DOI: 10.19783/j.cnki.pspc.240555

基于约束强化学习的综合能源系统优化调度研究

李天明，王小君，窦嘉铭，刘 婴，司方远，和敬涵

（北京交通大学电气工程学院，北京 100044）

摘要：“双碳”目标下，分布式能源高比例渗透与异质能源耦合加剧迫使综合能源系统(integrated energy system, IES)优化调度问题的求解难度提升，深度强化学习为解决上述问题提供了有效手段。然而，传统深度强化学习通常将安全约束以惩罚项形式加权添加至奖励函数，加权系数一般由人工确定且在迭代过程中保持固定，一定程度上影响了算法的收敛性能与约束处理能力。对此，提出一种基于约束强化学习的 IES 优化调度方法。首先，构建了基于 IES 机组运行与系统潮流约束的安全价值网络，并通过拉格朗日乘子与经济价值网络动态并行协同，分别评估智能体决策的安全与经济价值。其次，利用原始对偶的思路，交替更新智能体策略与拉格朗日乘子，以规避人工设置惩罚系数引起的主观偏差对 IES 调度决策的影响。同时，利用专家知识引导智能体开展训练，防止其盲目寻优造成算力浪费。最后，基于电-热耦合系统开展仿真算例对比分析，验证了所提方法的安全性与高效性。

关键词：综合能源系统；优化调度；深度强化学习；约束强化学习

Research on optimal dispatch of integrated energy systems based on constrained reinforcement learning

LI Tianming, WANG Xiaojun, DOU Jiaming, LIU Zhao, SI Fangyuan, HE Jinghan

(School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: With the “dual-carbon” goal, the high penetration of distributed energy and intensified coupling of heterogeneous energy sources have made it difficult to solve the optimal dispatch problem in integrated energy systems (IES). Deep reinforcement learning provides an effective means to address this challenge. However, traditional deep reinforcement learning usually weights the safety constraints to the reward function in the form of penalty terms, and the weighting coefficients are usually determined manually and remain fixed during iterations, affecting the convergence performance and constraint handling capability of the algorithm to some extent. To address this issue, this paper proposes an IES optimal dispatch method based on constrained reinforcement learning. First, a safety value network based on IES unit operation and system power flow constraints is constructed. The safety and an economic value of agent decisions are evaluated respectively through the dynamic parallel synergy of Lagrange multipliers and an economic value network. Second, the primal-dual approach is used to update the agent policy and Lagrange multipliers alternately to circumvent the influence of subjective bias caused by manually set penalty coefficients in IES scheduling decisions. Additionally, expert knowledge is leveraged to guide the training process to prevent computational resource waste due to blind optimization. Finally, simulation case studies are carried out based on an electric-thermal coupled system to verify the safety and efficiency of the proposed method.

This work is supported by the National Natural Science Foundation of China (No. 51977005).

Key words: integrated energy system; optimal dispatch; deep reinforcement learning; constrained reinforcement learning

0 引言

“双碳”目标下，综合能源系统(integrated energy system, IES)因其多种能源在时空维度上的互补互济，已成为促进分布式能源就地消纳、提质增效、低碳减排的典型代表[1-2]。IES 的安全稳定运行是实现能源的高效利用与绿色转型的重要支撑。然而，分布式能源高比例渗透导致电网潮流越限、电压波动等安全问题日益显著，系统调度调控难度提升[3]。

基金项目：国家自然科学基金项目资助(51977005)；国家自然科学基金青年基金项目资助(52207112)

因此，如何在保证系统安全稳定运行的前提下，提升系统经济效益、实现高效决策已成为 IES 亟需解决的关键难题。

目前，IES 传统运行调度方法大多为解析方法[4]和启发式方法[5]。随着多样化负荷需求增多与能源耦合形式日趋复杂，实时调度安全约束条件与决策变量随之增多[6]。面对高维非线性问题时，启发式方法寻优速度较慢，易陷入局部最优解；解析方法往往需精确建模，利用凸松弛或线性化方法简化模型求解[7]，致使结果难以满足实际调度需求。因此，IES 传统调度方法在安全运行与高效决策方面面临巨大挑战。

随着电力系统数字化、智能化的发展，以深度强化学习(deep reinforcement learning, DRL)为代表的智能调度方法可有效利用海量运行数据直接挖掘系统状态与调度决策之间的复杂映射[8]，凭借“去模型化”优势在无功优化[9]、电网频率控制[10]、能源调度[11]等多个领域广泛应用。其中，文献[12]基于离散动作的深度 Q 网络求解了多主体间的能量协同优化问题。考虑到上述离散动作可引发次优决策，文献[13]引入连续动作空间的多智能体深度确定性策略梯度算法，实现了园区的最优能量管理。鉴于源荷双侧不确定性逐步增强，文献[14]利用近端策略优化算法实现了 IES 的动态调度决策。上述研究为实现 IES 高效决策提供了良好的思路，但仍面临以下挑战：1) 无模型 DRL 普遍将约束越限以惩罚形式加权纳入奖励函数，约束信息与权重系数的主观映射致使调度智能体难以对安全稳定边界准确把握[15]；2) IES 优化空间复杂，智能体初始探索空间大，寻优过程漫长进而导致训练效率低下。

为解决传统 DRL 对约束处理不足的问题，已有学者采用控制障碍函数法[16]、约束策略优化[17]、安全校正[18]等多种安全强化学习方法进行研究。文献[16]将控制障碍函数与 DRL 结合避免了训练过程中的不安全探索。文献[17]构建约束马尔可夫决策过程(constrained Markov decision process, CMDP)，并利用约束策略优化(constrained policy optimization, CPO)算法进行求解以保证策略安全性。文献[18]引入安全校正环节改变智能体原有输出决策实现系统实时响应。上述约束方法均具有一定的效果，但其在处理含源荷不确定性、电网潮流等非线性约束问题时仍存在训练效率低下、计算成本昂贵等问题[6]。

考虑到实际调度需满足相应安全约束并提高计算效率，本文提出一种基于约束强化学习的 IES 优化调度方法，主要研究内容如下：1) 基于模型知识增设安全价值网络以实现目标函数与约束条件解耦，将拉格朗日乘子与双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)算法结合，进而提出面向 IES 安全经济调度的约束强化学习优化方法；2) 将模仿学习预训练引入上述约束强化学习优化方法中，利用专家经验加速智能体学习进程；3) 仿真算例表明，所提方法相较于传统 DRL 有更高的安全性与时效性。

1 基于约束强化学习的 IES 优化调度框架

为便于读者理解，本节首先介绍物理对象，随后概述所提出的基于约束强化学习的 IES 优化调度整体框架。

1.1 区域综合能源系统

本文研究对象为区域 IES。区域 IES 具备多种异质能源产销供一体化功能[19]，作为社会能源消耗的重要单元，其能源结构与用能需求日趋复杂，能源调度需要更加精细、安全的管理[20]，其涉及电、热、气 3 种能源形式。源侧由上级电网、上级热网和外部气网为系统持续供能；荷侧以电、热负荷为主要用能形式；系统内部异质能源耦合复杂，主要能源转换设备包括热电联产机组(combined heat and power, CHP)、电锅炉(electric boiler, EB)和燃气锅炉(gas boiler, GB)；内部设有光伏阵列组(photovoltaic, PV)，利用分布式能源满足区域能源供给；系统中还包括电储能(battery energy storage, BES)装置，以确保系统高效、灵活运行。区域 IES 供能结构如附录 A 图 A1 所示。

1.2 基于约束强化学习的 IES 优化调度框架

区域 IES 负荷需求较大、能源形式多，系统的安全运行与高效决策是可靠供能、提质增效的前提。故而，本文构建了基于约束强化学习的 IES 优化调度框架，如图 1 所示。该框架涉及三个关键阶段，即模仿学习预训练阶段、离线训练阶段和在线应用部署阶段。

1) 模仿学习预训练阶段：该阶段缩小智能体初始探索空间至最优策略邻域，进而提升算法性能。

2) 离线训练阶段：该阶段分别构建经济价值网络与安全价值网络实现经济目标与安全约束解耦，并利用原始对偶思想自适应更新智能体调度策略，平衡最优性与安全性。

3) 在线应用部署阶段：在该阶段中，将 IES 当前运行状态信息作为输入，调度智能体根据状态数据快速制定相应调度方案。

后文将以此框架展开，详细介绍 IES 能源调度模型、求解算法及整体实现流程。

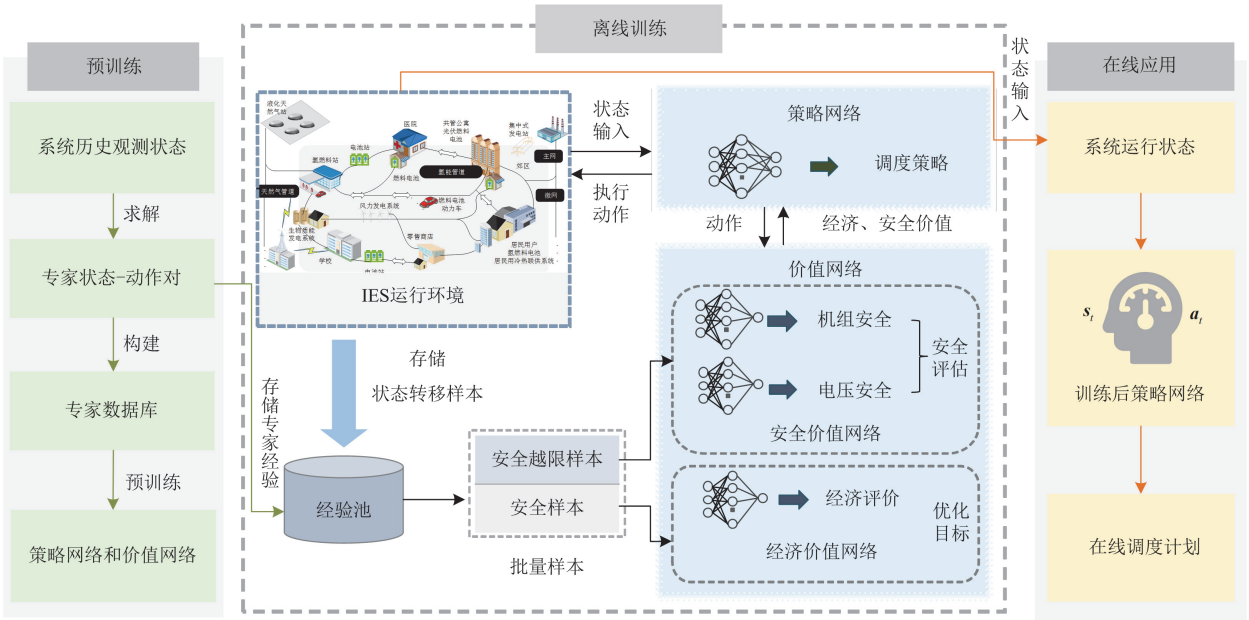


图 1　基于约束强化学习的 IES 优化调度框架

Fig. 1 Framework for IES optimal dispatch based on constrained reinforcement learning

2　基于约束马尔可夫决策过程的 IES 能源调度模型

本节在第 1 节框架之下对 IES 能源调度模型进行构建。首先构建了含优化目标与安全约束的 IES 能源调度数学模型，并进一步将其转化为 CMDP，为约束强化学习算法部署奠定基础。

2.1 IES 能源调度数学模型

2.1.1 IES 能源调度目标函数

本文调度目标为在 IES 安全运行前提下降低运行成本。因此，以总调度成本最小为目标函数，如式(1)所示。

$$\min F=\sum_{t=1}^{T}\left(F_{\mathrm{NG}}(t)+F_{\mathrm{grid}}(t)+F_{\mathrm{BES}}(t)\right) \tag{1}$$

式中：F 为系统运行成本；T 为运行周期的时间段数；$F_{\mathrm{NG}}(t)$、$F_{\mathrm{grid}}(t)$ 和 $F_{\mathrm{BES}}(t)$ 分别为 t 时刻的购气成本、购电成本和电储能折旧成本，分别如式(2)—式(4)所示。

$$F_{\mathrm{NG}}(t)=\sum_{i\in\Omega_{\mathrm{NG}}}\kappa_{\mathrm{NG}}^{t}\left(M_{\mathrm{CHP}}^{i,t}+M_{\mathrm{GB}}^{i,t}\right) \tag{2}$$

$$F_{\mathrm{grid}}(t)=\sum_{i\in\Omega_{\mathrm{grid}}}\kappa_{\mathrm{E}}^{t}P_{\mathrm{grid}}^{i,t} \tag{3}$$

$$F_{\mathrm{BES}}(t)=\sum_{i\in\Omega_{\mathrm{BES}}}\alpha_{\mathrm{BES}}\left|P_{\mathrm{BES}}^{i,t}\right| \tag{4}$$

式中：Ω_{NG}、Ω_{grid} 和 Ω_{BES} 分别为购气、购电和电储能的节点集合；κ_{NG}^{t}、κ_{E}^{t} 分别为 t 时刻的购气和购电单价；$M_{\mathrm{GB}}^{i,t}$、$M_{\mathrm{CHP}}^{i,t}$ 分别为 t 时刻节点 i 处 GB 和 CHP 消耗天然气的总量；$P_{\mathrm{grid}}^{i,t}$ 为节点 i 在 t 时刻向上级电网购电总量；α_{BES} 为 BES 的折旧成本系数；$P_{\mathrm{BES}}^{i,t}$ 为节点 i 在 t 时刻 BES 的充、放电功率，正值为放电，负值为充电。

2.1.2 IES 能源调度约束条件

IES 能源调度需在系统运行安全边界内开展，因此，设定网络安全约束[21]包括电/热网的潮流约束、节点电压约束和机组运行约束。

1) 电网潮流约束

建立节点注入型的潮流模型与节点功率平衡约束，如式(5)—式(7)所示。

$$\begin{cases}P^{i,t}=U^{i,t}\sum_{j=1}^{N}U^{j,t}\left(G_{ij}\cos\delta_{ij}^{t}+B_{ij}\sin\delta_{ij}^{t}\right)\\ Q^{i,t}=U^{i,t}\sum_{j=1}^{N}U^{j,t}\left(G_{ij}\sin\delta_{ij}^{t}-B_{ij}\cos\delta_{ij}^{t}\right)\end{cases} \tag{5}$$

$$P^{i,t}=P_{\mathrm{grid}}^{i,t}+P_{\mathrm{PV}}^{i,t}+P_{\mathrm{CHP}}^{i,t}+P_{\mathrm{BES}}^{i,t}-P_{\mathrm{EB}}^{i,t}-P_{\mathrm{load}}^{i,t} \tag{6}$$

$$Q^{i,t}=Q_{\mathrm{grid}}^{i,t}+Q_{\mathrm{PV}}^{i,t}+Q_{\mathrm{CHP}}^{i,t}-Q_{\mathrm{load}}^{i,t} \tag{7}$$

式中：$P^{i,t}$、$Q^{i,t}$ 分别为 t 时刻流入节点 i 的有功功率和无功功率；$U^{i,t}$、$U^{j,t}$ 分别为节点 i、j 在 t 时刻的电压；N 为电网节点数目；G_{ij}、B_{ij} 分别为两节点间的电导和电纳；δ_{ij}^{t} 为 t 时刻两节点间的相角差；$P_{\mathrm{PV}}^{i,t}$、$Q_{\mathrm{PV}}^{i,t}$ 分别为 t 时刻节点 i 处 PV 输出的有功功

率和无功功率；$P_{\mathrm{CHP}}^{i,t}$ 和 $Q_{\mathrm{CHP}}^{i,t}$ 分别为 t 时刻节点 i 处 CHP 输出的有功功率和无功功率；$P_{\mathrm{EB}}^{i,t}$ 为 t 时刻节点 i 处 EB 消耗的有功功率；$P_{\mathrm{load}}^{i,t}$ 和 $Q_{\mathrm{load}}^{i,t}$ 分别为 t 时刻节点 i 处负荷的有功功率和无功功率；$Q_{\mathrm{grid}}^{i,t}$ 为 t 时刻节点 i 与上级电网交互的无功功率。

2) 节点电压约束

为保证系统安全、改善电能质量，t 时刻节点 i 的电压 $U^{i,t}$ 满足约束如式(8)所示。

$$U_i^{\min} \leqslant U^{i,t} \leqslant U_i^{\max} \tag{8}$$

式中，$U_i^{\max}$ 和 $U_i^{\min}$ 分别为节点 i 安全电压上、下限。

3) 热网潮流模型

建立热网潮流模型，采用固定水流量的质调节方式，模型约束如式(9)—式(13)所示。

$$C_{\mathrm{p}}\boldsymbol{A}_{\mathrm{down}}\boldsymbol{M}\boldsymbol{A}_{\mathrm{down}}^{\mathrm{T}}\boldsymbol{T}_{\mathrm{n}}^{t} - C_{\mathrm{p}}\boldsymbol{A}_{\mathrm{up}}\boldsymbol{M}\boldsymbol{T}_{\mathrm{e}}^{t} = \boldsymbol{H}_{\mathrm{n}}^{t} \tag{9}$$

$$\boldsymbol{H}_{\mathrm{n}}^{t} = \boldsymbol{H}_{\mathrm{EB}}^{t} + \boldsymbol{H}_{\mathrm{CHP}}^{t} + \boldsymbol{H}_{\mathrm{GB}}^{t} - \boldsymbol{H}_{\mathrm{load}}^{t} \tag{10}$$

$$\boldsymbol{E}(\boldsymbol{A}_{\mathrm{down}}^{\mathrm{T}}\boldsymbol{T}_{\mathrm{n}}^{t} - \boldsymbol{T}_{\mathrm{a}}^{t}) + \boldsymbol{T}_{\mathrm{a}}^{t} = \boldsymbol{T}_{\mathrm{e}}^{t} \tag{11}$$

$$\boldsymbol{M} = \mathrm{diag}(m_1, m_2, \cdots, m_i) \tag{12}$$

$$\boldsymbol{E} = \mathrm{diag}\left(\mathrm{e}^{-\frac{\lambda_1 L_1}{C_{\mathrm{p}} m_1}}, \mathrm{e}^{-\frac{\lambda_2 L_2}{C_{\mathrm{p}} m_2}}, \cdots, \mathrm{e}^{-\frac{\lambda_i L_i}{C_{\mathrm{p}} m_i}}\right) \tag{13}$$

式中：C_{p} 为水的比热容；$\boldsymbol{A}_{\mathrm{up}}$ 和 $\boldsymbol{A}_{\mathrm{down}}$ 分别为上、下关联矩阵；$\boldsymbol{A}_{\mathrm{down}}^{\mathrm{T}}$ 为下关联矩阵的转置矩阵；$\boldsymbol{M}$ 为管道流量矩阵；$\boldsymbol{H}_{\mathrm{n}}^{t}$ 为 t 时刻注入节点的热功率矩阵；$\boldsymbol{H}_{\mathrm{EB}}^{t}$、$\boldsymbol{H}_{\mathrm{CHP}}^{t}$、$\boldsymbol{H}_{\mathrm{GB}}^{t}$ 和 $\boldsymbol{H}_{\mathrm{load}}^{t}$ 分别为 EB、CHP、GB 和负荷在 t 时刻的热功率矩阵；$\boldsymbol{E}$ 为温度衰减系数矩阵；$\boldsymbol{T}_{\mathrm{n}}^{t}$、$\boldsymbol{T}_{\mathrm{e}}^{t}$ 和 $\boldsymbol{T}_{\mathrm{a}}^{t}$ 分别为 t 时刻节点温度、管道末端温度与环境温度矩阵；m_i、λ_i 和 L_i 分别为第 i 条管道的流量、导热率与长度。

4) 机组运行约束

各机组需满足的输出功率如式(14)所示。

$$P_x^{\min} \leqslant P_x^{i,t} \leqslant P_x^{\max} \tag{14}$$

式中：$P_x^{i,t}$ 为 i 节点处机组 x 在 t 时刻的有功出力；$P_x^{\max}$ 和 $P_x^{\min}$ 分别为任一时刻机组 x 出力上、下限。各机组具体运行特性见附录 B。

2.2 基于约束马尔可夫决策过程的IES能源调度模型

本节进一步将上述含多约束的 IES 能源调度问题建模成 CMDP，旨在满足网络安全约束并最大化经济效益。具体来说，首先构造了基于无约束马尔可夫决策过程(Markov decision process, MDP)的 IES 能源调度模型，并在此基础上嵌入安全约束条件得到 CMDP，以用于后续求解。CMDP 与数学模型之间的变量对应关系如表 1 所示。

表 1 CMDP 与数学模型之间的变量对应关系

Table 1 Correspondence between variables of CMDP and mathematical model

CMDP 过程要素	IES 优化调度变量
环境	IES 系统运行环境
状态空间	系统运行状态与负荷需求
动作空间	IES 能源调度决策变量
奖励函数	系统运行总成本
安全约束函数	安全约束条件，包含电压偏移量、机组越限程度
状态转移概率	与环境有关，具有随机性

2.2.1 无约束 MDP 模型

本节将深入阐述不考虑安全约束的 MDP 结构中的状态空间、动作空间、奖励函数和动作价值函数及最优策略设计过程，并在 2.2.2 节引入约束部分得到 CMDP 模型。

1) 状态空间设计

调度决策过程中，系统运行状态受到诸如电热负荷、设备输出功率与环境随机性以及整体网络拓扑等多种因素的共同影响[8]。因此，状态空间 $\boldsymbol{s}_t$ 设计如式(15)所示。

$$\begin{aligned}\boldsymbol{s}_t = [&P_{\mathrm{load}}^{i,t}, Q_{\mathrm{load}}^{i,t}, H_{\mathrm{load}}^{i,t}, P_{\mathrm{PV}}^{i,t}, C_{\mathrm{SOC}}^{i,t-1},\\ &P_{\mathrm{BES}}^{i,t-1}, P_{\mathrm{CHP}}^{i,t-1}, H_{\mathrm{EB}}^{i,t-1}, H_{\mathrm{GB}}^{i,t-1}]\end{aligned} \tag{15}$$

式中：$H_{\mathrm{load}}^{i,t}$ 为节点 i 在 t 时刻的热负荷；$C_{\mathrm{SOC}}^{i,t-1}$ 为 t−1 时刻节点 i 处 BES 的荷电状态；$P_{\mathrm{BES}}^{i,t-1}$、$P_{\mathrm{CHP}}^{i,t-1}$ 分别为 t−1 时刻节点 i 处 BES 的充放电功率、CHP 机组的有功功率；$H_{\mathrm{EB}}^{i,t-1}$、$H_{\mathrm{GB}}^{i,t-1}$ 分别为 t−1 时刻节点 i 处 EB 机组、GB 机组的热功率。

进一步考虑源荷不确定性以增强模型对环境变化的应对能力，本文在光伏与电热负荷历史数据上叠加随机性作为系统状态输入，其概率密度函数表示为

$$f(N_t) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{N}}}\exp\left[-\frac{(N_t - \mu_{\mathrm{N}})^2}{2\sigma_{\mathrm{N}}^2}\right] \tag{16}$$

式中：N_t 为电、热负荷及光伏功率的波动偏差；μ_{N} 和 σ_{N} 分别为电、热负荷及光伏功率波动偏差的期望和标准差。

2) 动作空间设计

调度模型中决策变量[13]包括 CHP 出力增量 $\Delta P_{\mathrm{CHP}}^{i,t}$ 和 EB 出力增量 $\Delta H_{\mathrm{EB}}^{i,t}$、当前时刻 BES 的充放电功率和 PV 的无功功率，上级电网考虑为平衡节点，剩余决策变量可通过推导获得。因此，动作空间 $\boldsymbol{a}_t$ 如式(17)所示。

$$\boldsymbol{a}_t = [\Delta P_{\mathrm{CHP}}^{i,t}, \Delta H_{\mathrm{EB}}^{i,t}, P_{\mathrm{BES}}^{i,t}, Q_{\mathrm{PV}}^{i,t}] \tag{17}$$

通过动作空间决策的机组出力增量和状态空间的前一时刻机组出力，可以得到当前时刻的机组出力功率，具体如式(18)—式(21)所示。

$$\Delta P_{\mathrm{CHP}}^{i,t}=\rho_{\mathrm{CHP}}^{i,t}R_{\mathrm{CHP}} \tag{18}$$

$$\Delta H_{\mathrm{EB}}^{i,t}=\rho_{\mathrm{EB}}^{i,t}R_{\mathrm{EB}} \tag{19}$$

$$P_{\mathrm{CHP}}^{i,t}=P_{\mathrm{CHP}}^{i,t-1}+\Delta P_{\mathrm{CHP}}^{i,t} \tag{20}$$

$$H_{\mathrm{EB}}^{i,t}=H_{\mathrm{EB}}^{i,t-1}+\Delta H_{\mathrm{EB}}^{i,t} \tag{21}$$

式中：$\rho_{\mathrm{CHP}}^{i,t}$、$\rho_{\mathrm{EB}}^{i,t}$为调度智能体执行的动作系数，$\rho_{\mathrm{CHP}}^{i,t}$、$\rho_{\mathrm{EB}}^{i,t}\in[-1,1]$；$R_{\mathrm{CHP}}$和$R_{\mathrm{EB}}$分别为 CHP 和 EB 的爬坡率。

3) 奖励函数设计

IES 能源调度旨在降低运行成本，因此将 t 时段调度成本的最小化问题转化为奖励函数的最大化形式，如式(22)所示。

$$r(\boldsymbol{s}_t,\boldsymbol{a}_t)=-\xi_0(F_{\mathrm{NG}}(t)+F_{\mathrm{grid}}(t)+F_{\mathrm{BES}}(t)) \tag{22}$$

式中：$r(\boldsymbol{s}_t,\boldsymbol{a}_t)$为即时奖励；$\xi_0$为缩放系数。

4) 动作价值函数及最优策略设计

强化学习旨在智能体与环境随机交互，学习最大化总期望折扣奖励$Q(\pi)$的最优策略π'，具体表示为

$$\pi'=\arg\max Q(\pi) \tag{23}$$

$$Q(\pi)=E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{s}_t,\boldsymbol{a}_t)\right] \tag{24}$$

式中：$E_{\pi}[\cdot]$为策略π下的期望函数；γ为折扣因子。

2.2.2 嵌入安全约束的 CMDP 模型

在上述无约束 MDP 中，智能体策略学习往往以目标奖励与安全约束的加权形式开展。约束权重过大可能过度惩罚导致区域 IES 经济性能下降，而权重过小则可能违反相关约束进而引发安全问题[22]，这与 IES 实际调度需求存在偏差。

因此，为避免人工调节权重引起主观偏差并保证系统安全合理运行，本节在 MDP 的基础上集成相应安全约束条件进而构成 CMDP。CMDP 的结构化设计可考虑多种约束，将安全性条件纳入决策过程，有效缩小探索空间[23]，进而更准确地把握系统安全运行边界，满足其安全稳定运行需求。考虑到网络潮流约束是系统运行可靠性的重要支撑[7]，而机组的合理运行也是 IES 提质增效的先决条件。本文将上述两个约束条件作为安全指标，进而构建 CMDP。具体来说，首先定义两个安全限制函数，分别为 t 时刻电压偏移安全限制函数$c_1(\boldsymbol{s}_t,\boldsymbol{a}_t)$和机组安全限制函数$c_2(\boldsymbol{s}_t,\boldsymbol{a}_t)$，如式(25)和式(26)所示。

$$c_1(\boldsymbol{s}_t,\boldsymbol{a}_t)=\sum_{i=1}^{N}\xi_1\left|U^{i,t}-U_i^*\right| \tag{25}$$

$$c_2(\boldsymbol{s}_t,\boldsymbol{a}_t)=\sum\nolimits_x\xi_2\left(\left|P_x^{i,t}-P_x^{\max}\right|^{+}+\left|P_x^{\min}-P_x^{i,t}\right|^{+}\right) \tag{26}$$

式中：ξ_1和ξ_2为安全限制缩放系数；U_i^*为基准电压，取值为 1.0 p.u.；算子$|a|^{+}=\max(0,a)$。

定义上述安全限制函数后，在当前策略π下未来的折扣安全成本$C_k(\pi)$和需要满足的安全约束如式(27)和式(28)所示。

$$C_k(\pi)=E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t c_k(\boldsymbol{s}_t,\boldsymbol{a}_t)\right] \tag{27}$$

$$C_k(\pi)\leqslant d_k,\forall k\in[1,2] \tag{28}$$

式中，d_k为约束容差。

结合 MDP 与安全限制函数进而得到基于 CMDP 的优化模型，具体描述如式(29)所示。

$$\begin{aligned}&Q(\pi)=E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{s}_t,\boldsymbol{a}_t)\right]\\&\text{s.t. }C_k(\pi)\leqslant d_k,\forall k\in[1,2]\end{aligned} \tag{29}$$

3　面向 IES 能源调度的约束强化学习算法

在第 1 节框架之下，本节提出一种约束强化学习算法以高效求解上述 CMDP 模型，即专家经验(以模仿学习嵌入)引导的原始对偶双延迟深度确定性策略梯度(primal-dual twin delayed deep deterministic policy gradient, PD-TD3)算法。具体来说，构建 PD-TD3 引导智能体安全训练，并利用模仿学习辅助其快速趋优。

3.1 PD-TD3 算法

TD3 在面对复杂波动性和不确定性环境时具有较好的算法稳定性[24]，因此，本文延用 TD3 算法架构。为进一步保证调度安全，受原始对偶优化[25]和数模融合策略[26]的启发，本文构建了 PD-TD3 算法。该算法在 TD3 基础上增设两个安全价值网络(为加以区分，本文将 TD3 中原有的价值网络称为经济价值网络)以评价机组与电压安全性能。通过引入拉格朗日乘子将约束优化问题转化为无约束问题，并利用原始对偶思想对拉格朗日乘子与调度策略自适应更新以确保调度决策安全，实现算法的有效收敛。原始对偶优化问题建立和参数更新方式如下。

3.1.1 原始对偶优化问题建立

基于 CMDP 的 PD-TD3 算法的优化目标如式(30)所示。

$$\begin{aligned}&\pi^*=\arg\max Q(\pi)\\&\text{s.t. }C_k(\pi)\leqslant d_k,\forall k\in[1,2]\end{aligned} \tag{30}$$

式中，π^*为最优策略。

引入拉格朗日乘子将该过程的安全约束内嵌至

策略网络，进而将式(30)转换为无约束极小-极大问题，具体如式(31)和式(32)所示。

$$\pi^* = \underset{\lambda_k \geqslant 0}{\operatorname{argmin}} \max L(\pi, \lambda_k) \tag{31}$$

$$L(\pi, \lambda_k) = Q(\pi) - \sum\nolimits_k \lambda_k [C_k(\pi) - d_k],\ \forall k \in [1,2] \tag{32}$$

式中，λ_k 为拉格朗日乘子。

3.1.2 参数更新

针对上述优化目标，本文利用原始对偶思想交替更新拉格朗日乘子与策略网络参数，确保在追求经济效益的同时满足安全性能。

1) 策略网络原始对偶更新过程

当智能体输出策略严格满足物理约束时，式(30)与式(31)呈强对偶，式(30)与式(31)将会同时达到最优[27]。策略的寻优过程可视为策略网络参数与拉格朗日乘子的交替更新过程，其中拉格朗日乘子 λ_k 的更新方式如式(33)所示。

$$\lambda_k \leftarrow \Gamma_k' \left\{ \lambda_k + \Gamma_k [\eta_k (C_k(\pi) - d_k)] \right\}^+ \tag{33}$$

式中：η_k 为拉格朗日乘子的学习率；Γ_k、Γ_k' 为乘子系数，避免乘子数值过大或过小对决策结果产生影响。

策略网络参数 θ^π 的更新方式如式(34)所示。

$$\theta^\pi \leftarrow \theta^\pi + l^\pi \nabla_{\theta^\pi} L(\pi, \lambda_k) \tag{34}$$

式中：l^π 为策略网络的学习率；∇ 为梯度计算函数。

策略的寻优过程如上所示，λ_k 与 θ^π 交替迭代更新，直至智能体停止更新调度策略。

2) 价值网络更新过程

神经网络设计方面，本文价值网络均采用前馈神经网络架构。受篇幅限制本文不做赘述，具体更新过程见文献[24]。

3.2 模仿学习

为加速算法收敛进程，缩小智能体初始探索空间至最优策略邻域以避免盲目探索造成算力浪费，调度智能体利用模仿学习算法学习专家经验[11]。具体来说，智能体通过模仿专家轨迹数据集进行初始策略优化，其策略通过式(35)和式(36)进行更新。

$$L^{\mathrm{IL}}(\theta^\mu) = \frac{1}{M} \sum_{m=1}^{M} \left[\boldsymbol{a}_m^{\mathrm{e}} - \boldsymbol{a}(\boldsymbol{s}_m \left| \theta^\mu \right.) \right]^2 \tag{35}$$

$$\theta^\mu \leftarrow \theta^\mu - \alpha_\mu \nabla_{\theta^\mu} L^{\mathrm{IL}}(\theta^\mu) \tag{36}$$

式中：$L^{\mathrm{IL}}(\cdot)$ 为模仿学习损失函数；θ^μ 为预训练过程中策略网络参数；M 为动作个数；$\boldsymbol{a}_m^{\mathrm{e}}$ 为状态 $\boldsymbol{s}_m$ 对应的专家动作；$\boldsymbol{a}(\boldsymbol{s}_m \left| \theta^\mu \right.)$ 为模仿专家的智能体动作；α_μ 为模仿学习学习率。智能体模仿学习的具体流程见3.3.1 节预训练部分。

3.3 算法求解流程

所提算法包含价值网络(分为经济价值网络、机组安全价值网络和电压安全价值网络)、策略网络、目标策略网络和目标价值网络。TD3 在深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法基础上采用截断双Q学习以更好地解决算法的高估问题[24]，故本文延用 TD3 算法的双网络结构(在图 2 中以符号#区分)，并采用延迟策略更新和目标策略噪声干扰手段，以增强算法的稳定性。算法采用经验回放技术[8]，提高样本利用效率。基于约束强化学习的 IES 优化调度方法的整体实现流程包括预训练、离线训练和在线应用三部分。

3.3.1 预训练+离线训练

为缩小智能体初始探索空间进而提高计算效率，首先采用模仿学习对策略网络和价值网络开展预训练，具体做法为：首先，基于历史数据利用模型驱动方法求解获取 IES 能源调度问题的专家经验并存储到经验池中；其次，利用式(35)和式(36)预训练策略网络，策略网络在状态 $\boldsymbol{s}_t$ 下执行动作 $\boldsymbol{a}_t$，得到 $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$、$c_1(\boldsymbol{s}_t, \boldsymbol{a}_t)$ 和 $c_2(\boldsymbol{s}_t, \boldsymbol{a}_t)$，并连同 $\boldsymbol{s}_{t+1}$ 形成序列元组存储至经验池；随后，价值网络利用该序列元组输出经济 Q 值和安全值，并对价值网络进行反向更新，采用延迟策略更新目标网络；最后，循环上述步骤，直至损失函数收敛，预训练过程结束。

经预训练后的调度智能体已初步具备快速学习 IES 调度策略的能力。为进一步提升 IES 调度过程的安全性和经济性，利用预训练后的能源调度模型开展离线训练，离线训练流程如图 2 所示，可分为 5 个步骤。

1) 智能体与环境交互，并根据当前系统观测状态 $\boldsymbol{s}_t$ 执行动作 $\boldsymbol{a}_t$，同时进行潮流计算，进而得到 $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$、$c_1(\boldsymbol{s}_t, \boldsymbol{a}_t)$ 和 $c_2(\boldsymbol{s}_t, \boldsymbol{a}_t)$，并将其作为经验元组存储到经验池中(但不覆盖专家经验)，循环步骤 1)，直至达到经验池存储最大经验数；

2) 将经验池存储的状态-动作对输入安全价值网络、经济价值网络及其相应目标网络，分别输出计算安全值、经济 Q 值及其相应目标值，以此进行价值网络权重更新；

3) 根据价值网络所得安全值更新拉格朗日乘子 λ_1 和 λ_2，同时更新策略梯度；

4) 软更新目标价值网络与目标策略网络，并清空经验池；

5) 循环步骤 1)—步骤 4)，直至达到最大训练回合数，保存神经网络参数和当前拉格朗日乘子 λ_1 和 λ_2，结束训练。

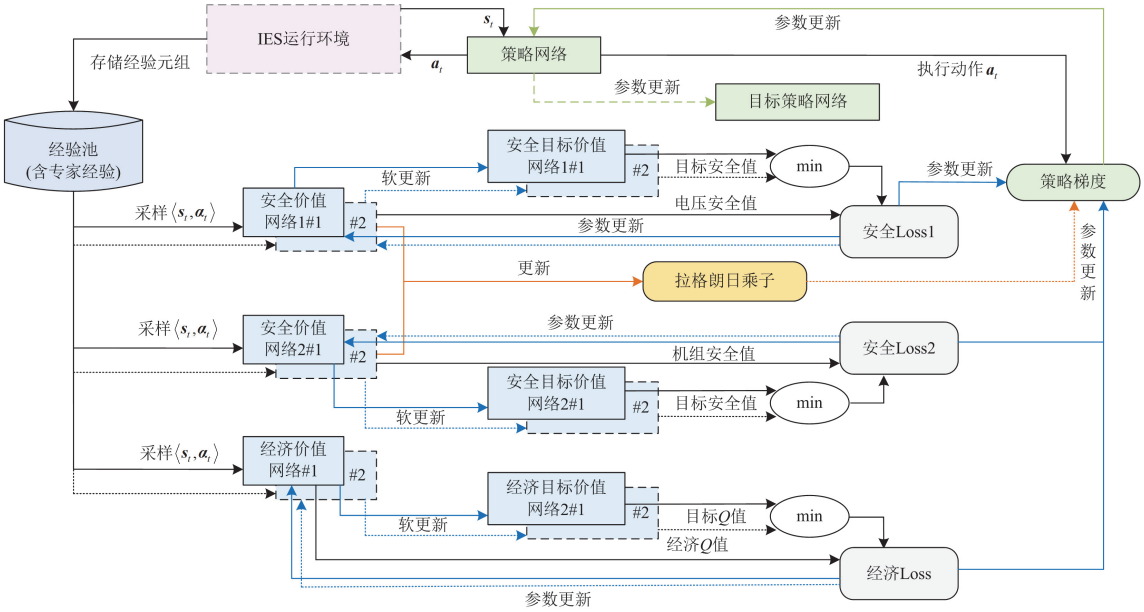


图 2　离线训练流程图

Fig. 2 Flow chart of offline training

3.3.2 在线应用

训练并验证完毕后，神经网络参数将保持不变，训练好的调度模型将被部署到实际 IES 中。在在线应用阶段，IES 仅需已训练完成的策略网络执行最优动作。具体过程可简述为：首先，系统持续监控包括运行状况、能源消耗及外部环境条件等的实时状态 $\boldsymbol{s}_t$，并由调度中心执行调度动作 $\boldsymbol{a}_t$；其次，执行动作 $\boldsymbol{a}_t$ 后系统在下一时刻转移至状态 $\boldsymbol{s}_{t+1}$；最后，系统在 t+1 时刻执行动作同时得到奖励值 $r(\boldsymbol{s}_t,\boldsymbol{a}_t)$ 和安全值 $c_1(\boldsymbol{s}_t,\boldsymbol{a}_t)$、$c_2(\boldsymbol{s}_t,\boldsymbol{a}_t)$。以此类推，直至完成全天 T 个调度时刻的决策信息。

4　算例分析

4.1　算例设置

本文采用附录 A 图 A2 所示的 IES 系统作为仿真算例，其网络结构基于文献[28]修改。电力系统采用标幺值计算，在节点 1 处接入主电网，基准容量为 100 MVA，基准电压为 23 kV，其分时电价如附录 C 表 C1 所示。热力系统供水温度恒定为 100 ℃，环境温度恒定为 10 ℃。系统参数基于文献[21]修改，如附录 C 表 C2、表 C3 所示。天然气购买单价为 4.0 元/m^3，热值为 9.98 kW/m^3。算法中网络架构如 3.3 节，并基于文献[25]调节算法超参数，神经网络激活函数为 ReLu，隐藏层神经元个数为(128, 64, 64)，其他参数如附录 C 表 C4 所示。

4.2　离线训练结果分析

采用所提算法对 IES 能源调度模型进行离线训练，分别从算法安全越限情况和拉格朗日乘子变化两方面展开分析。

4.2.1 安全越限情况分析

为展现本文所提算法的安全与收敛性能，电压越限和机组越限情况如图 3 所示。

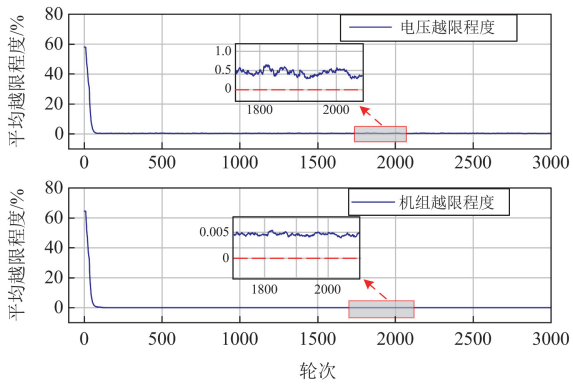


图 3　越限程度曲线

Fig. 3 Overrun degree curves

由图 3 可以看出：训练初期，智能体主要依靠模仿专家策略执行调度计划，且其与系统环境交互经验尚少，未形成对 IES 系统运行安全边界的准确认知，因此系统的早期决策行为造成机组出力越限和电压越限现象较为明显。随后，在专家经验指引下，智能体与环境持续动态交互，其可根据越限信息判断输出动作是否可行，进而快速捕捉系统安全运行特性。经过 125 轮，电压越限程度由最初的 58.12%下降至 0.48%，机组出力越限程度由 64.49%

下降至 0.0049%，并分别趋于稳定。可以看出，经过本文算法训练后，系统在经济效益最大化的同时也基本保证了电压稳定及机组出力安全合理。

4.2.2 拉格朗日乘子变化分析

为展现拉格朗日乘子在智能体训练过程中起到安全限制作用，拉格朗日乘子变化曲线如图 4 所示。

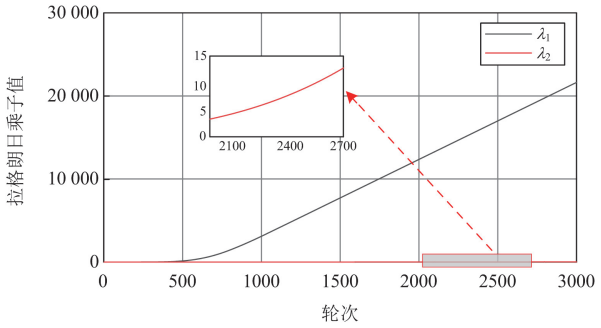


图 4 拉格朗日乘子变化曲线

Fig. 4 Change curves of Lagrange multipliers

由图 4 可以看出：拉格朗日乘子在训练过程中总体呈增长趋势，与安全越限程度呈负相关，表明其能够自适应更新并合理引导智能体在训练过程中学习安全策略。另外，受安全限制函数设计、算法超参数与缩放系数等因素影响，两个拉格朗日乘子的增长趋势也存在一定差异。随着训练进行，拉格朗日乘子值逐步增加，即使是很小的安全越限情况也可加以避免。

4. 3 在线应用仿真结果分析

离线训练结束后，其策略网络可用于 IES 能源调度问题实时决策。本阶段将从电、热功率调度结果两方面展开分析。

4.3.1 电功率调度结果分析

本文所提方案应用在 IES 能源调度问题中的电功率调度优化结果如图 5 所示。

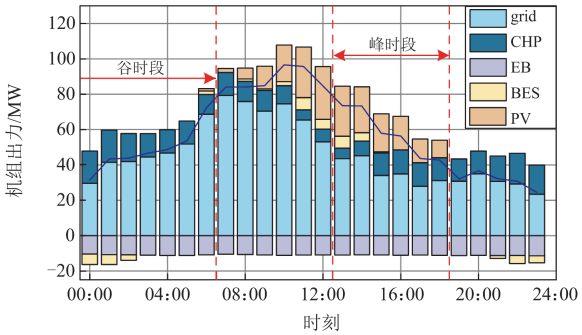


图 5 电功率调度结果

Fig. 5 Dispatch results of electric power

由图 5 可以看出：系统与上级电网的交互功率随电价变化而调整，并且在各调度时刻保证了电能供需平衡。13：00—18：00 期间为电价峰时段，系统尽可能减少向上级电网购电，同时增加电储能装置的放电量。10：00—19：00 受光照影响 PV 发电量明显增多，但不能完全平衡电负荷需求，因而需要 CHP 机组和主电网持续供电，储能装置也在电价较高时段放电以保证用电平衡。而 00：00—12：00 这一电价相对较低时段，系统明显增大向电网购电量，07：00 时刻向上级电网购电量达 79.33 MW，占此时刻电负荷需求的 94.33%。

4.3.2 热功率调度结果分析

本文所提方案应用在 IES 能源调度问题中的热功率调度优化结果如图 6 所示。

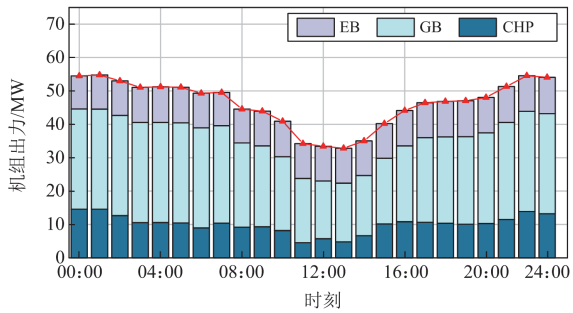


图 6 热功率调度结果

Fig. 6 Dispatch results of heat power

由图 6 可以看出：该系统受电价波动和网络安全约束影响对机组的热功率输出进行了相应调整，在各个调度时刻可有效满足热负荷需求。在 16：00—24：00 热负荷需求高峰时段，在网络安全约束的影响下，EB 和 CHP 机组的出力变化限制在一定范围内，主要热量由 GB 提供。如 14：00 时刻热负荷需求较低，GB 出力为 18.06 MW，仅占这一时刻热负荷需求的 40.71%；而随后热负荷显著增加，到 23：00 时 GB 出力为 30 MW，占这一时刻热负荷需求的 55.48%。为保证系统潮流在安全范围内，CHP 和 EB 出力增加并不明显，GB 出力明显增多以保证系统安全稳定。

4. 4 不同安全调度方法结果对比分析

为进一步验证本文所提优化策略的可行性与有效性，设计了 3 种优化方案进行综合对比分析。方案 1：TD3；方案 2：CPO；方案 3：本文所提方法。

4.4.1 奖励函数收敛性对比分析

为验证本文所提方法在 IES 能源调度问题中的训练效果，本文针对以上 3 种不同方案的奖励函数收敛性能展开对比分析，收敛曲线如图 7 所示。

由图 7 可以看出：本文算法在训练前期奖励值较低，这是由于智能体在某些超出专家策略范围的场景中需要通过探索性试错得到决策动作。随着训

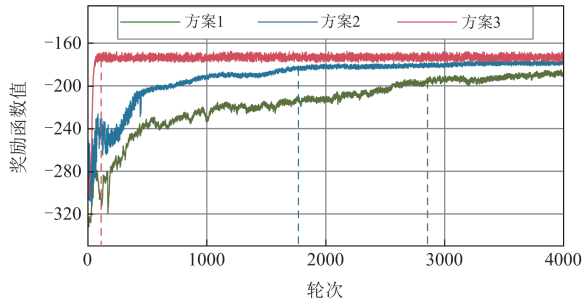


图 7　3 种方案下的奖励函数值

Fig. 7 Value of the reward function in three scenarios

练次数的增加，本文方案相较于方案 1 和方案 2 收敛速度更快、收敛曲线更加平滑。最终，3 种方案的奖励值分别于 2860 轮(方案 1)、1780 轮(方案 2)和 125 轮(本文方案)训练后趋于稳定，本文方案奖励函数的收敛速度相较于方案 1 和方案 2 分别提升了约 95.63%和 92.98%。在训练后期，本文方案奖励函数收敛值最高。相比之下，方案 1 在安全性与经济性之间可能存在权衡处理不当，方案 2 则因对 CMDP 进行简化处理而导致调度决策性能下降。这一结果表明，在 IES 能源调度问题中，拉格朗日乘子引入与专家经验引导可提升算法的收敛效果与收敛速度。

4.4.2 经济效益分析

本节将测试数据应用到上述 3 种方案训练完成的模型中进行在线性能验证。园区一天内不同方案的机组出力运行成本如表 2 所示。

表 2　不同方案下的优化结果对比

Table 2 Comparison of optimization results under different schemes

元

方案	购电成本	CHP 成本	GB 成本	BES 成本	运行成本
方案 1	579 521.65	640 255.92	191 626.22	3.15	1 411 406.94
方案 2	686 698.44	486 812.67	252 704.46	2.95	1 426 218.52
方案 3	796 110.28	346 602.74	291 885.46	2.80	1 434 601.28

由表 2 可以看出：在全面考虑安全约束的背景下，方案 3 相较于方案 1 和方案 2 呈现出更高的整体运行成本。具体来看，方案 3 相较于方案 1 和方案 2，购电成本分别增加了 37.37%和 15.93%，CHP 成本分别减少了 45.86%和 28.80%。这表明，由于网络安全约束限制，CHP 机组出力受到制约，而相应电、热负荷功率分别由主电网和 GB 提供，从而确保电能和热能需求平衡。

总体而言，方案 3 的运行策略体现了一个更加平衡的视角，通过精细的能源管理策略来提升经济效益，同时保障系统运行的可靠性和安全性。

4.4.3 安全性能分析

对 3 种优化方案的电压稳定性从全局和局部分别展开分析。考虑到 CHP 机组同时供给电、热两种负荷需求，是电热耦合系统中的关键设备，因此以节点 6 展开局部分析。全局优化结果对比如表 3 所示，局部电压对比结果如图 8 所示。

由表 3 和图 8 可以看出：总体来说，方案 1 和方案 2 都存在电压越限的情况，而方案 3 没有出现

表 3　不同方案下的全局优化结果对比

Table 3 Comparison of global optimization results under different schemes

方案	越限时段数/个	平均电压偏差/p.u.	最大电压偏差/p.u.
方案 1	6	0.047	0.101
方案 2	3	0.026	0.066
方案 3	0	0.015	0.045

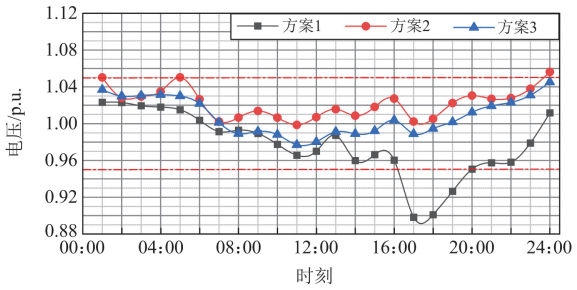


图 8　节点 6 电压水平对比

Fig. 8 Comparison of voltage level of node 6

电压越限，且最大电压偏差值最小；方案 3 的平均电压偏差仅为 0.015 p.u.，明显低于方案 1 和方案 2，这表明方案 3 在整个优化时段内保持了稳定的电压水平。从局部分析，方案 1 与方案 2 的 CHP 机组所在节点因电热耦合，机组出力调节不当，进而导致越限时段数较多且电压偏差较大。相对而言，方案 3 受安全约束制约，其节点 6 机组输出波动受到较大限制，进而可维持该节点电压稳定。

总体来说，上述指标反映出方案 3 在确保系统电压安全稳定方面具有显著优势，对于保障设备和用户的安全至关重要。

4.5　预训练前后结果对比分析

4.5.1 奖励函数对比分析

为证明本文引入模仿学习的高效性，本文将未经模仿学习预训练的 PD-TD3 算法与本文算法两种

场景的奖励函数进行对比分析，结果如图 9 所示。

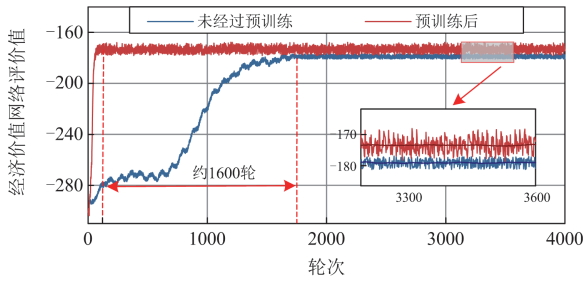


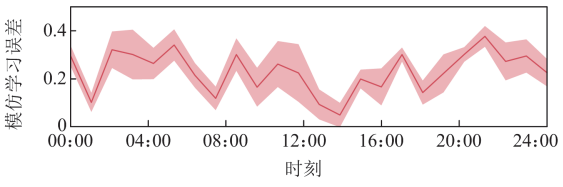
图 9 经济价值网络评价值收敛曲线

Fig. 9 Convergence curves of the economic value network evaluation value

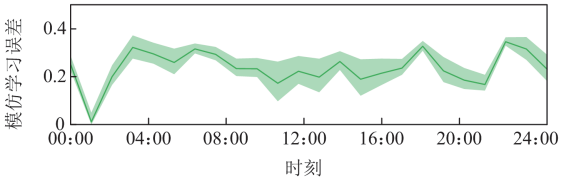
由图 9 可以看出：训练初期，未经预训练算法的收敛曲线上升缓慢，说明此时仅依靠智能体与环境交互探索学习最优策略，导致训练初期探索空间较大，收敛缓慢；本文算法在训练初期与系统环境交互，其收敛曲线呈现振荡趋势，但在专家经验的引导下可快速趋于收敛。最终，两种算法分别在 1760 轮和 125 轮后趋于稳定。这表明专家经验加快了智能体学习进程，进而提升了算法的收敛速度。

4.5.2 误差分析

为进一步检验预训练方法的有效性，从测试集数据中随机抽取 10 个测试日进行算法对比验证，结果如图 10 所示。所抽取的测试日中，对电热耦合设备 CHP 和 EB 机组利用模仿学习误差[29](如式(37))量化算法结果与专家经验解之间的差距。



(a) CHP机组功率



(b) EB机组功率

图 10 模仿学习误差曲线

Fig. 10 Curves of imitation learning error

$$R=\sqrt{\frac{1}{n}\sum_{l=1}^{n}[(x_l-y_l)/y_l]^2} \tag{37}$$

式中：R 为模仿学习误差；n 为动作空间维度；x_l 为本文算法得到的机组出力值；y_l 为对应期望值，即专家经验解。

由图 10 可以看出，PD-TD3 给出的解与专家经验解略有差异，测试过程全天调度的模仿学习误差保持在 0.4 以下。这说明在所提方法中，专家经验初始化策略在有效引导智能体动作的同时，智能体能够进行一定程度的自主探索。

4.6 与经典 DRL 对比分析

为进一步验证本文方法的有效性，将所提方法与近端策略优化(proximal policy optimization, PPO)算法和 DDPG 进行比较。算法收敛后的电压越限情况如图 11 所示，优化结果对比如表 4 所示。

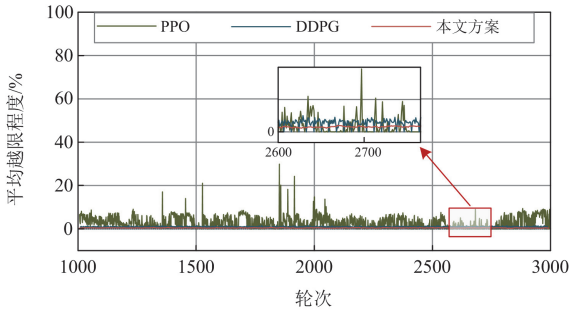


图 11 平均越限程度变化曲线

Fig. 11 Change curves of average overrun degree

表 4 不同方法的优化结果对比

Table 4 Comparison of optimization results with different methods

方法	在线决策时间/s	平均电压偏差/p.u.	运行总成本/元
DDPG(奖励罚函数法)	0.267	0.048	1 407 202.39
PPO(奖励罚函数法)	0.316	0.039	1 423 654.21
本文方法	0.135	0.015	1 434 601.28

由表 4 可以看出，本文所提方法求解速度快、电压偏移量小，总运行成本相比两种经典强化学习算法分别略有增加(分别为 0.019%和 0.007%)。由图 11 可以看出，本文算法经训练已收敛后电压越限程度更小，收敛曲线更加平稳。总体而言，本文所提算法更能保证电力系统的稳定性与安全性，符合文献[30]的经济-安全-环境的矛盾三角观点。

5 结论

在日益复杂的 IES 能源调度问题中，传统 DRL 对安全约束考虑不足且收敛能力较差。为解决该问题，本文将拉格朗日法与传统 DRL 算法结合，提出了一种约束强化学习优化调度方法。主要结论如下：

1) 所提方法将 IES 调度问题构建为约束马尔可

夫决策过程，通过拉格朗日乘子将约束优化问题转换为无约束极小-极大问题，借助原始对偶思想自适应更新训练，避免了传统 DRL 手动调整惩罚系数引起的主观偏差，实现了智能体在安全边界内自适应寻优。

2) 所提算法包括经济和安全两种价值网络，弥补了单一价值网络拟合能力的不足。在保持良好经济性能的基础上，所提算法与传统 DRL 相比具有更高的训练效率。

3) 调度智能体通过模仿专家经验，将初始决策空间收敛至最优策略邻域，大幅缩短了学习过程。相比于传统 TD3 和 CPO 算法，所提方法训练速度分别提升了 95.63%和 92.98%。

随着 IES 内部环境日益复杂，本文方法可保障能源调度的安全性与训练效率，为同类问题提供了高效、安全的解决方案。

附录 A

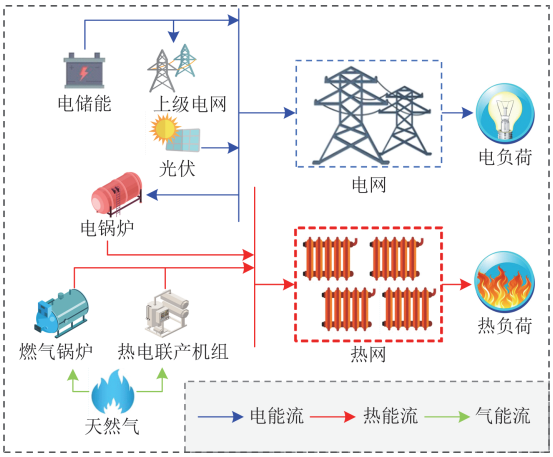


图 A1 IES 供能结构

Fig. A1 Energy supply structure of IES

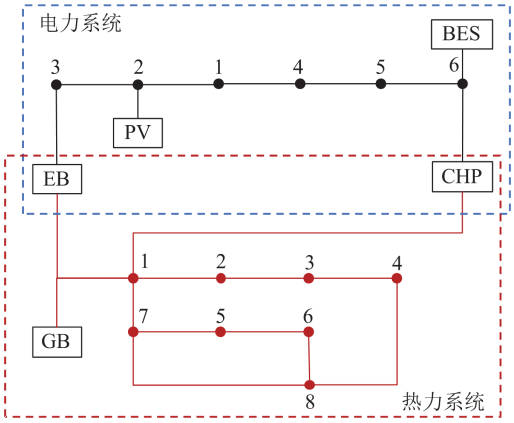


图 A2 IES 拓扑图

Fig. A2 Topological diagram of IES

附录 B

1 机组运行特性及约束条件

1.1 CHP运行约束

采用定热电比的背压式CHP机组，其运行特性与约束条件为

$$H_{\mathrm{CHP}}^{i,t}=b_{\mathrm{CHP}}P_{\mathrm{CHP}}^{i,t} \tag{B1}$$

$$P_{\mathrm{CHP}}^{i,t}=M_{\mathrm{CHP}}^{i,t}\eta_{\mathrm{CHP}}H_{\mathrm{NG}} \tag{B2}$$

$$P_{\mathrm{CHP}}^{\min}\leqslant P_{\mathrm{CHP}}^{i,t}\leqslant P_{\mathrm{CHP}}^{\max} \tag{B3}$$

$$-R_{\mathrm{CHP}}\leqslant P_{\mathrm{CHP}}^{i,t}-P_{\mathrm{CHP}}^{i,t-1}\leqslant R_{\mathrm{CHP}} \tag{B4}$$

式中：$H_{\mathrm{CHP}}^{i,t}$ 为t时刻节点i处CHP输出的热功率；b_{CHP} 为CHP热电比；η_{CHP} 为CHP的电转化效率；H_{NG} 为天然气热值；$P_{\mathrm{CHP}}^{\max}$ 和 $P_{\mathrm{CHP}}^{\min}$ 分别为CHP输出有功功率的上、下限；R_{CHP} 为CHP的爬坡率。

1.2 BES运行约束

BES的运行特性与荷电状态(state of charge, SOC)可表示为

$$C_{\mathrm{SOC}}^{i,t}=C_{\mathrm{SOC}}^{i,t-1}-\frac{\eta_{\mathrm{BES}}P_{\mathrm{BES}}^{i,t}\Delta t}{Q_{\mathrm{BES}}} \tag{B5}$$

$$C_{\mathrm{SOC}}^{\min}\leqslant C_{\mathrm{SOC}}^{i,t}\leqslant C_{\mathrm{SOC}}^{\max} \tag{B6}$$

$$\eta_{\mathrm{BES}}=\begin{cases}\eta_{\mathrm{ch}} & P_{\mathrm{BES}}^{i,t}<0\\ \dfrac{1}{\eta_{\mathrm{dis}}} & P_{\mathrm{BES}}^{i,t}\geqslant 0\end{cases} \tag{B7}$$

式中：$C_{\mathrm{SOC}}^{i,t}$、$C_{\mathrm{SOC}}^{i,t-1}$ 分别为t、$t-1$时刻节点i处BES的荷电状态；η_{BES} 为充放电系数；Δt 为一个调度时段，数值为1 h；Q_{BES} 为BES的容量；$C_{\mathrm{SOC}}^{\max}$ 和 $C_{\mathrm{SOC}}^{\max}$ 分别为BES荷电状态的上、下限；η_{ch} 和 η_{dis} 分别为BES的充、放电效率系数。

1.3 PV运行约束

PV的运行约束如式(B8)所示。

$$(P_{\mathrm{PV}}^{i,t})^2+(Q_{\mathrm{PV}}^{i,t})^2\leqslant(S_{\mathrm{PV}})^2 \tag{B8}$$

式中，S_{PV} 为PV的视在功率。

1.4 GB运行约束

GB运行可靠性高[21]，可消耗天然气产生热功率进而维持功率平衡，运行约束为

$$H_{\mathrm{GB}}^{i,t}=M_{\mathrm{GB}}^{i,t}\eta_{\mathrm{GB}}H_{\mathrm{NG}} \tag{B9}$$

$$H_{\mathrm{GB}}^{\min}\leqslant H_{\mathrm{GB}}^{i,t}\leqslant H_{\mathrm{GB}}^{\max} \tag{B10}$$

式中：$H_{\mathrm{GB}}^{i,t}$ 为 t 时刻节点 i 处 GB 输出的热功率；η_{GB} 为 GB 的热转化效率；$H_{\mathrm{GB}}^{\max}$ 和 $H_{\mathrm{GB}}^{\min}$ 分别为 GB 输出热功率上、下限。

1.5 EB 运行约束

EB将电能转化为热能，其运行约束为

$$H_{\mathrm{EB}}^{i,t}=P_{\mathrm{EB}}^{i,t}\eta_{\mathrm{EB}} \tag{B11}$$

$$H_{\mathrm{EB}}^{\min}\leqslant H_{\mathrm{EB}}^{i,t}\leqslant H_{\mathrm{EB}}^{\max} \tag{B12}$$

$$-R_{\mathrm{EB}}\leqslant H_{\mathrm{EB}}^{i,t}-H_{\mathrm{EB}}^{i,t-1}\leqslant R_{\mathrm{EB}} \tag{B13}$$

式中：$H_{\mathrm{EB}}^{i,t}$、$H_{\mathrm{EB}}^{i,t-1}$ 分别为 t、$t-1$时刻节点 i 处 EB 输出的热功率；η_{EB} 为 EB 的电热转化效率；$H_{\mathrm{EB}}^{\max}$ 和 $H_{\mathrm{EB}}^{\min}$ 分别为 EB 输出热功率的上、下限；R_{EB} 为 EB 的爬坡率。

附录 C

表 C1 分时电价

Table C1 Time-of-use electric price

时段	购电电价/(元/kWh)	售电电价/(元/kWh)
谷时段(23:00—07:00)	0.15	0.05
平时段(07:00—12:00, 19:00—23:00)	0.8	0.2
峰时段(12:00—19:00)	1.3	0.5

表 C2 设备运行参数

Table C2 Operational parameters of device

设备类型	功率下限/MW	功率上限/MW	下爬坡限制/(MW/h)	上爬坡限制/(MW/h)
CHP	0	30	−12	12
GB	0	30	−10	10
EB	0	50	−10	10

表 C3 其他参数

Table C3 Other parameters

参数	取值	参数	取值
$\eta_{\mathrm{CHP_p}}$	0.35	$C_{\mathrm{SOC}}^{\min}$	0.2
φ_{CHP}	0.9	$C_{\mathrm{SOC}}^{\max}$	0.8
b_{CHP}	0.8	η_{ch}	0.95
η_{GB}	0.85	η_{dis}	0.95
η_{EB}	0.95	ρ_{BES}	0.05

表 C4 神经网络其他参数

Table C4 Other parameters of neural network

参数	数值
探索率	0.2
软更新系数	0.05
折现率	0.98
批量大小	256
总训练回合数	5000

参考文献

[1] 王蓬蓬, 宋运忠. 计及最恶劣场景概率和供需灵活性的综合能源系统分布鲁棒低碳优化调度[J]. 电力系统保护与控制, 2024, 52(13): 78-89.
WANG Pengpeng, SONG Yunzhong. Distributed robust low-carbon optimal scheduling of an integrated energy system considering worst-case scenario probability and flexibility of supply and demand[J]. Power System Protection and Control, 2024, 52(13): 78-89.

[2] 李欣, 陈英彰, 李涵文, 等. 考虑碳交易的电-热综合能源系统两阶段鲁棒优化低碳经济调度[J]. 电力建设, 2024, 45(6): 58-69.
LI Xin, CHEN Yingzhang, LI Hanwen, et al. Two-stage robust optimization of low-carbon economic dispatch for electricity-thermal integrated energy system considering carbon trade[J]. Electric Power Construction, 2024, 45(6): 58-69.

[3] 冯斌, 胡轶婕, 黄刚, 等. 基于深度强化学习的新型电力系统调度优化方法综述[J]. 电力系统自动化, 2023, 47(17): 187-199.
FENG Bin, HU Yijie, HUANG Gang, et al. Review on optimization methods for new power system dispatch based on deep reinforcement learning[J]. Automation of Electric Power Systems, 2023, 47(17): 187-199.

[4] CHEN Maozhi, LU Hao, CHANG Xiqiang, et al. An optimization on an integrated energy system of combined heat and power, carbon capture system and power to gas by considering flexible load[J]. Energy, 2023, 273.

[5] XIONG Guojiang, SHUAI Maohang, HU Xiao. Combined heat and power economic emission dispatch using improved bare-bone multi-objective particle swarm optimization[J]. Energy, 2023, 244.

[6] 徐业琰, 姚良忠, 廖思阳, 等. 基于多智能体 Actor-Double-Critic 深度强化学习的源－网－荷－储实时优化调度方法研究[J/OL]. 中国电机工程学报: 1-15[2023-08-15]. https://doi.org/10.13334/j.0258-8013.pcsee.231054.
XU Yeyan, YAO Liangzhong, LIAO Siyang, et al. Studies on real-time optimal dispatch method of source-grid-load-storage based on multi-agent Actor-Double-Critic deep reinforcement learning[J/OL]. Proceedings of the CSEE: 1-15[2023-08-15]. https://doi.org/10.13334/j.0258-8013.pcsee.231054.

[7] 杨挺, 刘豪, 王静, 等. 基于深度强化学习的园区综合能源系统低碳经济调度[J/OL]. 电网技术: 1-11[2023-11-02]. https://doi.org/10.13335/j.1000-3673.pst.2023.1555.
YANG Ting, LIU Hao, WANG Jing, et al. Deep reinforcement learning-based low-carbon economic dispatch of park integrated energy system[J/OL]. Power System Technology: 1-11[2023-11-02]. https://doi.org/10.13335/j.1000-3673.pst.2023.1555.

[8] 史一茹, 张大海, 李立新, 等. 基于生成对抗模仿学习的综合能源系统能量优化调度[J/OL]. 高电压技术: 1-

13[2023-08-04]. https://doi.org/10.13336/j.1003-6520.hve.20230537.
SHI Yiru, ZHANG Dahai, LI Lixin, et al. Optimal energy dispatch for integrated energy systems based on generative adversarial imitation learning[J/OL]. High Voltage Engineering: 1-13[2023-08-04]. https://doi.org/10.13336/j.1003-6520.hve.20230537.

[9] 李付强, 张文朝, 潘艳, 等. 基于改进深度确定性策略梯度算法的电压无功优化策略[J]. 智慧电力, 2024, 52(5): 1-7, 30.
LI Fuqiang, ZHANG Wenchao, PAN Yan, et al. Reactive voltage optimization strategy based on improved depth deterministic strategy gradient algorithm[J]. Smart Power, 2024, 52(5): 1-7, 30.

[10] 王力成, 邓宝华, 黄刚, 等. 知识–数据混合驱动的电网频率协同控制算法[J]. 中国电机工程学报, 2022, 42(23): 8523-8534.
WANG Licheng, DENG Baohua, HUANG Gang, et al. Coordinated system frequency control with a hybrid knowledge-data driven algorithm[J]. Proceedings of the CSEE, 2022, 42(23): 8523-8534.

[11] GARMROODI A D, NASIRI F, HAGHIGHAT F. Optimal dispatch of an energy hub with compressed air energy storage: a safe reinforcement learning approach[J]. Journal of Energy Storage, 2023, 57.

[12] 陈明昊, 孙毅, 胡亚杰, 等. 基于纵向联邦强化学习的居民社区综合能源系统协同训练与优化管理方法[J]. 中国电机工程学报, 2022, 42(15): 5535-5550.
CHEN Minghao, SUN Yi, HU Yajie, et al. The collaborative training and management-optimized method for residential integrated energy system based on vertical federated reinforcement learning[J]. Proceedings of the CSEE, 2022, 42(15): 5535-5550.

[13] SUN Qingkai, WANG Xiaojun, LIU Zhao, et al. Multi-agent energy management optimization for integrated energy systems under the energy and carbon co-trading market[J]. Applied Energy, 2022, 324.

[14] 蔺伟山, 王小君, 孙庆凯, 等. 不确定性环境下基于深度强化学习的综合能源系统动态调度[J]. 电力系统保护与控制, 2022, 50(18): 50-60.
LIN Weishan, WANG Xiaojun, SUN Qingkai, et al. Dynamic dispatch of an integrated energy system based on deep reinforcement learning in an uncertain environment[J]. Power System Protection and Control, 2022, 50(18): 50-60.

[15] 胡维昊, 曹迪, 黄琦, 等. 深度强化学习在配电网优化运行中的应用[J]. 电力系统自动化, 2023, 47(14): 174-191.
HU Weihao, CAO Di, HUANG Qi, et al. Application of deep reinforcement learning in optimal operation of distribution network[J]. Automation of Electric Power Systems, 2023, 47(14): 174-191.

[16] YU Peipei, ZHANG Hongcai, SONG Yonghua. District cooling system control for providing regulation services based on safe reinforcement learning with barrier functions[J]. Applied Energy, 2023, 347.

[17] ZHAO Rui, LI Yun, GAO Fei, et al. Multi-agent constrained policy optimization for conflict-free management of connected autonomous vehicles at unsignalized intersections[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(6): 5374-5388.

[18] SUN Qingkai, WANG Xiaojun, LIU Zhao, et al. Dynamic energy dispatch strategy for integrated energy system based on constrained reinforcement learning[J]. CSEE Journal of Power and Energy Systems, 2021, 36(5): 4848-4851.

[19] 李畸勇, 赵新哲, 郑一飞, 等. 基于纳什谈判考虑能源共享的区域综合能源系统优化配置[J]. 电力系统保护与控制, 2023, 51(5): 22-32.
LI Jiyong, ZHAO Xinzhe, ZHENG Yifei, et al. Optimal configuration of a regional integrated energy system considering energy sharing based on Nash negotiation[J]. Power System Protection and Control, 2023, 51(5): 22-32.

[20] 赵振宇, 任旭. 考虑动态能价及碳证交易的综合能源系统零碳优化[J]. 电力建设, 2024, 45(8): 36-50.
ZHAO Zhenyu, REN Xu. Zero-carbon optimization of integrated energy system considering dynamic energy prices and carbon certificate trading[J]. Electric Power Construction, 2024, 45(8): 36-50.

[21] 蔺伟山, 王小君, 孙庆凯, 等. 计及安全约束的综合能源系统深度强化学习优化调度策略研究[J]. 电网技术, 2023, 47(5): 1970-1983.
LIN Weishan, WANG Xiaojun, SUN Qingkai, et al. Optimal dispatch strategy of integrated energy system based on deep reinforcement learning considering security constraints[J]. Power System Technology, 2023, 47(5): 1970-1983.

[22] 焦昊, 殷岩岩, 吴晨, 等. 基于安全强化学习的主动配电网有功-无功协调优化调度[J]. 中国电力, 2024, 57(3): 43-50.
JIAO Hao, YIN Yanyan, WU Chen, et al. Coordinated optimization of active and reactive power of active distribution network based on safety reinforcement learning[J]. Electric Power, 2024, 57(3): 43-50.

[23] 张昌昕, 张兴龙, 徐昕, 等. 安全强化学习及其在机器人系统中的应用综述[J]. 控制理论与应用, 2023, 40(12): 2090-2103.

ZHANG Changxin, ZHANG Xinglong, XU Xin, et al. Safe reinforcement learning and its applications in robotics: a survey[J]. Control Theory & Applications, 2023, 40(12): 2090-2103.

[24] 顾雪平, 刘彤, 李少岩, 等. 基于改进双延迟深度确定性策略梯度算法的电网有功安全校正控制[J]. 电工技术学报, 2023, 38(8): 2162-2177.
GU Xueping, LIU Tong, LI Shaoyan, et al. Active power correction control of power grid based on improved twin delayed deep deterministic policy gradient algorithm[J]. Transactions of China Electrotechnical Society, 2023, 38(8): 2162-2177.

[25] 仪忠凯, 梁寿愚, 王伟, 等. 电力系统调度决策：一种示教学习辅助加速的安全强化学习方法[J]. 中国电机工程学报, 2024, 44(13): 5084-5097.
YI Zhongkai, LIANG Shouyu, WANG Wei, et al. Power system dispatch: an accelerated safe reinforcement learning approach by incorporating learning from demonstration[J]. Proceedings of the CSEE, 2024, 44(13): 5084-5097.

[26] 李峰, 王琦, 胡健雄, 等. 数据与知识联合驱动方法研究进展及其在电力系统中应用展望[J]. 中国电机工程学报, 2021, 41(13): 4377-4390.
LI Feng, WANG Qi, HU Jianxiong, et al. Combined data-driven and knowledge-driven methodology research advances and its applied prospect in power systems[J]. Proceedings of the CSEE, 2021, 41(13): 4377-4390.

[27] 董雷, 杨子民, 乔骥, 等. 基于分层约束强化学习的综合能源多微网系统优化调度[J]. 电工技术学报, 2024, 39(5): 1436-1453.
DONG Lei, YANG Zimin, QIAO Ji, et al. Optimal scheduling of integrated energy multi-microgrid system based on hierarchical constraint reinforcement learning[J]. Transactions of China Electrotechnical Society, 2024, 39(5): 1436-1453.

[28] 吕金玲, 王小君, 窦嘉铭, 等. 考虑运行状态信息的综合能源系统图强化学习优化调度[J]. 电力系统保护与控制, 2024, 52(2): 1-14.
LÜ Jinling, WANG Xiaojun, DOU Jiaming, et al. Optimal dispatch of an integrated energy system based on graph reinforcement learning considering operation state information[J]. Power System Protection and Control, 2024, 52(2): 1-14.

[29] 陈海东, 蒙飞, 张越, 等. 基于生成对抗模仿学习的电力系统动态经济调度[J]. 电网技术, 2022, 46(11): 4373-4380.
CHEN Haidong, MENG Fei, ZHANG Yue, et al. Dynamic economic dispatch of power system based on generative adversarial imitation learning[J]. Power System Technology, 2022, 46(11): 4373-4380.

[30] 郭剑波. 构建新型电力系统是实现能源转型、达成“双碳”目标的有效途径[N]. 国家电网报, 2021-09-07(005).
GUO Jianbo. Building a new type of power system is an effective way to achieve energy transformation and achieve the “dual carbon” goal[N]. State Grid News, 2021-09-07(005).

收稿日期：2024-05-08；　　修回日期：2024-08-21

作者简介：

李天明(2001—)，女，硕士研究生，研究方向为人工智能在综合能源系统中的应用；E-mail: 22121469@ bjtu.edu.cn

王小君(1978—)，男，通信作者，教授，博士生导师，研究方向为电力系统分析与控制、综合能源系统优化等。E-mail: xjwang1@bjtu.edu.cn

（编辑　魏小丽）