Improved Proximal Policy Optimization Algorithm for Sequential Security-constrained Optimal Power Flow Based on Expert Knowledge and Safety Layer

Yanbo Chen, Qintao Du, Honghai Liu, Liangcheng Cheng, and Muhammad Shahzad Younis

***Abstract*—In recent years, reinforcement learning (RL) has emerged as a solution for model-free dynamic programming problem that cannot be effectively solved by traditional optimization methods. It has gradually been applied in the fields such as economic dispatch of power systems due to its strong self-learning and self-optimizing capabilities. However, existing economic scheduling methods based on RL ignore security risks that the agent may bring during exploration, which poses a risk of issuing instructions that threaten the safe operation of power system. Therefore, we propose an improved proximal policy optimization algorithm for sequential security-constrained optimal power flow (SCOPF) based on expert knowledge and safety layer to determine active power dispatch strategy, voltage optimization scheme of the units, and charging/discharging dispatch of energy storage systems. The expert experience is introduced to improve the ability to enforce constraints such as power balance in training process while guiding agent to effectively improve the utilization rate of renewable energy. Additionally, to avoid line overload, we add a safety layer at the end of the policy network by introducing transmission constraints to avoid dangerous actions and tackle sequential SCOPF problem. Simulation results on an improved IEEE 118-bus system verify the effectiveness of the proposed algorithm.**

***Index Terms*—Sequential security-constrained optimal power flow (SCOPF), expert experience, safety layer, renewable energy, safe reinforcement learning.**

I. Introduction

POWER system operation and control is encountering increasing challenges as renewable energy is integrated into the power system on a big scale. On the one hand, the large-scale access of wind and photovoltaic power requires the dispatch operation to rapidly adjust to the unpredictability of renewable energy output. On the other hand, the capacity of renewable energy that the power system can accomodate is also limited by the transmission capacity of the power lines. The main challenge in increasing the capacity of the power system to accomodate renewable energy is how to combine the characteristics of the power system and load demand with the output characteristics of wind and photovoltaic power, which helps establish a safe and reliable security-constrained optimal power flow (SCOPF) model of power systems [1], [2].

Mathematically, the sequential SCOPF problem is a complex nonlinear mixed-integer programming problem, which is traditionally solved by interior point method, Dantzig-Wolf algorithm, interior-point algorithm, and cross-entropy algorithm [3] - [5], etc., which are efficient and precise in solving small-scale problems, but their computational efficiency is relatively low in large-scale systems. Moreover, the control ability requirements and random characteristics of the renewable energy station greatly increase the difficulty sequential SCOPF modeling.

With strong self-learning and self-optimizing capabilities, reinforcement learning (RL) has gained attraction in the areas like the economic dispatch of power systems because it can solve model-free dynamic programming problem that cannot be effectively solved by conventional optimization decision-making methods [6], [7]. An RL algorithm based on policy iteration is proposed in [8] to solve the economic dispatch problem, where the economic dispatch problem is modelled as a multi-stage decision-making issue. A pre-training scheme based on simulation learning is designed in [9] to investigate the active power dispatch in power systems with large-scale renewable energy sources, significantly enhancing the application ability of the deep RL (DRL) algorithm. A self-adaptive uncertain economic dispatch model based on DRL is adopted in [10] to deal with uncertain vari-

Manuscript received: April 14, 2023; revised: June 21, 2023; accepted: September 12, 2023. Date of CrossCheck: September 12, 2023. Date of online publication: November 13, 2023.

This work was supported in part by National Natural Science Foundation of China (No. 52077076) and in part by the National Key R&D Plan (No. 2021YFB2601502).

Y. Chen (corresponding author), Q. Du, and H. Liu are with State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources and School of Electrical & Electronic Engineering, North China Electric Power University, Beijing 102206, China, and Y. Chen is also with School of Engineering, Xining University, Xining 810008, China, and Key Laboratory of Smart Operation of New Energy Power System, Ministry of Education, Qinghai University, Xining 810016, China (e-mail: chenyanbo@ncepu.edu.cn; 1791543766@qq.com; hbdl_liuhonghai@163.com).

L. Cheng is with China Electric Power Research Institute, Nanjing 210003, China (e-mail: 479146353@qq.com).

M. S. Younis is with National University of Sciences and Technology, Islamabad, 44000, Pakistan (e-mail: muhammad.shahzad@seecs.edu.pk).

DOI: 10.35833/MPCE.2023.000232

ations in load and renewable energy. The model adopts multiple experience pool replay strategies to eliminate the correlation of sample data and improve the utilization of high-quality data, which is instrumental in speeding up the training process. Overall, the model is based on RL, which maximizes economic benefits by reducing power generation costs through sensible unit combination and load allocation. The aforementioned economic dispatch methods prioritize economic benefits and do not consider the safety and reliability of the power system sufficiently.

As far as we know, there has been limited research on RL-based security-constrained economic dispatch problem [11]. In order to avoid voltage limit violation at nodes, a node voltage penalty term is added to the global reward in [12], creating a negative reward mechanism. To accomplish safe and efficient dispatch, the optimization is carried out through interactive iteration to limit the voltage amplitude of each node within a safe range. In [13], the reward function is improved to introduce risk penalty for power flow convergence and power balance penalty. The mechanism is conducive to guide the agent to generate a dispatch scheme that satisfies the safety constraints. Safe RL (SRL) is suggested in [14] to solve the problem of optimal operation of distribution network. To explore the best dispatch policy while ensuring safety, the expected cumulative cost function is used to represent the environmental constraints, and the neural network is trained based on the constrained policy optimization (CPO) algorithm. A holomorphic embedding-based soft actor-critic (HE-SAC) algorithm is developed in [15] to find fast optimal operable power flow (OOPF) by leveraging DRL and advanced complex analysis techniques. The reward function is modified to improve the feasibility by considering constraint violations and policy entropy.

The above studies have made significant explorations, but the following issues have not been solved.

1) Most of current RL-based SCOPF studies are in favor of a small-scale power systems, but are not validated in the large-scale ones. When the power system is large and the decision-making action space is high-dimensional and continuous, it is possible to cause the agent to slow down the convergence rate and even fail to effectively explore the optimal strategy because the training effect of RL is sensitive to the number of actions.

2) The prior studies do not pay enough attention to the potential security risks that DRL poses while conducting exploration. In order to improve the convergence and security of agent, [12]-[15] add constraint penalty terms in training and combine them with objective function or build expected cumulative cost function to constrain action. These methods necessitate the manual design of punishment terms or the adjustment of penalty coefficients, which calls for strong modeling abilities.

3) It is challenging to ensure the practicability of agent because it cannot guarantee that the required safety constraints can be simultaneously satisfied in practice. For instance, some studies have neglected the capacity limit of transmission line by only considering power balance and voltage optimization. It is difficult to implement load allocation while making sure that all security requirements are met in the actual operation of power system.

4) The existing dispatch methods based on RL are not efficient enough in considering the targeted guidance of the agent to improve the utilization rate of renewable energy. One of the key objectives of electricity market construction is the promotion of renewable energy utilization. Addressing the ambiguity and unpredictability of renewable energy and minimizing energy waste brought on by the curtailment of wind and photovoltaic resources are presently two of the most important issues that require attention.

To this end, we propose an improved proximal policy optimization (PPO) algorithm for sequential SCOPF based on expert knowledge and safety layer (called EK-CPPO) to overcome the shortfalls of the existing RL-based algorithms. The major contributions of this paper lie in three perspectives.

1) We propose a constrained Markov decision process (CMDP) formulation for the sequential SCOPF problem to determine the active power dispatch strategy, the voltage optimization scheme of the units, and the charging/discharging dispatch of the energy storage systems, which can adapt to the uncertain changes of intermittent power and load. Compared with existing RL-based algorithms, this formulation does not require the design of specific reward functions for constrained problems.

2) We embed the expert knowledge to guide the training of the agent and improve the execution effectiveness of the agent to deal with constraints such as power balance. At the same time, the policy enables us to efficiently consume renewable energy, and the utilization rate of renewable energy is maximized while ensuring the economy of the scheduling strategy.

3) We employ a safety layer to the end of the policy network to introduce transmission constraints to the power systems. Different from traditional DRL algorithms, the designed policy can effectively ensure that the dispatch actions meet the transmission constraints, and guarantee the safety of the agent in the exploration process, which can contribute to tackling the sequential SCOPF problem.

The rest of this paper is organized as follows. Section II presents the safety-constrained economic dispatch model. Section III proposes guided training based on expert experience. Section IV introduces re-constrain actions based on the safety layer. Case studies are given in Section V. Finally, conclusions are drawn in Section VI.

II. Safety-constrained Economic Dispatch Model

The OPF problems in AC environments aim to search for the control policy to minimize the cost of power purchase while satisfying the security constraints of power systems [16]. In this section, the optimal objective of sequential OPF is introduced firstly, and then the limit and constraints of grid operation are defined. Finally, the sequential SCOPF problem is expressed as a CMDP and solved by SRL.

A. Optimization Objectives

Dispatch planning requires adjusting the operation plan of power generation strategies to ensure continuous balance and

stability of the power systems based on actual power demand and changes in power supply. In the power systems with large-scale renewable energy sources, the dispatch planning should minimize the operation cost and maximize the utilization rate of renewable energy to achieve the goal of energy conservation and emission reduction while ensuring the safe and stable operation of the power system [17]. The dispatch objective delineated in this paper encompasses the voltage and reactive power optimization, the active power optimization and the strategic dispatch of the energy storage system for charging and discharging operations. We use subscript $t \in T$ to index interval, $n \in N$ to index thermal power units, $m \in M$ to index renewable energy units, $s \in S$ to index the energy storage systems, where T, N, M, and S represent the sets of time intervals, thermal power units, renewable energy units, and energy storage systems, respectively. The operation costs and the utilization function of renewable energy can be expressed by (1) and (2), respectively. The operation costs include the costs of each unit and the energy storage system.

$$f_1(P_{i,t}^G)=\sum_{t\in T}\sum_{n\in N}(a_n(P_{n,t}^G)^2+b_nP_{n,t}^G+c_n)+\sum_{t\in T}\sum_{s\in S}(a_s\left|P_{s,t}^{ES}\right|+b_s) \tag{1}$$

$$f_2(P_{i,t}^G)=\sum_{t\in T}\left(1-\frac{\sum_{m\in M}P_{m,t}^G}{\sum_{m\in M}P_{m,t}^{G,\max}}\right) \tag{2}$$

where $P_{n,t}^G$ and $P_{m,t}^G$ are the outputs of thermal power unit n and renewable energy unit m at time t, respectively; $P_{s,t}^{ES}$ is the charging and discharging power of the energy storage system at time t; a_n, b_n, c_n, a_s, and b_s are the cost coefficients; and $P_{m,t}^{G,\max}$ is the maximum output power of the renewable energy unit m at time t.

B. Limits and Constraints

We use the subscript $k \in K$ to index network nodes, $l \in L$ to index lines, and $q \in Q$ to index the load, where K, L, and Q represent the sets of network nodes, lines, and loads, respectively. For sequential SCOPF problem, the constraints shown in (3)-(10) must be satisfied.

$$\sum_{n\in N}P_{n,t}^G+\sum_{m\in M}P_{m,t}^G=\sum_{q\in Q}P_{q,t}^D \tag{3}$$

$$P_i^{G,\min}\le P_{i,t}^G\le P_i^{G,\max}\quad \forall i\in N\bigcup M \tag{4}$$

$$-\sigma_n^d\Delta t\le P_{n,t}^G-P_{n,t-1}^G\le\sigma_n^u\Delta t\quad \forall n\in N \tag{5}$$

$$V_k^{B,\min}\le V_{k,t}^B\le V_k^{B,\max}\quad \forall k\in K \tag{6}$$

$$\underline{P}_l\le P_{l,t}^L\le\bar{P}_l\quad \forall l\in L \tag{7}$$

$$P_{l,t}^L=\sum_{n\in N}S_{nk}P_{n,t}^G+\sum_{m\in M}S_{mk}P_{m,t}^G-\sum_{q\in Q}S_{qk}P_{q,t}^D\quad \forall l\in L \tag{8}$$

$$\underline{P}_s^{ES}\le P_{s,t}^{ES}\le\bar{P}_s^{ES}\quad \forall s\in S \tag{9}$$

$$C_s^{ES,\min}\le C_{s,t}^{ES}\le C_s^{ES,\max}\quad \forall s\in S \tag{10}$$

where $P_{q,t}^D$ is the load demand of load q at time t; $P_{i,t}^G$ is the output of unit i at time t; $P_i^{G,\min}$ and $P_i^{G,\max}$ are the lower and upper limits of the output of unit i at time t, respectively; σ_n^d and σ_n^u are the maximum downward and upward adjustment rates of thermal power unit n, respectively; $P_{l,t}^L$ is the transmission power of line l at time t; $\bar{P}_l$ and $\underline{P}_l$ are the forward and reverse transmission power of line l, respectively; S_{nk}, S_{mk}, and S_{qk} are the sensitivities of n, m, and q to node k, respectively; $V_{k,t}^B$ is the voltage limit of node k at time t; $V_k^{B,\min}$ and $V_k^{B,\max}$ are the minimum and maximum voltage limits of node k, respectively; $\underline{P}_s^{ES}$ and $\bar{P}_s^{ES}$ are the maximum discharging and charging efficiencies of the energy storage system, respectively; $C_{s,t}^{ES}$ is the capacity limit of the energy storage system at time t; and $C_s^{ES,\min}$ and $C_s^{ES,\max}$ are the minimum and maximum capacity limits of the energy storage system, respectively.

Equation (3) is the power balance constraint; (4) and (5) are the lower and upper bound limits on output and the ramp rate constraint, respectively; (6) is the bus voltage constraint; (7) and (8) are the power constraints of transmission line; and (9) and (10) represent the charging and discharging rates and energy storage capacity limit of the energy storage system, respectively.

C. CMDP Formulation

One of the main challenges of Markov modeling for sequential SCOPF problem is how to handle constraints [18]. In most model-free methods, constraints are modeled as negative reward in the Markov decision process (MDP) [19]. However, as discussed in [20], it is difficult to determine a penalty coefficient to balance constraint violations and reward. Considering large-scale optimization problem, effective constraints are often not satisfied, and may even lead to inability of the agent to converge. To address this issue, we propose a CMDP formulation for the sequential SCOPF problem. The basic elements of the proposed CMDP formulation will be elaborated as follows.

1) Action Space and Observation Space

PPO is regarded as a state-of-the-art policy gradient algorithm [21], which essentially represents the strategy π as a neural network with a parameter $\boldsymbol{\theta}$. Through the interaction of neural network and environment, the sequence τ containing H steps is formed as:

$$\boldsymbol{\tau}=\{\boldsymbol{s}_1,\boldsymbol{a}_1,\boldsymbol{s}_2,\boldsymbol{a}_2,\ldots,\boldsymbol{s}_H,\boldsymbol{a}_H\} \tag{11}$$

where $\boldsymbol{s}_t$ ($t=1,2,\ldots,H$) is the state vector of the current environment; $\boldsymbol{a}_t$ ($t=1,2,\ldots,H$) is the decision vector output by the neural network; and H is the length of the sequence $\boldsymbol{\tau}$.

In the same state, the output of the neural network satisfies the probability distribution of parameter $\boldsymbol{\theta}$, so the sequence $\boldsymbol{\tau}$ is uncertain and the occurring probability of sequence $\boldsymbol{\tau}$ is calculated by:

$$\begin{aligned}p_{\boldsymbol{\theta}}(\boldsymbol{\tau})=&p(\boldsymbol{s}_1)p_{\boldsymbol{\theta}}(\boldsymbol{a}_1|\boldsymbol{s}_1)p(\boldsymbol{s}_2|\boldsymbol{s}_1,\boldsymbol{a}_1)p_{\boldsymbol{\theta}}(\boldsymbol{a}_2|\boldsymbol{s}_2)p(\boldsymbol{s}_3|\boldsymbol{s}_2,\boldsymbol{a}_2)\ldots=\\&p(\boldsymbol{s}_1)\prod_{t\in T}p_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)\end{aligned} \tag{12}$$

where $p(\boldsymbol{s}_1)$ is the probability that the initial environment is in the state $\boldsymbol{s}_1$; $p_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ is the probability of $\boldsymbol{a}_t$ output by the neural network with parameter $\boldsymbol{\theta}$ in the state $\boldsymbol{s}_t$; and $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$ is the probability that the next state is $\boldsymbol{s}_{t+1}$ when $\boldsymbol{a}_t$ is performed under state $\boldsymbol{s}_t$.

The selected $\boldsymbol{s}_t$ is shown in (13), which includes the active power $P_{n,t-1}^G$ and $P_{m,t-1}^G$ of the thermal power units and the renewable energy units in the previous period, the voltage

$V_{n,t-1}^{G}$ of the nodes where the thermal power units are located and the voltage $V_{m,t-1}^{G}$ of the nodes where the renewable energy units are located, the charging and discharging dispatch $P_{s,t-1}^{ES}$, the residual battery capacity $B_{s,t}^{ES}$ of the energy storage system during the previous period, the upper limit $\bar{P}_t$ and the lower limit $\underline{P}_t$ of the active power adjustment of the units during the current period, the loading rate of transmission line during the current period $\rho_{l,t}^{L}$, and the forecasted active power load $P_{q,t+1}^{D}$ during the next period.

$$\boldsymbol{s}_t=[P_{n,t-1}^{G}, P_{m,t-1}^{G}, V_{n,t-1}^{G}, V_{m,t-1}^{G}, P_{s,t-1}^{ES}, B_{s,t}^{ES}, \bar{P}_t, \underline{P}_t, \rho_{l,t}^{L}, P_{q,t+1}^{D}] \tag{13}$$

$$\underline{P}_{i,t}=\begin{cases}\max(P_i^{G,\min}, P_{i,t}^{G}-\sigma_n^{d}\Delta t) & \forall i\in N\\ 0 & \forall i\in M\end{cases} \tag{14}$$

$$\bar{P}_{i,t}=\begin{cases}\min(P_i^{G,\min}, P_{i,t}^{G}+\sigma_n^{u}\Delta t) & \forall i\in N\\ P_i^{G,\max} & \forall i\in M\end{cases} \tag{15}$$

In the sequential SCOPF model established in this paper, $\boldsymbol{a}_t$ is given as:

$$\boldsymbol{a}_t=[P_{n,t}^{G}, P_{m,t}^{G}, V_{n,t}^{G}, V_{m,t}^{G}, P_{s,t}^{ES}] \tag{16}$$

It should be noted that the policy network outputs non-normalized action through the activation function tanh, with the output ranging from −1 to 1. In order to correspond the output of the policy network with the actual range of actions, the output of the policy network needs to be de-normalized according to the action space of (6), (9), (14), and (15), so that the real dispatch policy can be obtained.

2) Reward

Each stage can obtain a specific reward, and the reward of sequence $\boldsymbol{\tau}$ can be expressed by the expected cumulative reward obtained by the neural network in the case of strategy π. The training goal of DRL is to find an optimal strategy to maximize the expected reward of the sequence $\boldsymbol{\tau}$.

In this method, in order to quickly find the optimal strategy, we can define the reward generated after interacting with the environment during the t^{th} time period, as shown in (17).

$$r(\boldsymbol{s}_t, \boldsymbol{a}_t)=c_1 r_1(\boldsymbol{s}_t, \boldsymbol{a}_t)+c_2 r_2(\boldsymbol{s}_t, \boldsymbol{a}_t) \tag{17}$$

$$r_1(\boldsymbol{s}_t, \boldsymbol{a}_t)=\mathrm{e}^{-\left(\sum_{n\in N}(a_n(P_{n,t}^{G})^2+b_n P_{n,t}^{G}+c_n)+\sum_{s\in S}(a_s\left|P_{s,t}^{ES}\right|+b_s)\right)}-1 \tag{18}$$

$$r_2(\boldsymbol{s}_t, \boldsymbol{a}_t)=\frac{\sum_{m\in M}(P_{m,t-1}^{G}+\Delta P_{m,t}^{G})}{\sum_{m\in M}P_{m,t}^{G,\max}} \tag{19}$$

where c_1 and c_2 are the weights of dispatch requirements.

3) Termination Condition

The actions output by an agent must meet a series of operational rules. For example, the input power of any generator cannot exceed the specified upper and lower limits, the active power adjustment of the generator must be less than the set ramp rate, and the total output of all units must maintain a dynamic balance with the load demand. At the same time, when the agent adjusts the terminal voltage, the reactive power output of the units must be within its upper and lower limits. If these rules are violated, the emulator will prompt "illegal action" and forcibly end the current episode. In addition, if there is a "soft overload" (the line current exceeds the limit but does not exceed 135% of the loading rate) in four consecutive time steps of any line, the line will be shut down. If a "hard overload" (the line current exceeds 135% of the loading rate) occurs, the line will be immediately shut down. The outage of transmission lines in the power system will lead to varying degrees of power flow transfer, which may lead to cascading failures, ultimately leading to system disconnection or collapse [22]. Therefore, it is necessary to avoid outage events caused by transmission line overload as much as possible [23], [24].

D. SRL Method

Based on modeling the sequential SCOPF problem as CMDP, the optimization objective can be expressed as solving the constrained RL problem. For this purpose, we establish an SRL model to express the sequential SCOPF model. The mapping relationship between state $\boldsymbol{s}_t$ and the cost is defined as cost function c_i, which is used to measure the violation degree of constraints when agent chooses $\boldsymbol{a}_t$. Different cost functions represent different types of damages. The cumulative reward function can be expressed as $G_t=\sum_{h=0}^{H}\gamma_{cost}^{h} r_i(\boldsymbol{s}_{t+h}, \boldsymbol{a}_{t+h})$ and the cumulative cost function can be expressed as $G_t^{c_i}=\sum_{h=0}^{H}\gamma_{cost}^{h} c_i(\boldsymbol{s}_{t+h}, \boldsymbol{a}_{t+h})$. The goal of SRL can be written in (20). The agent will receive a reward r_i and a cost c_i for each step. To achieve a trade-off between optimal reward and safety, it is necessary to optimize the reward function that satisfies safety constraints to maximize the long-term reward while satisfying the cost threshold [25]:

$$\begin{cases}\pi=\arg\max\limits_{\pi}\mathop{\mathbb{E}}\limits_{\tau\sim\pi}[G_t]=\arg\max\limits_{\pi}\mathop{\mathbb{E}}\limits_{\tau\sim\pi}\left[\sum\limits_{h=0}^{H}\gamma_{cost}^{h} r_i(\boldsymbol{s}_{t+h}, \boldsymbol{a}_{t+h})\right]\\ \text{s.t. } \bar{c}_i(\boldsymbol{s}_t)\le C_i \quad \forall i\in K\end{cases} \tag{20}$$

where $\mathop{\mathbb{E}}\limits_{\tau\sim\pi}$ is the expectation value; γ_{cost}^{h} is the discount factor; $\bar{c}_i(\cdot)$ is the single-step cost function; and C_i is the threshold of the cost function.

It is a challenging task to solve the constraints in (20). In the initial training stage, it is difficult to meet all state constraints for the agent initialized using random strategies because of the lack of prior knowledge. Only when the number of violations is sufficient can the agent gradually avoid dangerous actions. This situation leads to low exploration efficiency and slow convergence speed of the agent, and may also bring risks.

Therefore, as shown in Fig. 1, we establish a dispatch model based on SRL to determine the load allocation, voltage optimization, and charging and discharging schemes of the energy storage systems. The dataset indicated in Fig. 1 meets the requirements of typical scenarios for grid operation, including the network congestion, severe load fluctuation, and renewable energy curtailment. This paper introduces expert experience to improve the execution of constraints such as power balance in the training process, while guiding the agent to effectively improve the utilization rate of renewable energy. In addition, in order to avoid line overload, we also add a safety layer introducing line capacity constraints, which is instrumental in avoiding dangerous actions and solving the sequential SCOPF problem.

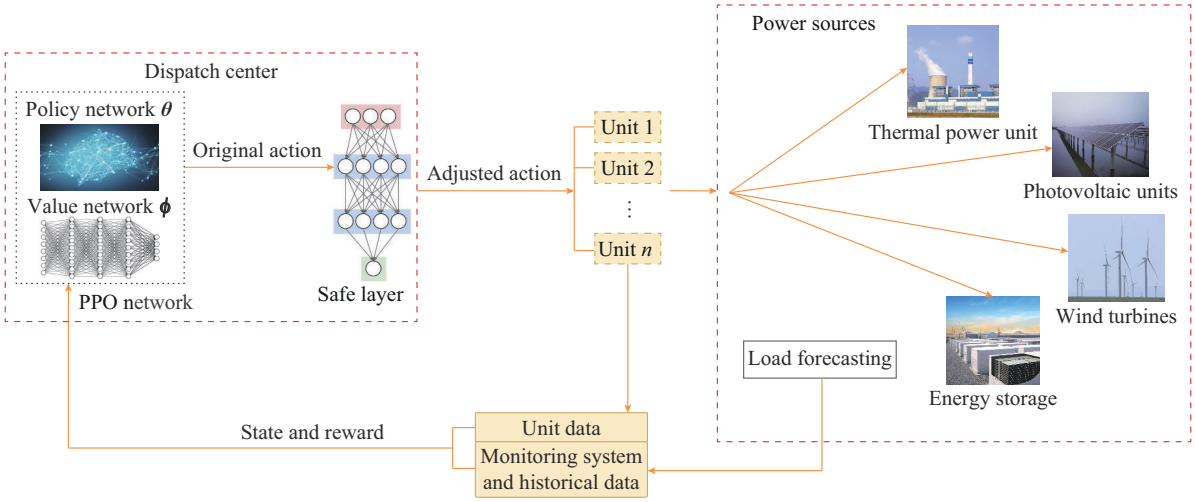

Fig. 1. Sequential SCOPF based on SRL.

III. Guided Training Based on Expert Experience

A. Parameter Update Method

The advantage function is used to calculate the relative advantage of a particular action $\boldsymbol{a}_t$ compared with other possible actions in a given state $\boldsymbol{s}_t$. It can map the state-action value function to the same baseline as the value function, achieving normalization of the state-action value function, which can contribute to reducing variance, and avoiding overfitting caused by excessive variance. In this paper, the GAE is adopted as the estimation method for the advantage function δ_t, which is defined by:

$$\hat{A}_t = \sum_{h=0}^{H}(\gamma\lambda)^h \delta_{t+h} = \delta_t + \sum_{h=1}^{H}(\gamma\lambda)^h \delta_{t+h} = \delta_t + \gamma\lambda \sum_{h=0}^{H}(\gamma\lambda)^h \delta_{t+h+1} = \delta_t + \gamma\lambda\hat{A}_{t+1} \tag{21}$$

$$\delta_t = r_t + \gamma\lambda(\boldsymbol{s}_{t+1}) - V_{\boldsymbol{\phi}}(\boldsymbol{s}_t) \tag{22}$$

where λ is a hyperparameter which is used to balance the trade-off between variance and bias; γ is the discount factor; and $V_{\boldsymbol{\phi}}(\cdot)$ is the state value function.

In this paper, the variable β is chosen to control the weight relationship between the constraint term and the objective term. The KL divergence is used as a penalty term and added to the objective function. The combined objective function is also referred to as the loss function of the actor network, which is calculated by:

$$J(\boldsymbol{\theta}) = \frac{1}{H}\sum_{t=1}^{H}\left\{\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\boldsymbol{\theta}_{old}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}\hat{A}_t - \beta y_{KL}[\pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s}_t), \pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}_t)]\right\} \tag{23}$$

where $\dfrac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\boldsymbol{\theta}_{old}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}$ is the ratio of probability between the new policy and old policy; $\boldsymbol{\theta}_{old}$ is the parameter of the policy before the policy update; and $y_{KL}(\pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s}_t), \pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}_t))$ is the KL divergence term that measures the difference between the new policy and old policy, mainly limiting the magnitude of policy updates.

The Critic network is used to evaluate the value function of the current state. The PPO algorithm updates the parameters $\boldsymbol{\phi}$ of the Critic network using the loss function shown in (24).

$$L(\boldsymbol{\phi}) = -\sum_{t=1}^{H}\left(\sum_{t'>t}\gamma^{t'-t} r_{t'} - V_{\boldsymbol{\phi}}(\boldsymbol{s}_t)\right)^2 \tag{24}$$

B. Introducing Expert Experience

To guide the agent to give priority to the renewable energy utilization when allocating loads, and to restrict the search direction within the feasible operation region of the power systems, this subsection embeds expert knowledge into the training process to guide the training of the agent, which is instrumental in reducing the curtailment rate of renewable energy, and improving the convergence.

We introduce regularization terms based on expert experience into the loss function of the Actor network, which include power balance and renewable energy utilization. In summary, the loss function of Actor is updated as:

$$J'(\boldsymbol{\theta}) = w_q J(\boldsymbol{\theta}) + \frac{1}{H}\sum_{t=1}^{H}\sum_{r=1}^{2} w_r \cdot reg_{r,t} \tag{25}$$

where w_q, w_1, and w_2 are the weights of the regularization terms; and $reg_{1,t}$ and $reg_{2,t}$ are the regularization terms related to power balance constraint and renewable energy utilization, respectively. $reg_{1,t}$ indicates the square of the difference between the total load demand and the total output of units, which is calculated by (26). By introducing $reg_{1,t}$, the agent can be guided to achieve a balance between power generation and load demand in the power system.

$$reg_{1,t} = \left(\sum_{q\in Q} P_{q,t}^{D} - \sum_{i\in N\cup M} P_{i,t}^{G}\right)^2 \tag{26}$$

where ΔP_t^D is the difference between the total load at time t and the total load at time $t-1$.

$reg_{2,t}$ is conducive to guiding the agent to maximize the renewable energy utilization while ensuring the safe operation of the power system, which is defined by:

$$reg_{2,t}=1-\frac{\sum_{m\in M}(P^{G}_{m,t-1}+\Delta P^{G}_{m,t})}{\sum_{m\in M}P^{G,\max}_{m,t}} \tag{27}$$

When training the Actor-Critic, the Actor network can be updated according to the optimized loss function, which is shown in (25).

IV. Re-constrain Actions Based on Safety Layer

Whether the dispatch action can strictly meet the line capacity constraints is the key to solving the sequential SCOPF problem [26], [27]. Existing economic dispatch methods based on RL often ignore the key constraint of transmission constraints. Therefore, this section proposes an action constraint method with safety layer based on the principle of fast sensitivity method, which can realize effective constraint on line capacity.

A. Principle of Fast Sensitivity Method

We simplify nonlinear elements such as generators and loads in the power system to a constant current source, and simplify static elements such as transformers, transmission lines, capacitors, and reactors to equivalent lines connected in series or parallel for simulation, so that the power system can be simplified into a linear network [28], [29]. We make the following reasonable assumptions for further analysis:

1) The power flow calculation does not consider the influence of grounding lines.

2) The resistance of high-voltage transmission lines is much smaller than their reactance.

3) The voltage at both ends of the transmission lines is close to the rated voltage and the voltage phase angle difference is small.

Supporting evidence suggests that line l connects nodes i and j, therefore, line power flow P_l is calculated by:

$$P_l=\frac{\psi_i-\psi_j}{x_{ij}} \tag{28}$$

where ψ_i and ψ_j are the phase angles of nodes i and j, respectively; and x_{ij} is the impedance between nodes i and j. We use P_k^{SP} to represent the injection power of node k, which is the difference between the total active power of generators and the total demand of loads at node i. Then, the active power flow of node k can be defined by:

$$P_l=\frac{\psi_i-\psi_j}{x_{ij}}=\sum_{k=1}^{N-1}\frac{X_{ik}-X_{jk}}{x_{ij}}P_k^{SP} \tag{29}$$

where X_{ik} and X_{jk} are the elements of $\boldsymbol{X}$, and $\boldsymbol{X}$ represents the inverse matrix of $\boldsymbol{B}$, which is defined as:

$$\begin{cases} B_{ii}=\sum_{j\in i}\frac{1}{x_{ij}} \\ B_{ij}=-\frac{1}{x_{ij}} \end{cases} \tag{30}$$

Therefore, the sensitivity coefficient S_{lk} of line l to the injection power of node k can be calculated by:

$$S_{lk}=\frac{X_{ik}-X_{jk}}{x_{ij}} \tag{31}$$

We can conclude that there is a linear relationship between line power flow and injection power of each node. When line power flow exceeds the line capacity limit, it is possible to effectively reduce the active power flow of the line by adjusting the injection power of related nodes in proportion to the overload degree [30], [31].

B. Action Correction Mechanism Based on Safety Layer

Based on the characteristic that there is a linear relationship between the line power flow and the injection power of each node, we propose a corrective method based on single-step linear transformation. The method will undergo linearization processing, as shown in (32), which modifies c_i to ensure that the violation of transmission constraints satisfies the cost threshold.

$$\bar{c}_i(\boldsymbol{s}_{t+1})=c_i(\boldsymbol{s}_t,\boldsymbol{a}_t)\approx\bar{c}_i(\boldsymbol{s}_t)+(\boldsymbol{g}(\boldsymbol{s}_t;w_i))^{\mathrm{T}}\boldsymbol{a}_t \tag{32}$$

where $\boldsymbol{g}(\boldsymbol{s}_t;w_i)$ is designed as an action correction function with a neural network structure, and $\boldsymbol{g}(\boldsymbol{s}_t;w_i)$ is a first-order approximation of $c_i(\boldsymbol{s}_t,\boldsymbol{a}_t)$ with respect to $\boldsymbol{a}_t$; and w_i represents the weights of the neural network. $\boldsymbol{g}(\boldsymbol{s}_t;w_i)$ takes $\boldsymbol{s}_t$ as input and outputs a vector of the same dimension as $\boldsymbol{a}_t$. The meaning of $\boldsymbol{g}(\boldsymbol{s}_t;w_i)$ is to explicitly represent the sensitivity of action changes to the power flow changes. Due to the constraint on the transmission line capacity, the number of correction functions is equal to the number of transmission lines, and thus $i\in L$. We can consult Fig. 2 to gain a more intuitive understanding.

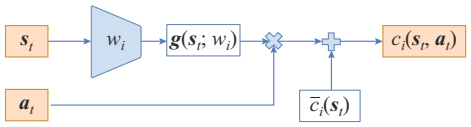

Fig. 2. Motion correction method based on single-step linear transformation.

It should be noted that $\boldsymbol{g}(\boldsymbol{s}_t;w_i)$ serves as the pre-training model prior to training the agent, and only requires one training execution [32]. In addition, it is optional to conduct additional continuous training of $\boldsymbol{g}(\boldsymbol{s}_t;w_i)$ during the agent training process. The experimental results indicate that continuous training improves the performance of the RL compared with pre-training alone. Therefore, we only present the results with continuous training.

We will demonstrate how to employ the method of locally modifying the policy to solve problem shown in (32). $\pi_{\boldsymbol{\theta}}(\boldsymbol{s}_t)$ represents the deterministic action selected by the policy network. Subsequently, a safety layer is added at the end of the policy network to improve the policy through local modification, with the aim of solving the problem, as shown in (33).

$$\begin{cases} \pi^{*}_{\boldsymbol{\theta}}(\boldsymbol{s}_t)=\arg\min_{\boldsymbol{a}_t} f(\boldsymbol{s}_t,\boldsymbol{a}_t,\pi_{\boldsymbol{\theta}}(\boldsymbol{s}_t))=\arg\min_{\boldsymbol{a}_t}\frac{1}{2}\left\|\boldsymbol{a}_t-\pi_{\boldsymbol{\theta}}(\boldsymbol{s}_t)\right\|^2 \\ \text{s.t. } c_i(\boldsymbol{s}_t,\boldsymbol{a}_t)\le C_i \quad \forall i\in L \end{cases} \tag{33}$$

It is necessary for the safety protection layer to disturb the original action as little as possible while ensuring that the modified policy satisfies necessary constraints. A safety layer is constructed on top of the policy network, and the ac-

tion is optimized through local modification during each forward propagation. Figure 3 illustrates the action correction mechanism based on the safety layer.

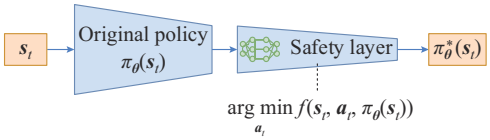

Fig. 3. Action correction mechanism based on safety layer.

As previously mentioned, we can substitute the linear model shown in (34) for $c_i(\boldsymbol{s}_t, \boldsymbol{a}_t)$.

$$\begin{cases} \pi_\theta^*(\boldsymbol{s}_t) = \arg\min_{\boldsymbol{a}_t} \dfrac{1}{2}\left\| \boldsymbol{a}_t - \pi_\theta(\boldsymbol{s}_t) \right\|^2 \\ \text{s.t. } \bar{c}_i(\boldsymbol{s}_t) + (\boldsymbol{g}(\boldsymbol{s}_t; w_i))^{\mathrm{T}} \boldsymbol{a}_t \le C_i \quad \forall i \in L \end{cases} \tag{34}$$

In the formulation, the feasible solution for (34) is represented as $(\pi_\theta^*, \{\lambda_i^*\}_{i=1}^L)$, where λ_i^* is the optimal Lagrangian multiplier associated with the constraints, which is defined as:

$$\lambda_i^* = \left[\frac{(\boldsymbol{g}(\boldsymbol{s}_t; w_i))^{\mathrm{T}} \pi_\theta(\boldsymbol{s}_t) + \bar{c}_i(\boldsymbol{s}_t) - C_i}{(\boldsymbol{g}(\boldsymbol{s}_t; w_i))^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{s}_t; w_i)} \right]^+ \tag{35a}$$

$$\pi_\theta^*(\boldsymbol{s}_t) = \pi_\theta(\boldsymbol{s}_t) - \lambda_{i^*}^* \boldsymbol{g}(\boldsymbol{s}_t; w_{i^*}) \tag{35b}$$

where $i^* = \arg\max_i \lambda_i^*$.

As both the objective function and constraints in (35) are convex functions, the optimality condition for the feasible solution $(\pi_\theta^*, \{\lambda_i^*\}_{i=1}^L)$ is satisfied by the Karush-Kuhn-Tucker (KKT) conditions, which is calculated by:

$$\nabla_{\boldsymbol{a}_t} L = \pi_\theta^*(\boldsymbol{s}_t) - \pi_\theta(\boldsymbol{s}_t) + \sum_{i \in L} \lambda_i^* \boldsymbol{g}(\boldsymbol{s}_t; w_i) = \boldsymbol{0} \tag{36a}$$

$$\lambda_i^* (\bar{c}_i(\boldsymbol{s}_t) + (\boldsymbol{g}(\boldsymbol{s}_t; w_i))^{\mathrm{T}} \pi_\theta^*(\boldsymbol{s}_t) - C_i) = 0 \quad \forall i \in L \tag{36b}$$

C. Training Process

DRL uses dynamic information perceived in the environment to make sequential decisions. The agent interacts continuously with the environment to improve the decision-making effect. Every time the agent completes an episode with the environment, we proceed with the training of parameters. The number of training episodes is denoted as *Epochs*. The training process of the proposed EK-CPPO algorithm is summarized in Algorithm 1, and its main steps are explained as follows.

1) We initialize the training environment by initializing the parameters $\boldsymbol{\theta}^\mu$ of policy network and parameters $\boldsymbol{\theta}^Q$ of value function network. During the interaction between the agent and the environment, the observation is standardized to obtain the state information of $\boldsymbol{s}_t$, which is then used as the input to the policy network to output the action information of $\pi_\theta(\boldsymbol{s}_t)$.

2) Based on the safety layer, the actions are constrained further. After the policy network outputs the action $\pi_\theta(\boldsymbol{s}_t)$, we use $\boldsymbol{g}(\boldsymbol{s}_t; w_i)$ to modify c_i. By using the safety layer, we obtain the adjusted dispatch policy $\pi_\theta^*(\boldsymbol{s}_t)$, namely $\boldsymbol{a}_t$, which strictly satisfies the transmission constraint. Finally, the agent interacts with the environment using the dispatch policy and receives the corresponding reward r_t and the next state $\boldsymbol{s}_{t+1}$. The sampled experience $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})$ is stored in the experience replay pool $\mathcal{B}$.

3) We conduct guided training based on expert experience. We add the two regularization terms shown in (25) to the gradient of the policy network with respect to the parameter $\boldsymbol{\theta}^\mu$. The transfer processes, with the number of J randomly selected from the experience replay buffer, are extracted for use in training the agent based on expert experience, with the aim of maximizing the reward obtained during each interaction.

Algorithm 1: EK-CPPO algorithm for sequential SCOPF

Initialize network parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$

for $t = 1$ **to** *Epochs* **do**

for $h = 1$ **to** H **do**

Initialize training environment, and sample an initial state $\boldsymbol{s}_t$

Choose action $\pi_\theta(\boldsymbol{s}_t)$

Find feasible solution $(\pi_\theta^*, \{\lambda_i^*\}_{i=1}^L)$ by the gradient $\nabla_{\boldsymbol{a}_t} L$

$$\nabla_{\boldsymbol{a}_t} L = \pi_\theta^*(\boldsymbol{s}_t) - \pi_\theta(\boldsymbol{s}_t) + \sum_{i \in L} \lambda_i^* \boldsymbol{g}(\boldsymbol{s}_t; w_i)$$

$$\lambda_i^* (\bar{c}_i(\boldsymbol{s}_t) + (\boldsymbol{g}(\boldsymbol{s}_t; w_i))^{\mathrm{T}} \pi_\theta^*(\boldsymbol{s}_t) - C_i) = 0 \quad \forall i \in L$$

$$\pi_\theta^*(\boldsymbol{s}_t) = \pi_\theta(\boldsymbol{s}_t) - \lambda_{i^*}^* \boldsymbol{g}(\boldsymbol{s}_t; w_{i^*}),\ i^* = \arg\max_i \lambda_i^*$$

Observe $\boldsymbol{s}_{t+1}, r_t$

Store the tuple $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})$ in $\mathcal{B}$

end if

Sample mini-batch of J transitions from $\mathcal{B}$

Calculate the loss function of Critic and Actor

$$L(\boldsymbol{\phi}) \leftarrow -\sum_{t=1}^{H} \left(\sum_{t' > t} \gamma^{t'-t} r_{t'} - V_{\boldsymbol{\phi}}(\boldsymbol{s}_t) \right)^2$$

Update $\boldsymbol{\phi}$ by the gradient $\nabla_{\boldsymbol{\phi}} L$

$$J(\boldsymbol{\theta}) \leftarrow \frac{1}{H} \sum_{t=1}^{H} \left(\frac{\pi_\theta(\boldsymbol{a}_t | \boldsymbol{s}_t)}{\pi_{\theta_{old}}(\boldsymbol{a}_t | \boldsymbol{s}_t)} \hat{A}_t - \beta y_{KL}(\pi_{\theta_{old}}(\cdot | \boldsymbol{s}_t), \pi_\theta(\cdot | \boldsymbol{s}_t)) \right)$$

$$J'(\boldsymbol{\theta}) \leftarrow w_q J(\boldsymbol{\theta}) + \frac{1}{H} \sum_{t=1}^{H} \sum_{r=1}^{2} w_r \cdot reg_{r,t}$$

Update $\boldsymbol{\theta}$ by the gradient $\nabla_{\boldsymbol{\theta}} J'(\boldsymbol{\theta})$

end for

V. Case Studies

The modified IEEE 118-bus system [33] is used as a case to carry out the case studies, which has 54 units, including 36 thermal power units and 18 renewable energy units, and the topology is shown in Fig. 4. The system parameters are shown in Table I. With 5-min time intervals, the constructed dataset contains 106000 intervals in one year. This dataset meets the requirements of typical scenarios for grid operation, including tie network congestion, severe load fluctuation, and renewable energy curtailment. The code is written in Python 3.6.8 based on the deep learning package PyTorch.

The simulation is carried out on a workstation with an Intel Core i7-1165 CPU, 2.8 GHz. The total number of episodes *Epochs* is 5×10^4, and each episode contains 288 sched-

uling intervals, corresponding to a single day with 5-min interval.

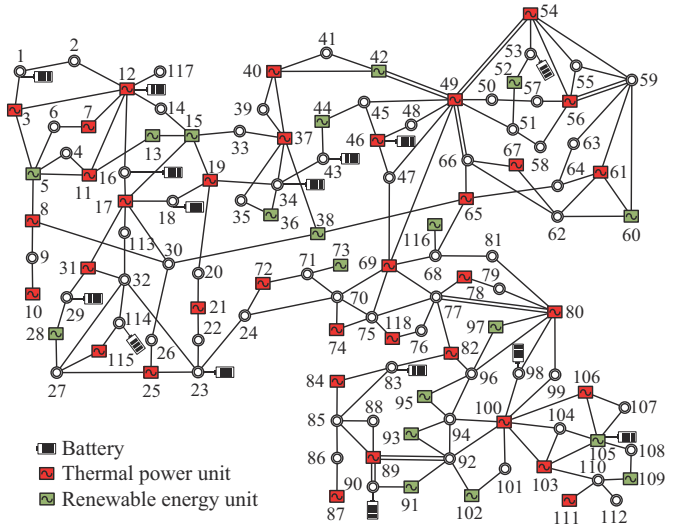

Fig. 4. Modified IEEE 118-bus system topology.

TABLE I
PARAMETERS OF IEEE 118-BUS SYSTEM AFTER MODIFICATION

Parameter	Value	Parameter	Value
Number of areas	3	Number of units	54
Number of buses	126	Number of thermal power units	35
Number of loads	91	Number of renewable energy units	18
Number of lines	185	Peak load (MW)	3687

A. Parameter Setting

Considering that the dimensions of the state and action spaces are 438 and 54, respectively, both the Actor and Critic networks are set to have 3 layers of neurons, and the number of neurons in each layer is 1024, 512, and 256, respectively. Except for the last layer of the Actor network, which uses the tanh activation function, all other neural layers in both the Actor and Critic networks use the ReLU activation function. In addition, the training effect of the neural network is influenced by hyperparameters, and different hyperparameters are suitable for different grid sizes. This paper selects a set of parameters with better training results, as shown in Table II.

TABLE II
PARAMETERS

Parameter	Value
Cost normalization factor (M)	2×10^5
Replay buffer ($\mathcal{B}$)	10^6
Weight of scheduling requirements (c_1, c_2)	2, 3
Regular term weights (w_q, w_1, w_2)	5, 1, 40
Soft update parameter (τ)	0.001
Batch size in training (J)	64
Discount factor (γ)	0.9
Learning rate of Critic (lr^Q)	0.001
Learning rate of Actor (lr^μ)	0.0001

B. Performance of EK-CPPO

1) Training Performance

To highlight the superiority of the proposed EK-CPPO algorithm, we compare its dispatch performance with several state-of-the-art DRL algorithms, including deep deterministic policy gradient (DDPG) [34], twin delayed deep deterministic policy gradient (TD3) [35], PPO, and expert knowledge driven PPO (EK-PPO). In Fig. 5, we run each algorithm for 15 times with different random seeds to compare the dispatch performance of these algorithms. The initial training error curves of policy network are plotted in Fig. 5.

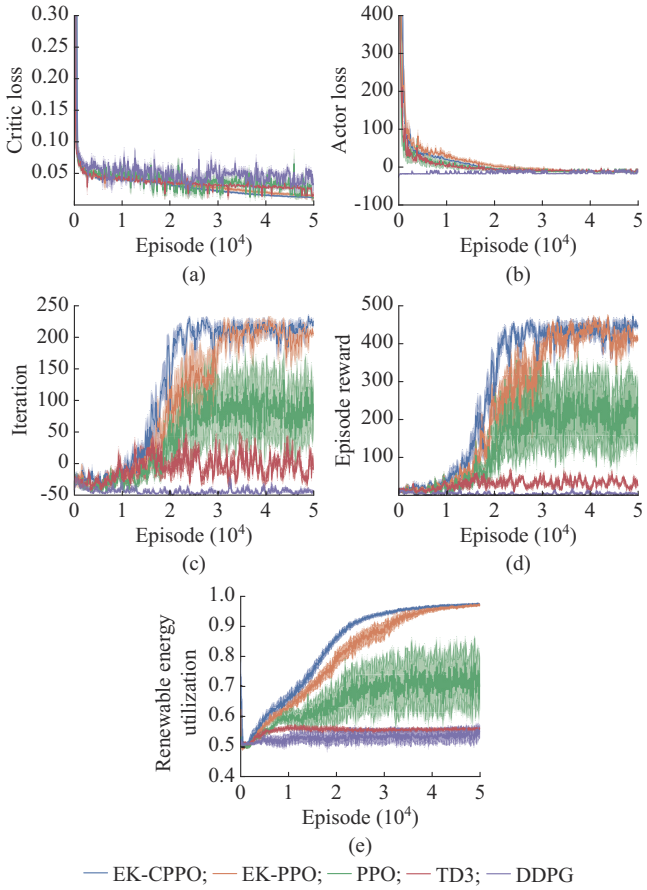

Fig. 5. Initial training error curves of policy network. (a) Critic loss. (b) Actor loss. (c) Iteration. (d) Episode reward. (e) Renewable energy utilization.

The results indicate that: ① DDPG and TD3 fail to achieve effective gradient training, resulting in the failure of the agent to converge. Although the PPO without knowledge guidance can converge, the dispatch effectiveness fluctuates greatly, and the action output by the agent is easy to end the round early because it violates the security policy. ② EK-PPO and EK-CPPO, which can perform sufficient gradient training, have a gradually improved dispatch effectiveness. ③ Compared with EK-PPO (which only meets the constraints of power balance), EK-CPPO has more reliability of power supply and higher robustness due to learning a better security operation strategy for large power system (which meets the constraints of power balance and transmission line capacity). ④ EK-PPO and EK-CPPO can effectively priori-

tize renewable energy utilization, achieving utilization rates of 96.3% and 97.1%, respectively, which are significantly higher than those of the DDPG, TD3, and PPO whose rates are 53.8%, 55.2%, and 71.2%, respectively.

To demonstrate the execution of safety constraints by each algorithm, multiple indicators are recorded during the training process, including the cumulative number of violations for violated power balance constraint (DoV1), the cumulative number of violations for violated transmission line capacity constraint (DoV2), and the cumulative number of episodes that end prematurely due to violated constraint (DoV3). The probability distribution of actions violating constraints is shown in Fig. 6.

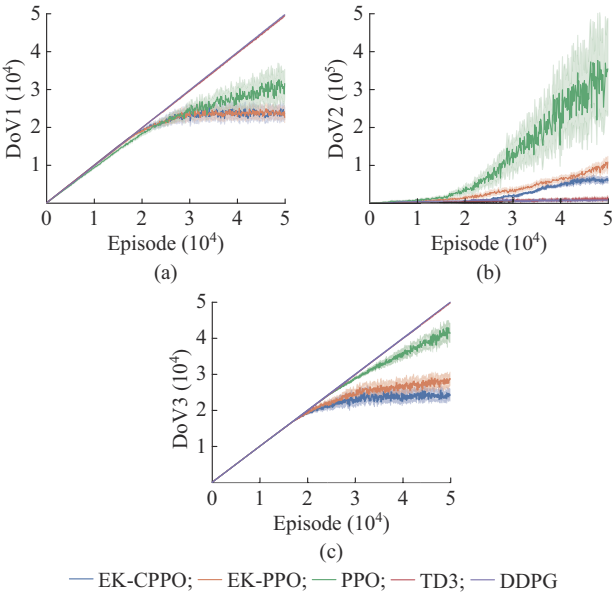

Fig. 6. Probability distribution of actions violating constraints. (a) Change in DoV1. (b) Change in DoV2. (c) Change in DoV3.

1) As shown in Fig. 6(a), without the guidance of expert experience, DDPG, TD3, and PPO fail to learn the correct action, which results in that the agents violate the power balance constraint in most episodes, leading to the early termination of the episode.

2) Under the knowledge guidance of expert experience, EK-PPO and EK-CPPO can quickly learn the safe operation to meet the constraints of power balance. After 300 episodes of training, the agent almost no longer makes actions that violate the constraints of power balance.

3) Figure 6(b) shows that by adding a safety layer to readjust the action, EK-CPPO can effectively learn a dispatch policy that satisfies the transmission constraints through continuous interaction with the environment. It should be noted that DDPG and TD3 greatly reduce the number of interactions with the environment (much less than $5\times10^4\times288$) due to the early end of the episode, which results in a smaller value of DoV2.

After the training has converged, the slope of the EK-CPPO in Fig. 6(c) approaches 0, and the value of DoV3 no longer increases. This indicates that EK-CPPO is able to output the actions that strictly satisfy all the constraints, with very few occurrences of constraint violations. The curves obtained by DDPG, TD3, PPO, and EK-PPO will maintain a certain slope, indicating that the algorithms have not fully learned action policies that satisfy all the constraints. Premature termination of an episode may still occur due to constraint violations.

2) Testing Performance

Figure 7 shows the results of the continuous operation of the EK-CPPO on the IEEE 118-bus system for four days, where we plot in 5-min time intervals. The results indicate that even in the case of a tested system with a large number of generating units, the EK-CPPO can still successfully learn accurate strategies to coordinate the operation of thermal power units, renewable energy units, and energy storage systems, ensuring the safe operation and economic benefits of the power system. Besides, the energy storage device is charged during non-peak electricity demand periods or when the renewable energy resources are abundant, and discharged during peak electricity demand periods or when the renewable energy resources are scarce, which is consistent with actual situation. Furthermore, it can be observed that the voltage output by the agent designed in this paper can effectively maintain the voltage of each node in the power system within a reasonable range of [0.95,1.05]p.u..

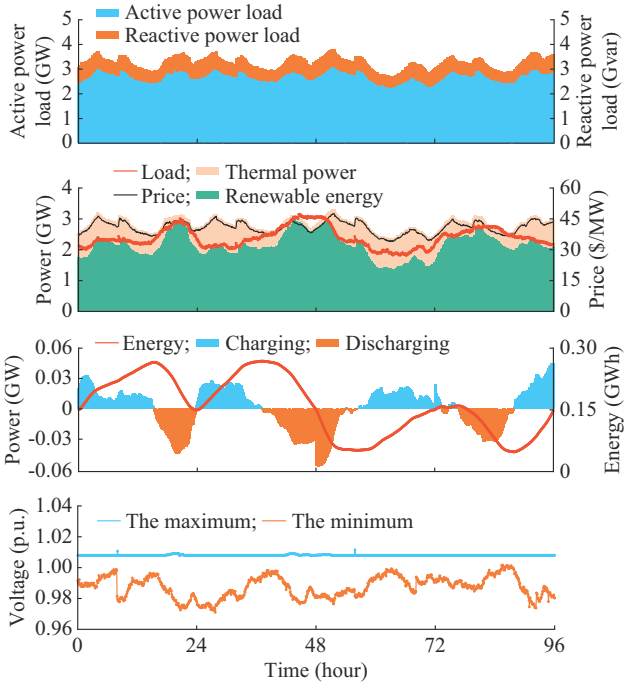

Fig. 7. Results of continuous operation of EK-CPPO on IEEE 118-bus systems.

The probability distribution of the five algorithms triggering the simulator to prompt "illegal action" due to constraint violations is calculated, as shown in Table III. The main reasons for the prompt "illegal action" include the violated power balance and transmission line overloads ("soft overloads" for more than four time steps or "hard overloads").

To evaluate the effectiveness of EK-CPPO in handling random line outage, we conduct experiments in five different scenarios on the test set, as shown in Table IV.

TABLE III
PROBABILITY OF TRIGGERING SIMULATOR TO PROMPT "ILLEGAL ACTION"

Algorithm	Violated power balance (%)	Transmission line overload (%)	Total
DDPG	97.8	2.2	100.0
TD3	96.5	2.1	98.6
PPO	25.2	40.1	65.3
EK-PPO	0.5	4.0	4.5
EK-CPPO	0.1	0.3	0.4

TABLE IV
POWER CHANGE OF LINES

Scenario	Disconnected line	Overloaded line	The maximum allowable power (p.u.)	Power flow (EK-PPO) (p.u.)	Power flow (EK-CPPO) (p.u.)
1	Line 44	Line 46	0.85	1.22	0.65
2	Line 66	Line 61	1.10	1.21	0.75
3	Line 85	Line 71	0.85	0.98	0.45
		Line 89	1.00	1.34	0.67
4	Line 114	Line 115	1.20	1.48	0.75
		Line 118	1.00	1.36	0.52
		Line 121	0.70	1.02	0.38
5	Line 123	Line 138	0.85	1.18	0.52

We simulate the situation where random line disconnections occur in the power system to test the responsiveness of the agent. A comparison is conducted between the line power generated by the EK-CPPO considering line capacity constraints and the allowed maximum line power. Compared with EK-PPO and due to the influence of the safety layer, EK-CPPO effectively eliminates the occurrence of line overcurrents, which is shown in Table IV. As for scenario 1, we set the simulation to run until the 100th interval when a line disconnection occurs. The voltage variation curves of the five units (G-36, G-38, G-46, G-65, G-116) near the disconnected line are shown in Fig. 8. It can be observed that when the line disconnects, there will be a brief and severe fluctuation in node voltage, but it subsequently stabilizes. Throughout the process, none of the node voltages exceed the threshold, which fully verifies the strong enforcement capability of the intelligent agent towards the constraints.

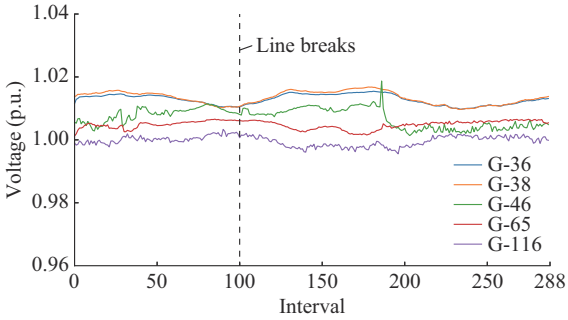

Fig. 8. Voltage variation curves of five units near disconnected line.

The comparison of comprehensive evaluation results is shown in Table V. IPOPT is based on SCOPF with the minimum adjustment amount of unit output [36]. IPOPT converts the MIP problem to linear programming problem by adding the relaxation constraint of the output limit of the units. Finally, the solver IPOPT [37] is called to solve the problem. Results show that the IPOPT adjusts the unit output based on the principle of the minimum adjustment amount. Since it does not activate enough units to provide reserves to cope with the uncertainty of renewable energy, it has the lowest operation cost. EK-CPPO based on SRL can dynamically respond to the stochastic fluctuations of load and new energy resources, achieving the highest utilization rate of renewable energy among all algorithms (up to 97.8%). Moreover, in terms of solution efficiency, EK-CPPO is significantly faster than the IPOP.

TABLE V
COMPARISON OF COMPREHENSIVE EVALUATION RESULTS

Algorithm	Iteration	Operation cost (¥)	Renewable energy utilization (%)	Average reward per step	Average reward per episode	Online computation time (s)
IPOPT	287	53890.2	96.9	1.491	428.2	270
PPO	120	61548.3	72.1	0.720	86.4	<20
EK-PPO	274	57547.3	97.2	1.470	402.8	<20
EK-CPPO	287	54013.4	97.8	1.490	427.6	<22

VI. CONCLUSION

The existing SCOPF problem based on RL ignores the security risks that may be brought by the agent in the exploration process, and there is a risk of issuing instructions that may threaten the safe operation of the power system. Aiming at the sequential SCOPF problem, we propose a CMDP formulation, which has been tested on the improved IEEE 118-bus system. The main innovations and conclusions of this paper are summarized as follows.

1) EK-CPPO is used to determine the output scheme of the units, the reactive power and voltage optimization scheme of the units, and the charging and discharging dispatch of energy storage systems, which can adapt to the uncertain changes of intermittent power sources and load. The proposed method is completely based on data-driven implementation and does not depend on any physical model, statistical model, or mathematical programming optimizer.

2) The proposed method plays a significant role in ensuring the safe, stable, and economic operation of the power systems. By introducing expert experience and safety layers, the agent can learn cost-effective operation through a safe action exploration mechanism. Experimental results show that

the proposed method can make the dispatch strategy strictly abide by the safety constraints, and significantly improve the execution effect of capacity constraint of transmission line.

3) While ensuring the dispatch economy of the power system and ensuring the normal operation of the power system, the proposed method maximizes the renewable energy utilization and effectively improves the accomodation capacity of renewable energy.

REFERENCES

[1] X. Wei, Y. Xiang, J. Li *et al.*, "Self-dispatch of wind-storage integrated system: a deep reinforcement learning approach," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 3, pp. 1861-1864, Jul. 2022.

[2] E. B. Fisher, R. P. O'Neill, and M. C. Ferris, "Optimal transmission switching," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1346-1355, Aug. 2008.

[3] D. Phan and J. Kalagnanam, "Some efficient optimization methods for solving the security-constrained optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 863-872, Mar. 2014.

[4] J. Kardoš, D. Kourounis, and O. Schenk, "Two-level parallel augmented Schur complement interior-point algorithms for the solution of security constrained optimal power flow problems," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1340-1350, Mar. 2020.

[5] L. de M. Carvalho, A. M. L. da Silva, and V. Miranda, "Security-constrained optimal power flow via cross-entropy method," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6621-6629, Jun. 2018.

[6] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: reinforcement learning framework," *IEEE Transactions on Power Systems*, vol. 19, no. 1, pp. 427-435, Feb. 2004.

[7] J. G. Vlachogiannis and N. D. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Transactions on Power Systems*, vol. 19, no. 3, pp. 1317-1325, Aug. 2004.

[8] E. A. Jasmin, T. P. I. Ahamed, and V. P. J. Raj, "Reinforcement learning approaches to economic dispatch problem," *International Journal of Electrical Power & Energy Systems*, vol. 33, no. 4, pp. 836-845, May 2011.

[9] X. Han, C. Mu, J. Yan *et al.*, "An autonomous control technology based on deep reinforcement learning for optimal active power dispatch," *International Journal of Electrical Power & Energy Systems*, vol. 145, p. 108686, Feb. 2023.

[10] J. Li, T. Yu, X. Zhang *et al.*, "Efficient experience replay based deep deterministic policy gradient for AGC dispatch in integrated energy system," *Applied Energy*, vol. 285, p. 116386, Mar. 2021.

[11] T. Wang, Y. Tang, B. Wang *et al.*, "Traditional methods and artificial intelligence: current status, challenges and future directions of power flow control optimization algorithms," *Proceedings of the CSEE*, vol. 43, no. 5, pp. 1799-1818, Mar. 2023.

[12] J. Zhang, T. Pu, Y. Li *et al.*, "Distributed power supply optimal scheduling strategy based on multi-agent deep reinforcement learning," *Power System Technology*, vol. 46, no. 9, pp. 3496-3504, Sept. 2022.

[13] Z. Yang, Z. Ren, Z. Sun *et al.*, "A security-constrained economic dispatch method for renewable energy power system based on near-end strategy optimization algorithm," *Power System Technology*, vol. 47, no. 3, pp. 988-998, Mar. 2023.

[14] H. Li and H. He, "Learning to operate distribution networks with safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 1860-1872, May 2022.

[15] A. R. Sayed, X. Zhang, G. Wang *et al.*, "Optimal operable power flow: sample-efficient holomorphic embedding-based reinforcement learning," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1739-1751, Apr. 2023.

[16] Z. Yi, X. Wang, C. Yang *et al.*, "Real-time sequential security-constrained optimal power flow: a hybrid knowledge-data-driven reinforcement learning approach," *IEEE Transactions on Power Systems*, vo. 39, no. 1, pp. 1664-1680, Apr. 2023.

[17] H. Chen and X. Wang, "An overview of optimization methods for unit commitment problems," *Automation of Electric Power Systems*, vol. 23, no. 5, pp. 51-56, Mar. 1999.

[18] W. Wang, N. Yu, Y. Gao *et al.*, "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Jul. 2020.

[19] N. Heess, G. Wayne, D. Silver *et al.*, "Learning continuous control policies by stochastic value gradients," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Cambridge, USA, Mar. 2015, pp. 2944-2952.

[20] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2427-2439, May 2020.

[21] J. Schulman, S. Levine, P. Abbeel *et al.*, "Trust region policy optimization," in *Proceedings of the 32th International Conference on Machine Learning*, Lille, France, Feb. 2015, pp. 1889-1897.

[22] T. Wang, Y. Liu, X. Gu *et al.*, "Identification of vulnerable power lines based on cascading fault spatiotemporal diagram," *Proceedings of the CSEE*, vol. 39, no. 20, pp. 5962-5972, Sept. 2019.

[23] R. A. Jabr, A. H. Coonick, and B. J. Cory, "A homogeneous linear programming algorithm for the security constrained economic dispatch problem," *IEEE Transactions on Power Systems*, vol. 15, no. 3, pp. 930-936, Aug. 2000.

[24] K. E. Van Horn, A. D. Domínguez-García, and P. W. Sauer, "Measurement-based real-time security-constrained economic dispatch," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3548-3560, Sept. 2016.

[25] C. C. White and D. J. White, "Markov decision processes," *Eurpean Journal of Operational Research*, vol. 39, no 1, pp. 1-16, Mar. 1989.

[26] Z. Xiao, S. Jia, J. Zhu *et al.*, "Power control strategy of tie-line in wind microgrid," *Journal of Electrical Technology*, vol. 32, no. 15, pp. 169-179, Aug. 2017.

[27] Y. Chen, C. Wu, and J. Qi, "Data-driven power flow method based on exact linear regression equations," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 800-804, May 2022.

[28] M. U. Qureshi, S. Grijalva, M. J. Reno *et al.*, "A fast scalable quasi-static time series analysis method for PV impact studies using linear sensitivity model," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 1, pp. 301-310, Jan. 2019.

[29] S. Wang, Q. Liu, and X. Ji, "A fast sensitivity method for determining line loss and node voltages in active distribution network," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1148-1150, Jan. 2018.

[30] Z. Xu, Y. Xiao, Q. Li *et al.*, "A comparative study of power crossing control methods based on sensitivity and particle swarm optimization," *Power System Protection and Control*, vol. 48, no. 15, pp. 177-186, Aug. 2020.

[31] Y. Chen, H. Chen, Y. Jiao *et al.*, "Data-driven robust state estimation through off-line learning and on-line matching," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 897-909, Jul. 2021

[32] Y. Chen, A. D. Dominguez-Garcia, and P. W. Sauer, "Measurement-based estimation of linear sensitivity distribution factors and applications," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1372-1382, May 2014.

[33] I. Peña, C. B. Martinez-Anido, and B.-M. Hodge, "An extended IEEE 118-bus test system with high renewable penetration," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 281-289, Jan. 2018.

[34] Y. Duan, X. Chen, R. Houthooft *et al.*, "Benchmarking deep reinforcement learning for continuous control," in *Proceeding of the 33rd International Conference on Machine Learning*, New York, USA, Jun. 2016, pp. 1329-1338.

[35] H. Liu, F. Liang, T. Hu *et al.*, "Multi-scale fusion model based on gated recurrent unit for enhancing prediction accuracy of state-of-charge in battery energy storage systems," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 2, pp. 405-414, Mar. 2024.

[36] J. Geng, L. Li, J. Yao *et al.*, "New SCED based on minimum bias objective for large-scale wind integrated power systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 14881-14885, Jan. 2011.

[37] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25-57, Mar. 2006.

Yanbo Chen received the B.S., M.S., and Ph.D. degrees in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, China Electric Power Research Institute, Beijing, China, and Tsinghua University, Beijing, China, in 2007, 2010, and 2013, respectively. He is currently a Professor of North China Electric Power University, Beijing, China. He is also with School of Engineering, Xining University, Xining, China, and with New Energy (Photovoltaic) Industry Research Center, Qinghai University, Qinghai, China. His research interests include state estimation and power system analysis and control.

Qintao Du received the B.E. degree in electrical engineering from Guangdong University of Technology, Guangzhou, China, in 2021. He is currently pursuing the M.S. degree in electrical engineering at the North China Electric Power University, Beijing, China. His research interests include state estimation and deep reinforcement learning applications in power systems.

Honghai Liu received the B.E. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2022. He is currently pursuing the Ph.D. degree in electrical engineering at the North China Electric Power University, Beijing, China. His research interests include optimization of virtual power plant and deep reinforcement learning applications in power systems.

Liangcheng Cheng received the B.E. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in electrical engineering at the Graduate School of China Electric Power Research Institute, Nanjing, China. His research interests include power system analysis and deep reinforcement learning applications in power systems.

Muhammad Shahzad Younis received the bachelor degree from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, the master degree from the University of Engineering and Technology, and the Ph.D. degree from University Technology PETRONAS, Malaysia. He is currently a Tenured Professor at NUST. His research interests include spanning multidisciplinary areas such as smart grid, digital twin for industrial applications, embedded systems, and signal processing and artificial intelligence.