基于物理约束梯度引导的微电网能量调度安全强化学习策略

沈佳俊[1,2]，杨 渌[1,2]，陈英豪[1,2]，郭方洪[1,2†]

(1. 浙江工业大学 信息工程学院，杭州 310023；2. 全省复杂系统智能感知与控制重点实验室，杭州 310023)

摘 要: 当前微电网能量调度面临的挑战在于时序耦合约束导致决策空间维度显著提升, 以及交流潮流方程引入非线性约束, 增加了计算复杂度, 使得整体优化模型具有较强的非凸性, 从而大幅增加了问题的求解难度. 针对上述问题, 提出一种基于安全强化学习和物理约束梯度引导的微电网能量调度方法. 该方法构建基于深度学习的动作修正安全层网络, 在环境交互过程中对智能体动作进行投影, 以保障动作满足物理可行性并有效提升探索效率. 进一步地, 将该安全层嵌入至网络训练过程, 从而提升了强化学习 Critic 网络Q值估计精度以及 Actor 网络对物理约束的学习效率. 基于 IEEE 14 节点模型构建的微电网电-氢耦合潮流系统实验表明, 所提方法在调度决策性能上优于拉格朗日乘子法 (TD3-Lag) 和惩罚项法 (TD3-Pen). 同时与基于数值优化的安全层方法相比, 保持了相近的性能表现, 但部署速度提升了约 3 个数量级.

关键词: 交流潮流；非凸性；物理约束梯度；安全强化学习；微电网

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1231

引用格式: 沈佳俊, 杨渌, 陈英豪, 等. 基于物理约束梯度引导的微电网能量调度安全强化学习策略 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

Physics-constrained and gradient-guided reinforcement learning for secure energy dispatch in microgrids

SHEN Jia-jun[1,2]，*YANG Hao*[1,2]，*CHEN Ying-hao*[1,2]，*GUO Fang-hong*[1,2†]

(1. College of Information Engineering， Zhejiang University of Technology， Hangzhou 310023， China；2. Zhejiang Key Laboratory of Intelligent Perception and Control for Complex Systems，Hangzhou 310023，China)

Abstract: Current microgrid energy scheduling faces critical challenges, where temporal coupling constraints lead to a significant expansion of the decision space, and AC power flow equations introduce nonlinear constraints that increase computational complexity, resulting in an overall optimization model with strong non-convexity and substantially increased solving difficulty.To address these issues, this paper proposes a microgrid energy scheduling method based on safe reinforcement learning and physics-constrained gradient guidance. The method constructs a deep learning-based action correction safety layer that projects agent actions into the feasible domain during environment interaction, ensuring operational physical feasibility while effectively improving exploration efficiency.Furthermore, by embedding this safety layer into the network training process, it enhances the Q-value estimation accuracy of the Critic network and improves the physical constraint learning efficiency of the Actor network. Experimental results on an IEEE 14-bus model-based microgrid with electro-hydrogen coupled power flow demonstrate that the proposed method outperforms the Lagrangian multiplier method (TD3-Lag) and the penalty-based method (TD3-Pen) in scheduling decision performance. Compared to the numerically optimized safety layer approach, it achieves approximately three orders of magnitude faster deployment speed while maintaining similar performance levels.

Keywords: AC power flow；non-convexity；physics-constrained gradient；safe reinforcement learning；microgrid

0 引 言

在“碳达峰、碳中和”目标的推动下, 电力系统正经历着由传统化石能源主导向高比例可再生能源接入的深度转型[1-2]. 风电、光伏等可再生能源以其

收稿日期: 2025-11-27；**录用日期:** 2026-03-03.

基金项目: 国家自然科学基金项目 (62373328)；浙江省自然科学基金项目 (LR25F030003).

责任编委: 邢兰涛.

†通信作者. E-mail: fhguo@zjut.edu.cn

清洁、可持续的特点被广泛接入电网, 但其固有的间歇性与随机性也显著增加了系统运行的不确定性, 对电力系统调度带来严峻挑战. 为提升系统灵活性与可再生能源消纳能力, 微电网 (Microgrid, MG) 作为一种能够实现局部能源自治与协调控制的分布式组织形态, 日益受到广泛关注[3-5]. 在此背景下, 如何实现微电网的高效、经济与安全能量调度, 已成为可再生能源充分利用与系统稳定运行的关键问题.

在传统经济调度问题中, 通常以成本最小化为目标, 并基于直流潮流或简化网络模型进行求解[6-9]. 这类方法虽然能够获得经济性较优的方案, 但由于忽略了交流潮流约束, 所得方案在实际电网中可能无法满足安全运行要求. 交流最优潮流 (Alternating Current Optimal Power Flow, ACOPF) 在电力系统运行与控制中发挥着核心作用. 该模型通过考虑电压幅值、相角及无功功率约束, 能够准确地反映电力系统的物理特性, 从而在建模层面更贴近实际系统的运行规律. 然而, 其强非凸性也为求解带来了显著挑战. 近年来, 深度学习因其在复杂非线性映射建模方面展现出的强大能力, 为高效求解单时刻交流潮流约束问题提供了新的可行路径. 一些研究采用深度神经网络直接拟合系统中的非线性映射关系以实现快速求解[10], 另有研究通过引入拉格朗日动态惩罚机制[11-12] 或构建预测-重构框架[13] 来提升约束处理能力; 还有工作将物理约束嵌入网络结构, 如通过拓扑嵌入技术构建统一模型[14] 或设计功率平衡修复层[15-16]. 尽管上述方法在提升单时刻交流潮流问题的求解效率方面取得了显著进展, 然而当应用于多时刻 ACOPF 问题时, 却难以解决时序耦合与非凸潮流约束并存的难题. 该问题根源在于, 当从单时刻优化扩展到多时段序列决策时, 时序耦合导致了决策空间的维度剧增, 而各个时段的非凸潮流约束在时间维度上叠加, 共同形成了一个高维且结构复杂的可行域. 在此类高维序列决策问题中, 基于深度学习的优化方法极易陷入局部最优解, 或产生不可行的调度方案.

针对以上问题, 深度强化学习 (Deep Reinforcement Learning, DRL) 凭借其在复杂环境中自主探索与学习的能力, 已成为电力系统优化调度领域的研究热点[17-22]. 其中, PPO[18](Proximal Policy Optimization)、DDPG[19](Deep Deterministic Policy Gradient) 与 TD3[20](Twin Delayed Deep Deterministic Policy Gradient) 等算法因在处理连续动作空间方面表现优异而获得广泛应用. 尽管这些算法在经济效益方面表现良好, 但其决策在训练与实际应用中仍存在违反系统安全约束的风险, 如破坏有功-无功平衡, 甚至引发节点电压或线路潮流越限等安全问题. 为提升深度强化学习决策的安全性, 现有研究主要遵循三种技术路径. 第一类基于惩罚函数, 通过在奖励函数中引入惩罚项以规避不安全动作[23], 然而, 各类约束在量级上存在差异, 导致智能体难以学习到稳健的安全策略. 第二类基于约束强化学习, 采用拉格朗日松弛法[24] 或将约束违反值显式加入目标函数[25], 旨在将原问题转化为无约束优化. 但这类方法在优化目标与安全约束之间常难以平衡, 且通常依赖复杂精细的参数整定. 相比之下, 基于安全层的方法通过引入独立的安全保护模块, 能更有效地规避上述调参困境. 该方法核心思想是利用一个安全层对智能体原始动作进行修正, 具体实现手段包括构建基于惩罚凸凹过程的凸安全层[26]、通过数值优化方法直接修正不安全动作[27]、或采用融合知识-数据驱动的线性规划进行可行域投影[28]. 这类方法为严格满足安全约束提供了更具潜力的解决途径, 但其存在求解速度慢、安全层的修正无法保留梯度信息的缺陷, 从而阻碍策略网络高效学习约束.

综上所述, 针对微电网能量调度中因决策空间维度剧增及可行域高度复杂所导致的易陷局部最优与方案不可行问题, 本文提出了一种基于物理约束梯度引导的微电网能量调度安全强化学习方法. 该方法的核心在于构建一个可微的动作修正安全层, 通过将智能体的动作投影至满足物理约束的可行域内, 有效缩减了由时序耦合带来的高维无效搜索空间, 从而大幅减少了不可行调度方案的产生; 进一步地, 针对非凸约束导致的局部最优陷阱, 本文将该可微安全层嵌入至强化学习网络训练过程中, 利用其可导特性将物理约束的梯度信息直接反向传播至策略网络. 这种物理梯度的显式引导机制, 替代了传统盲目试错的惩罚机制, 使得智能体能够在复杂的非凸可行域表面获得明确的优化方向指引, 从而在保障调度安全性的同时, 高效收敛至兼具经济性的全局最优策略.

1 问题描述

本节基于 IEEE 14 节点拓扑结构, 考虑了一种面向微电网的电-氢耦合潮流优化问题. 在满足设备约束和潮流约束的前提下, 通过优化系统中各设备的调度策略, 实现微电网经济运行与安全稳定之间的协同优化.图 1 展示了系统的整体建模框架, 后续将详细阐述建模中各类设备的运行约束、潮流约束及目标函数.

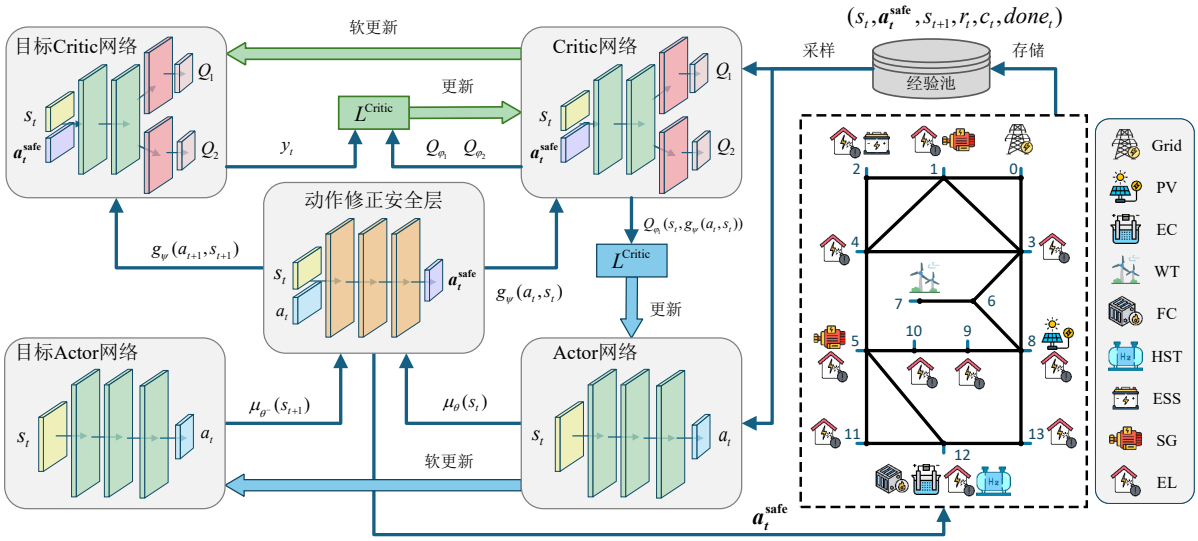

图1　基于物理约束梯度引导的安全强化学习框架

1.1　设备约束

本研究关注一个互联多能源微电网, 包含主电网和以下 8 种设备类型:

1) 主电网 (Grid): 将节点 0 设置为与主电网连接的松弛节点, 其通过调节功率输出确保系统有功和无功功率平衡, 且在时刻t向主电网购买的有功功率定义为$P^{\text{grid}}_{i,t,\text{buy}}$, 向主电网出售的有功功率定义为$P^{\text{grid}}_{i,t,\text{sell}}$, 该节点交易需满足以下约束条件:

$$P^{\text{Grid}}_{i,t,\text{buy}} P^{\text{Grid}}_{i,t,\text{sell}} = 0, \tag{1}$$

$$P^{\text{Grid}}_{i,\min} \leqslant P^{\text{Grid}}_{i,t,\text{buy}},\ P^{\text{Grid}}_{i,t,\text{sell}} \leqslant P^{\text{Grid}}_{i,\max}. \tag{2}$$

公式 (1) 确保了主电网的购电与售电操作不能同时进行, 公式 (2) 则限定了与电网交换功率的上、下限, 以保证系统的安全运行范围.

2) 电负荷 (Electrical Load, EL): 在节点 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13 设置电负荷. 将节点i在时刻t的有功功率负荷定义为$P^{\text{EL}}_{i,t}$, 节点i在时刻t的无功功率负荷定义为$Q^{\text{EL}}_{i,t}$.

3) 风力发电 (Wind Turbine, WT): 在节点 7 设置风力发电, 将节点 7 在时刻t的风力发电有功功率定义为$P^{\text{WT}}_{7,t}$

4) 光伏发电 (Photovoltaic, PV): 在节点 8 设置风力发电, 将节点 8 在时刻t的光伏发电有功功率定义为$P^{\text{PV}}_{8,t}$.

5) 同步发电机 (Synchronous Generator, SG): SG 能够提供稳定的有功功率和无功功率, 用于电网电压调节和功率平衡. 在节点 1, 5 分别配置SG_1和SG_2, 将节点 1, 5 在时刻t提供的有功功率分别为$P^{\text{SG}_1}_{1,t}$和$P^{\text{SG}_2}_{5,t}$, 提供的无功功率分别为$Q^{\text{SG}_1}_{1,t}$和$Q^{\text{SG}_2}_{5,t}$, 其运行约束包括功率约束和爬坡约束:

$$P^{\text{SG}}_{i,t+1} = P^{\text{SG}}_{i,t} + \Delta P^{\text{SG}}_{i,t}, \tag{3}$$

$$Q^{\text{SG}}_{i,t+1} = Q^{\text{SG}}_{i,t} + \Delta Q^{\text{SG}}_{i,t}, \tag{4}$$

$$P^{\text{SG}}_{i,\min} \leqslant P^{\text{SG}}_{i,t} \leqslant P^{\text{SG}}_{i,\max}, \tag{5}$$

$$Q^{\text{SG}}_{i,\min} \leqslant Q^{\text{SG}}_{i,t} \leqslant Q^{\text{SG}}_{i,\max}, \tag{6}$$

$$\Delta P^{\text{SG}}_{i,\min} \leqslant \Delta P^{\text{SG}}_{i,t} \leqslant \Delta P^{\text{SG}}_{i,\max}, \tag{7}$$

$$\Delta Q^{\text{SG}}_{i,\min} \leqslant \Delta Q^{\text{SG}}_{i,t} \leqslant \Delta Q^{\text{SG}}_{i,\max}. \tag{8}$$

其中, 公式 (3) 和 (4) 分别描述了第i个 SG 在相邻时段之间$P^{\text{SG}}_{i,t}$和$Q^{\text{SG}}_{i,t}$的更新关系, 其变化量由对应的爬坡功率$\Delta P^{\text{SG}}_{i,t}$与$\Delta Q^{\text{SG}}_{i,t}$决定. 公式 (5) 和 (6) 给出了设备在任一时刻的有功和无功输出必须满足的物理边界条件. 公式 (7) 和 (8) 则限制了爬坡约束的上下界, 确保每个时段的调节幅度不超过设备的可调节能力.

6) 能源储存系统 (Energy Storage System, ESS): ESS 可以在可再生能源出力波动或负荷变化时储存和释放电能, 维持系统功率平衡与运行稳定. 在节点 2 设置 ESS, 将节点 2 在时刻t提供的有功功率定义为$P^{\text{dis}}_{2,t}$, 消耗的有功功率为$P^{\text{ch}}_{2,t}$, 其约束包括状态更新约束和容量约束:

$$E^{\text{ESS}}_{i,t+1} = E^{\text{ESS}}_{i,t} + (P^{\text{ch}}_{i,t}\eta^{\text{ch}} - \frac{P^{\text{dis}}_{i,t}}{\eta^{\text{dis}}})\Delta t, \tag{9}$$

$$P^{\text{ch}}_{i,\min} \leqslant P^{\text{ch}}_{i,t} \leqslant P^{\text{ch}}_{i,\max}, \tag{10}$$

$$P^{\text{dis}}_{i,\min} \leqslant P^{\text{dis}}_{i,t} \leqslant P^{\text{dis}}_{i,\max}, \tag{11}$$

$$E^{\text{ESS}}_{i,\min} \leqslant E^{\text{ESS}}_{i,t} \leqslant E^{\text{ESS}}_{i,\max}, \tag{12}$$

$$P^{\text{ch}}_{i,t} P^{\text{dis}}_{i,t} = 0. \tag{13}$$

其中, 公式 (9) 给出了 ESS 在相邻时段的能量状态转移关系, $E^{\text{ESS}}_{i,t}$表示t时刻 ESS 的容量, η^{ch}和η^{dis}分别表示充放电效率. 公式 (10) 和 (11) 分别限制了

ESS 在充电与放电模式下的功率输出必须处于允许的最小与最大范围之间, 确保充放电过程不超过设备额定能力. 公式 (12) 给出了 ESS 容量的上下界, 防止储能过充或过放. 公式 (13) 规定了同一时刻 ESS 不能同时充电与放电, 从而保证其运行方式的物理合理性.

7) 氢能源链 (Hydrogen Energy Chain, HEC): 氢能源链由电解槽 (Electrolytic Cell, EC), 储氢罐 (Hydrogen Storage Tank, HST) 和燃料电池 (Fuel Cell, FC) 三部分组成. 通过电解槽将多余电能转化为氢气储存, 并在需要时通过燃料电池将储存的氢能转换为电能, 为电网提供灵活调度能力, 提升可再生能源的消纳能力和系统稳定性. 在节点 12 设置氢能源链, 将节点 12 在时刻t提供和消耗的有功功率分别定义为$P_{12,t}^{\mathrm{FC}}$和$P_{12,t}^{\mathrm{EC}}$. 其约束包括状态更新约束和容量约束:

$$P_{i,t,\mathrm{H_2}}^{\mathrm{EC}} = P_{i,t}^{\mathrm{EC}}\eta^{\mathrm{EC}}, \tag{14}$$

$$P_{i,\min}^{\mathrm{EC}} \leqslant P_{i,t}^{\mathrm{EC}} \leqslant P_{i,\max}^{\mathrm{EC}}, \tag{15}$$

$$P_{i,t}^{\mathrm{FC}} = P_{i,t,\mathrm{H_2}}^{\mathrm{FC}}\eta^{\mathrm{FC}}, \tag{16}$$

$$P_{i,\min}^{\mathrm{FC}} \leqslant P_{i,t}^{\mathrm{FC}} \leqslant P_{i,\max}^{\mathrm{FC}}, \tag{17}$$

$$E_{i,t+1}^{\mathrm{HST}} = E_{i,t}^{\mathrm{HST}} + (P_{i,t,\mathrm{H_2}}^{\mathrm{EC}}\eta^{\mathrm{HST}} - \frac{P_{i,t,\mathrm{H_2}}^{\mathrm{FC}}}{\eta^{\mathrm{HST}}})\Delta t, \tag{18}$$

$$E_{i,\min}^{\mathrm{HST}} \leqslant E_{i,t}^{\mathrm{HST}} \leqslant E_{i,\max}^{\mathrm{HST}}, \tag{19}$$

$$P_{i,t,\mathrm{H_2}}^{\mathrm{EC}} P_{i,t,\mathrm{H_2}}^{\mathrm{FC}} = 0. \tag{20}$$

其中, 公式 (14) 定义了 EC 以制氢效率系数η^{EC}将电功率转化为氢能输出, 公式 (16) 则描述了 FC 以制电效率系数η^{FC}将氢能转化为电能输出. 公式 (15) 和 (17) 分别限制了 EC 和 FC 在运行过程中的功率必须处于允许的最小与最大范围内. 公式 (18) 进一步描述了储氢罐能量的动态更新机制, 其能量状态受 EC 产氢与 FC 耗氢影响, 并随时间步长变化. 为避免储氢过量或不足, 公式 (19) 对储氢量设置了物理上下界. 在公式 (20) 中, 对 EC 的制氢过程与 FC 的耗氢过程施加互斥约束, 规定两者不得在同一时刻进行.

1.2 潮流约束

在电力系统的时序运行过程中, 为保证其在每个时间步t的安全性与可行性, 需要满足下列潮流约束条件:

$$P_{i,t} = \sum_{j\in\Omega_i} V_{i,t}V_{j,t}(G_{ij}\cos\theta_{ij,t} + B_{ij}\sin\theta_{ij,t}), \tag{21}$$

$$Q_{i,t} = \sum_{j\in\Omega_i} V_{i,t}V_{j,t}(G_{ij}\sin\theta_{ij,t} - B_{ij}\cos\theta_{ij,t}), \tag{22}$$

$$P_{ij,t} = V_{i,t}^2 G_{ij} - V_{i,t}V_{j,t}(G_{ij}\cos\theta_{ij,t} + B_{ij}\sin\theta_{ij,t}), \tag{23}$$

$$Q_{ij,t} = -V_{i,t}^2 B_{ij} + V_{i,t}V_{j,t}(B_{ij}\cos\theta_{ij,t} - G_{ij}\sin\theta_{ij,t}), \tag{24}$$

$$P_{ij,t}^2 + Q_{ij,t}^2 \leqslant (S_{ij,t}^{\max})^2, \tag{25}$$

$$V_{i,\min} \leqslant V_{i,t} \leqslant V_{i,\max}, \tag{26}$$

$$P_{i,t}^{\mathrm{NL}} = P_{i,t}^{\mathrm{EL}} - P_{i,t}^{\mathrm{WT}} - P_{i,t}^{\mathrm{PV}}, \tag{27}$$

$$P_{i,t} = P_{i,t}^{\mathrm{SG}_1} + P_{i,t}^{\mathrm{SG}_2} + P_{i,t}^{\mathrm{dis}} + P_{i,t}^{\mathrm{FC}} + P_{i,t,\mathrm{buy}}^{\mathrm{Grid}} - (P_{i,t}^{\mathrm{EC}} + P_{i,t}^{\mathrm{NL}} + P_{i,t}^{\mathrm{ch}} + P_{i,t,\mathrm{sell}}^{\mathrm{Grid}}), \tag{28}$$

$$Q_{i,t}^{\mathrm{NL}} = Q_{i,t}^{\mathrm{EL}} - Q_{i,t}^{\mathrm{WT}} - Q_{i,t}^{\mathrm{PV}}, \tag{29}$$

$$Q_{i,t} = Q_{i,t}^{\mathrm{SG}_1} + Q_{i,t}^{\mathrm{SG}_2} - Q_{i,t}^{\mathrm{NL}}. \tag{30}$$

其中i和j为母线索引, 支路ij标识了连接母线i和j的输电线路, $j\in\Omega_i$为与节点i相连的节点集合. 节点i在时刻t的电压幅值和相角分别记为$V_{i,t}$和$\theta_{i,t}$, 支路ij两端的相角差定义为$\theta_{ij,t} = \theta_{i,t} - \theta_{j,t}$. 支路$ij$的导纳参数为电导$G_{ij}$和电纳$B_{ij}$. 节点$ij$的净有功和无功注入功率分别为$P_{i,t}$和$Q_{i,t}$; 支路$ij$上流过的有功和无功功率分别为$P_{ij,t}$和$Q_{ij,t}$; 支路$ij$的传输容量上限由最大复功率$S_{ij,t}^{\max}$表示; 节点$i$的净有功负荷和净无功负荷则分别由$P_{i,t}^{\mathrm{NL}}$和$Q_{i,t}^{\mathrm{NL}}$表示.

1.3 目标函数

$$C_{i,t}^{\mathrm{SG}_1} = a^{\mathrm{SG}_1}(P_{i,t}^{\mathrm{SG}_1})^2 + b^{\mathrm{SG}_1}P_{i,t}^{\mathrm{SG}_1} + c^{\mathrm{SG}_1}, \tag{31}$$

$$C_{i,t}^{\mathrm{SG}_2} = a^{\mathrm{SG}_2}(P_{i,t}^{\mathrm{SG}_2})^2 + b^{\mathrm{SG}_2}P_{i,t}^{\mathrm{SG}_2} + c^{\mathrm{SG}_2}, \tag{32}$$

$$C_{i,t}^{\mathrm{EC}} = a^{\mathrm{EC}}(P_{i,t}^{\mathrm{EC}})^2 + b^{\mathrm{EC}}P_{i,t}^{\mathrm{EC}} + c^{\mathrm{EC}}, \tag{33}$$

$$C_{i,t}^{\mathrm{FC}} = a^{\mathrm{FC}}(P_{i,t}^{\mathrm{FC}})^2 + b^{\mathrm{FC}}P_{i,t}^{\mathrm{FC}} + c^{\mathrm{FC}} - \beta^{\mathrm{H_2}}P_{i,t}^{\mathrm{FC}}, \tag{34}$$

$$C_{i,t}^{\mathrm{Grid}} = -\rho_{\mathrm{buy}}P_{i,t,\mathrm{buy}}^{\mathrm{Grid}} + \rho_{\mathrm{sell}}P_{i,t,\mathrm{sell}}^{\mathrm{Grid}}. \tag{35}$$

其中, $C_{i,t}^{\mathrm{SG}_1}$、$C_{i,t}^{\mathrm{SG}_2}$、$C_{i,t}^{\mathrm{EC}}$、$C_{i,t}^{\mathrm{FC}}$分别表示SG_1、SG_2、EC、FC 的成本函数, $\beta^{\mathrm{H_2}}$是对制氢用氢的补贴系数, ρ_{buy}和ρ_{sell}分别表示向主电网购买和出售的电价, $C_{i,t}^{\mathrm{Grid}}$表示与主电网交易后的净收益. 由于训练过程中难保系统绝对安全, 需对越限设定相应惩罚, 具体如下:

$$C_{i,t,\mathrm{pen}}^{\mathrm{SG}_1} = \max(0, P_{i,\min}^{\mathrm{SG}_1} - P_{i,t}^{\mathrm{SG}_1}, P_{i,t}^{\mathrm{SG}_1} - P_{i,\max}^{\mathrm{SG}_1}) + \max(0, Q_{i,\min}^{\mathrm{SG}_1} - Q_{i,t}^{\mathrm{SG}_1}, Q_{i,t}^{\mathrm{SG}_1} - Q_{i,\max}^{\mathrm{SG}_1}), \tag{36}$$

$$C_{i,t,\mathrm{pen}}^{\mathrm{SG}_2} = \max(0, P_{i,\min}^{\mathrm{SG}_2} - P_{i,t}^{\mathrm{SG}_2}, P_{i,t}^{\mathrm{SG}_2} - P_{i,\max}^{\mathrm{SG}_2}) + \max(0, Q_{i,\min}^{\mathrm{SG}_2} - Q_{i,t}^{\mathrm{SG}_2}, Q_{i,t}^{\mathrm{SG}_2} - Q_{i,\max}^{\mathrm{SG}_2}), \tag{37}$$

$$C_{i,t,\mathrm{pen}}^{P_{ij,t}^2+Q_{ij,t}^2} = \max(0, P_{ij,t}^2 + Q_{ij,t}^2 - (S_{ij,t}^{\max})^2), \tag{38}$$

$$C_{i,t,\text{pen}}^{\text{ESS}} = \max(0, E_{i,\min}^{\text{ESS}} - E_{i,t}^{\text{ESS}}, E_{i,t}^{\text{ESS}} - E_{i,\max}^{\text{ESS}}), \quad (39)$$

$$C_{i,t,\text{pen}}^{\text{HST}} = \max(0, E_{i,\min}^{\text{HST}} - E_{i,t}^{\text{HST}}, E_{i,t}^{\text{HST}} - E_{i,\max}^{\text{HST}}), \quad (40)$$

$$C_{i,t,\text{pen}}^{\text{V}} = \max(0, 0.95 - V_{i,t}, V_{i,t} - 1.05). \quad (41)$$

其中, $C_{i,t,\text{pen}}^{\text{SG}_1}$和$C_{i,t,\text{pen}}^{\text{SG}_2}$分别是对超过$\text{SG}_1$和$\text{SG}_2$容量的惩罚, $C_{i,t,\text{pen}}^{P_{ij,t}^2+Q_{ij,t}^2}$是对超过最大支路传输功率的惩罚, $C_{i,t,\text{pen}}^{\text{ESS}}$和$C_{i,t,\text{pen}}^{\text{HST}}$是对超过 ESS 和 HST 容量的惩罚, $C_{i,t,\text{pen}}^{\text{V}}$是对超过额定电压的惩罚.

综上分析, 在满足上述约束 (1)-(30) 的前提下, 该微电网优化调度的目标函数定义为系统内所有节点设备运行成本与越限惩罚之和:

$$C_t^{\text{cost}} = \sum_{i=1}^{N}(C_{i,t}^{\text{SG}_1} + C_{i,t}^{\text{SG}_2} + C_{i,t}^{\text{EC}} + C_{i,t}^{\text{FC}} + C_{i,t}^{\text{Grid}}), \quad (42)$$

$$C_t^{\text{pen}} = \sum_{i=1}^{N}(C_{i,t,\text{pen}}^{\text{SG}_1} + C_{i,t,\text{pen}}^{\text{SG}_2} + C_{i,t,\text{pen}}^{P_{ij,t}^2+Q_{ij,t}^2} + C_{i,t,\text{pen}}^{\text{ESS}} + C_{i,t,\text{pen}}^{\text{HST}} + C_{i,t,\text{pen}}^{V}), \quad (43)$$

$$\begin{cases} \min \sum_{t=1}^{T}(C_t^{\text{cost}} + \alpha_{\text{pen}} C_t^{\text{pen}}) \\ \text{s.t. } (1)-(43) \end{cases}. \quad (44)$$

其中, N表示节点数量, T为表示优化时域的总时间步数, α_{pen}为平衡成本和惩罚项的系数.

2 基于物理约束梯度引导的微电网能量调度强化学习方法

针对上节构建的含储能单元的微电网多时段交流最优潮流问题, 其决策空间维度随调度时段增加而显著扩大, 储能单元的动态约束与发电机爬坡约束深度耦合, 同时各时段的非凸潮流可行域在时间维度上相互制约, 共同形成了一个高维非凸优化问题. 为此, 本节提出一种基于物理约束梯度引导的微电网能量调度方法. 该方法首先将微电网能量管理问题建模为马尔可夫决策过程, 并采用嵌有安全层的 TD3 算法架构; 进而, 训练一个基于深度学习的动作修正网络, 将其同步嵌入至智能体的环境探索与网络训练过程中, 以实现对原始动作的实时可行性校正与训练阶段的物理梯度引导.

2.1 马尔可夫决策过程

MDP 由五元组(S, A, P, R, γ)构成, 其中S和A分别表示状态的集合和动作的集合; $P(s'|s,a)$: $S \times A \to S$表示从当前动作对(s,a)到下一个状态(s')的转移概率; $R(s,a)$表示在状态s下执行动作a时获得的即时奖励, γ为折扣因子.

2.1.1 状态空间

在t时刻微电网的状态空间$s_t \in S$可以描述为:

$$s_t = (t, P_{i,t}^{\text{SG}_1}, Q_{i,t}^{\text{SG}_1}, P_{i,t}^{\text{SG}_2}, Q_{i,t}^{\text{SG}_2}, E_{i,t}^{\text{ESS}}, E_{i,t}^{\text{HST}}, P_{i,t}^{\text{EL}}, Q_{i,t}^{\text{EL}}, P_{i,t}^{\text{PV}}, P_{i,t}^{\text{WT}}). \quad (45)$$

2.1.2 动作空间

在t时刻微电网的动作空间$a_t \in A$可以描述为:

$$a_t = (\Delta P_{i,t}^{\text{SG}_1}, \Delta Q_{i,t}^{\text{SG}_1}, \Delta P_{i,t}^{\text{SG}_2}, \Delta Q_{i,t}^{\text{SG}_2}, P_{i,t}^{\text{ch}}, P_{i,t}^{\text{dis}}, P_{i,t}^{\text{EC}}, P_{i,t}^{\text{FC}}). \quad (46)$$

2.1.3 奖励函数

在t时刻的奖励函数定义为:

$$r_t = -(C_t^{\text{cost}} + \alpha_{\text{pen}} C_t^{\text{pen}}). \quad (47)$$

2.2 TD3 算法框架

TD3 算法是基于 DDPG 算法提出的改进型深度强化学习算法, 采用 Actor-Critic 框架, 适用于连续动作空间中的学习任务[17]. 在该算法中, Actor 策略网络记为$\mu_{\boldsymbol{\theta}}(s_t)$. 针对 DDPG 中$Q$值估计易出现过高的问题, TD3 引入两个相互独立的Q网络, 分别记为$Q_{\boldsymbol{\varphi}_1}(s_t, a_t)$和$Q_{\boldsymbol{\varphi}_2}(s_t, a_t)$; 在线 Critic 网络通过最小化相应的损失函数进行更新:

$$Q_{1,t} = Q_{\boldsymbol{\varphi}_1}(s_t, a_t), Q_{2,t} = Q_{\boldsymbol{\varphi}2}(s_t, a_t), \quad (48)$$

$$\mathcal{L}^{\text{Critic}} = \mathbb{E}_{s_t \sim \mathcal{D}_2}[(Q_{1,t} - y_t)^2 + (Q_{2,t} - y_t)^2]. \quad (49)$$

TD3 算法通过引入经验回放机制来提升训练效率并稳定Q值的更新过程, 将环境中采样得到的经验元组$(s_t, a_t, s_{t+1}, r_t, done_t)$存储于经验池$\mathcal{D}_2$中, 同时, 算法构建目标 Actor 网络$\mu_{\boldsymbol{\theta}^-}$以及目标 Critic 网络$Q_{\boldsymbol{\varphi}_1^-}$与$Q_{\boldsymbol{\varphi}_2^-}$, 并用于计算目标$Q$值$y_t$:

$$y_t = r_t + \gamma(1 - done_t) \min(Q_{\boldsymbol{\varphi}_1^-}(s_t, \mu_{\boldsymbol{\theta}^-}(s_t)), Q_{\boldsymbol{\varphi}_2^-}(s_t, \mu_{\boldsymbol{\theta}^-}(s_t))). \quad (50)$$

在此基础上, Actor 网络利用策略梯度进行更新, 以实现长期期望回报最大化:

$$\nabla_\theta L^{\text{Actor}} = -\mathbb{E}_{s_t \sim \mathcal{D}_2}[\nabla_{a_t} Q_{\boldsymbol{\varphi}_1}(s_t, \mu_{\boldsymbol{\theta}}(s_t)) \nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(s_t)]. \quad (51)$$

基于上述公式, 可以通过梯度下降来求得在线 Critic 网络和在线 Actor 网络的参数$\boldsymbol{\varphi}$和$\boldsymbol{\theta}$:

$$\boldsymbol{\varphi} \leftarrow \boldsymbol{\varphi} - \lambda_{\boldsymbol{\varphi}} \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{\text{Critic}}, \quad (52)$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathcal{L}^{\text{Actor}}. \quad (53)$$

其中$\lambda_{\boldsymbol{\varphi}}$和$\lambda_{\boldsymbol{\theta}}$分别表示在线 Critic 网络和在线 Actor 网络的学习率. 目标网络通过软更新方式进行更新, 从而保持更新过程的稳定 ($\tau \ll 1$):

$$\boldsymbol{\varphi}^- \leftarrow \tau\boldsymbol{\varphi} + (1-\tau)\boldsymbol{\varphi}^-, \quad (54)$$

$$\boldsymbol{\theta}^- \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta}^-. \quad (55)$$

2.3 基于深度学习的动作修正网络

在微电网能量调度策略的训练过程中, 动作探

索通常需要加入噪声以提升探索效率, 这可能导致训练阶段出现违反约束的行为. 为保证动作的安全性, 本节设计了一个基于深度学习的动作修正安全层, 将不满足约束的动作投影回可行域, 目标是在欧氏范数意义下对原始动作进行最小幅度的修正, 其数学形式如下所示.

$$\begin{cases} \underset{a_t^{\text{safe}}}{\operatorname{argmin}} \|a_t^{\text{safe}} - a_t\| \\ \text{s.t.} \quad (1)-(43) \end{cases} \tag{56}$$

由于交流潮流方程具有明显的非线性特征, 其约束条件本质上呈现非凸性, 从而使传统基于数值优化的方法难以实现快速求解, 难以满足实际应用中的实时性要求. 针对这一问题, 本节提出了一种基于深度学习的动作修正策略, 以替代传统的数值优化求解方式. 该策略由三层全连接神经网络构成, 其结构如图 2 所示.

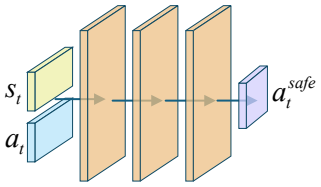

图2 基于深度学习的动作修正网络

在强化学习的预训练阶段, 采用数值优化方法对智能体产生的不安全动作进行修正. 为确保监督标签的质量, 在数据生成阶段实施了严格的可行性筛选机制, 仅当数值求解器成功收敛并返回满足所有物理约束的可行解时, 该样本才会被存入安全层经验池$\mathcal{D}_1$; 对于不可行的样本直接丢弃. 此外, 需要指出的是, 此处数值求解的目标是将动作投影至可行域, 而非直接求解非凸 ACOPF 的全局最优成本, 因此其求解难度相对较低, 能够为安全层提供稳定且高质量的监督信号. 预训练结束后, 利用所收集的全部样本对安全层进行监督学习, 其中 85% 的数据用于训练集, 其余 15% 作为验证集. 训练过程中, 使用均方误差作为损失函数, 其定义如下:

$$\mathcal{L} = MSE(a_i^{\text{pre}} - a_i^{\text{safe}}). \tag{57}$$

其中, a_i^{pre}表示深度学习模型对第i个样本预测的原始动作, a_i^{safe}则表示该样本对应的经数值优化修正后的安全动作. 基于预训练样本得到的安全层网络, 能够在强化学习的早期训练阶段及存在扰动的场景下, 有效提升智能体的动作安全性, 进而改善探索效率. 然而, 在强化学习过程中, 智能体与环境持续交互所产生的数据分布会随策略更新而发生改变, 从而导致与预训练阶段的数据分布产生差异, 为适应此类由策略迭代引起的数据分布漂移, 本节在强化学习训练阶段对基于深度学习的动作修正网络参数进行微调. 具体流程如算法 1 第 33-48 行所示, 在完成 Actor 与 Critic 网络的训练后, 利用指定典型日的运行数据对策略进行测试, 以获取直观的性能评估, 并通过如下不等式判断是否需启动微调:

$$\begin{cases} R_p = \sum\limits_{t=0}^{T} r_t, \quad C_p = \sum\limits_{t=0}^{T} C_t^{\text{pen}} \\ \Delta_R = \frac{|R_p - R_{p-n}|}{|R_p|} \\ \Delta_R > \tau_{\text{R}}, \quad C_p > \tau_{\text{pen}} \end{cases} \tag{58}$$

算法1 物理约束梯度引导的安全强化学习

1 begin
2 初始化Actor-Critic网络 μ_θ,μ_{θ^-},Q_φ,Q_{φ^-};
3 初始化安全层网络g_ψ;
4 初始化安全层经验池$\mathcal{D}_1$和经验回放缓冲区$\mathcal{D}_2$;
5 **for** $i = 1 : N_{\text{pre}}$ **do**
6 获得初始状态s_t;
7 **for** $t = 1 : T$ **do**
8 初始化高斯噪声ε;
9 Actor网络输出动作$a_t = \mu_{\boldsymbol{\theta}}(s_t) + \varepsilon$;
10 数值优化求解(56)得a_t^{safe};
11 执行a_t^{safe}得s_{t+1}和r_t;
12 存储元组$(s_t, a_t, a_t^{\text{safe}})$至$\mathcal{D}_1$;
13 **end**
14 **end**
15 从$\mathcal{D}_1$中随机采样M个元组$(s_t, a_t, a_t^{\text{safe}})$;
16 基于(57)更新安全层参数$\boldsymbol{\psi}$;
17 **for** $episode = 1 : N_{\text{total}}$ **do**
18 **for** $explore = 1 : N_{\text{explore}}$ **do**
19 获得初始状态s_t;
20 **for** $t = 1 : T$ **do**
21 同步骤8-9;
22 经安全层处理得a_t^{safe};
23 执行a_t^{safe}得s_{t+1},r_t和$done_t$;
24 存储元组$(s_t, a_t^{\text{safe}}, s_{t+1}, r_t, done_t)$至$\mathcal{D}_2$;
25 **end**
26 **end**
27 **for** $update = 1 : N_{\text{train}}$ **do**
28 从$\mathcal{D}_2$中随机采样M个元组$(s_t, a_t^{\text{safe}}, s_{t+1}, r_t, done_t)$;
29 基于(61)计算目标Q值y_t;
30 基于(62)和(66)计算Actor-Critic网络梯度;
31 基于(54)和(55)更新目标网络参数;
32 **end**
33 **for** $t = 1 : T$ **do**
34 Actor网络输出动作$a_t = \mu_{\boldsymbol{\theta}}(s_t)$;

35　　同步骤22-23;
36　**end**
37　**if** (58) 不等式组成立 **then**
38　　$count+=1$;
39　　**if** $count>K$ **then**
40　　**for** $t=1:T$ **do**
41　　Actor网络输出动作$a_t=\mu_{\boldsymbol{\theta}}(s_t)$;
42　　同步骤10-12;
43　　**end**
44　　通过最小化损失函数来更新安全层网络:
45　　$\mathcal{L}(\boldsymbol{\psi})=\mathbb{E}_{s_t\sim\mathcal{D}_1}[\|g_{\boldsymbol{\psi}}(s_t,a_t)-a_t^{\text{safe}}\|^2]$;
46　　$\boldsymbol{\psi}\leftarrow\boldsymbol{\psi}-\lambda_{\boldsymbol{\psi}}\nabla_{\boldsymbol{\psi}}\mathcal{L}(\boldsymbol{\psi})$;
47　　**end**
48　**end**
49 **end**
50 **end**

其中, R_p和C_p分别表示第p轮的奖励函数值与违反约束的惩罚值, Δ_R表示第p轮相较于第$p-n$轮的策略变化量, τ_R和τ_{pen}为策略变化量和惩罚值的相应阈值, n为常数. 当 (59) 不等式组成立时, 表明策略在近期出现较大波动或不安全行为显著增多, 计数变量$count$随之累加. 当这种波动在一定训练轮数内持续出现时, 表明策略更新已导致状态分布发生偏移, 当前动作修正网络难以适应新的数据分布, 此时进入微调阶段. 在微调阶段, 基于当前策略采样少量数据, 对动作修正网络的参数进行更新, 以增强其对新状态分布的适应性和安全性.

2.4　Actor-Critic 网络训练

在基于数值优化的安全层方法中, 核心问题在于修正前后的动作之间缺乏可导的梯度关联, 从而阻碍了强化学习中神经网络梯度的有效反向传播. 为缓解这一问题, 现有方法通常在经验回放缓冲区中同时存储原始动作及其对应的安全动作所获得的即时奖励. 然而, 这两个动作之间的映射关系并非直接对应, 而是经历了一个从“不安全动作→安全动作→奖励值”的隐式转换过程. 忽略这一中间转换环节会导致 Critic 网络对Q值的估计出现偏差, 进而影响策略学习的最终精度.

为解决上述问题, 本节将上一节所提出的基于深度学习的动作修正网络嵌入至 Actor-Critic 训练框架中, 从而在训练过程中引入物理约束的梯度引导. 该方法能够有效提升 Critic 网络对Q值估计的准确性, 并增强 Actor 网络对物理约束的学习效率.

1) 环境探索和经验存储: 首先, 智能体观测当前环境状态s_t并输入至 Actor 网络, 得到原始动作a_t. 随后, 在a_t中添加高斯噪声以增强探索. 为保证动作安全, 将a_t与状态s_t共同输入至安全层, 得到安全动作a_t^{safe}. 智能体执行a_t^{safe}后, 环境返回奖励r_t和s_{t+1}. 最终, 将完整转移元组$(s_t,a_t^{\text{safe}},s_{t+1},r_t,done_t)$存入经验回放缓冲区$\mathcal{D}_2$, 供后续网络训练使用.

2) 网络更新: 在更新 Actor-Critic 网络时, 从经验池中均匀采样M个经验元组, 记为$\{(s_t^m,a_t^m,s_{t+1}^m,r_t^m,done_t^m)\}_{m=1,\dots,M}$; 由于经验回放缓冲区$\mathcal{D}_2$中存储的动作为安全动作, 为保持动作空间在时序上的一致性, 目标 Actor 网络所输出的下一时刻动作a_{t+1}^m也需满足相同的安全性约束. 为此, 将其输入至安全层g_{ψ}中进行处理, 得到对应的安全动作$a_{t+1}^{\text{safe},m}$, 具体的更新公式如下:

$$a_{t+1}^m=\mu_{\boldsymbol{\theta}^-}(s_{t+1}^m)+\varepsilon, \tag{59}$$

$$a_{t+1}^{\text{safe},m}=g_{\boldsymbol{\psi}}(a_{t+1}^m,s_{t+1}^m). \tag{60}$$

同理目标Q值y_t也进行更新:

$$y_t^m=r_t^m+\gamma(1-done_t^m)\min(Q_{\boldsymbol{\varphi}_1^-}(s_{t+1}^m,a_{t+1}^{\text{safe},m}),Q_{\boldsymbol{\varphi}_2^-}(s_{t+1}^m,a_{t+1}^{\text{safe},m})). \tag{61}$$

其中, m表示第m个采样得到的经验样本, $done$为终止状态判断标志, 用于决定目标Q值y_t是否包含未来奖励, 在线 Critic 网络的损失函数更新方式与公式 (49) 一致, 由于 Critic 仅负责评估给定状态-动作对的价值, 不直接参与动作的生成和修正过程, 因此其梯度的更新过程中不引入安全层的梯度信息, 具体更新方式如下:

$$\nabla_{\varphi_1,\varphi_2}L^{\text{Critic}}=\mathbb{E}_{s_t\sim\mathcal{D}_2}[2(Q_{1,t}-y_t)\nabla_{\varphi_1}Q_{1,t}]+\mathbb{E}_{s_t\sim\mathcal{D}_2}[2(Q_{2,t}-y_t)\nabla_{\varphi_2}Q_{2,t}]. \tag{62}$$

更新 Actor 网络时, 为使其输出的动作与 Critic 网络的评估标准相一致, 需将在线 Actor 网络生成的原始动作a_t^m同样经由安全层g_{ψ}映射为安全动作$a_t^{\text{safe},m}$, 其计算公式如下:

$$a_t^m=\mu_{\boldsymbol{\theta}}(s_t^m)+\varepsilon, \tag{63}$$

$$a_t^{\text{safe},m}=g_{\boldsymbol{\psi}}(a_t^m,s_t^m). \tag{64}$$

在线 Actor 网络的损失函数也相应进行更新:

$$\mathcal{L}^{\text{Actor}}=-\mathbb{E}_{s_t\sim\mathcal{D}_2}[Q_{\boldsymbol{\varphi}_1}(s_t,a_t^{\text{safe}})]. \tag{65}$$

在计算 Actor 网络损失$\mathcal{L}^{\text{Actor}}$的过程中, 由于动作需经安全层修正之后才参与前向计算, 依据链式法则, 安全层的梯度将通过反向传播直接回传至在线 Actor 网络. 此时, Actor 参数的更新将同时受到安全层梯度与 Critic 网络奖励信号Q_{φ_1}的共同驱动. 该机制确保了即使安全层仅输出满足约束的次优解, 策略网络仍能在奖励信号的指引下向全局最优收敛.

$$\nabla_\theta \mathcal{L}^{\text{Actor}} = -\mathbb{E}_{s_t\sim\mathcal{D}_2}[\nabla_{a_t^{\text{safe}}} Q_{\varphi_1}(s_t, a_t^{\text{safe}}) \frac{\partial g_\psi(a_t, s_t)}{\partial a_t} \nabla_\theta \mu_\theta(s_t)]. \tag{66}$$

在线 Actor 网络与在线 Critic 网络的参数更新分别依据公式 (52) 与 (53) 进行, 其对应目标网络的参数则遵循公式 (54) 与 (55) 所定义的软更新方式. 综上所述, 本节所设计的安全层通过在环境探索中实现可行域投影以保障操作安全性, 并在网络训练中引入物理约束梯度以引导策略更新方向, 从而提升了智能体在复杂调度空间中的学习效率与策略性能.

3 实验结果与分析

本节基于第一节所建立的模型开展实验, 旨在验证所提出方法的有效性. 实验将所提算法与 TD3-Pen[23]、TD3-Lag[24] 以及作为最优基准的 Gurobi 求解器进行对比, 重点分析了各方法在约束违反程度与奖励函数收敛性方面的表现. 在此基础上, 对本文方法训练生成的智能体调度策略进行了合理性评估; 同时, 对比了基于深度学习的动作修正安全层与基于数值优化的安全层[27] 在部署时间与调度性能上的差异; 此外, 通过消融实验来验证安全层在网络训练过程中的作用.

3.1 数据集和仿真设置

本实验选取包含光伏发电, 风力发电及电负荷的全年 365 天数据进行训练, 如图 3 所示. 为了充分评估模型在不同季节与气候条件下的泛化能力, 将数据集按约 8:2 的比例划分为训练集与测试集, 并采用分层随机采样策略, 从全年每个月中各随机抽取 6 天数据 (共计 72 天) 构成测试集, 其余数据用于强化学习训练. 实验所用的设备参数和电价数据分别如表 1 和表 2 所示. 在强化学习训练中, Actor 网络和安全层均采用三层隐藏层结构, 输出层使用 Tanh 激活函数以约束动作空间范围; Critic 网络采用共享特征层与双分支输出层的架构设计, 其中共享特征层包括两层隐藏层, 每个分支输出层各包含一层隐藏层, 所有隐藏层均设置 256 个神经元; 训练所用超参数的设置如下: 预训练轮数 N_{pre} 为 100, 训练总轮数 N_{total} 为 2000, 每次探索轮数 N_{explore} 为 50, 折扣因子 γ 为 0.995, 惩罚因子 α_{pen} 为 100, 软更新因子为 5×10^{-3}, Actor 网络和 Critic 网络学习率分别设置为 1×10^{-4} 和 1.5×10^{-4}, 安全层微调的学习率设为 1×10^{-5}, 经验回放缓冲区设置为 48 000, 每次网络更新采样批次大小为 4 800, 每轮网络更新次数 N_{train} 为 20, 策略延迟更新参数设为 2, 即每两次更新 Critic 网络参数后更新一次 Actor 网络参数. 微调参数 n 为 20, K 为 4, 阈值 τ_{R} 和 τ_{pen} 分别为 0.002 和 0.05. 实验在 Python 3.8 和 pytorch 2.4.1 环境下完成, 实验平台配置包括 Intel(R) Core(TM) i7-10750H CPU, 16 GB RAM 以及 NVIDIA GTX 2 070 GPU.

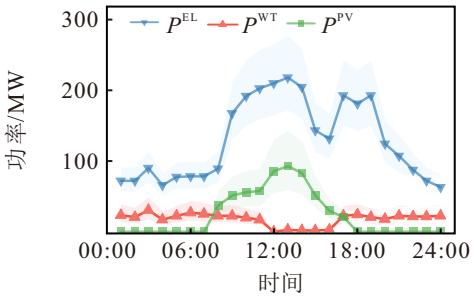

图3 历史风、光、负荷分布图

表1 设备参数

SG_1运行参数							HEC运行参数					
$P_{\max}$ (MW)	$\Delta P_{\min}$ (MW/h)	$\Delta P_{\max}$ (MW/h)	$Q_{\min}$ (Mvar)	$Q_{\max}$ (Mvar)	$\Delta Q_{\min}$ (Mvar/h)	$\Delta Q_{\max}$ (Mvar/h)	$E_{\min}^{\text{HST}}$ (MWh)	$E_{\max}^{\text{HST}}$ (MWh)	$P_{\max}^{\text{EC/FC}}$ (MW)	η^{EC} (—)	η^{FC} (—)	η^{HST} (—)
100	−50	50	−100	100	−80	80	35	85	15	0.9	0.8	0.97
运行参数							ESS运行参数					
$P_{\max}$ (MW)	$\Delta P_{\min}$ (MW/h)	$\Delta P_{\max}$ (MW/h)	$Q_{\min}$ (Mvar)	$Q_{\max}$ (Mvar)	$\Delta Q_{\min}$ (Mvar/h)	$\Delta Q_{\max}$ (Mvar/h)	$E_{\min}^{\text{ESS}}$ (MWh)	$E_{\max}^{\text{ESS}}$ (MWh)	$P_{\max}^{\text{ch}}$ (MW)	$P_{\max}^{\text{dis}}$ (MW)	η^{ch} (—)	η^{dis} (—)
90	−45	45	−90	90	−75	75	20	100	20	20	0.95	0.95
SG_1成本参数			SG_2成本参数			EC成本参数			FC成本参数			
a (元/MW^2)	b (元/MW)	c (元)	a (元/MW^2)	b (元/MW)	c (元)	a (元/MW^2)	b (元/MW)	c (元)	a (元/MW^2)	b (元/MW)	c (元)	β^{H_2} (元/MW)
2.88	60	100	1.88	80	100	1.5	40	80	2.06	20	60	30

3.2 调度性能分析

本节将对训练所得最优模型的调度性能进行评估. 考虑到第 181 天 (6 月末) 呈现出显著的高光伏出力与高负荷需求特征, 其导致的剧烈昼夜功率波

表2　分时电价

时段	购电价格 (元/kW·h)	售电价格 (元/kW·h)
0:00–8:00	0.3578	0.2
8:00–12:00	0.8325	0.4125
12:00–14:00	0.3578	0.2
14:00–20:00	0.8325	0.4125
20:00–22:00	1.2109	0.2
22:00–24:00	0.3578	0.2

动不仅对系统能量时移能力构成挑战, 也伴随着明显的节点电压波动风险. 因此, 选取该日作为典型运行工况, 对其 24 小时调度结果展开分析. 如图 4 所示, 图中不同颜色的图例对应第 1.1 节定义的各设备功率变量, 黑色点线表示系统总电负荷需求. 如图所示, 在 0-9 时段的低负荷时段, 发电机因其成本函数呈二次型且在低出力区间拥有较低的边际发电成本, 其供电成本显著低于外购电成本. 系统在此阶段优先利用发电机满足负荷需求, 并将剩余电能储存于蓄电池或通过电解槽转化为氢能存储于储氢罐, 同时向主电网出售部分电能. 在 9-12 时段, 随着负荷逐步上升, 系统通过蓄电池放电与燃料电池将储氢罐中氢能转换为电能来满足负荷需求, 并继续向主电网出售盈余电能以获取收益.12-15 时段光伏发电达到峰值, 系统将富余电能存储至蓄电池与储氢罐. 在 15-20 时段, 光伏出力下降, 系统转为依赖发电机与蓄电池放电以维持功率平衡, 并在电价较高时段出售电能以提高运行收益.20-24 时段负荷再次回落, 系统仅需依靠发电机与风力发电即可满足全部负荷需求.图 5 展示了各节点 24 小时内的电压变化曲线, 图中不同颜色的曲线分别代表系统中 14 个节点的电压波动轨迹; 所有节点电压始终稳定在 0.95p.u. 至 1.05p.u. 范围内, 有效保障了设备安全与系统稳定运行. 上述结果表明, 本文所提调度策略通过协同优化分布式机组, 储能系统与主电网交易, 实现了系统安全性与经济性的统一.

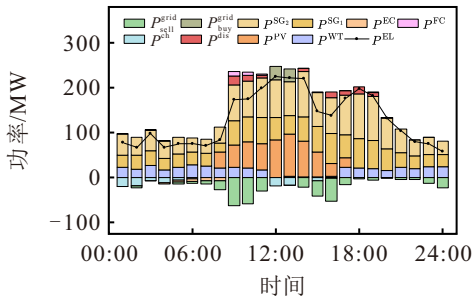

图4　系统优化调度图

3.3　安全层评估

本节对所提出的安全层进行性能评估, 首先将

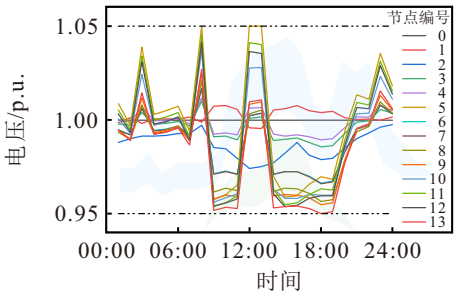

图5　系统节点电压图

其与 TD3-Lag, TD3-Pen 算法在训练过程及最终策略的约束满足性能方面进行对比; 此外, 还在部署时间与调度性能方面, 与基于数值优化的安全层方法进行了比较.图 6 展示了训练过程中各算法的约束违反值变化情况.TD3-Pen 算法约在 1000 轮训练后才将约束违反值降至 1 以下, 最终维持在 0.3 左右; TD3-Lag 在训练前期即可将约束违反值降至较低水平, 但训练过程中波动较大, 最终在 0.06 附近持续波动; 而本文所提的方法在安全层的作用下, 训练前期即可将约束违反值降至 0.07 左右, 并在后续训练中保持稳定, 最终约束违反值稳定在 0.04 左右. 训练结束后, 为评估模型在不同风光出力和负荷条件下的泛化能力, 采用测试集数据进行验证. 结果如表 3 所示, 本文所提方法的平均约束违反值为 0.044, 远低于 TD3-Pen 的 0.251 和 TD3-Lag 的 0.067.

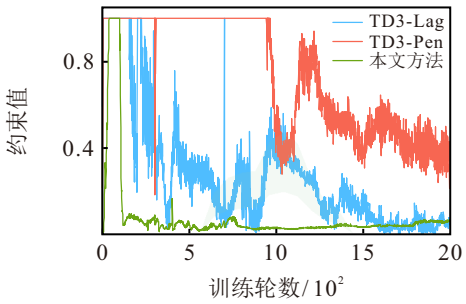

图6　对比实验约束值图

表3　算法在测试集上约束违反值对比

月份	TD3-Pen	TD3-Lag	本文方法
1	0.248	0.072	**0.042**
2	0.247	0.068	**0.048**
3	0.246	0.066	**0.046**
4	0.247	0.065	**0.046**
5	0.247	0.067	**0.045**
6	0.258	0.066	**0.045**
7	0.26	0.071	**0.045**
8	0.26	0.072	**0.044**
9	0.244	0.066	**0.043**
10	0.257	0.065	**0.042**
11	0.251	0.063	**0.043**
12	0.246	0.065	**0.044**
平均值	0.251	0.067	**0.044**

除算法性能外, 安全层的部署效率亦是其实际应用价值的关键体现. 为此, 本研究以 Gurobi 求解的数值优化安全层为基准, 对比分析了所提方法在部署阶段的时间消耗与经济成本, 如表 4 所示. 鉴于数值优化方法存在高昂的计算复杂度, 直接将其嵌入强化学习训练循环将导致无法承受的时间开销, 因此, 本实验在测试阶段利用训练好的最优模型进行 24 小时调度任务对比. 实验结果显示, 基于 Gurobi 的数值优化方案在测试集上完成单日调度的平均耗时为 530 秒, 调度成本为 377 922 元; 相比之下, 本文方法在单日调度的平均耗时仅需 0.48 秒, 调度成本为 378 456 元. 对比数据表明, 本文所提的安全层在经济成本上与数值优化安全层仅存在 0.14% 的微小偏差, 但在计算效率上实现了约三个数量级的显著提升. 综上所述, 本文方法在维持调度经济竞争力的同时, 显著降低了计算开销与部署时间, 展现出优越的实时响应能力与工程应用潜力.

表4 本文方法与数值优化方法的性能对比

月份	数值优化安全层		本文方法		误差(%)
	$r_{\text{数值优化}}$	时间(s)	$r_{\text{本文方法}}$	时间(s)	
1	**−376 528**	539	−377 159	**0.430**	0.17
2	**−372 681**	566	−372 940	**0.494**	0.07
3	**−370 202**	534	−370 663	**0.486**	0.12
4	**−374 911**	547	−375 395	**0.472**	0.13
5	**−377 171**	539	−377 717	**0.514**	0.14
6	**−390 252**	516	−390 958	**0.493**	0.18
7	**−387 884**	521	−388 281	**0.437**	0.10
8	**−385 196**	514	−385 655	**0.514**	0.12
9	**−379 450**	522	−380 143	**0.501**	0.18
10	**−380 858**	543	−381 421	**0.480**	0.15
11	**−371 553**	499	−372 088	**0.477**	0.14
12	**−368 377**	516	−369 055	**0.461**	0.18
平均值	**−377 922**	530	−378 456	**0.480**	0.14

3.4 关键参数分析与消融实验

为选取最优模型结构和超参数, 本节分析了安全层网络深度对策略学习的影响, 并对影响算法收敛性能的关键超参数进行测试, 最后通过梯度引导机制的消融实验进一步证实了所提框架的必要性.

3.4.1 安全层网络结构分析

安全层的深度直接决定了其非线性表征能力与梯度传播有效性, 并最终影响 Actor 网络的性能. 为确定最优网络结构, 本文对比了 1 至 4 层全连接网络在相同训练环境下的表现, 实验结果如图 7 所示. 单层与双层网络由于结构较浅, 在面对微电网高维, 非凸的安全约束边界时表征能力不足, 修正精度缺失导致向策略网络传递有偏的反馈信号, 严重阻碍了算法收敛至全局最优策略. 四层网络在初期虽具备较强拟合能力, 但过多参数引入计算冗余与训练震荡, 降低策略在后期学习的稳定性, 最终性能反略逊于三层网络. 相比之下, 三层全连接结构在模型容量与训练稳定性之间取得了最佳平衡, 既能精确拟合复杂安全边界, 又能为策略网络提供高质量, 低方差的修正引导, 最终获得优于其他层数结构的最高平均奖励值.

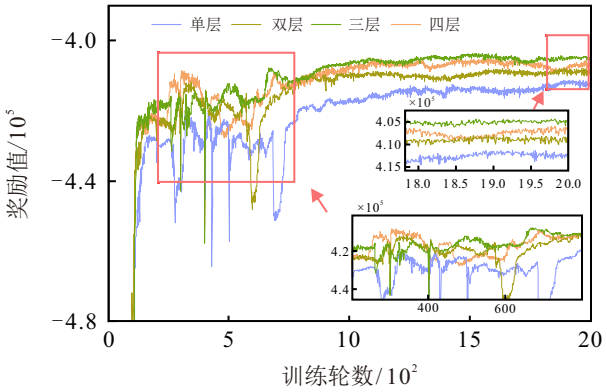

图7 不同安全层层数奖励值图

3.4.2 参数敏感性分析

为了探究学习率对模型收敛特性的影响, 本节依据表 5 设置了多组对比实验, 其对应的奖励值收敛曲线如图 8 所示. 结果表明, 较低的基准学习率显著限制了策略优化的速率, 导致收敛周期延长; 反之, 过高的学习率虽在训练初期加速了性能提升, 但后期因参数更新步长过大, 易引发训练震荡, 难以收敛至全局最优策略. 然而, 在排除上述极端情形后, 从收敛曲线的趋势来看, 所提方法表现出良好的参数适应性, 当学习率设定在1×10^{-4}的数量级范围内, 模型均能实现稳定收敛并获得较高的奖励回报.

表5 各实验组网络学习率参数设置

网络	参数组1	参数组2	参数组3	参数组4	参数组5	参数组6
Actor	5×10^{-5}	7×10^{-5}	1×10^{-4}	3×10^{-4}	5×10^{-4}	1×10^{-3}
Critic	7.5×10^{-5}	1×10^{-4}	1.5×10^{-4}	4.5×10^{-4}	7.5×10^{-4}	1.5×10^{-3}
安全层	1×10^{-6}	7×10^{-6}	1×10^{-5}	3×10^{-5}	5×10^{-5}	1×10^{-4}

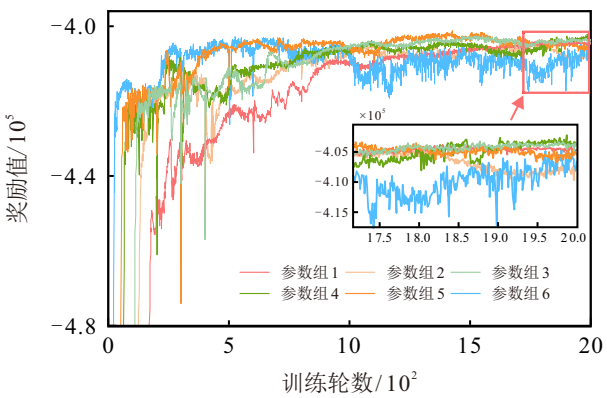

图8　不同学习率奖励值图

由于基于深度学习的动作修正安全层在本质上属于非线性函数逼近器, 在拟合原始动作与安全动作之间的映射时, 不可避免地存在一定的数值逼近误差, 因此, 本文引入惩罚项作为辅助正则化手段, 旨在弥补安全层输出的偏差, 图 9 对比了不同惩罚系数设定下的模型表现, 当α_{pen}=10 时较弱的惩罚力度无法提供足够的梯度信号, 导致策略收敛滞后; 而当α_{pen}=200 时, 过强的惩罚力度则引入了剧烈的梯度波动, 导致训练后期出现震荡并降低了拟合精度. 然而, 当惩罚系数处于适中的取值范围内时, 模型均表现出优异的收敛稳定性与策略性能. 这一结果证实, 本文方法对于惩罚项系数的变化具有良好的参数鲁棒性, 仅作为辅助校正机制的惩罚项无需精细调节即可保证模型性能.

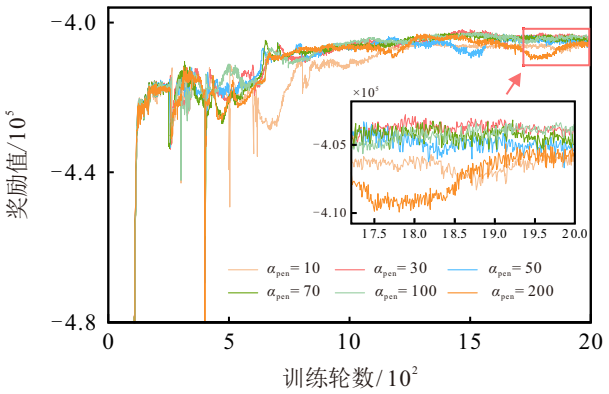

图9　不同惩罚项系数奖励值图

3.4.3　梯度引导机制的消融实验

本节通过设计消融实验, 评估基于深度学习的动作修正安全层在 Actor-Critic 网络训练过程中的作用. 实验对比两种设置: 一是仅在探索阶段引入安全层并存储原始动作用于网络训练; 二是采用本文提出的完整方法, 即在探索和训练阶段均嵌入安全层. 两种设置在网络结构与超参数配置上保持一致, 重点从奖励函数表现与约束违反情况两方面进行对比分析. 本文方法能够有效避免 Critic 网络学习“不安全动作→安全动作→奖励值”这一隐式转换过程导致的估计偏差, 同时使 Actor 网络在物理约束梯度的引导下更有效地学习安全动作. 图 10 与图 11 的结果表明, 相较于仅在探索阶段使用安全层的方案, 本文所提方法在约束违反量上表现出更快的收敛速度, 且在训练全程保持更高的奖励水平, 验证了将安全层嵌入 Actor-Critic 训练全过程的有效性.

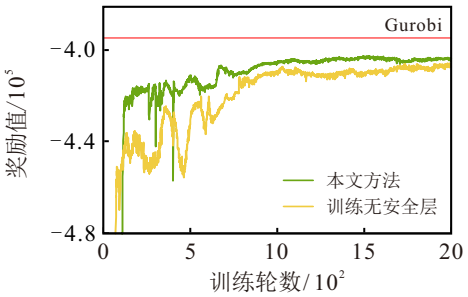

图10　消融实验奖励值图

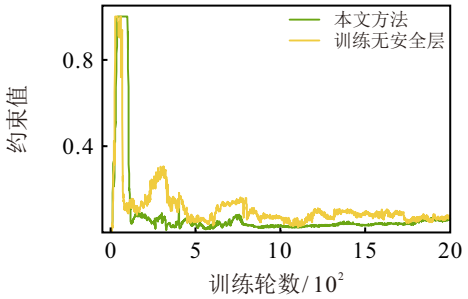

图11　消融实验约束值图

3.5　对比分析

本节将所提方法与 TD3-Lag 及 TD3-Pen 在模型性能方面进行比较. 如图 12 所示, 各算法在训练过程中的奖励变化曲线清晰展示了其性能差异. 在考虑交流潮流约束的非凸优化问题中, TD3-Pen 方法在奖励函数中除运行成本外还引入了约束违反惩罚项. 由于惩罚项梯度变化剧烈, 训练前期奖励信号呈现显著不稳定性, 表现为大幅波动; 随着训练推进, 策略探索过程中仍出现多次小幅震荡, 直至约 1 500 轮后奖励曲线方逐渐收敛.TD3-Lag 方法通过拉格朗日乘子将约束条件融入目标函数, 使策略在更新时可同步降低约束违反与提升奖励, 因而收敛速度优于 TD3-Pen. 然而早期探索阶段仍存在较大波动, 至中后期曲线才趋于稳定上升. 上述两种方法在奖励信号或约束反馈不充分时, 优化过程易陷入局部极值. 相比之下, 本文方法在物理约束梯度的引导下能够高效学习安全策略, 训练初期即获得较高且稳定的奖励水平, 约 800 轮即达到稳定, 后续持续提升性能, 展现出更优的稳定性与学习效率.

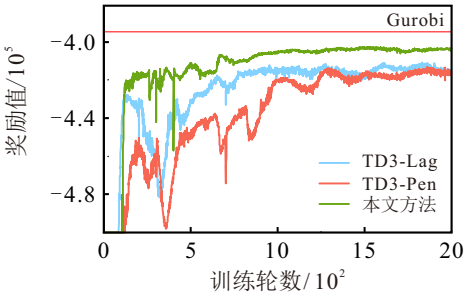

图12　对比实验奖励值图

在完成模型训练后, 利用测试集对具有不同风,

光与负荷特征的数据对三种算法进行测试, 以评估其在多样化运行条件下的泛化性能. 由于输入数据的波动会直接影响奖励值的绝对大小, 为更客观地比较各模型的调度表现, 采用如下定义的相对误差作为评估指标:

$$\varepsilon = \frac{|r - r_{\text{Gurobi}}|}{|r_{\text{Gurobi}}|}. \tag{67}$$

其中r表示模型得到的奖励值, r_{Gurobi}表示使用Gurobi在24小时调度得到的奖励值, 并以此作为最优基准. 相对误差ε越接近0表示模型的性能越接近于最优解. 如表6所示, 本文提出的方法与最优解的相对误差为2.12%, 远小于TD3-Pen的4.55%和TD3-Lag的4.56%, 体现出更优的调度性能.

表6 算法在测试集上奖励值对比

月份	r_{Gurobi}	TD3-Pen		TD3-Lag		本文方法	
		$r_{\text{TD3-Pen}}$	ε (%)	$r_{\text{TD3-Lag}}$	ε (%)	r_{TD3}	ε (%)
1	**−369945**	−386485	4.47	−386693	4.53	−377159	**1.95**
2	**−365011**	−382364	4.75	−382164	4.70	−372940	**2.17**
3	**−362711**	−379979	4.76	−379815	4.72	−370663	**2.19**
4	**−367659**	−384056	4.46	−384386	4.55	−375395	**2.10**
5	**−369554**	−386398	4.56	−386677	4.63	−377717	**2.21**
6	**−383166**	−400198	4.45	−399790	4.34	−390958	**2.03**
7	**−380108**	−396799	4.39	−396595	4.34	−388281	**2.15**
8	**−377702**	−394298	4.39	−394619	4.48	−385655	**2.11**
9	**−372308**	−389544	4.63	−389054	4.50	−380143	**2.10**
10	**−373368**	−389928	4.44	−390408	4.56	−381421	**2.16**
11	**−364618**	−381209	4.55	−381519	4.64	−372088	**2.05**
12	**−361227**	−378173	4.69	−378366	4.74	−369055	**2.17**
平均值	**−370615**	−387453	4.55	−387507	4.56	−378456	**2.12**

4 结 论

本研究证实了所提出的基于物理约束梯度引导的微电网能量调度安全强化学习策略, 是解决非凸潮流约束下策略优化难题的有效途径. 在此框架下, Critic网络能够直接针对修正后的安全动作进行价值评估, 从而有效消除了因评价目标不一致导致的估计偏差; 同时, 物理约束梯度的反向传播机制则直接引导Actor网络向可行域更新, 从根本上克服了传统数值修正法导致的梯度阻断问题. 基于改进的IEEE 14节点电力系统实验表明, 本文方法以2.12%的微小误差逼近全局最优解的同时, 相比于传统基于数值优化的安全层方法, 实现了决策速度约三个数量级的显著提升. 上述结果表明, 所提方法在保障系统物理安全性的基础上, 有效解决了考虑潮流计算下微电网能量调度模型中计算效率与优化精度难以平衡的难题, 为实现复杂非线性场景下的实时安全调度提供了新思路.

然而, 该方法目前仍存在一定的局限性, 模型的训练与推理过程高度依赖于预设的微电网拓扑结构, 缺乏跨拓扑的泛化能力, 当电网物理连接发生显著变化时往往需要重新训练. 针对这一不足, 未来的研究将致力于提升模型对变拓扑场景适应性的研究, 从而实现调度策略在不同微电网架构间的高效迁移与泛化.

参考文献 (References)

[1] 国家电网公司. "碳达峰、碳中和"行动方案[J]. 国家电网, 2021(03): 50-52.
(State Grid Corporation of China. Action Plan for Carbon Peaking and Carbon Neutrality[J]. State Grid, 2021(03): 50-52.)

[2] 刘鑫蕊, 李新宇, 郭亮亮, 等. 考虑自治-互济的多园区端网协同弹性控制[J]. 控制与决策, 2025, 40(9): 2693-2700.
(Liu X R, Li X Y, Guo L L, et al. Resilience cooperative control for MMG considering autonomy-mutual aid of edge-network[J]. Control and Decision, 2025, 40(9): 2693-2700.)

[3] 郭方洪, 徐博文, 张文安, 等. 基于学习优化的智能电网能量管理研究综述[J]. 控制与决策, 2022, 37(05): 1089-1101.
(Guo F H, Xu B W, Zhang W A, et al. Learning-to-optimize based energy management in smart grid: A survey[J]. Control and Decision, 2022, 37(05): 1089-1101.)

[4] Su T, Wu T, Zhao J, et al. A review of safe reinforcement learning methods for modern power systems[J]. Proceedings of the IEEE, 2025, 113(3): 213-255.

[5] Liu X K, Jiang H, Wang Y W, et al. A distributed iterative learning framework for DC microgrids: current sharing and voltage regulation[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2020, 4(2): 119-129.

[6] Hao G K, Li Y Z, Li Y, et al. Lyapunov-based safe reinforcement learning for microgrid energy management[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025.

[7] Yang H, Tian H, Guo F, et al. Safety-Guaranteed Energy Management in Networked Multienergy Microgrids: A Multi-Actor Single-Critic Deep Reinforcement Learning Approach[J]. IEEE Transactions on Industrial Informatics, 2026. DOI: 10.1109/TII.2025.3649145.

[8] 郭方洪, 冯秀荣, 杨溟, 等. 基于数据模型双驱动的新能源微电网分布鲁棒优化调度[J]. 电力系统自动化, 2024, 48(20): 36-47.
(Guo F H, Feng X R, Yang H, et al. Dual-data-model-driven Distributionally Robust Optimal Scheduling of Renewable Energy Microgrid[J]. Automation of Electric Power Systems, 2024, 48(20): 36-47.)

[9] 郭亮亮, 刘鑫蕊, 李新宇, 等. 一种提高均流精度的直

流微电网分布式储能 SOC 加速均衡控制策略[J]. 控制与决策, 2025, 40(8): 2383-2390.
(Guo L L, Liu X R, Li X Y, et al. An accelerated SOC balancing control strategy for DC microgrid distributed energy storage with improved current sharing accuracy[J]. Control and Decision, 2025, 40(8): 2383-2390.)

[10] Zamzam A S, Baker K. Learning optimal solutions for extremely fast AC optimal power flow[C]. In: Proceedings of the 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). Tempe, USA: IEEE, 2020: 1-6.

[11] Chen K, Bose S, Zhang Y. Unsupervised deep learning for AC optimal power flow via Lagrangian duality[C]. In: Proceedings of GLOBECOM 2022 — IEEE Global Communications Conference. Rio de Janeiro, Brazil: IEEE, 2022: 5305–5310.

[12] Park S, Van Hentenryck P. Self-supervised primal-dual learning for constrained optimization[C]. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington DC, USA: AAAI Press, 2023, 37(4): 4052–4060.

[13] Pan X, Chen M, Zhao T, et al. DeepOPF: A feasibility-optimized deep neural network approach for AC optimal power flow problems[J]. IEEE Systems Journal, 2023, 17(1): 673-683.

[14] Zhou M, Chen M H, Low S H. DeepOPF-FT: One deep neural network for multiple AC-OPF problems with flexible topology[J]. IEEE Transactions on Power Systems, 2023, 38(1): 964-967.

[15] Chen W, Tanneau M, Van Hentenryck P. End-to-end feasible optimization proxies for large-scale economic dispatch[J]. IEEE Transactions on Power Systems, 2023, 39(2): 4723-4734.

[16] Han J, Wang W, Yang C, et al. Frmnet: A feasibility restoration mapping deep neural network for AC optimal power flow[J]. IEEE Transactions on Power Systems, 2024, 39(5): 6566-6577.

[17] Yang H, Liu S, Guo F H, Wu X. Two-Stage Power Scheduling for More Electric Aircraft: A Hybrid Deep Reinforcement Learning-Based Strategy[J]. IEEE Transactions on Industrial Electronics, 2025, 72(11): 11747-11757.

[18] Lee S, Seon J, Sun Y G, et al. Novel architecture of energy management systems based on deep reinforcement learning in microgrid[J]. IEEE Transactions on Smart Grid, 2024, 15(2): 1646-1658.

[19] Domínguez-Barbero D, García-González J, Sanz-Bobi M Á, et al. Energy management of a microgrid considering nonlinear losses in batteries through Deep Reinforcement Learning[J]. Applied Energy, 2024, 368: 123435.

[20] Tu Z, Zhang W, Liu W. Deep reinforcement learning-based optimal control of DC shipboard power systems for pulsed power load accommodation[J]. IEEE Transactions on Smart Grid, 2023, 14(1): 29-40.

[21] 郭方洪, 伍泽芃, 杨淏, 等. 基于个性化联邦强化学习的异构多微网能量调度[J]. 自动化学报, 2025, 51(09): 2072-2084.
(Guo F H, Wu Z P, Yang H, et al. Energy Scheduling of Heterogeneous Multi-microgrid Based on Personalized Federated Reinforcement Learning[J]. ACTA AUTOMATICA SINICA, 2025, 51(09): 2072-2084.)

[22] 王丹璐, 孙秋野, 苏涵光. 多微网系统端网协同分布式实时智能优化[J]. 控制与决策, 2024, 39(11): 3801-3809.
(Wang D L, Sun Q Y, Su H G. Collaborative distributed real-time intelligent optimization of multi-microgrid system[J]. Control and Decision, 2024, 39(11): 3801-3809.)

[23] Cao D, Hu W, Xu X, et al. Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices[J]. Journal of Modern Power Systems and Clean Energy, 2021, 9(5): 1101-1110.

[24] 季颖, 王建辉. 基于深度强化学习的微电网在线优化调度[J]. 控制与决策, 2022, 37(07): 1675-1684.
(Ji Y, Wang J H. Online optimal scheduling of a microgrid based on deep reinforcement learning[J]. Control and Decision, 2022, 37(07): 1675-1684.)

[25] Wu Z, Zhang M, Gao S, et al. Physics-informed reinforcement learning for real-time optimal power flow with renewable energy resources[J]. IEEE Transactions on Sustainable Energy, 2025, 16(1): 216-226.

[26] Sayed A R, Wang C, Anis H I, et al. Feasibility constrained online calculation for real-time optimal power flow: A convex constrained deep reinforcement learning approach[J]. IEEE Transactions on Power Systems, 2023, 38(6): 5215-5227.

[27] Wang Y, Qiu D, Sun M, et al. Secure energy management of multi-energy microgrid: A physics-informed safe reinforcement learning approach[J]. Applied Energy, 2023, 335: 120759.

[28] Yi Z, Wang X, Yang C, et al. Real-time sequential security-constrained optimal power flow: A hybrid knowledge-data-driven reinforcement learning approach[J]. IEEE Transactions on Power Systems, 2024, 39(1): 1664-1680.

作者简介

沈佳俊 (2002—), 男, 硕士生, 主要研究方向为微电网优化调度, E-mail: 211124030090@zjut.edu.cn;

杨淏 (1999—), 男, 博士生, 主要研究方向为微电网能量管理, E-mail: haoyang@zjut.edu.cn;

陈英豪 (2001—), 男, 硕士生, 主要研究方向为微电网优化调度, E-mail: yinghaochan@zjut.edu.cn;

郭方洪 (1987—), 男, 副教授, 博士生导师, 主要研究智能电网、分布式控制与优化, E-mail: fhguo@zjut.edu.cn.