Research on Optimal Power Grid Scheduling Based on Transfer Reinforcement Learning

Qihui Dai, Xiang Hu, Jianlong Li, Aomei Jiang, Wei Xiong
Control and Computer Engineering, North China Electric Power University, Chang'ping, Beijing 102200, China
Corresponding author: Qihui Dai (e-mail: dqh7889@163.com)

ABSTRACT To enhance power grid adaptability amid rising renewable energy integration, this paper proposes M^3-PPO, a meta-reinforcement learning algorithm that enables efficient the strategy transfer and rapid adaptation across tasks with varying energy mixes. Built upon a base framework (M-PPO) that integrates PPO and MAML, M^3-PPO introduces two key innovations to overcome MAML's training instability: a Mamba-based context encoder for richer task representation in the inner loop, and a global-local momentum update mechanism for smoother meta-parameter optimization in the outer loop. Experiments on the Grid2Op platform demonstrate that M^3-PPO significantly outperforms baseline algorithms in generalization and scheduling efficiency, achieving robust performance even when simulating complex energy environments. The approach is particularly suitable for integration with antenna-enabled smart grid monitoring, wireless data acquisition, and edge-computing platforms, providing an engineering-oriented solution for adaptive, real-time, and robust power grid scheduling in modern renewable-rich energy systems.

INDEX TERMS power grid scheduling, transfer reinforcement learning, M^3-PPO, wireless smart grid

I. INTRODUCTION

THE stable control of the power grid, a critical infrastructure, is of utmost importance. Traditional grid control relies on manual operations based on standard procedures and historical experience, which often struggle to accommodate the sudden, high-magnitude load fluctuations frequently imposed by major industries. Driven by the pursuit of the "dual-carbon goal", the increasing proportion of renewable energy has intensified volatility on both the source and load sides, making it difficult for traditional grid control methods to cope. Therefore, there is an urgent need to develop automated grid control methods that can adapt to the dynamic characteristics of renewable energy.

Research on automated power grid control can be mainly divided into two categories: model-based and model-free methods. Model-based methods transform the automated power grid control problem into a mathematically constrained optimization problem by constructing a mathematical model of the operation of the power system [1,2], and the model is then solved using optimization techniques such as dynamic programming and linear programming. Although having a solid theoretical foundation, model-based methods often face limitations such as severe dependence on domain expertise and poor scalability in practical applications [3], thus prompting researchers to turn to explore model-free methods [4]. Model-free methods (such as methods based on cybernetics or reinforcement learning) do not rely on accurate mathematical models, but learn power grid control decisions from historical power grid control data in a data-driven manner, and thus have received extensive attention in the field of power grid control tasks. Among them, reinforcement learning methods combined with deep learning techniques can effectively handle high-dimensional state spaces and complex system dynamics, and autonomously learn and optimize strategy by interacting with the simulation environment and in a trial-and-error manner [5].

Reinforcement learning has currently shown great potential in automated grid control tasks such as power grid energy management [6], voltage control [7], and power system emergency control [8]. Related research mainly focuses on how to optimize the control strategy of different tasks through the reinforcement learning framework, so as to improve control performance, exploration efficiency, and real-time response ability. Although reinforcement learning has shown excellent potential for automated grid control, as the grid's energy composition evolves toward higher proportions of renewables, a strategy trained on one specific mix lacks generalization ability for different, albeit related, tasks. Consequently, these strategies struggle to handle new control tasks that emerge as the grid environment evolves into new states with differing energy mixes. To overcome this limitation and endow RL agents with the necessary

generalization capability, it is imperative to enable the effective transfer of common decision-making knowledge from multiple historical tasks. This capability guides agents to efficiently explore and rapidly adapt to new tasks. Figure 1 shows the knowledge transfer and few-shot rapid adaptation tasks of grid scheduling strategy in different energy combination environments.

By enabling knowledge transfer across tasks, meta-reinforcement learning has established itself as an important approach for enhancing RL adaptability, showing considerable promise in grid control applications [8-10]. This paper focuses on the problem of cross-task transfer and rapid adaptation of strategy in automated grid control, and proposes a meta-reinforcement learning method that combines context enhancement and momentum update mechanisms to improve the adaptation efficiency and training stability of strategy in new tasks. This method is based on the Proximal Policy Optimization algorithm (PPO) and combines the model-agnostic meta-learning method (MAML), and is built on the M-PPO framework to inherit the advantages of both in policy optimization and rapid adaptation. Aiming at the problem that the double-layer optimization framework of MAML often leads to unstable training due to higher-order differentiation, we innovate from two aspects of inner and outer layer optimization: in the inner layer optimization, inspired by the context-based meta-reinforcement learning method, by introducing a context encoder to enhance the ability to express task context features, it not only improves the exploration and adaptation efficiency of the algorithm for new tasks, but also enhances the consistency of the inner layer weight adjustment direction and reduces the gradient volatility; in the outer layer optimization, inspired by the exponential moving average (EMA) mechanism (such as Adam [11]), a global-local momentum update mechanism is designed to smooth the meta-parameter update process, thereby improving the stability of meta-gradient updates and making the meta-gradient estimation more accurate and reliable.

Our main contributions include the following three aspects:

1. A novel M^3-PPO algorithm for transferable grid scheduling.

We develop a meta-reinforcement learning architecture, termed M^3-PPO, built upon PPO and MAML. A global-local momentum update mechanism is introduced to mitigate the training instability inherent in MAML. By transferring scheduling knowledge across power grid environments with different energy compositions, M^3-PPO enables efficient exploration and rapid policy adaptation in unseen tasks, thereby significantly enhancing generalization performance.

2. A Mamba-based context encoder for enhanced task representation.

To address the limitations of conventional sequence modeling methods in capturing complex task contexts, the Mamba model [12] is employed as the context encoder. Leveraging its long-range dependency modeling capability, the proposed approach strengthens task feature extraction and representation, thereby improving the agent's adaptability to related new tasks and enhancing the consistency of inner-loop policy updates.

3. Comprehensive validation of performance superiority in grid scheduling tasks.

Extensive experiments conducted on the Grid2Op simulation platform using the L2RPN public dataset demonstrate that M^3-PPO consistently outperforms benchmark algorithms across multiple power grid scheduling scenarios. In environments characterized by complex energy structures, the proposed method achieves faster adaptation and more stable policy performance, confirming its practical effectiveness.

II. RELATED WORKS

A. MODEL-BASED AUTOMATED POWER GRID CONTROL METHODS

Model-based automated power grid control methods rely on accurate mathematical modeling of power grid operation, transforming complex power grid control problems into constrained optimization problems. Sridhar [1] et al. proposed a model-based anomaly detection and attack mitigation algorithm to enhance the resilience of power systems against cyberattacks. Li et al. [2] constructed a cyber-physical model for the LFC system and proposed a GAN-based FDIA detection and defense mechanism to improve the ability of power systems to cope with false data injection attacks. Although model-based methods provide a solid foundation for power grid control in theory and methodology, their limitations in highly complex and dynamically changing scenarios are prompting researchers to seek integration with data-driven methods, such as combining deep reinforcement learning techniques to alleviate modeling dependence and improve adaptability.

B. REINFORCEMENT LEARNING-BASED AUTOMATED POWER GRID CONTROL METHODS

In recent years, with the continuous improvement of the intelligence level of power grids, reinforcement learning has shown broad application prospects in the field of power grid control, covering multiple key directions such as energy management [6], voltage regulation [7], and emergency control [13]. Samadi et al. [6] proposed a multi-agent distributed energy management method based on reinforcement learning to optimize the behavior and operating costs of distributed energy. Duan et al. [7] proposed an autonomous voltage control framework "Grid Mind" based on deep reinforcement learning, combining DQN and DDPG algorithms to achieve fast response and timely control actions. Chen et al. [13] proposed a model-free emergency control method based on reinforcement learning, combining multi-Q learning and DDPG algorithms to handle limited scenarios and general multi-scenario emergency control respectively, and prevent the power system from collapsing.

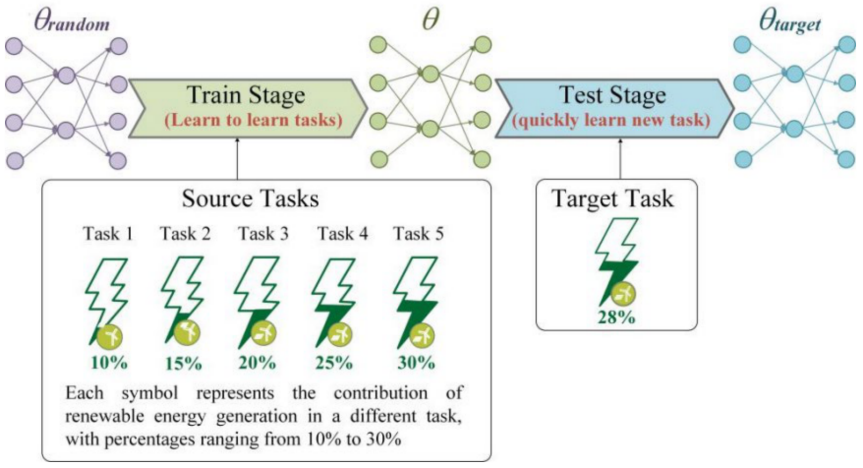

FIGURE 1. Schematic diagram of cross-task knowledge transfer and rapid adaptation of power grid scheduling strategy

In addition, the application of reinforcement learning in fields such as frequency regulation [14] and power grid scheduling [15] has further broadened its application scenarios in power grid control. Meanwhile, competition platforms such as L2RPN [16] and CityLearn [17] have also promoted the practical application research of reinforcement learning control methods by focusing on specific issues such as robustness, adaptability, and microgrid energy coordination. Although reinforcement learning has demonstrated excellent performance in the field of power grid control, it still faces many challenges in practical applications, especially deep reinforcement learning models often struggle to adapt to complex scenarios outside the training environment due to overfitting and lack of generalization ability [18]. Therefore, how to use prior knowledge to improve model adaptability has become one of the important directions in the field of power grid control [19]. In this context, meta-reinforcement learning provides a new idea for solving the generalization problem by learning strategy to quickly adapt to new tasks in the meta-training stage.

C. META-REINFORCEMENT LEARNING-BASED AUTOMATED POWER GRID CONTROL METHODS

Meta-reinforcement learning aims to improve the cross-task adaptation ability of reinforcement learning models and has been widely studied in multiple fields in recent years. Its core idea is to learn the initial parameters or context generation mechanism of the model through the meta-training stage, enabling it to quickly adapt to similar new tasks and thus enhancing the generalization performance. Existing meta-reinforcement learning methods are mainly divided into two categories: context-based methods [20] and gradient-based methods [21]. In fact, the excellent cross-task adaptation ability of meta-reinforcement learning has been demonstrated in many fields [22-24]. For example, in the field of robot control, Belmonte-Baeza et al. [22] proposed a design optimization framework based on meta-reinforcement learning, and its application to the optimizing kinematics and actuator parameters of quadrupedal robots. By training a locomotion policy that can quickly adapt to different designs, it is possible to control robots of different designs to track random velocity commands over various rough terrains. In the field of autonomous driving, Ye et al. [23] proposed an adaptive lane-changing strategy based on meta-reinforcement learning. By integrating deep reinforcement learning and the MAML framework, it achieved the rapid adaptation and safe decision-making of autonomous vehicles in different traffic conditions and dynamic environments, thus significantly enhancing the decision-making generalization ability and safety in complex driving scenarios. The above research shows the effectiveness and universality of meta-reinforcement learning in improving the generalization and adaptability of models.

The successful experiences of meta-reinforcement learning in multiple fields provide a reference for its application in other complex scenarios. Especially in power grid control tasks, its cross-task adaptation ability shows great potential, and many innovative studies have emerged. For example, Huang et al. [8] proposed a deep meta-reinforcement learning algorithm for training power system emergency control policies against voltage stability problems and fast adapting them to unseen operation conditions and/or contingency scenarios. Xiong et al. [9] proposed a Meta-RL-based transferable scheduling strategy for Home Energy Management Systems, which is designed to maintain performance across seasons and user types while reducing energy costs with less training data and time. Jakobeit et al. [10] proposed a control method based on meta-reinforcement learning for cross-task current control of permanent magnet synchronous motors. Through meta-learning, it achieved rapid adaptation to motors with different power ratings and significantly improved the generality and transfer ability of the control policy. These studies together show that meta-reinforcement learning provides an innovative solution to the adaptation problem in complex dynamic power system scenarios.

III. PROBLEM FORMULATION

This section begins by detailing the power grid under study, including its basic configuration and operational principles. Subsequently, we elaborate on how to model the power grid scheduling problem as a Markov decision process (MDP).

A. POWER GRID OPTIMIZATION SCHEDULING MODEL

1. Basic Structure of the Power Grid

As shown in Figure 2, the power grid usually consists of the following key components: the generators $g \in G$, the loads $c \in C$, the substations $s \in S$, and the transmission lines $l \in L$. The output power of a generator is defined as active power $P_G \in P_G$ and reactive power $Q_G \in Q_G$, while the load demand is similarly defined by $P_C \in P_C$ and $Q_C \in Q_C$. The power flow on a transmission line is defined as active power $P_{ij} \in P$ and reactive power $Q_{ij} \in Q$, where each transmission line connects substations s_i and s_j. The current on transmission line $l \in L$ is represented as I_l. The topology of the power grid is represented by $\kappa \in K$, where κ defines the node connectivity of the power grid, and K is the set of all possible topologies. The conductance and susceptance between nodes are represented by the admittance matrices M_G^κ and M_B^κ respectively.

2. Objective Function

The objective of power grid optimization scheduling is to minimize the total operating cost by optimizing the scheduling policy while meeting the requirements of safe operation and supply-demand balance. This objective function consists of two parts: the transmission loss cost $\zeta_{\text{trans}}(t)$ and the rescheduling cost $\zeta_{\text{adjust}}(t)$, and its mathematical expression is as follows:

$$\min_{P_G, Q_G, \kappa} C_{\text{grid}} = \sum_{t=1}^{T} \left(\zeta_{\text{trans}}(t) + \zeta_{\text{adjust}}(t)\right) \tag{1}$$

Transmission loss cost $\zeta_{\text{trans}}(t)$*:* The energy loss caused by the Joule effect in the transmission line.

Rescheduling cost $\zeta_{\text{adjust}}(t)$*:* The cost of increasing or decreasing the generator output power to maintain the balance of the power grid. Its value is positively correlated with the adjustment amplitude of the generator output power and is expressed as:

$$\zeta_{\text{adjust}}(t) \propto \sum_{g \in G} |P_{G,g}(t) - P_{G,g}(t-1)| \tag{2}$$

3. Model Constraints

(1) Topology Control Constraint

To optimize the operation of the power grid, the network structure of the system can be dynamically adjusted through topology control actions κ_t. Its update rule is:

$$M_G^{t+1}, M_B^{t+1} = f(M_G^t, M_B^t, \kappa_t), \quad \forall t \tag{3}$$

(2) Power flow equation constraints

The power flow equation is the basic equation for power flow calculation, used to express the power balance relationship among nodes in the power grid. The power flow equation is expressed as:

$$f(\theta_t, v_t, P_t, Q_t; M_G^t, G_B^t) = 0, \quad \forall t \tag{4}$$

(3) Variable range constraints

Variables such as node voltage, generator output power, and line power flow need to be within the specified range, specifically expressed as:

$$\begin{cases} \theta_t \in \Theta, \quad \forall t, \\ v_t \in V, \quad \forall t, \\ P_G(t) \in P_G, \quad Q_G(t) \in Q_G, \quad \forall t, \\ P_{ij}(t) \in P, \quad Q_{ij}(t) \in Q, \quad \forall t. \end{cases} \tag{5}$$

B. MODELING OF POWER GRID SCHEDULING PROBLEM BASED ON MARKOV DECISION PROCESS

Given that Equation (1) constitutes a non-linear mixed-integer optimization problem, traditional methods like dynamic programming often struggle with real-time performance and model dependency. To address this, we model the grid scheduling task as a MDP, defined by the tuple (S, A, P, R). The components of this tuple are detailed below.

1. state space S

The composition of the state space S is crucial for solving the power grid scheduling problem. We define the state at time t as s_t, it integrates key features including the power values of generators and loads $(S_{PG}, S_{QG}, S_{PC}, S_{QC})$, generator dispatch targets and their actual implementations $(S_{\text{dispatch,target}}, S_{\text{dispatch,actual}})$, operational constraints (line/substation cooling times $S_{\text{tcool,line}}, S_{\text{tcool,sub}}$ and line overflow duration $S_{\text{toverflow}}$), and topological information (bus connectivity S_κ, line capacity utilization S_ρ, and line status S_s).

2. action space A

The action space A is composite, comprising: (1) discrete topological actions a_κ that reconfigure line-to-bus connections within substations, and (2) continuous rescheduling actions a_G that adjust the active and reactive power setpoints of generators to optimize power flow and cost.

3. transition probability $P : S \times A \times S \to R$

The transition probability P is governed by the power flow and topology Equations (3) and (4), which deterministically define the next state s' in our simulation environment. The dynamics of the grid environment are inherently stochastic, primarily due to the continuous fluctuations in load demand and renewable power output. Different energy mix scenarios, with renewable energy penetration rates varying from 10% to 30%, are each associated with a distinct time series. At each time step, after the agent takes an action, the next state s' is computed deterministically by the built-in power flow solver based on the current state, the action, and the specific load and generation values at that time

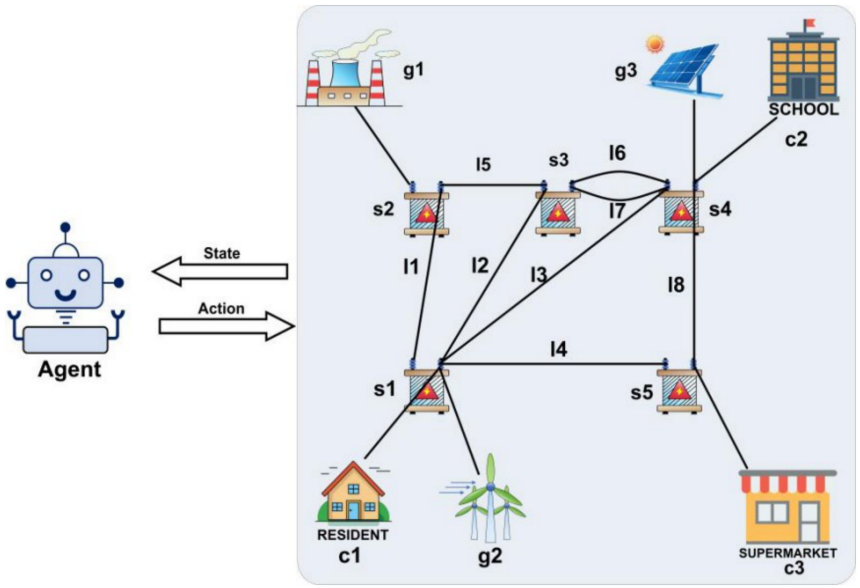

FIGURE 2. Schematic diagram of the basic structure of the power grid

step, using equations (3) and (4). Therefore, the transition probability P captures the fact that the simulator generates state transition paths based on deterministic physical rules, which are driven by a sequence of stochastic inputs.

4. reward function $R : S \times A \to R$

The design of the reward function R aims to guide the agent to balance the economy and system stability in the scheduling process. At each time step t, the reward is calculated based on the definition of the L2RPN environment [16], and the specific form is as follows:

$$r(s_t, a_t) = \begin{cases} C_{\text{const}} - C_{\text{grid}}(t), & \text{normal operation} \\ -1, & \text{system crash} \end{cases} \tag{6}$$

As shown in formula (6), when the system is operating normally, positive feedback is provided according to the value of $C_{\text{const}} - C_{\text{grid}}(t)$. Among them, C_{const} is a constant, and $C_{\text{grid}}(t)$ is the actual operating cost at time t. When the system collapses, the reward value is set to -1 to impose a severe penalty and explicitly avoid the risk of instability.

5. objective function

The agent's objective is to find a policy π_θ that maximizes the expected cumulative discounted reward:

$$\max_{\theta} \; W(\theta) = E_{\pi_\theta}\left[\sum_{t=1}^{T} \gamma^t r(s_t, a_t)\right] \quad \text{s.t.} \quad a_t = \pi_\theta(s_t), \quad a \in A, \quad s \in S \tag{7}$$

Where γ is the discount factor balancing short-term and long-term rewards.

IV. METHODOLOGY

We first introduce the basic knowledge of gradient-based meta-learning methods and the Mamba model, laying a theoretical foundation for combining gradient-based and context-based meta-reinforcement learning methods to achieve rapid adaptation. Then, we present the key steps of our proposed M^3-PPO (a Momentum- and Mamba-empowered Meta-RL framework based on PPO) algorithm and clarify how they fit in with existing power grid operation procedures to promote the agent's rapid exploration and efficient adaptation in new environments.

A. GRADIENT-BASED META-LEARNING METHODS

Gradient-based meta-learning methods optimize the initial parameters of the model so that it can quickly adapt in new tasks through few gradient descent steps, usually implemented by two optimization processes: inner-layer optimization and outer-layer optimization.

1. Inner-layer optimization process

The model parameters θ are iteratively updated according to the loss function L_{T_i} for each task T_i to obtain new parameters θ_i' adapted to the task. The typical iterative method is single-step gradient descent, that is:

$$\theta_i' = \theta - \alpha \nabla_\theta L_{T_i}(f_\theta) \tag{8}$$

where α is the inner-loop learning rate.

2. Outer-layer optimization process

The initial parameters θ of the model are optimized by integrating the learning experiences of each task in the inner loop. Its goal is to make θ perform well on all tasks after inner fine-tuning. The update formula is as follows:

$$\theta = \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} L_{T_i}(f_{\theta_i'}) \tag{9}$$

where β is the meta-learning rate.

B. MAMBA MODEL

Traditional linear time-invariant state space models (SSMs) are computationally efficient but limited in sequence context

representation. To address this, Mamba [12] introduces a selective state space model (S6) with an input-dependent selection mechanism. As shown in Figure 3, the core innovation is dynamic parameterization: instead of static parameters, the key matrices B, C, and the time scale parameter Δ are projected from the input sequence x. Specifically, the parameters B, C and Δ are dynamically projected from the input sequence $x \in \mathbb{R}^{B\times L\times D}$, as follows:

$$B = S_B(x) \tag{10}$$

$$C = S_C(x) \tag{11}$$

$$\Delta = \tau_\Delta\left(\Delta + S_\Delta(x)\right) \tag{12}$$

Among them, $B \in \mathbb{R}^{B\times L\times N}$, $C \in \mathbb{R}^{B\times L\times N}$, $\Delta \in \mathbb{R}^{B\times L\times D}$. The projection functions S_B and S_C perform linear projection to the N-dimensional space. The symbol τ_Δ represents the Softplus activation function [25], and $\text{Projection}_\Delta = \text{Broadcast}_\Delta(\text{Linear}_1(x))$ converts x to the size $\mathbb{R}^{B\times L\times 1}$ and broadcasts it to the D dimension. This allows the model to selectively retain or discard information based on the current input, enabling effective long-range dependency modeling. A hardware-aware algorithm ensures computational efficiency during this selective scanning process.

C. PROPOSED METHOD

We construct the foundation of M^3-PPO by integrating the stability of PPO [26] with the rapid adaptation capability of MAML [21] within a meta-reinforcement learning framework (denoted as M-PPO). Building upon this M-PPO base, we introduce two key innovations to address the training instability of MAML: (1) a Mamba-enhanced context encoder for the inner loop, and (2) a global-local momentum update mechanism for the outer loop. The overall framework of the M^3-PPO algorithm is described in Figure 4.

1. Context Encoder Module

Context encoders, which extract task representations from past experiences, are crucial for meta-reinforcement learning [9,27]. While LSTM [9] and Transformer [28] are commonly used, they face limitations in long-sequence modeling: LSTM's recurrent structure dilutes long-range dependencies, and Transformer's attention becomes sparse with long sequences.

To address these limitations, we adopt the Mamba model [12] as the context encoder. The encoder processes a trajectory of past interactions, represented as a sequence of state-action-reward tuples $D = \{(s_1, a_1, r_1), \ldots, (s_T, a_T, r_T)\}$, as detailed in Algorithm 1. Mamba's selective state space mechanism provides superior dynamic representation capabilities for long sequences with complex temporal dependencies, such as those in power grid scheduling. This enhances task representation quality and accelerates adaptation to new tasks.

2. Global-Local Momentum Update Meta-Parameter Mechanism

We draw inspiration from the EMA mechanism and propose a global-local momentum update mechanism to establish a more balanced dynamic parameter adjustment method between global knowledge and local tasks. As illustrated in Figure 5, while standard MAML initializes each task with the global meta-parameter θ (left), our approach (right) computes the initial parameters for task $i > 1$ as:

$$\theta_i = \mu\theta + (1-\mu)\theta_{i-1} \tag{13}$$

where μ is an adjustable trade-off factor used to balance the influence of the global parameter θ and the previous task parameter θ_{i-1} on the current task. This design offers four key advantages:

Enhanced adaptation efficiency: θ_{i-1} provides a more appropriate start for tasks with a gradually increasing proportion of renewable energy generation, reducing the adaptation cost.

Smooth task transition: blending the local experience θ_{i-1} and the global prior θ enables gradual knowledge transfer between sequentially related tasks.

Improved task collaboration: creates a knowledge chain across tasks, promoting progressive learning.

Stabilized meta-optimization: smoothens the gradient direction across tasks, buffering the impact of task variations.

In summary, the complete process of the M^3-PPO algorithm is detailed in Algorithm 1.

Algorithm 1: M^3-PPO Algorithm

Require: $p(T)$: distribution over tasks
Require: α, β, μ: hyperparameters
Randomly initialize meta parameter θ of the actor and critic networks, where $\theta = (\theta_a, \theta_v)$.
While not done do
Sample a batch of tasks $T_i \sim p(T)$
for all T_i do
if $(i > 1)$ then
$\theta_i = \mu\theta + (1-\mu)\theta_{i-1}$
else
$\theta_i = \theta$
Sample a single trajectory $D_i = \{(s_1, a_1, r_1), \ldots, (s_T, a_T, r_T)\}$ using policy $\pi(a_t|s_t, a_{t-1}, r_{t-1}; \theta_a)$ in T_i
EncodedInput1 = Mamba(D_i)
Evaluate ∇L_{actor} and ∇L_V using EncodedInput1
Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha\nabla_\theta L_{T_i}(f_\theta)$
Sample a single trajectory $D_i' = \{(s_1, a_1, r_1), \ldots, (s_T, a_T, r_T)\}$ using policy $\pi(a_t|s_t, a_{t-1}, r_{t-1}; \theta_a)$ in T_i
end for
EncodedInput2 = Mamba(D_i')
Update $\theta \leftarrow \theta - \beta\nabla_\theta \sum_{T_i\sim p(T)} L_{T_i}(f_{\theta_i'})$ using each EncodedInput2
end while

3. Overall Training Process

(a) Architecture and Loss Function Gradient of Mamba Network in RL

We employ Mamba as a shared feature extractor, followed by independent linear output layers forming the actor and critic networks. Generally, the RL process adopts the PPO algorithm [26]. The objective loss of the actor network includes the proximal ratio clipping loss L_π^{clip} and the entropy L_H, and the objective loss function of the critic network is the state value function loss L_V.

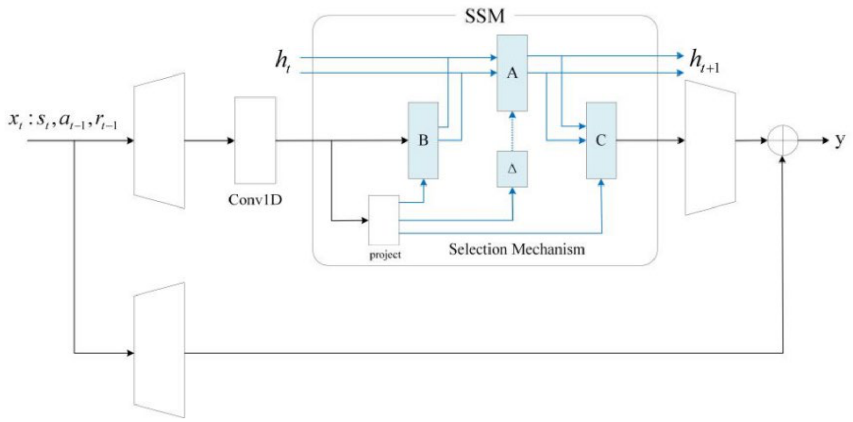

FIGURE 3. The Mamba Module (the blue boxes denote the projection functions S_B, S_C, and S_Δ that generate the input-dependent parameters. The blue arrows illustrate the data flow and the pathway of the selective information scanning mechanism within the SSM module.)

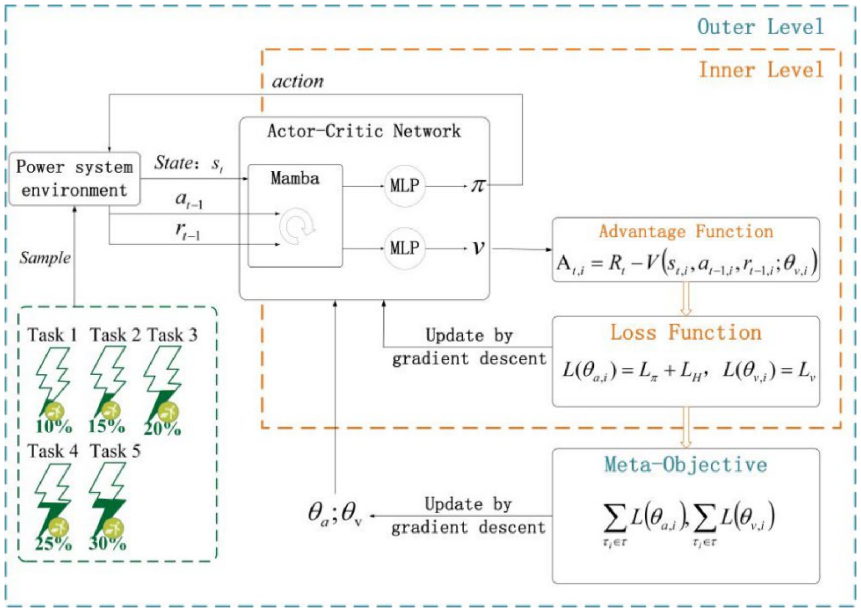

FIGURE 4. Overall Framework of the M^3-PPO Algorithm

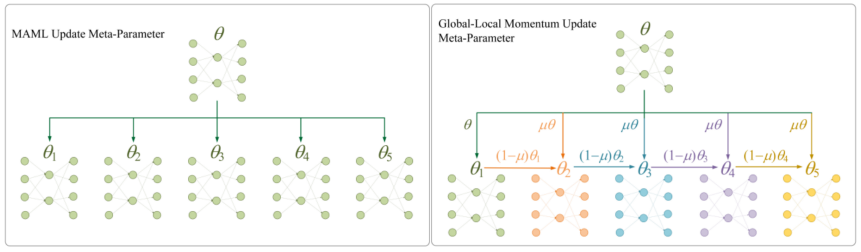

FIGURE 5. (left) parameter update of the MAML framework; (right) parameter update of the global-local momentum update mechanism

The detailed gradient of the objective loss function of the actor network is divided into two parts: $\nabla L_\pi^{\text{clip}}$ and ∇L_H, and the formula is as follows:

$$\nabla L_{\text{actor}} = \nabla L_\pi^{\text{clip}} + \nabla L_H \tag{14}$$

where $\nabla L_\pi^{\text{clip}}$ is derived from the PPO clipped objective and ∇L_H is the entropy gradient. The gradient ∇L_V of the objective loss function of the critic network is:

$$\nabla L_V = [R_t - V(s_t; \theta_v)] \frac{\partial V(s_t; \theta_v)}{\partial \theta_v} \tag{15}$$

where R_t is the discounted return.

(b) Meta-training and Meta-testing Phases

The training follows a standard meta-RL paradigm:

Meta-training: The model is trained over a distribution of tasks $p(T)$. The inner-loop adapts parameters per task via gradient descent, while the outer-loop updates the meta-parameters by aggregating gradients from all tasks.
Meta-testing: On a novel task, the pre-trained Mamba encoder remains fixed, providing a generalized task representation. Only the actor and critic are fine-tuned with few samples, enabling rapid adaptation.

V. EXPERIMENTS

In this section, we introduce the experimental design and result analysis for validating the effectiveness of the algorithm. We first describe the experimental environment, baseline algorithms, and model parameter settings. Secondly, we design comparison experiments and ablation experiments to prove the effectiveness of the Mamba module in power grid environments with different renewable energy proportions. Finally, we demonstrate the transferability under five different energy mix distributions through behavioral analysis experiments.

A. EXPERIMENTAL SETUP

1. Dataset & Environment: Experiments are built on the publicly available grid dataset from the L2RPN competition (l2rpn_neurips_2020_track2_small [16]), which provides real-world grid topology and timeseries operational data. We utilize the Grid2Op simulator to create the training and testing environments. The simulator calculates the next system state based on actions and the current state via an integrated power flow solver. The model is trained on a distribution of tasks generated from five different energy mixes, with renewable penetration rates ranging from 10% to 30%. During testing, the agent adapts to novel, unseen energy mixes within the same 10%-30% range to evaluate its generalization capability. The parameter details of the power grid structure are shown in Table 1.

TABLE 1. Power Grid Topology Parameters

Grid	N_{sub}	N_{line}	N_{gen}	N_{load}
l2rpn_neurips_2020_track2	118	186	62	99

2. Baseline Algorithms: To prove the effectiveness of the proposed method, the baseline algorithms we selected cover from random actions, classic deep reinforcement learning algorithms (DoubleDuelingRDQN, A2C, PPO), to the meta-reinforcement learning framework M-PPO algorithm which is the cornerstone of the M^3-PPO algorithm.

(1) Random Algorithm: A power grid control algorithm that uses random actions.

(2) DoubleDuelingRDQN Algorithm: A power grid control algorithm that uses D3QN and is optimized using the lstm network.

(3) A2C Algorithm: A power grid control algorithm that uses A2C.

(4) PPO Algorithm: A power grid control algorithm that uses PPO.

(5) M-PPO Algorithm: A power grid control algorithm based on M-PPO.

3. Evaluation Metrics:
ARPE (Accumulated Reward Per Episode): The ARPE indicates the comprehensive performance, integrating the economy and security metric. A higher ARPE value means a better comprehensive performance of algorithms.
ASPE (Accumulated Survival Per Episode): The ASPE measures the total number of steps an agent survives in an episode, reflecting the cumulative time during which the agent maintains normal grid operation without a power outage. A higher ASPE indicates greater operational stability of the agent.

4. Hyperparameter Settings:

(1) The Mamba network is the shared layer with an input dimension of observation_size. The actor network contains 2 hidden layers, with the number of neurons in each layer being 4*observation_size, 32 in sequence, and the number of neurons in the output layer is equal to action_size. The critic network contains 2 hidden layers, with the number of neurons in each layer being 2*observation_size, 32 in sequence, and there is only 1 neuron in the output layer. All networks apply the ReLU activation function.

(2) The discount factor γ is 0.95, and the hyperparameter weighing the contribution of the entropy regularization term is 0.01. A total of 100,000 steps are trained, and the maximum number of steps for the agent in each round is set to 2000.

VI. EXPERIMENTAL RESULTS

A. PERFORMANCE COMPARISON

We compared the ARPE values and ASPE values of the proposed M^3-PPO algorithm with the baseline algorithms Random, DoubleDuelingRDQN, A2C, PPO, M-PPO during the training phase in the IEEE-118 environment to verify the effectiveness of the proposed algorithm. Based on the results of 4 rounds of training phase experiments, we calculated the mean and 0.95 confidence intervals of each metric and evaluated and compared the performance of each algorithm accordingly. Table 2 and Figure 6 detail the comparison results between the algorithms.

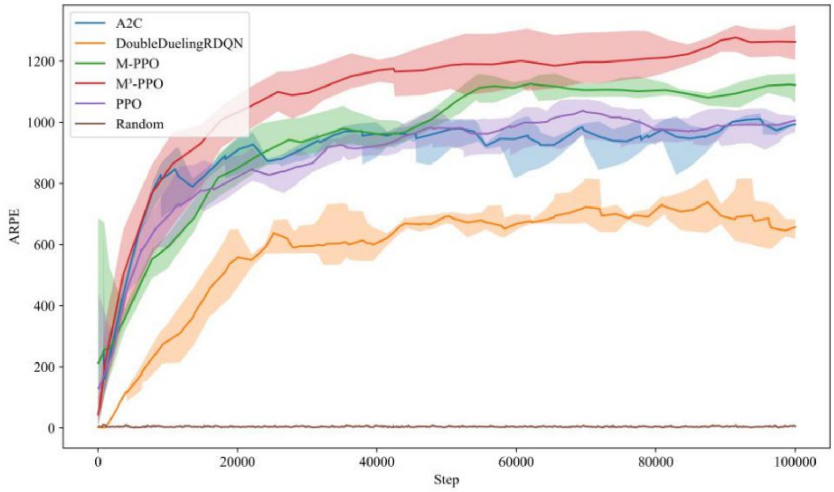

FIGURE 6. Comparative Analysis of ARPE Values During the Training Phase

TABLE 2. Performance Comparison During Training

	Random	DoubleDuelingRDQN	A2C	PPO	M-PPO	M^3-PPO
ARPE	5.09±2.65	657.03±32.34	993.70±2.19	1006.06±25.72	1121.71±47.45	1263.36±56.49
ASPE	8.19±3.00	775.46±35.83	1135.84±6.59	1152.57±24.35	1283.17±63.77	1442.99±62.33

From Table 2 and Figure 6, we can observe the following:

(1) The metric values of these algorithms all converge to a specific interval, while the metric values of the random algorithm are close to zero.

(2) The average ARPE and ASPE values of the A2C and PPO algorithms are significantly higher than those of the Random and DoubleDuelingRDQN algorithms. The trends of the A2C and PPO curves are similar, both rising steadily with the increase in the number of steps. The PPO curve is slightly higher than the A2C for most of the time, and the final performance of PPO is slightly better with a smoother curve.

(3) Although the M-PPO algorithm is better than the standard PPO, its curve fluctuates more compared to the M^3-PPO.

(4) For the M^3-PPO algorithm based on the M-PPO architecture, its average ARPE and ASPE values are 1263.36 $\pm$ 56.49 and 1442.99 $\pm$ 62.33, respectively, significantly exceeding all other algorithms. Although its confidence interval is slightly wider, the overall value is higher, and the curve smoothness is better than that of M-PPO.

These results indicate that all reinforcement learning algorithms except the Random algorithm achieve the basic objectives of the grid scheduling task. The superior performance of A2C and PPO over DoubleDuelingRDQN suggests that the actor-critic framework more effectively handles high-dimensional state spaces and long-term dependencies. Moreover, M^3-PPO, incorporating a global-local momentum mechanism and a Mamba-based context encoder, demonstrates clear advantages in both performance and training stability, highlighting its potential for complex decision-making tasks such as power grid scheduling.

B. ANALYSIS OF THE TRANSFERABILITY OF POWER GRID SCHEDULING STRATEGY

To verify the effectiveness of our proposed innovative algorithm M^3-PPO in enhancing the cross-task generalization ability and adaptability of power grid scheduling strategy, we compared the test results of the M^3-PPO algorithm and the benchmark algorithm in a similar but unknown energy portfolio environment. The specific evaluation metrics include ARPE and ASPE values. Table 3 and Figure 7 show the comparison results of different algorithms during the test phase.

TABLE 3. Performance Comparison in Testing Phase

	Random	DoubleDuelingRDQN	A2C	PPO	M-PPO	M^3-PPO
ARPE	4.07±2.44	601.06±28.93	944.19±32.4	948.66±41.48	1031.20±66.5	1192.44±47.5
ASPE	7.34±2.14	714.40±34.36	1074.79±36.0	1093.02±47.78	1197.85±70.2	1371.65±47.22

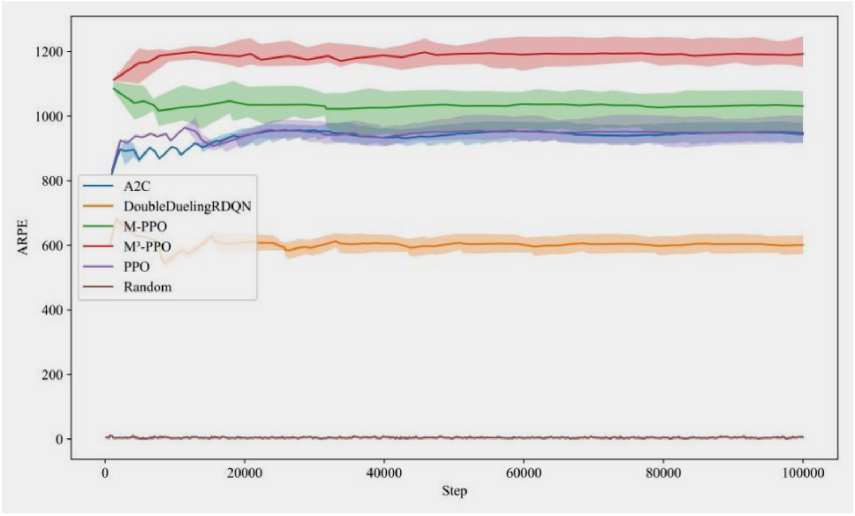

FIGURE 7. Comparative Analysis of ARPE Values in the Testing Phase

From Table 3 and Figure 7, we can directly observe the following phenomena:

(1) The ARPE value of the Random algorithm is always close to zero throughout the test process.

(2) The curve of the DoubleDuelingRDQN algorithm is overall flat, and the ARPE value is stable at a relatively low level.

(3) The curve trends of the A2C and PPO algorithms are similar, converging to similar ARPE values, and PPO is slightly better than A2C.

(4) M-PPO algorithm: Its overall ARPE value is higher than that of A2C and PPO, but its confidence interval is relatively wide.

(5) The curve of the M^3-PPO algorithm is always at the highest position. Its ARPE value significantly exceeds other algorithms. The ARPE reaches 1192.44 ± 47.51, while the ASPE achieves 1371.65 ± 47.22. Moreover, the confidence interval is relatively narrow, and the overall curve is relatively smooth, with less fluctuation than M-PPO.

These results confirm the strong transferability of M^3-PPO. The global-local momentum mechanism enhances robustness to state distribution shifts, while the Mamba-based encoder improves sequential dependency modeling, enabling effective policy adaptation in new scenarios. In contrast, although M-PPO has a certain transfer foundation, the large fluctuations in its performance curve indicate that its policy stability is still insufficient when facing new energy combinations. Overall, M^3-PPO offers a reliable solution for scheduling in complex grid environments.

C. ANALYSIS OF THE EFFECTIVENESS OF THE MAMBA MODULE

We compared the ARPE and ASPE values of the M-PPO algorithm without a representation module and the M^2-PPO algorithm with Mamba used as the representation module in this paper to verify the effectiveness of the representation module used. The comparison results in the training and testing phases are shown in Table 4, Table 5 and Figure 8.

TABLE 4. Performance Comparison During Training

	M-PPO	M^2-PPO
ARPE	1121.71 ± 47.45	1208.58 ± 55.07
ASPE	1283.17 ± 63.77	1379.56 ± 62.44

TABLE 5. Performance Comparison in Testing Phase

	M-PPO	M^2-PPO
ARPE	1031.20 ± 66.54	1144.40 ± 18.53
ASPE	1197.85 ± 70.22	1319.66 ±19.73

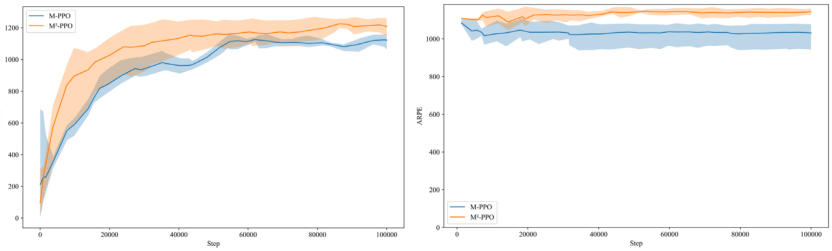

FIGURE 8. Comparative Analysis of ARPE Values (Training Phase: Left Subfigure; Testing Phase: Right Subfigure)

From the results in Table 4, Table 5 and Figure 8, it can be seen that introducing the Mamba representation module based on M-PPO can significantly improve the basic performance of the algorithm.

Specifically, by enhancing the feature representation of the task context, the M^2-PPO algorithm improves adaptability and generalization ability of the algorithm to different tasks, thus significantly improving the overall performance. This result fully verifies the effectiveness of the Mamba encoder introduced in this paper and its key role in enhancing the algorithm performance.

VII. CONCLUSION

This paper proposes an M^3-PPO algorithm based on meta-reinforcement learning, aiming to improve the generalization ability and adaptability of the power grid scheduling strategy in cross-task scenarios. The algorithm realizes the efficient exploration and rapid strategy adaptation of structurally similar but unseen tasks by migrating the scheduling strategy learned by the agent in the power grid environment with diverse energy combinations. M^3-PPO takes PPO as the basic framework, integrates the idea of MAML, and on the basis of the meta-reinforcement learning infrastructure M-PPO, takes into account both the policy optimization efficiency and the cross-task adaptation ability. To further improve the training stability and new task adaptation performance, the algorithm introduces the Mamba model as a context encoder in the inner-layer update and combines the global-local momentum optimization mechanism in the outer-layer update. Experimental results show that compared with the existing benchmark algorithms, M^3-PPO exhibits better generalization performance and comprehensive scheduling efficiency in various power grid scheduling scenarios. Especially in the complex energy structure environment, its adaptation speed and policy stability are significantly better than the existing methods.

Future research will focus on two key directions to further enhance the practicality of M^3-PPO: (1) advancing meta-generalization by integrating cross-domain multi-task learning mechanisms to strengthen its adaptability in diverse energy environments, and (2) optimizing computational efficiency through lightweight or distributed computing techniques to facilitate real-time deployment in large-scale grid operations.

AUTHOR CONTRIBUTIONS

Conceptualization –Qihui Dai, Xiang Hu, Jianlong Li, Aomei Jiang and Wei Xiong; methodology – Qihui Dai, Xiang Hu, Aomei Jiang and Wei Xiong; investigation – Qihui Dai, Xiang Hu, Jianlong Li, Aomei Jiang and Wei Xiong; writing-original draft preparation – Qihui Dai, Xiang Hu, Jianlong Li, Aomei Jiang and Wei Xiong. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Sridhar and M. Govindarasu, "Model-Based Attack Detection and Mitigation for Automatic Generation Control," IEEE Transactions on Smart Grid, vol. 5, no. 2, pp. 580-591, 2014, doi: 10.1109/TSG.2014.2298195.

[2] Y. Li, R. Huang, and L. Ma, "False Data Injection Attack and Defense Method on Load Frequency Control," IEEE Internet of Things Journal, vol. 8, no. 4, pp. 2910-2919, 2021, doi: 10.1109/JIOT.2020.3021429.

[3] A. O. Erick and K. A. Folly, "Reinforcement Learning Approaches to Power Management in Grid-tied Microgrids: A Review," Proceedings of the 2020 Clemson University Power Systems Conference (PSC); 10-13 March 2020; Clemson, SC, USA, doi: 10.1109/PSC50246.2020.9131138.

[4] L. Lei, Y. Tan, G. Dahlenburg, W. Xiang, and K. Zheng, "Dynamic Energy Dispatch Based on Deep Reinforcement Learning in IoT-Driven Smart Isolated Microgrids," IEEE Internet of Things Journal, vol. 8, no. 10, pp. 7938-7953, 2021, doi: 10.1109/JIOT.2020.3042007.

[5] X. Zhou, J. Wang, X. Wang, and S. Chen, "Deep Reinforcement Learning for Microgrid Operation Optimization: A Review," Proceedings of the 2023 8th Asia Conference on Power and Electrical Engineering (ACPEE); 14-16 April 2023; Tianjin, China, doi: 10.1109/ACPEE56931.2023.10135713.

[6] E. Samadi, A. Badri, and R. Ebrahimpour, "Decentralized multi-agent based energy management of microgrid using reinforcement learning," International Journal of Electrical Power & Energy Systems, vol. 122, Art. no. 106211, 2020, doi: 10.1016/j.ijepes.2020.106211.

[7] J. Duan, "Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations," IEEE Transactions on Power Systems, vol. 35, no. 1, pp. 814-817, 2020, doi: 10.1109/TPWRS.2019.2941134.

[8] R. Huang, "Learning and Fast Adaptation for Grid Emergency Control via Deep Meta Reinforcement Learning," IEEE Transactions on Power Systems, vol. 37, no. 6, pp. 4168-4178, 2022, doi: 10.1109/TPWRS.2022.3155117.

[9] L. Xiong, "Meta-Reinforcement Learning-Based Transferable Scheduling Strategy for Energy Management," IEEE Trans. Circuits Syst. I, vol. 70, no. 4, pp. 1685-1695, 2023, doi: 10.1109/TCSI.2023.3240702.

[10] D. Jakobeit, M. Schenke, and O. Wallscheid, "Meta-Reinforcement-Learning-Based Current Control of Permanent Magnet Synchronous Motor Drives for a Wide Range of Power Classes," IEEE Transactions on Power Electronics, vol. 38, no. 7, pp. 8062-8074, 2023, doi: 10.1109/TPEL.2023.3256424.
[11] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014. arXiv preprint. arXiv:1412.6980, doi: 10.48550/arXiv.1412.6980.
[12] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," 2024. arXiv: arXiv:2312.00752, doi: 10.48550/arXiv.2312.00752.
[13] C. Chen, M. Cui, F. Li, S. Yin, and X. Wang, "Model-Free Emergency Frequency Control Based on Reinforcement Learning," IEEE Transactions on Industrial Informatics, vol. 17, no. 4, pp. 2336-2346, 2020, doi: 10.1109/TII.2020.3001095.
[14] Z. Yan and Y. Xu, "A Multi-Agent Deep Reinforcement Learning Method for Cooperative Load Frequency Control of a Multi-Area Power System," IEEE Transactions on Power Systems, vol. 35, no. 6, pp. 4599-4608, 2020, doi: 10.1109/TPWRS.2020.2999890.
[15] X. Hu, Y. Zhang, H. Xia, W. Wei, Q. Dai, and J. Li, "Toward Fair Power Grid Control: A Hierarchical Multiobjective Reinforcement Learning Approach," IEEE Internet of Things Journal, vol. 11, no. 4, pp. 6582-6595, 2024, doi: 10.1109/JIOT.2023.3314522.
[16] A. Marot, I. Guyon, B. Donnot, G. Dulac-Arnold, P. Panciatici, and M. Awad, "L2RPN: Learning to Run a Power Network in a Sustainable World NeurIPS2020 challenge design," Réseau de Transport d'Électricité. Paris, France: White Paper; 2020.
[17] J. R. Vázquez-Canteli, J. Kämpf, G. Henze, and Z. Nagy, "CityLearn v1.0: An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning," Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '19; 13-14 November 2019; New York, NY, USA. New York, NY, USA: Association for Computing Machinery; 2019, doi: 10.1145/3360322.3360998.
[18] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman, "Quantifying Generalization in Reinforcement Learning," Proceedings of the 36th International Conference on Machine Learning, PMLR, 202, [Online]. Available: https://proceedings.mlr.press/v97/cobbe19a.html.
[19] T. Wu and J. Wang, "Artificial intelligence for operation and control: The case of microgrids," The Electricity Journal, vol. 34, no. 1, Art. no. 106890, 2021, doi: 10.1016/j.tej.2020.106890.
[20] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning," 2016. arXiv: arXiv:1611.02779, doi: 10.48550/arXiv.1611.02779.
[21] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, [Online]. Available: https://proceedings.mlr.press/v70/finn17a.html.
[22] Á. Belmonte-Baeza, J. Lee, G. Valsecchi, and M. Hutter, "Meta Reinforcement Learning for Optimal Design of Legged Robots," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 12134-12141, 2022, doi: 10.1109/LRA.2022.3211785.
[23] F. Ye, P. Wang, C.-Y. Chan, and J. Zhang, "Meta Reinforcement Learning-Based Lane Change Strategy for Autonomous Vehicles," Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV); Nagoya, Japan; 1 November 2021. New York, NY, USA: IEEE; 2021, doi: 10.1109/IV48863.2021.9575379.
[24] Q. Liu, "A few-shot disease diagnosis decision making model based on meta-learning for general practice," Artificial Intelligence in Medicine, vol. 147, Art. no. 102718, 2024, doi: 10.1016/j.artmed.2023.102718.
[25] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, "Improving deep neural networks using softplus units," Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN); Killarney; 12-17 July 2015; New York, NY, USA: IEEE; 2015, doi: 10.1109/IJCNN.2015.7280459.
[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," 2017. arXiv: arXiv:1707.06347, doi: 10.48550/arXiv.1707.06347.
[27] H. Fu, "Towards Effective Context for Meta-Reinforcement Learning: An Approach based on Contrastive Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 7457-7465, 2021, doi: 10.1609/aaai.v35i8.16914.
[28] L. Kirsch, J. Harrison, J. Sohl-Dickstein, and L. Metz, "General-Purpose In-Context Learning by Meta-Learning Transformers," 2024. arXiv: arXiv:2212.04458, doi: 10.48550/arXiv.2212.04458.