTROL EFFECT OF IES

The optimized scheduling result is shown in Fig. 7. During the valley electricity price period (23:00 to 05:00), the wind power output is relatively high and CHP does not work. The battery storage charges from wind power and external power grid to satisfy the electric load. The gas production cost of the PtG unit is lower than that of gas purchase from the gas station. As the gas load demand is lower during this period, the gas load is supplied by the PtG unit and GS. Since CHP stops working, the heat load is supplied by the EB, and the HS stores excessive energy at the same time.

During the flat electricity price period (05:00-07:00, 09:00-13:00, 14:00-17:00, and 20:00-23:00), the electricity price is higher than that of the valley period. CHP starts to work. The electric load is supplied by WT, PV, CHP, and external power grid, and BS charges or discharges to obtain a higher reward. For example, during 05:00-07:00, since the electric load demand is higher than WT and PV output, and the SOC of BS is relatively high, the BS chooses not to charge or discharge. The gas load and CHP are supplied by the gas network. The CHP adopts the ‘determining electricity by heat’ mode to meet the heat load demand.

During the peak electricity price period (07:00-09:00, 13:00-14:00, and 17:00-20:00), the CHP adopts the ‘determining heat by electricity’ mode. The electric load is mainly provided by wind power, PV, CHP, and BS to reduce the cost of electricity purchase. The insufficient part is purchased from the external power grid. The gas load demand is supplied by the gas station, and the function of GS is to stabilize the gas load fluctuation.

With the proposed A3C-based optimization strategy, each energy storage device selects charging or discharging modes at each scheduling time. Fig. 8 shows the charge-discharge...
power and the SOC of the energy storage system. Since the penalties for exceeding the limit of actions are added in the reward function, the capacity of the energy storage and the climbing is limited within the constraints.

### C. COMPARISON OF CONVERGENCE SPEEDS AND SCHEDULING EFFECTS OF DIFFERENT ALGORITHMS

To verify the effectiveness of the A3C-based scheduling strategy proposed in this paper, DDPG and DQN algorithms are used for comparative analysis. The parameters of DDPG are the same as the proposed method. The parameters of DQN are selected from [18].

The training time and rewards of the DQN algorithm, DDPG algorithm and the A3C algorithm are shown in Table 6. It can be seen that, compared with DDPG and DQN algorithms, the training time of the A3C algorithm with the asynchronous mechanism is reduced by 30% and 37%, and the reward is increased by 25.6% and 30.8%.

The reward curves of the three DRL methods in the learning process are shown in Fig. 11. The A3C algorithm consumes the lowest cost. During the grid electricity price peak period, the total cost using the A3C algorithm is less than the other two methods. By scheduling the output of each unit, the electricity consumption with the A3C algorithm is shifted from the peak electricity price period to the valley electricity price period. The total costs with DDPG and DQN algorithms are higher.

---

**Figures:**
- **Fig. 9:** The optimized scheduling result of various energy storage equipment.
- **Fig. 10:** The reward curves of the three DRL methods.
- **Fig. 11:** Reward curve in the training process of different algorithms.

---

**Table 6.** Reward and training time of different algorithms.

| Algorithm | Training time/h | Reward |
|-----------|-----------------|--------|
| A3C       | 3.3             | 72.6   |
| DDPG      | 5.4             | 57.8   |
| DQN       | 4.9             | 55.5   |

---

**References:**
- [18]
To further verify the stability and generalization ability of different DRL and MPC methods in solving the economic dispatch of IES, 20 days of data are randomly selected from the test set. MPC is widely used as a classical optimization method in the optimal scheduling of integrated energy systems [34]. The parameters of MPC are selected from [16]. In Table 7, compared with DDPG, DQN and MPC methods, the average cost with the A3C method is reduced by 3%, 5.2% and 8.7%, respectively.

V. CONCLUSION

In this paper, an A3C-based optimal scheduling framework is proposed for the electricity-gas-heat integrated energy system economic dispatch. We described the IES optimal scheduling problems as an MDP, which does not rely on expert knowledge of the system modeling and the supply-demand forecast of the IES. The proposed A3C-based optimal scheduling framework can optimize the control decision for supply and demand sides by asynchronous learning of agents, and to reduce the relevance of parameters update of neural networks by the multi-agent utilization of the CPU multi-threading function. Compared with the DQN and DDPG algorithms, the training time of the proposed methods is shortened by 30% and 37%. And the daily average operating cost is reduced by 5.2% and 3% compared with the DQN, DDPG and MPC.

In this paper, the economic dispatch of the integrated energy system is mainly realized by coordinating the output of each unit on the energy supply side, without considering the influence of the energy network and the demand side. In future research, the source-load coordination optimization method considering comprehensive demand response will be further studied.

REFERENCES

[1] H. Wang, J. Yang, Z. Chen, G. Li, J. Liang, Y. Ma, H. Dong, H. Ji, and J. Feng, “Optimal dispatch based on prediction of distributed electric heating storages in combined electricity and heat networks,” Appl. Energy, vol. 267, pp. 1–10, Jun. 2020.

[2] S. Lu, W. Gu, K. Meng, S. Yao, B. Liu, and Z. Y. Dong, “Thermal inertial aggregation model for integrated energy systems,” IEEE Trans. Power Syst., vol. 35, no. 3, pp. 2374–2387, May 2020.

[3] S. Lu, W. Gu, C. Zhang, K. Meng, and Z. Dong, “Hydraulic-thermal cooperative optimization of integrated energy systems: A convex optimization approach,” IEEE Trans. Smart Grid, vol. 11, no. 6, pp. 4814–4823, Nov. 2020.

[4] Y. Ma, H. Wang, F. Hong, J. Yang, Z. Chen, H. Cui, and J. Feng, “Modelling and optimization of combined heat and power with power-to-gas and carbon capture system in integrated energy system,” Energy, vol. 236, Dec. 2021, Art. no. 121392.

[5] J. Yang, N. Zhang, Y. Cheng, C. Kang, and Q. Xia, “Modeling the operation mechanism of combined P2G and gas-fired plant with CO2 recycling,” IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 1111–1121, Jun. 2019.

[6] S. Lu, W. Gu, K. Meng, and Z. Dong, “Economic dispatch of integrated energy systems with robust thermal comfort management,” IEEE Trans. Sustain. Energy, vol. 12, no. 1, pp. 222–233, Jan. 2021.

[7] X. Liu, F. Zhang, Q. Sun, and W. Zhong, “Multi-objective optimization strategy of integrated electric-heat system based on energy storage situation division,” IEEE Access, vol. 9, pp. 19004–19024, 2021.

[8] A. Elshen and Q. Yang, “Scenario-based investment planning of isolated multi-energy microgrids considering electricity, heating and cooling demand,” Appl. Energy, vol. 235, pp. 1277–1288, Feb. 2019.

[9] R. Wang, P. Wang, and G. Xiao, “A robust optimization approach for energy generation scheduling in microgrids,” Energy Convers. Manage., vol. 106, pp. 597–607, Dec. 2015.

[10] M. Qadran, J. Wu, N. Jenkins, and J. Ekamaeye, “Operating strategies for a GB integrated gas and electricity network considering the uncertainty in wind power forecasts,” IEEE Trans. Sustain. Energy, vol. 5, no. 1, pp. 128–138, Jan. 2014.

[11] Y. Zhao, Y. Lu, C. Yan, and S. Wang, “MPC-based optimal scheduling of grid-connected low energy buildings with thermal energy storages,” Energy Buildings, vol. 86, pp. 415–426, Jan. 2015.

[12] Z. Li, C. Zang, P. Zeng, and H. Yu, “Combined two-stage stochastic programming and receding horizon control strategy for microgrid energy management considering uncertainty,” Energies, vol. 9, no. 7, p. 499, Jun. 2016.

[13] C. Lv, H. Yu, P. Li, C. Wang, X. Xu, S. Li, and J. Wu, “Model predictive control based robust scheduling of community integrated energy system with operational flexibility,” Appl. Energy, vol. 243, pp. 250–265, Jun. 2019.

[14] M. Patterson, N. F. Macia, and A. M. Kannan, “Hybrid microgrid model based on solar photovoltaic battery fuel cell system for intermittent load applications,” IEEE Trans. Energy Convers., vol. 30, no. 1, pp. 359–366, Mar. 2015.

[15] M. Petrollese, L. Valverde, D. Cocco, G. Cau, and J. Guerra, “Real-time integration of optimal generation scheduling with MPC for the energy management of a renewable hydrogen-based microgrid,” Appl. Energy, vol. 166, pp. 96–106, Mar. 2016.

[16] X. Dou, J. Wang, Z. Wang, L. Li, L. Bai, S. Ren, and M. Gao, “A dispatching method for integrated energy system based on dynamic time-interval of model predictive control,” J. Modern Power Syst. Clean Energy, vol. 8, no. 5, pp. 841–852, 2020.

[17] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, “Reinforcement learning-based microgrid energy trading with a reduced power plant schedule,” IEEE Internet Things J., vol. 6, no. 6, pp. 10728–10737, Dec. 2019.

[18] Z. Wan, H. Li, H. He, and D. Prokhorov, “Model-free real-time EV charging scheduling based on deep reinforcement learning,” IEEE Trans. Smart Grid, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[19] Y. Du and F. Li, “Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning,” IEEE Trans. Smart Grid, vol. 11, no. 2, pp. 1066–1076, Mar. 2020.

[20] T. T. Nguyen, D. N. Nguyen, and S. Nahavandi, “Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications,” IEEE Trans. Cybern., vol. 50, no. 9, pp. 3826–3839, Sep. 2020.

[21] Y. Keneshlou, T. Shi, N. Ramakrishnan, and C. K. Reddy, “Deep reinforcement learning for sequence-to-sequence models,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 7, pp. 2469–2489, Jul. 2020.

[22] Y. Liu, D. Zhang, and H. B. Goot, “Optimization strategy based on deep reinforcement learning for home energy management,” CSEE J. Power Syst., vol. 6, no. 3, pp. 572–582, Sep. 2020.

[23] Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang, “Real-time energy management of a microgrid using deep reinforcement learning,” Energies, vol. 12, no. 12, p. 2291, Jun. 2019.

[24] S. Brandi, M. S. Fiscelli, M. Martellacci, and A. Capozzoli, “Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings,” Energy Buildings, vol. 224, Oct. 2020, Art. no. 110225.

[25] F. S. Gorostiza and F. M. Gonzalez-Longatt, “Deep reinforcement learning-based controller for SOC management of multi-electrical energy storage system,” IEEE Trans. Smart Grid, vol. 11, no. 6, pp. 5039–5050, Nov. 2020.

[26] Y. Jiang and L. Guo, “Research on wind power accommodation for an electricity-gas-heat integrated microgrid system with power-to-gas,” IEEE Access, vol. 7, pp. 87118–87126, 2019.
[27] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1928–1937.

[28] M. Chen, T. Wang, K. Ota, M. Dong, M. Zhao, and A. Liu, “Intelligent resource allocation management for vehicles network: An A3C learning approach,” Comput. Commun., vol. 151, pp. 485–494, Feb. 2020.

[29] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. M. Leung, “Power control based on deep reinforcement learning for spectrum sharing,” IEEE Trans. Wireless Commun., vol. 19, no. 6, pp. 4209–4219, Jun. 2020.

[30] J. Zhai, X. Wu, S. Zhu, and H. Liu, “Low carbon economic dispatch of regional integrated energy system considering load uncertainty,” in Proc. 34rd Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC), Jun. 2019, pp. 642–647.

[31] JIAN DONG received the B.S. and M.S. degrees in electrical engineering from Shenyang University of Technology, Shenyang, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include power system stability analysis and optimal scheduling of the integrated energy systems.

[32] Y. Yingqi, G. Yifang, H. Fenglu, H. Yi, Z. Qingbin, and X. Jie, “Day-ahead dynamic economic dispatch of incremental distribution network considering P2G,” in Proc. Int. Conf. Power Syst. Technol. (POWERCON), Nov. 2018, pp. 272–278.

[33] L. Li and S. Yu, “Optimal management of multi-stakeholder distributed energy systems in low-carbon communities considering demand response resources and carbon tax,” Sustain. Cities Soc., vol. 61, Oct. 2020, Art. no. 102230.

[34] D. Tavernini, M. Metzler, P. Graber, and A. Sorniotti, “Explicit nonlinear model predictive control for electric vehicle traction control,” IEEE Trans. Control Syst. Technol., vol. 27, no. 4, pp. 1438–1451, Jul. 2019.

JUNYOU YANG (Member, IEEE) received the B.Eng. degree from Jilin University of Technology, Jilin, China, the M.Sc. degree from Shenyang University of Technology, Shenyang, China, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China. He was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of Toronto, Canada, from 1999 to 2000. He is currently the Head of the School of Electrical Engineering, Shenyang University of Technology. He is also a Distinguished Professor in the Province of Liaoning and the first 100 level candidates in the BaiQianWan Talents Program. He has led more than 50 research projects and has more than 200 publications in his technical field. His research interests include wind energy, special motor, and its control.

JUNYOU YANG received the B.Eng. degree from Jilin University of Technology, Jilin, China, in 2012 and 2016, respectively. She is currently an Intermediate Engineer with the Electricity Intensive Control Department, State Grid Liaoning Marketing Service Center, Shenyang. She is mainly involved in integrated energy system stability analysis.

LIU GAO received the bachelor’s degree from Heilongjiang University of Technology, in 2020. She is currently pursuing the master’s degree in electrical engineering with Shenyang University of Technology. Her research interest includes scheduling optimization for integrated energy systems.

XIRAN ZHOU received the bachelor’s degree in electrical engineering from Shenyang University of Technology, in 2020, where she is currently pursuing the master’s degree in electrical engineering. Her research interest includes scheduling optimization for integrated energy systems.

XINYI LU received the B.S. and M.S. degrees in electrical engineering from Shenyang Agricultural University, Shenyang, China, in 2012 and 2016, respectively. She is currently an Intermediate Engineer with the Electricity Intensive Control Department, State Grid Liaoning Marketing Service Center, Shenyang. She is mainly involved in integrated energy system stability analysis.

HAIXIN WANG (Member, IEEE) received the B.Eng. degree in electrical engineering from Inner Mongolia Agricultural University, Hohhot, China, in 2011, and the M.Sc. and Ph.D. degrees in electrical engineering from Shenyang University of Technology, Shenyang, China, in 2014 and 2017, respectively. He did his Postdoctoral Research at the Shenyang University of Technology, in 2018, and the School of Engineering, Cardiff University, Cardiff, U.K., in 2019. He is currently an Associate Professor with the School of Electrical Engineering, Shenyang University of Technology. His research interests include integrated energy systems, wind power generation, and cyber-physical systems.