

基于深度强化学习的建筑综合能源系统主从博弈 优化调度

申晓宁^{1,2,3,4}, 陈星晖¹, 陈文言¹, 许新苏¹

(1. 南京信息工程大学自动化学院, 南京 210044; 2. 南京信息工程大学江苏省大气环境与装备技术协同创新中心, 南京 210044; 3. 南京信息工程大学江苏省大数据分析技术重点实验室, 南京 210044; 4. 江苏省气象能源利用与控制工程技术研究中心, 南京 210044)

摘要: 随着全球社会日益关注可持续能源的实践过渡, 建筑综合能源系统优化对提高能源消费的低碳性及经济性具有重要意义。为此对建筑能源运营商的调度策略和定价策略展开了研究。首先, 同时考虑供给侧和需求侧的信息交互特性, 以供给侧为领导者、需求侧为追随者建立一种基于主从博弈框架的建筑综合能源系统双侧优化模型。其次, 针对主从博弈框架双侧的多次信息交互提出了一种基于自适应动作探索机制的深度确定性策略梯度算法以有效求解所建模型。自适应动作探索机制根据累计奖励的方差与 critic 网络平均损失值构建自适应探索系数改进算法的动作选择策略, 保证算法的精度和稳定性。最后, 通过实例验证了所提算法的有效性。实验结果表明, 与其他深度强化学习算法相比, 所提算法可以提高算法的收敛精度、稳定性和能源运营商的总收益, 从而辅助能源供应商做出更好的决策。

关键词: 深度强化学习; 建筑综合能源系统; 智能调度; 主从博弈; 能源定价

Stackelberg Game Optimization Scheduling in Building Integrated Energy Systems Based on Deep Reinforcement Learning

SHEN Xiaoning^{1,2,3,4}, CHEN Xinghui¹, CHEN Wenyan¹, XU Xinsu¹

(1. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 3. Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 4. Jiangsu Meteorological Energy Utilization and Control Engineering Technology Research Center, Nanjing 210044, China)

Abstract: As the global society is increasingly concerned about the transition to sustainable energy practices, building integrated energy systems optimization is significant in improving low-carbon and economic energy consumption. Therefore, research is conducted on the scheduling and pricing strategies of building energy operators. Firstly, the information interaction characteristics of both the supply side and the demand side are considered. A two-side optimization model of the building integrated energy system based on the Stackelberg game framework is established with the supply side as the leader and the demand side as the follower. Secondly, a deep deterministic strategy gradient algorithm is proposed based on the adaptive action exploration mechanism to solve the constructed model efficiently given the multiple information interactions between the two sides of the Stackelberg game framework. The adaptive action exploration mechanism constructs the action selection strategy of the adaptive exploration coefficient improvement algorithm based on the variance of the cumulative rewards and the average loss value of the critic network, ensuring the algorithm's accuracy and stability. Finally, the effectiveness of the proposed algorithm is verified by examples. The experimental results show that compared with other deep reinforcement learning algorithms, the proposed algorithm can improve the convergence accuracy and stability of the algorithm, as well as the total revenue of the energy operator, thus assisting the energy supply side in making better decisions.

Key words: deep reinforcement learning; building integrated energy system; intelligent scheduling; Stackelberg game; energy pricing

0 引言

随着全球越来越关注建筑区域中传统能源向绿色能源的转型,建筑综合能源系统(building integrated energy system, BIES)优化对提高建筑的节能减排效果起着重要作用^[1]。BIES是指建筑区域内整合煤炭、石油、天然气、电能、热能等多种能源,实现多种能源子系统之间的优化运行、协同管理和交互响应。BIES优化问题可以细分为建筑能源调度问题和建筑能源定价问题。解决建筑能源调度问题旨在提供一种合理的能源调度方案以降低建筑能耗。根据调查,全球建筑能耗占总能耗的33%,在未来20年,建筑能耗预计将激增50%^[2]。因此,建筑能源调度问题是BIES中关键的一环。此外,考虑到零售能源价格的不确定性,解决建筑能源定价问题对于维护消费者和建筑能源运营商(building energy operator, BEO)的利益至关重要^[3]。为了创造安全、可靠、经济和舒适的环境,有必要研究新的BIES优化方法。

目前,国内外学者对BIES的模型做了大量研究。从调度时间角度出发,早期的研究人员大多研究日前调度,即提前一日进行调度。文献[4]介绍了关于电能需求管理机制的日前调度模型。文献[5]在获得日前负荷值和可再生能源发电量的基础上提出一种日前经济调度计划。大多数日前调度模型都没有考虑能源需求的灵活性,而实时调度弥补了日前调度的缺点,它在调度当日每隔一段短暂的时间再进行一次调度,能够实时掌握能源需求的变动。文献[6]将实时阶段下的建筑能源管理问题根据实时环境参数做出自适应控制决策。文献[7]以实时调度的方式解决了各类机组设备与办公建筑相结合的新型近零能耗BIES系统优化模型。从优化目标角度出发,文献[8]构建了考虑建筑热平衡的经济优化单目标模型使系统的总运行成本最小。文献[9]在满足建筑能源需求基础上给出了实现低碳化目标的多时间尺度BIES。文献[10]除了考虑系统运行成本和CO₂排放量外,还将经济效益和环境

效益最大化作为优化目标。然而,上述目标忽略了用户对价格因素的满意度问题。从供需侧角度出发,文献[11]构建了考虑柔性负荷的建筑多能源系统单侧模型。文献[12]提出了基于节点-流量模型的BIES供需双侧优化模型,但它没有考虑到供给侧和需求侧间的信息交互问题。为此,本文结合了以供供给侧为领导者、需求侧为追随者的主从博弈框架,供给侧能够及时地收到需求侧的信息反馈并对应调整价格,制定更合理、有效的定价方案。

近年来,许多学者针对BIES的优化提出了多种方法,其中强化学习(reinforcement learning, RL)是通过智能体与环境的交互试错来逐步优化策略,最大化智能体从环境中获得的累计奖励值,从而自主学习到最优策略的一类算法。文献[13]提出了一种基于Q学习的建筑集群需求侧管理方法求解能源调度问题。文献[14]提出了基于Sarsa算法的小规模用户P2P交易决策分析方法以同时优化调度与定价方案。然而,当BIES中设备数量较多时,强化学习难以处理数据特征维度高的问题。深度强化学习算法结合了深度学习的高维特征提取能力,使用深度神经网络将复杂的状态动作空间映射为较低维度的特征空间,降低了算法的计算复杂度。文献[15]采用近端策略优化(proximal policy optimization, PPO)的方法获取了含电-气-热的多微网系统中微网服务商的定价策略。文献[16]设计了一种结合深度Q学习和风险价值的SQL深度强化学习方法处理具有不确定性极端天气事件的BIES。在众多深度强化学习算法中,深度确定性策略梯度(deep deterministic policy gradient, DDPG)是一种以确定性策略梯度为核心的深度强化学习算法,对于实时响应和计算资源有限的应用通常更高效,因此它适用于求解BIES动态的能源调度和能源定价问题。文献[17]提出了基于改进训练策略的增强型DDPG算法以获得BIES的最佳实时能耗控制策略。在深度强化学习中适当的动作策略选择机制对于平衡智能体的探索和利用能力非常重要。基础的DDPG算法在动作选择机制中添加高斯噪声以保证智能体的探索性,避免陷入局部最优。然而,它忽略了环境实时变化的问题。文献[18]基于进化策略对噪声权重进行扰动,以探索更丰富的动作集。文献[19]设计了一个与训练稳定性呈负相关的噪声权重系数以调整噪声幅度。然而,上述改进方法均忽略了智能

基金项目:国家自然科学基金资助项目(61502239);江苏省自然科学基金资助项目(BK20150924)。

Foundation item: Supported by the National Natural Science Foundation of China (61502239); the Natural Science Foundation of Jiangsu Province (BK20150924).

体训练代数对动作选择策略的影响,导致收敛速度较慢。

基于对现有 BIES 模型和求解方法不足之处的分析,本文构建了基于主从博弈框架的建筑综合能源系统优化模型。为了有效求解所提模型,提出了一种自适应探索 DDPG 算法(deep deterministic policy gradient with adaptive exploration, DDPG_AE)。最后,将本文算法与已有算法进行对比,结果表明该方法具有更高的求解精度和更快的收敛速度,验证了其有效性。

1 基于主从博弈框架的建筑综合能源系统优化模型

为了使建筑综合能源系统模型更贴合实际,本文建立了一种基于主从博弈框架的 BIES 优化模型。所提模型由系统研究框架、阶梯式奖惩碳交易模型、供给侧优化模型、需求侧优化模型和主从博弈模型组成。

1.1 系统架构

本文研究的 BIES 供需双侧架构如图 1 所示。

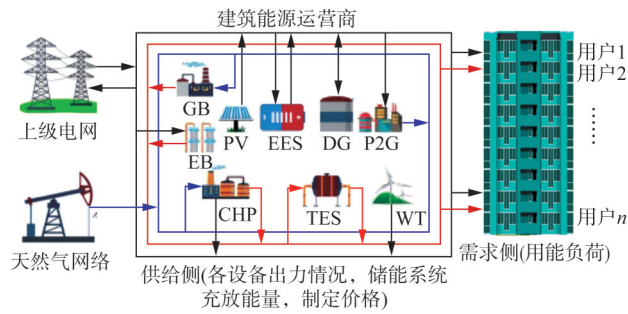


图 1 BIES 供需双侧架构图

Fig. 1 BIES supply and demand bilateral architecture diagram

在考虑到两侧之间的相互关系的基础上,建筑能源运营商 BEO 连接着用户和上级能源系统。上级能源系统由上级电网和天然气网络组成,而 BEO 内部各能源网络之间通过转换设备实现耦合。电力子系统直接与上级电网和用户相连,包括热电联产机组(cogeneration, combined heat and power, CHP)、分布式发电机组(distributed generation, DG)、电力储能系统(energy storage system, ESS)、光伏发电机组(photovoltaic, PV)和风力涡轮机(wind turbine, WT),为用户提供电能。与此同时,热力子系统作为一个独立系统,包含热电联产机组

(cogeneration, combined heat and power, CHP)、热量储能系统(thermal energy storage, TES)、电锅炉(thermal energy storage, EB)和燃气锅炉(gas fired boiler, GB),为用户提供热能。燃气子系统则通过连接天然气网络,通过天然气网络和电转气系统(power to gas, P2G)为用户提供燃气。BEO 通过销售电能和热能向用户提供服务,从而取得收益。

1.2 阶梯式奖惩碳交易模型

为了促进电力行业低碳技术的研发和推广以及新能源发电技术的应用,我国积极推动电力体制改革,并大力推行碳交易机制^[20],对企业的碳排放采取奖惩措施。

本文提出一种新的阶梯型奖惩碳交易模型,采用基准线法确定系统的无偿配额^[21],纳入配额管理的不同机组无偿配额分配标准不同,通过不同的碳配额系数确定每小时的碳配额。 t 时刻($t = 1, 2, \dots, T$)上级电网有功功率出力、DG 机组、CHP 机组的碳配额 $X_{\text{limit}}^{\text{grid},t}$ 、 $X_{\text{limit}}^{\text{DG},t}$ 、 $X_{\text{limit}}^{\text{CHP},t}$ 以及 BEO 的总碳配额 X_{limit}^t 如式(1)~(4)^[20]所示。

$$X_{\text{limit}}^{\text{grid},t} = \alpha_{\text{grid}} P_{\text{grid}}^t \quad (1)$$

$$X_{\text{limit}}^{\text{DG},t} = \alpha_{\text{DG}} P_{\text{DG}}^{e,t} \quad (2)$$

$$X_{\text{limit}}^{\text{CHP},t} = \alpha_{\text{CHP}}^e P_{\text{CHP}}^{e,t} + \alpha_{\text{CHP}}^h P_{\text{CHP}}^{h,t} \quad (3)$$

$$X_{\text{limit}}^t = X_{\text{limit}}^{\text{grid},t} + X_{\text{limit}}^{\text{DG},t} + X_{\text{limit}}^{\text{CHP},t} \quad (4)$$

式中: α_{grid} 、 α_{DG} 、 α_{CHP}^e 和 α_{CHP}^h 分别为上级电网有功功率出力、DG 机组输出功率、CHP 机组输出电功率和 CHP 机组输出热功率的碳配额系数^[20]; P_{grid}^t 、 $P_{\text{DG}}^{e,t}$ 、 $P_{\text{CHP}}^{e,t}$ 、 $P_{\text{CHP}}^{h,t}$ 分别为 t 时刻上级电网有功功率出力的电功率、DG 机组输出的电功率、CHP 机组输出的电功率和热功率。

为了简化碳交易模型,本文假设各电力机组的碳排放量与其发电产能成正比。 t 时刻上级电网有功功率出力、DG 机组、CHP 机组的实际碳排放量 $X_{\text{real}}^{\text{grid},t}$ 、 $X_{\text{real}}^{\text{DG},t}$ 、 $X_{\text{real}}^{\text{CHP},t}$ 以及 BEO 的实际碳排放总量 X_{real}^t 如式(5)~(8)^[20]所示。

$$X_{\text{real}}^{\text{grid},t} = \beta_{\text{grid}} P_{\text{grid}}^t \quad (5)$$

$$X_{\text{real}}^{\text{DG},t} = \beta_{\text{DG}} P_{\text{DG}}^{e,t} \quad (6)$$

$$X_{\text{real}}^{\text{CHP},t} = \beta_{\text{CHP}}^e P_{\text{CHP}}^{e,t} + \beta_{\text{CHP}}^h P_{\text{CHP}}^{h,t} \quad (7)$$

$$X_{\text{real}}^t = X_{\text{real}}^{\text{grid},t} + X_{\text{real}}^{\text{DG},t} + X_{\text{real}}^{\text{CHP},t} \quad (8)$$

式中: β_{grid} 、 β_{DG} 、 β_{CHP}^e 、 β_{CHP}^h 分别为上级电网有功功率出力、DG 机组输出功率、CHP 机组输出电功率

和CHP机组输出热功率的碳排放系数^[20]。

当整个系统的实际碳排放量大于碳配额时, BEO对超出部分的碳排放配额按照阶梯式价格进行购买, 且随着超额碳排放量的增加, 购买价格呈递增趋势。这意味着超额的碳排放量越多惩罚值越大。相反, 如果实际碳排放量小于碳配额时, 则对BEO进行阶梯式奖励, 碳排放越小, 奖励值越大, 从而形成一种奖惩机制。对单位碳交易价格的调整如式(9)—(10)所示。

若 $X'_{\text{limit}} \geq X'_{\text{real}}$, 即在 t 时刻分配的碳配额大于实际碳排放量, 则给予奖励。

$$E'_{\text{co}_2} = \begin{cases} a(1 + \theta), 0 \leq \frac{X'_{\text{limit}}}{X'_{\text{real}}} - 1 < 0.2 \\ a(1 + 2\theta), 0.2 \leq \frac{X'_{\text{limit}}}{X'_{\text{real}}} - 1 < 0.4 \\ a(1 + 3\theta), 0.4 \leq \frac{X'_{\text{limit}}}{X'_{\text{real}}} - 1 < 1 \end{cases} \quad (9)$$

若 $X'_{\text{limit}} < X'_{\text{real}}$, 即在 t 时刻分配的碳配额小于实际碳排放量, 则给予惩罚。

$$E'_{\text{co}_2} = \begin{cases} a(1 + \varepsilon), 0 \leq \frac{X'_{\text{real}}}{X'_{\text{limit}}} - 1 < 0.2 \\ a(1 + 2\varepsilon), 0.2 \leq \frac{X'_{\text{real}}}{X'_{\text{limit}}} - 1 < 0.4 \\ a(1 + 3\varepsilon), 0.4 \leq \frac{X'_{\text{real}}}{X'_{\text{limit}}} - 1 < 1 \end{cases} \quad (10)$$

式中: a 为原始单位碳交易价格; E'_{co_2} 为改进后的单位碳交易价格; θ 、 ε 分别为奖励系数和惩罚系数。

t 时刻碳交易成本 C'_{co_2} 的计算公式如式(11)所示。

$$C'_{\text{co}_2} = E'_{\text{co}_2} \times (X'_{\text{real}} - X'_{\text{limit}}) \quad (11)$$

当 $X'_{\text{limit}} < X'_{\text{real}}$ 时碳交易成本 C'_{co_2} 为正值, 表示对BEO进行奖励; 当 $X'_{\text{real}} > X'_{\text{limit}}$ 时 C'_{co_2} 为正值, 表示对BEO进行惩罚。

1.3 供给侧优化模型

供给侧将电、气、热耦合的BEO作为调度对象, 结合碳交易机制, 根据上级电网和天然气网络的特性分别与其交互并获取自身系统的能源。此外, BEO在各设备与能源环境约束下决策出各个机组的出力情况、储能系统的充放能量和向上级的购售能量, 同时制定用户的电价和热价实现自身的收益最大化。

1.3.1 目标函数

BEO综合考虑电、气和热的耦合, 结合阶梯式奖惩碳交易机制, 设计合理的能源调度策略和价格制定策略以实现最大化BEO的总收益, 目标函数如式(12)所示。

$$\max F_{\text{BEO}} = \sum_{t=1}^T (R'_{\text{sale}} + R'^{\text{e},t}_{\text{grid}} - C'_{\text{buy}} - C'_{\text{ope}} - C'_{\text{co}_2} - C'_{\text{es}}) \quad (12)$$

式中: T 为一天中的总时段数量; R'_{sale} 和 $R'^{\text{e},t}_{\text{grid}}$ 分别为 t 时刻BEO向用户出售电能、热能和向上级电网出售剩余电能的收益; C'_{co_2} 为 t 时刻BEO的碳交易成本; C'_{buy} 、 C'_{ope} 和 C'_{es} 分别为 t 时刻BEO向上级电网和气网的购能成本、运行和维护成本和储能成本。如式(13)—(17)所示。

$$R'_{\text{sale}} = L'_e \rho'_e + L'_h \rho'_h \quad (13)$$

$$R'^{\text{e},t}_{\text{grid}} = P'_{\text{sgrid}} \rho'_{\text{sgrid}} \quad (14)$$

$$C'_{\text{buy}} = P'_{\text{bgrid}} \rho'_{\text{bgrid}} + P'_{\text{bgas}} \rho'_{\text{bgas}} \quad (15)$$

$$C'_{\text{ope}} = \sum_i (KP'_i) \quad (16)$$

$$C'_{\text{es}} = N \times \begin{cases} P'_{\text{c,store}}, \text{充能} \\ P'_{\text{d,store}}, \text{放能} \end{cases} \quad (17)$$

式中: L'_e 和 L'_h 分别为 t 时刻需求侧响应的电负荷和热负荷; ρ'_e 、 ρ'_h 分别为 t 时刻的单位电价和单位热价; P'_{sgrid} 、 P'_{bgrid} 与 P'_{bgas} 分别为 t 时刻BEO向电网售电功率以及从电网和气网购电和购气的功率; ρ'_{sgrid} 、 ρ'_{bgrid} 与 ρ'_{bgas} 分别为 t 时刻BEO向电网售电的单位价格、购电的单位价格与购气的单位价格; K 为 t 时刻 i 设备的单位运维费用; P'_i 为 i 设备 t 时刻的输出功率; N 为储能设备的损耗系数; $P'_{\text{c,store}}$ 和 $P'_{\text{d,store}}$ 分别为 t 时刻储能设备的充能功率和放能功率。

1.3.2 约束条件

供给侧设计合理的能源调度策略和价格制定策略需要满足的约束条件如式(18)—(33)所示。

1) 能量与功率平衡约束: 供给侧需满足各类能源 t 时刻的实时供需平衡, 电和热子系统的能量与功率平衡约束如下。

$$L'_e = P'_{\text{DG}} + P'_{\text{e,chp}} + P'_{\text{pv}} + P'_{\text{wt}} + P'_{\text{bgrid}} + P'_{\text{d,store}} - P'_{\text{sgrid}} - P'_{\text{c,store}} - P'_{\text{c,EB}} \quad (18)$$

$$L'_h = P'_{\text{h,chp}} + P'_{\text{h,EB}} + P'_{\text{h,GB}} + P'_{\text{d,store}} - P'_{\text{c,store}} \quad (19)$$

$$L'_g = P'_{\text{bgas}} + P'_{\text{g,P2G}} - P'_{\text{g,chp}} - P'_{\text{g,GB}} \quad (20)$$

式中: L'_g 为 t 时刻需求侧响应的气负荷; P'_{DG} 、 $P'_{\text{e,chp}}$ 、 P'_{pv} 和 P'_{wt} 分别为 t 时刻DG机组、CHP机组、

PV 机组和 WT 机组的发电功率； $P_{c,store}^{ct}$ 、 $P_{d,store}^{ct}$ 、 $P_{c,store}^{hr}$ 和 $P_{d,store}^{hr}$ 分别为 t 时刻储电设备和储热设备的充放能功率； $P_{c,EB}^t$ 为 t 时刻 EB 机组电加热的功率； $P_{h,CHP}^t$ 、 $P_{h,EB}^t$ 和 $P_{h,GB}^t$ 分别为 t 时刻 CHP 机组、EB 机组和 GB 机组的发热功率； $P_{g,P2G}^t$ 为 t 时刻 P2G 机组产气功率； $P_{g,CHP}^t$ 和 $P_{g,GB}^t$ 分别为 t 时刻 CHP 机组和 GB 机组的耗气功率。

2) 设备功率约束：供给侧 t 时刻的设备需满足其输出功率不可超过其最大功率，每一次输出功率的变化也不能超过其最大爬坡功率。

$$0 \leq P_m^t \leq P_m^{\max} \quad (21)$$

$$|P_m^t - P_m^{t-1}| \leq P_{r,m}^{\max} \quad (22)$$

式中： P_m^t 为设备 m 在 t 时刻的运行功率； $P_m^{\max}$ 为设备 m 的最大运行功率； $P_{r,m}^{\max}$ 为设备 m 的运行最大爬坡功率。

3) 能源网络约束：供给侧 t 时刻的购电功率和购气功率不可超过其最大购电功率与最大购气功率。

$$0 \leq P_{bgrid}^t \leq P_{bgrid}^{\max} \quad (23)$$

$$0 \leq P_{bgas}^t \leq P_{bgas}^{\max} \quad (24)$$

式中 $P_{bgrid}^{\max}$ 和 $P_{bgas}^{\max}$ 分别为最大购电功率和购气功率。

4) 储能约束：储能设备包括储电设备蓄电池和储热设备热能存储系统。储能设备 t 时刻的充放能功率不可超过其最大充放能功率，且它的能源改变量与充能和放能效率相关。蓄电池在同一时刻只可存在充电、放电或不充不放其中一种状态，而热能存储系统则可以同时进行充放热量。

$$0 \leq P_{c,store}^{ct} \leq B_{c,store}^{ct} P_{c,store}^{e,max} \quad (25)$$

$$0 \leq P_{d,store}^{ct} \leq B_{d,store}^{ct} P_{d,store}^{e,max} \quad (26)$$

$$0 \leq P_{c,store}^{hr} \leq P_{c,store}^{h,max} \quad (27)$$

$$0 \leq P_{d,store}^{hr} \leq P_{d,store}^{h,max} \quad (28)$$

$$B_{c,store}^{ct} + B_{d,store}^{ct} \leq 1 \quad (29)$$

$$E_{ESS}^t = \begin{cases} E_{ESS}^{t-1} + P_{c,store}^{ct} \eta_{es}^c, & \text{充电} \\ E_{ESS}^{t-1} - \frac{P_{d,store}^{ct}}{\eta_{es}^d}, & \text{放电} \\ E_{ESS}^{t-1}, & \text{不充电不放电} \end{cases} \quad (30)$$

$$E_{TES}^t = E_{TES}^{t-1} + P_{c,store}^{hr} \eta_{es}^c - \frac{P_{d,store}^{hr}}{\eta_{es}^d} \quad (31)$$

式中： $P_{c,store}^{e,max}$ 和 $P_{d,store}^{e,max}$ 分别为蓄电池的最大充电功率和最大放电功率； $P_{c,store}^{h,max}$ 和 $P_{d,store}^{h,max}$ 分别为热能存储系统的最大充热功率和最大放热功率； $B_{c,store}^{ct}$ 和 $B_{d,store}^{ct}$ 为二进制变量，分别为蓄电池 t 时段充、放状态参数； E_{ESS}^t 和 E_{TES}^t 分别为 t 时刻蓄电池和热能存储系统

的储能量； η_{es}^c 和 η_{es}^d 分别为充能和放能效率系数。

5) 碳交易约束：碳交易成本由原始单位碳交易价格和奖惩系数组成，原始单位碳交易价格应为正值，如式(32)所示，且为了严格把控奖励和惩罚的程度，所设系数应满足式(33)。

$$a > 0 \quad (32)$$

$$0 < \theta < \varepsilon < 1 \quad (33)$$

1.4 需求侧优化模型

需求侧根据 BEO 制定的电价和热价对用户负荷进行优化，调整需求侧的电与热负荷响应以实现最大化消费者剩余。

1.4.1 目标函数

由于最大化消费者剩余中包括成本函数和效用函数^[22]，因此需求侧优化目标为最大化用户对于价格的效用函数与用户用能成本函数之差，如式(34)所示。

$$\max F_{user} = \sum_{t=1}^T [U_{user}^t - (f_{user}^{e,t} + f_{user}^{h,t})] \quad (34)$$

式中： U_{user}^t 为 t 时刻用户效用函数，如式(35)所示； $f_{user}^{e,t}$ 和 $f_{user}^{h,t}$ 分别为 t 时刻用户的购电成本和购热成本，如式(36)——(37)所示。

$$U_{user}^t = \left\{ \frac{L_{eb}^t \rho_{eb}^t}{\psi} \left[\left(\frac{L_{eb}^t}{L_{eb}^t} \right)^\psi - 1 \right] + \frac{L_{hb}^t \rho_{hb}^t}{\psi} \left[\left(\frac{L_{hb}^t}{L_{hb}^t} \right)^\psi - 1 \right] \right\} \quad (35)$$

$$f_{user}^{e,t} = (L_{eb}^t \rho_{eb}^t) \quad (36)$$

$$f_{user}^{h,t} = (L_{hb}^t \rho_{hb}^t) \quad (37)$$

式中： L_{eb}^t 和 L_{hb}^t 分别为 t 时刻用户的基线电负荷与热负荷（基线负荷定义为前一日的用户负荷）； ρ_{eb}^t 、 ρ_{hb}^t 分别为实际电价、实际热价； ρ_{eb}^t 和 ρ_{hb}^t 分别为 t 时刻基线电价和基线热价； $\psi \in (-\infty, 0) \cup (0, 1)$ 为用户负荷的调整系数。在 U_{user}^t 的第一项中，如果实际电负荷大于基线电负荷，则第一项的值为正。如果实际电负载小于基线电负荷，则第一项的值为负。当实际电负荷等于基线电负荷时，第一项为 0。而且，第一项值的增加随着实际电负荷的增大而减缓。第二项同理。对式(35)计算用户的最佳响应负荷，如式(38)——(39)所示。

$$L_{eb}^{t*} = \left(\frac{\rho_{eb}^t}{\rho_{eb}^t} \right)^\alpha L_{eb}^t \quad (38)$$

$$L_{hb}^{t*} = \left(\frac{\rho_{hb}^t}{\rho_{hb}^t} \right)^\alpha L_{hb}^t \quad (39)$$

式中: $\alpha = 1/(\psi - 1)$ 为用户负荷的弹性系数; L_e^* 、 L_h^* 分别为最佳响应电负荷、热负荷。当 $\rho_e^t > \rho_{eb}^t$ 时, 有 $L_e^* < L_{eb}^t$; 当 $\rho_e^t < \rho_{eb}^t$ 时, 有 $L_e^* > L_{eb}^t$ 。热的最佳响应负荷同理。

1.4.2 约束条件

用户需求响应负荷约束: 用户 t 时刻的用能负荷不可超过最大值, 如式(40)–(41)所示。

$$0 \leq L_e^t \leq L_e^{t, \max} \quad (40)$$

$$0 \leq L_h^t \leq L_h^{t, \max} \quad (41)$$

式中 $L_e^{t, \max}$ 和 $L_h^{t, \max}$ 分别为 t 时刻用户电负荷和热负荷的最大值。

1.5 主从博弈模型

主从博弈模型是一个两阶段的动态博弈, 博弈的时间是连续的。在该博弈模型中, 先做出决策的一方被称为领导者; 在领导者做出决策之后, 剩余的参与者根据领导者的决策进行决策, 被称为跟随者。接着领导者根据跟随者的决策对自己的决策进行调整, 如此往复, 直到达到纳什均衡。在本文的模型中, BEO作为领导者根据当前的实时状况制定价格策略, 用户作为跟随者通过BEO的决策进行用能负荷策略的调整, 时间步长为1 h, 体现了博弈模型的动态性。由于环境中同时存在多个用户, 因此构建一主多从的主从博弈模型。如图2所示。

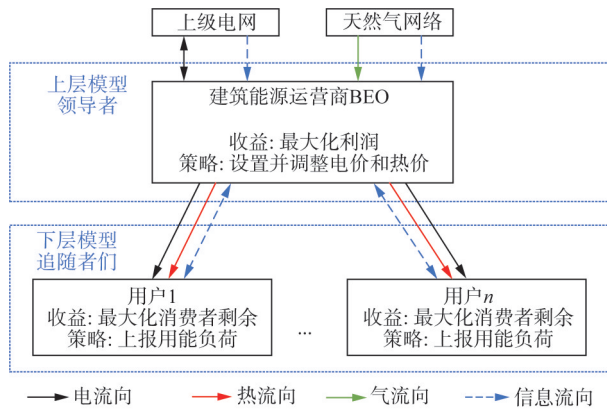


图2 主从博弈框架

Fig. 2 Stackelberg Game Framework

图2的博弈模型 G_p 中包含了5个元素, 分别为参与者 N , 供给侧策略 ρ_{price}^t 和需求侧策略 L_{load}^t , 供给侧收益 F_{BEO} 和需求侧收益 F_{user} , 如式(42)所示。

$$G_p = \{ N, \rho_{price}^t, L_{load}^t, F_{BEO}, F_{user} \} \quad (42)$$

1)参与者: 参与者 N 包括供给侧的BEO和需求侧的用户。其中, U_{BEO} 为该博弈的领导者,

$U_{ser1}, U_{ser2}, \dots, U_{sern}$ 表示 n 个用户为该博弈的跟随者。参与者集合 N 表示为:

$$N = \{ U_{BEO}, U_{ser1}, U_{ser2}, \dots, U_{sern} \} \quad (43)$$

2)策略: BEO的策略是决策各个机组的出力情况、储能系统的充放能状态和向上级的购售能计划, 同时向用户制定电价和热价, BEO策略集合 S_{BEO}^t 表示为:

$$S_{BEO}^t = \{ P_{DG}^t, P_{e, chp}^t, P_{h, chp}^t, P_{h, EB}^t, P_{h, GB}^t, P_{grid}^t, \rho_e^t, \rho_h^t, E_{ESS}^t, E_{TES}^t \} \quad (44)$$

用户的策略是根据BEO的定价策略调整自己的用能负荷, 用户策略集合 S_{user}^t 表示为:

$$S_{user}^t = \{ L_e^t, L_h^t \} \quad (45)$$

3)收益: 参与者们的收益分别对应于3.1.3节和3.1.4节所提两侧目标函数, BEO收益 F_{BEO} 见式(12), 用户的收益 F_{user} 见式(34)。

在该主从博弈模型中, 如果存在一种策略组合 (S_{BEO}^t, S_{user}^t) , 同时满足式(46)–(47), 则称此组合为优势策略, 视为达到Stackelberg均衡。

$$F_{BEO}(S_{BEO}^t, S_{user}^t) \geq F_{BEO}(S_{BEO}^t, S_{user}^t) \quad (46)$$

$$F_{user}(S_{BEO}^t, S_{user}^t) \geq F_{user}(S_{BEO}^t, S_{user}^t) \quad (47)$$

式中: $F_{BEO}(\cdot)$ 、 $F_{user}(\cdot)$ 分别为供给侧和需求侧的目标函数; S_{BEO}^t 、 S_{user}^t 分别为供给侧和需求侧的一种策略集合。该Stackelberg均衡中, 任何参与者如果单独改变自身策略都不会提高自身的收益。该主从博弈模型的解是存在且唯一的, 通过定理A证明如下。

定理A: 当主从博弈模型满足以下条件时, 则存在唯一的Stackelberg均衡^[23]。

1)博弈参与者的目标函数是关于博弈策略集的非空、连续函数。

2)跟随者目标函数是关于各自博弈策略集的连续凹/凸函数。

证明如下。

1)根据本文所提模型可知, 领导者BEO的策略集需要满足约束式(18)–(33); 跟随者用户的策略集需要满足约束式(40)–(41)。所以领导者与参与者的策略集都是非空且连续的。

2)对用户目标函数进行分析, 函数 F_{user} 分别对 L_e^t 和 L_h^t 求二阶偏导, 可得:

$$\frac{\partial^2 F_{user}}{(\partial L_e^t)^2} = (\psi - 1) L_{eb}^t \rho_{eb}^t \left(\frac{L_e^t}{L_{eb}^t} \right)^{\psi-2} < 0 \quad (48)$$

$$\frac{\partial^2 F_{\text{user}}}{(\partial L_h^t)^2} = (\psi - 1) L_{\text{hb}}^t \rho_{\text{hb}}^t \left(\frac{L_h^t}{L_{\text{hb}}^t} \right)^{\psi-2} < 0 \quad (49)$$

式中 ψ 为用户负荷的调整系数。由于除 $(\psi - 1)$ 外的式子均大于 0, 且 $\psi \in (-\infty, 0) \cup (0, 1)$, 则 $\frac{\partial^2 F_{\text{user}}}{(\partial L_c^t)^2}$ 和 $\frac{\partial^2 F_{\text{user}}}{(\partial L_h^t)^2}$ 均小于 0。同时, 对 F_{user} 进行交叉偏导, 均取值为 0。因此, F_{user} 的 Hessian 矩阵负定, F_{user} 存在唯一的极大值点。

显然, 用户目标函数是关于其策略集的多元线性函数, 根据凹凸性的定义, 线性函数既是凹函数也是凸函数。于是本文所提主从博弈模型存在唯一的 Stackelberg 均衡。

2 引入自适应动作探索机制的深度确定性策略梯度算法

针对基于主从博弈的 BIES 模型提出一种 DDPG_AE 算法, 给出 DDPG_AE 的算法框架并详细介绍算法的改进策略。

2.1 马尔可夫决策过程

智能体定义为 BEO, 环境定义为智能体观察到的世界, 包括上级电网、天然气网络、各机组设备和用户。假设环境可以与 BEO 多次交互, BEO 可以利用强化学习在这些交互中不断调整调度和定价策略, 以获得最优的调度和定价方案。该交互过程可以形式化为马尔可夫决策过程。其中马尔可夫决策过程包括用一个四元组 (S, A, R, P) , 分别表示状态空间、动作空间、奖励函数和状态转移概率。

1) 状态空间: 状态用于描述马尔可夫决策过程中环境的当前情况, 本文定义的状态空间由当前时段、储能设备的储能量、用户用能负荷、购电和购气价格以及 PV 与 WT 机组输出功率组成, 可以表示为一个九维向量 s_t , 如式(50)所示。

$$s_t = \{ t, E_{\text{ESS}}^t, E_{\text{TES}}^t, L_{\text{load}}^t, \rho_{\text{sgrid}}^t, \rho_{\text{bgrid}}^t, \rho_{\text{bgas}}^t, P_{\text{pv}}^t, P_{\text{wt}}^t \} \quad (50)$$

2) 动作空间: 动作空间组合了 BEO 的能源调度策略和定价策略。本文定义的动作空间由 t 时刻 DG、CHP (包含电和热)、EB、GB 机组的输出功率、BEO 的购售能计划, 制定的能源价格和储能设备的充能功率和放能功率组成, 可以表示为一个十维向量 a_t , 如式(51)所示。

$$a_t = \{ P_{\text{DG}}^t, P_{\text{e, chp}}^t, P_{\text{h, chp}}^t, P_{\text{h, EB}}^t, P_{\text{h, GB}}^t, P_{\text{grid}}^t, P_{\text{c}}^t, P_{\text{h}}^t, E_{\text{ESS}}^t, E_{\text{TES}}^t \} \quad (51)$$

3) 奖励函数: 奖励函数应直接反映优化问题的目标。本文定义在 t 时刻的即时奖励函数为 BEO 的净收益 r_t , 如式(52)所示。

$$r_t = (R_{\text{sale}}^t + R_{\text{grid}}^{\text{c}, t} - C_{\text{buy}}^t - C_{\text{ope}}^t - C_{\text{es}}^t) / 5 \quad (52)$$

4) 状态转移概率: 状态转移概率描述了状态 s_t 下执行动作 a_t 后, 环境受到影响进而 BEO 过渡到下一个状态 $s_{t+1} = s'$ 的概率, 如式(53)所示。

$$P_a(s, s') = P(s_{t+1} = s' | s_t = s, a_t = a) \quad (53)$$

式中: $P_a(s, s')$ 为状态 s 过渡到状态 s' 的状态转移概率; $s_{t+1} = s'$ 为 s_t 的下一时刻的状态; a 为状态 s 下执行的动作。

2.2 算法求解框架

本文提出的 DDPG_AE 算法求解框架如图 3 所示。首先构造 4 个神经网络, 分别是输入为状态 s 、参数为 θ 的 actor 网络 $\mu(s|\theta)$, 输入为状态 s 和 a 、参数为 φ 的 critic 网络 $Q(s, a|\varphi)$, 输入为状态 s 、参数为 θ' 的 actor 目标网络 $\mu(s|\theta')$ 以及输入为状态 s 和 a 、参数为 φ' 的 critic 目标网络 $Q(s, a|\varphi')$ 。采样阶段时, 随机取状态 s_t 输入到 actor 网络中, actor 网络根据状态 s_t 采用自适应动作探索机制 (见 2.3 节) 选择动作 a_t , 接着与环境交互输出相应的奖励 r_t 和下一状态 s_{t+1} , 构成四元组样本 (s_t, a_t, r_t, s_{t+1}) 并存放进经验回放池 D 中, 如此循环, 直至经验回放池中的样本数量达到某一规模 L 。此后进入训练阶段。

训练阶段开始时, 首先取状态 s_t 构成一条四元组样本存储进经验回放池 D 中, 若此时 D 已达最大容量, 则替换 D 中首条四元组并添加至末尾。然后, 在经验回放池模块中随机采样 M 条 (s_t, a_t, r_t, s_{t+1}) 四元组样本, 本文用下标 k 作为多条样本的表达, 即 (s_k, a_k, r_k, s_{k+1}) , 将其输入到 critic 网络并输出当前时刻状态-动作的 Q 值函数 $Q(s_k, a_k|\varphi)$, Q 值函数是在给定的状态下执行某一动作后所能获得的期望折扣奖励值, 如式(54)所示。接着, actor 目标网络给出下一状态 s_{k+1} 的估计最优动作 $\mu(s_{k+1}|\theta')$, critic 目标网络对 $\mu(s_{k+1}|\theta')$ 估计 $k+1$ 时刻状态-动作的 Q 值函数 $Q(s_{k+1}, \mu(s_{k+1}|\theta')|\varphi')$ 。最后, 根据贝尔曼方程^[19] 计算 k 时刻状态-动作的目标值 y_k , 如式(55)所示。

$$Q(s_k, a_k|\varphi) = E(r_k + \gamma r_{k+1} + \gamma^2 r_{k+2} + \cdots + \gamma^{n-k} r_T) \quad (54)$$

$$y_k = r(s_k, a_k) + \gamma Q(s_{k+1}, \mu(s_{k+1}|\theta')|\varphi') \quad (55)$$

式中: γ 为贝尔曼方程的折扣系数; r_T 为终止状态对应的奖励值。

当获取目标值 y_k 之后, 即可对4个网络参数进行更新。在 critic 网络中通过梯度下降法更新参数 φ , 如式(56)所示^[19]。

$$\varphi = \varphi + \omega_\varphi \frac{1}{M} \sum_{k=1}^M [\delta_k \nabla_\varphi Q(s_k, a_k|\varphi)] \quad (56)$$

式中: $\delta_k = y_k - Q(s_k, a_k|\varphi)$ 为样本的时序差分误差; ω_φ 为 critic 网络的学习率; ∇_φ 表示对参数 φ 求梯度向量。

actor 网络根据当前状态 s_k 产生最优动作 $\mu(s_k|\theta)$, 输入到 critic 网络后产生 Q 值函数反馈给 actor 网络, 用梯度上升法进行参数更新, 如式(57)所示。

$$\theta = \theta + \omega_\theta \frac{1}{M} \sum_{k=1}^M [\nabla_{\mu(s_k|\theta)} Q(s_k, \mu(s_k|\theta)) \cdot \nabla_\theta \mu(s_k|\theta)] \quad (57)$$

式中: ω_θ 为 actor 网络的学习率; ∇_θ 表示对参数 θ 求梯度向量; $\nabla_{\mu(s_k|\theta)}$ 表示在当前状态 s_k 下对 actor 网络的参数 θ 求梯度向量。

在 critic 目标网络和 actor 目标网络参数更新模块中, 采用软更新机制^[19], 将当前网络参数和目标网络参数加权平均, 以保证参数缓慢更新, 提高训练的稳定性, 如式(58)—(59)所示。

$$\varphi' \leftarrow \tau \varphi + (1 - \tau) \varphi' \quad (58)$$

$$\theta' \leftarrow \tau \theta + (1 - \tau) \theta' \quad (59)$$

式中: $\tau \in (0, 1)$ 为软更新因子; φ' 为 critic 目标网络的参数; θ' 为 actor 目标网络的参数。

2.3 自适应动作探索机制

在 DDPG 或其它强化学习算法中, 动作选择策略在具有连续动作空间的优化问题中至关重要, 它应保证智能体对动作空间进行大范围地全局探索。DDPG 原始算法中的动作选择策略采用的是正态分布抽样方式。具体而言, 智能体使用 actor 网络输出的当前最优动作 $a' = \mu(s|\theta)$ 作为正态分布的均值, 并引入标准差参数 σ , 构建一个正态分布。随后, 智能体从该正态分布中随机采样生成一个新的动作 a 代替 a' 并从 actor 网络输出。该实现方式如式(60)所示。

$$a = \mu_\varepsilon(s) = \mu(s|\theta) + \varepsilon, \varepsilon \sim N(0, \sigma) \quad (60)$$

式中: ε 为探索噪声; $N(0, \sigma)$ 为均值为0、方差为 σ 的正态分布。

在建筑综合能源系统优化过程中, 问题模型是一个持续多步的马尔可夫决策过程。即使将探索噪声 ε 的方差 σ 取得较小, 由于多步串联的作用, 最后产生的 (s, a) 分布方差仍可能过大, 使得动作选择过程结果过于偏向探索而缺乏开发, 最终将影响收敛稳定性等效果。

因此本文提出一种自适应探索的动作选择策略, 在式(60)的噪声 ε 前引入一个自适应探索系数

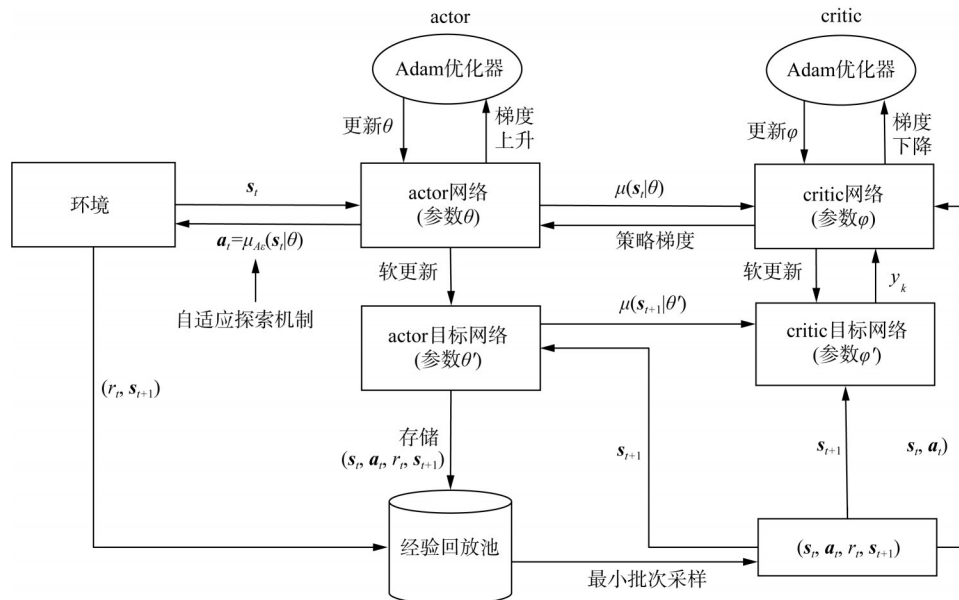


图3 DDPG_AE算法的求解框架

Fig. 3 Framework of the algorithm DDPG_AE

A, 如式(61)所示。

$$a = \mu_{A\epsilon}(s) = \mu(s|\theta) + A \cdot \epsilon, \epsilon \sim N(0, \sigma) \quad (61)$$

式中系数A结合了本次训练代 E_{pisode} 中critic网络的平均损失值与前 n 次训练代的累计奖励的方差, 如式(62)—(63)所示。

$$A = -\ln \left\{ \lambda \left[\frac{\text{var}(R^{E_{\text{pisode}}-1}, \dots, R^{E_{\text{pisode}}-i}, \dots, R^{E_{\text{pisode}}-n})}{T} + \frac{T}{(L_{\text{oss } 1}^{E_{\text{pisode}}} + \dots + L_{\text{oss } t}^{E_{\text{pisode}}} + \dots + L_{\text{oss } T}^{E_{\text{pisode}}})} \right] \right\} \quad (62)$$

$$R^{E_{\text{pisode}}-i} = r_1^{E_{\text{pisode}}-i} + \dots + r_t^{E_{\text{pisode}}-i} + \dots + r_T^{E_{\text{pisode}}-i} \quad (63)$$

式中: λ 为噪声系数; $\text{var}(\cdot)$ 为方差函数; $R^{E_{\text{pisode}}-i}$ 为第 $E_{\text{pisode}}-i$ 次训练代中智能体与环境互动后产生的累计奖励值, $i = 1, 2, \dots, n$; $r_t^{E_{\text{pisode}}-i}$ 为第 $E_{\text{pisode}}-i$ 次训练代中时间戳 t 的即时奖励; $L_{\text{oss } t}^{E_{\text{pisode}}}$ 为本次训练代 E_{pisode} 的时间戳 t 中由最后一条采样样本计算的critic网络损失值, $t = 1, 2, \dots, T$ 。A是一个关于奖励方差与critic网络平均损失值倒数之和的递减函数, 当两者之和越小时A的取值越大。

为了解决式(62)中两项的量纲问题, 首先需要从前 n 次训练代的累计奖励值作归一化处理。

$$R^{E_{\text{pisode}}-i} = \frac{R^{E_{\text{pisode}}-i} - R_{\min}}{R_{\max} - R_{\min}}, i = 1, 2, \dots, n \quad (64)$$

式中 $R_{\min}$ 和 $R_{\max}$ 分别为前 n 次训练代累计奖励中的最小值和最大值。

通过数学证明可得, 奖励归一化后的方差属于 $[0, 0.25]$ 。为了使critic网络损失值的倒数也属于该区间, 将critic网络的损失值归一化至 $[4, 100]$, 如式(65)所示。

$$L_{\text{oss } t}^{E_{\text{pisode}}} = \left[\frac{4 + \frac{96}{L_{\text{oss } \max}^{E_{\text{pisode}}} - L_{\text{oss } \min}^{E_{\text{pisode}}}} \times (L_{\text{oss } t}^{E_{\text{pisode}}} - L_{\text{oss } \min}^{E_{\text{pisode}}})}{(L_{\text{oss } t}^{E_{\text{pisode}}} - L_{\text{oss } \min}^{E_{\text{pisode}}})} \right] t = 1, 2, \dots, T \quad (65)$$

式中: $L_{\text{oss } \min}^{E_{\text{pisode}}}$ 和 $L_{\text{oss } \max}^{E_{\text{pisode}}}$ 分别为本次训练代 E_{pisode} 中critic网络损失值中的最小值和最大值; T 为采样周期。

由式(61)—(62)可得, 当智能体处于训练初期时, critic网络平均损失值较大, 若此时奖励的变化幅度较小(即奖励的方差小), 则增大自适应探索系数A, 以增强对动作空间的探索能力。在训练后期, critic网络平均损失值较小, 若此时奖励的变化幅度较大(即奖励的方差大), 则减小A, 以提高对当前最优动作的利用。

2.4 DDPG_AE算法的实现

DDPG_AE算法伪代码如表1所示。

DDPG_AE算法包括两层循环, 外层表示算法的一次迭代, 内层控制每次迭代在每个时间戳(如1h)中的参数训练。首先初始化网络参数、经验回放池和自适应探索系数(第1—3行)。然后开始采样阶段, 智能体获取初始观测状态, 根据高斯噪声选择合适的动作, 执行动作并根据状态转移概率跳转到下一状态并获得奖励(第5—8行), 并把 (s_t, a_t, r_t, s_{t+1}) 存储进经验回放池 D 的末尾(第9行)。

若 D 中的样本规模达到预设值, 则由采样阶段进入训练阶段。智能体重复采样阶段的操作(第11—14行), 若此时 D 中的样本规模达到最大规模, 则先删除 D 中首条样本再将新的四元组样本存储于 D 的末尾(第15—17行)。其次从 D 中采样批量样本分别计算critic网络目标值、样本的时序差分误差和各个梯度(第18—22行)。接着根据式(56)—(57)更新critic和actor网络参数(第23—24行), 并通过软更新机制式(58)—(59)规律地更新critic目标网络和actor目标网络(第25—26行), 最后式(62)计算自适应探索系数(第27—28行)。开启新一轮迭代, 直至迭代结束输出最终actor网络参数 θ 和critic网络参数 φ (第29行)。

3 实验仿真及结果分析

为了验证所提算法和改进策略的有效性, 采用Pycharm软件进行仿真实验, 计算机处理器参数为Intel(R) Core(TM)i5-10400 CPU @ 2.90 GHz, 16 GB RAM运行内存。选取2019年基于电负荷和热负荷的某建筑综合能源系统作为测试算例, 设计5组实验: 1)参数分析; 2)模型改进机制的有效性验证; 3)算法改进机制的有效性验证; 4)将所提算法DDPG_AE与4种代表性算法进行对比, 以验证所提算法的性能; 5)最优调度方案展示。

3.1 参数设置

本节对模型参数和算法参数分别设置与分析。

3.1.1 模型参数设置

本文模型中重要能源设备和机组的具体数据如表2所示^[24], 模型系数参数如表3所示^[24]。BEO从上级电网中购买电力和天然气的价格如图4^[24]。为了鼓励可再生能源的使用, 将BEO向上级电网售

表1 DDPG算法伪代码

Tab. 1 Pseudo-codes for the DDPG algorithm

算法1 DDPG_AE($E, T, L, V, \sigma, \gamma, M, \tau, \omega_\varphi, \omega_\theta, n$)

输入: E -训练迭代次数, T -采样周期(24 h), L -由采样阶段进入训练阶段时经验回放池 D 的规模, V - D 的最大容量, σ -标准差参数, γ -折扣系数, M -采样批次, τ -软更新学习率, ω_φ -critic 网络学习率, ω_θ -actor 网络学习率, n -计算累计奖励方差时迭代的数量

输出: θ -actor 网络的最优参数, φ -critic 网络的最优参数

```

1: 随机初始化当前 actor 网络和 critic 网络参数  $\theta$  和  $\varphi$ , 依据正态分布  $N(0, \sigma)$  初始化高斯噪声  $\varepsilon$ 
2:  $\theta' \leftarrow \theta, \varphi' \leftarrow \varphi, D \leftarrow \emptyset$ 
3:  $A = 1$ //初始化自适应探索系数  $A$ 
4: while size( $D$ ) <  $L$  do//采样阶段
5:   获取初始观测状态  $s_1$ 
6:   for  $t = 1$  to  $T$  do
7:      $a_t = \mu(s_t | \theta) + \varepsilon$ 
8:     执行动作  $a_t$  并获得奖励  $r_t$ , 观测下一状态  $s_{t+1}$ 
9:     将四元组  $(s_t, a_t, r_t, s_{t+1})$  存储于  $D$  的末尾
10: for episode = 1 to  $E$  do//训练阶段
11:   获取初始观测状态  $s_1$ 
12:   for  $t = 1$  to  $T$  do
13:      $a_t = \mu(s_t | \theta) + A \cdot \varepsilon$ //actor 网络根据自适应动作探索机制选择动作  $a_t$ 
14:     执行动作  $a_t$  并获得奖励  $r_t$ , 观测下一状态  $s_{t+1}$ 
15:     if size( $D$ )  $\geq V$  do
16:       删除  $D$  中首条四元组//更新  $D$ 
17:     将四元组  $(s_t, a_t, r_t, s_{t+1})$  存储于  $D$  的末尾
18:     在  $D$  中采样  $M$  个四元组  $(s_k, a_k, r_k, s_{k+1})$ 
19:     for  $k=1$  to  $M$  do
20:        $y_k = r(s_k, a_k) + \gamma Q(s_{k+1}, \mu(s_{k+1} | \theta') | \varphi')$ //计算 critic 网络目标值
21:        $\delta_k = y_k - Q(s_k, a_k | \varphi)$ //计算样本的时序差分误差
22:       计算  $\nabla_\varphi Q(s_k, a_k | \varphi)$ 、 $\nabla_{\mu(s_t | \theta)} Q(s_k, \mu(s_k | \theta))$  和  $\nabla_\theta \mu(s_k | \theta)$ 
23:        $\varphi = \varphi + \omega_\varphi \frac{1}{M} \sum_{k=1}^M [\delta_k \nabla_\varphi Q(s_k, a_k | \varphi)]$ //根据式(56)更新 critic 网络参数
24:        $\theta = \theta + \omega_\theta \frac{1}{M} \sum_{k=1}^M [\nabla_{\mu(s_t | \theta)} Q(s_k, \mu(s_k | \theta)) \cdot \nabla_\theta \mu(s_k | \theta)]$ //根据式(57)更新 actor 网络参数
25:        $\varphi' \leftarrow \tau \varphi + (1 - \tau) \varphi'$ //根据式(58)更新 critic 目标网络参数
26:        $\theta' \leftarrow \tau \theta + (1 - \tau) \theta'$ //根据式(59)更新 actor 目标网络参数
27:     if do
28:       根据式(62)计算  $A$ 
29: 输出  $\varphi, \theta$ 

```

电的价格设定为购电价格的一半^[25]。BEO 从天然气网络中购买的天然气价格为 0.25 元/kWh^[24]。碳交易原始价格为 0.08 元/kg^[20], 奖励系数 θ 为 0.2, 惩罚系数 ε 为 0.4。

3.1.2 算法参数设置与分析

应用 Python 中的 Adam 优化器实现神经网络的功能, 设置 actor 和 critic 神经网络的学习率 ω_θ 和 ω_φ 均为 0.01^[25], 软更新因子 τ 设为 0.05^[17]。贝尔曼方程的折扣系数 γ 设为 0.995^[25], 噪声中的标准差

参数 σ 为 1。经验回放池 D 的最大容量 V 为 50 000, 每次从经验回放池采样最小批次 M 设为 4 096, 当 D 中的样本数达到 $L=10\ 000$ 时由采样阶段进入训练阶段。设定采样周期 T 为 24 h, 时间步长为 1 h, 设置算法终止条件为迭代次数 E_{pisode} 达到 1 500 次, 自适应动作探索机制中的 n 设为 10。

本文引入了 2 个重要的超参数, 分别是 1.4.1 节用户负荷的弹性系数 α 和 2.3 节自适应动作探索机制中的噪声系数 λ 。采用 2 因素 4 水平的正交实

表 2 能源设备机组参数

Tab. 2 Energy equipment unit parameters

机组	容量/kW	爬坡功率/最大充放能功率/kWh
DG	500	200
CHP(电)	1 000	500
CHP(热)	1 000	500
GB	1 000	500
ESS	1 000	200
TES	1 000	200
EB	1 000	500

表 3 模型系数参数

Tab. 3 Model coefficient parameters

参数	数值	参数	数值
$\alpha_{\text{grid}}/(\text{g}\cdot\text{kWh}^{-1})$	0.79	$\beta_{\text{CHP}}^{\text{c}}/(\text{g}\cdot\text{kWh}^{-1})$	0.7
$\alpha_{\text{DG}}/(\text{g}\cdot\text{kWh}^{-1})$	0.4	$\beta_{\text{CHP}}^{\text{h}}/(\text{g}\cdot\text{kWh}^{-1})$	0.4
$\alpha_{\text{CHP}}^{\text{c}}/(\text{g}\cdot\text{kWh}^{-1})$	0.4	N	0.1
$\alpha_{\text{CHP}}^{\text{h}}/(\text{g}\cdot\text{kWh}^{-1})$	0.42	$\eta_{\text{es}}^{\text{c}}$	0.95
$\beta_{\text{grid}}/(\text{g}\cdot\text{kWh}^{-1})$	0.8	$\eta_{\text{es}}^{\text{d}}$	0.9
$\beta_{\text{DG}}/(\text{g}\cdot\text{kWh}^{-1})$	0.5	$K/(\text{元}\cdot\text{kW}^{-1})$	0.25

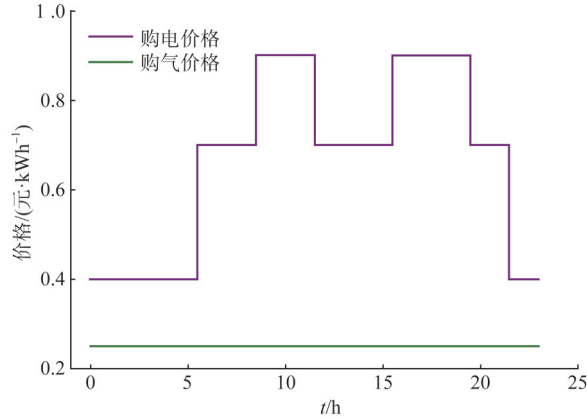


图 4 购电价格与购气价格随时段的变化曲线

Fig. 4 Curves of electricity purchase prices and gas purchase prices

验分析超参数的取值,选取数据 $\alpha \in (-1, -0.5, -0.2, -0.1)$, $\lambda \in (2, 2.5, 3, 3.5)$ 进行检测。将算法的累计奖励值作为性能评判指标,16种组合得到的结果见表4。由表4可知,当参数组合为 $\alpha = -0.5$, $\lambda = 3$ 时,算法能获得最高的累计奖励值,因此后续实验取 $\alpha = -0.5$, $\lambda = 3$ 。

3.2 模型改进机制的有效性验证

为了验证所提模型中阶梯式奖惩碳交易机制的有效性,将其从模型中移除并将碳价格固定为原始价格,且对BEO不采取碳配额限制,得到的模型

表 4 α 与 λ 的正交实验

Tab. 4 Orthogonal experiment of α and λ

λ	α			
	-1	-0.5	-0.2	-0.1
2	4 499	4 798	4 760	4 588
2.5	4 380	4 820	4 804	4 678
3	4 456	4 858	4 789	4 751
3.5	4 650	4 784	4 720	4 599

记为BEO_Stackelberg。为了验证所提模型中引入主从博弈框架的有效性,将其移除,即用户不再把响应负荷信息反馈给BEO,得到的模型记为BEO_carbon。将它们与所提模型BEO_proposed进行对比,以上3种模型采用相同的网络架构,以BEO的总收益、算法进入稳态的收敛代数和累计奖励值(算法最后一次训练代中各时间戳上奖励值之和)作为评判指标,如表5所示。

表 5 3种模型的比较结果

Tab. 5 Comparison results of the three models

项目	总收益/元	收敛代数	累计奖励值/元
BEO_proposed	24 288	750	4 858
BEO_stackelberg	22 480	800	4 496
BEO_carbon	19 015	900	3 804

由表5可知,所提模型的总收益24 288元,比BEO_Stackelberg模型提高了7.8%,比BEO_carbon模型提高了27.7%。导致这种现象的原因是所提阶梯式奖惩碳交易机制根据碳配额和实际碳排放的关系对BEO进行了不同程度的奖励和惩罚,鼓励算法决策出碳排放量更低的调度方案,从而提高了BEO的总收益和算法的累计奖励值。主从博弈框架使BEO和用户通过不断响应与交互,且在每一次交互中及时调整各自策略,最终达到Stackelberg均衡,获取最优方案,从而提高了BEO收益和累计奖励值。此外,BEO_proposed和BEO_stackelberg的收敛速度优于BEO_carbon,即引入主从博弈框架提升了智能体的收敛速度,原因是主从博弈使BEO获得用户的用能负荷反馈后能够及时调整能源价格,有助于算法更快地收敛到最优策略。上述实验结果表明,所提模型中的阶梯式奖惩碳交易机制和主从博弈框架是有效的。

3.3 算法改进机制的有效性验证

为了验证所提算法DDPG_AE中自适应动作探

索机制的有效性,分别将自适应动作探索机制用动作选择机制中的原始噪声策略和无噪声策略替代,形成了DDPG_normal和DDPG_nonoise两种对比算法。原始噪声策略的实现方法见式(60)。无噪声策略是指在选取动作时不添加噪声而直接选取actor网络中输出的当前最优动作。3种算法的累计奖励值随训练代数的变化曲线如图5所示,3种算法的总收益、收敛代数和累计奖励值如表6所示。

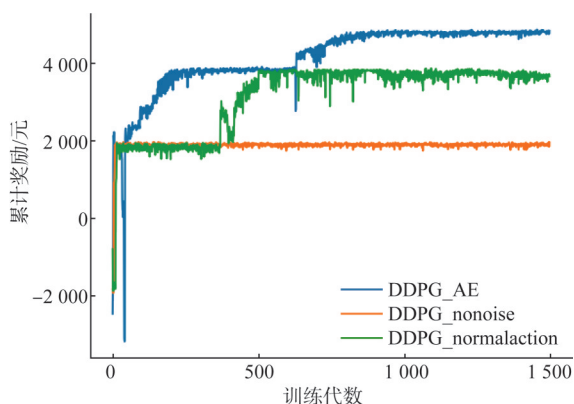


图5 3种算法的累计奖励值随训练代数的变化曲线
Fig. 5 Curves of cumulative reward values versus training generations for the three algorithms

表6 3种算法的比较结果

Tab. 6 Comparison results of the three algorithms

项目	总收益/元	收敛代数	累计奖励值/元
DDPG_AE	24 288	750	4 858
DDPG_nonoise	9 540	10	1 908
DDPG_normal	19 490	530	3 898

由图5和表6可见,所提算法DDPG_AE比其余2种算法获得了更好的收敛精度和总收益,DDPG_nonoise和DDPG_normal的收敛速度虽然更快,但它们均陷入了局部最优。所提算法在训练前期,DDPG_AE的累计奖励值与DDPG_normal无明显差异,但从650代开始,它获取了更高的累计奖励值并逐渐上升最后收敛到4858附近。

由于在训练前期,累计奖励值的方差和critic网络的损失值均较大,所提自适应动作探索机制与原始的动作选择策略作用相似;进入中期后,累计奖励值逐渐平稳,方差减小,而critic网络的损失值仍较大,使得自适应探索系数 A 增大,对动作的探索范围变广;训练后期,critic网络的损失值不断下降,系数 A 逐渐减小,对最优动作的利用增强,

最终收敛。DDPG_nonoise的曲线从初始阶段到结束均维持在很低的累计奖励值上,这是因为它每次只选取当前最优动作,智能体不具备探索性能,从而陷入局部最优。上述结果验证了所提自适应动作探索机制的有效性。

3.4 所提算法的性能验证

为了验证所提算法DDPG_AE的性能,文献[26–29]分别采用近端策略优化算法PPO和双延迟深度确定性策略梯度算法TD3实现综合能源系统的低碳经济调度,将所提算法与经典的PPO和TD3算法进行对比。所提算法结合了主从博弈框架,因此在PPO和TD3中也引入主从博弈框架,构成另两种对比算法:PPO_Stackelberg和TD3_Stackelberg。5种算法的比较结果如表7所示,种算法累计奖励值随训练代数的变化如图6所示。

表7 DDPG_AE与4种代表性算法的比较结果

Tab. 7 Comparison results of DDPG_AE with four representative algorithms

项目	总收益/元	收敛代数	训练时间/min	计算时间/s	累计奖励值/元
DDPG_AE	24 288	750	183	0.25	4 858
PPO	15 305	1 300	151	0.26	3 061
PPO_Stackelberg	19 080	800	160	0.28	3 816
TD3	11 300	900	188	0.22	2 660
TD3_Stackelberg	18 390	600	195	0.3	3 678

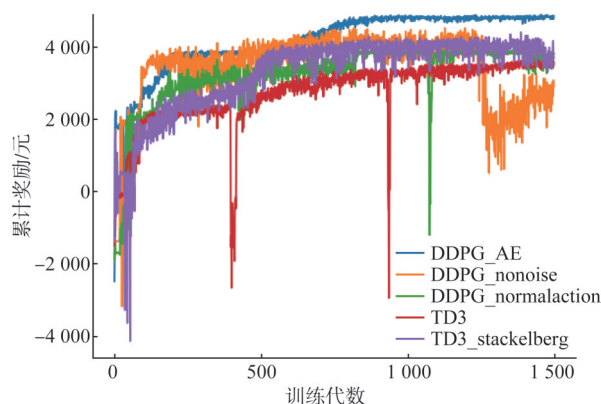


图6 DDPG_AE与4种代表性算法的累计奖励值随训练代数的变化曲线

Fig. 6 Curves of cumulative reward values versus training generations for DDPG_AE and four representative algorithms

由图6和表7可见,所提算法DDPG_AE的收敛精度、稳定性和总收益均优于4种对比算法,TD3_Stackelberg的收敛速度虽然更快,但它陷入了

局部最优。PPO_Stackelberg 和 TD3_Stackelberg 的总收益、累计奖励值和收敛代数均优于 PPO 和 TD3, 这表明了对于不同算法中 BEO 与用户均能在供需主从博弈框架的每一次交互中即时调整各自的策略(BEO 制定的价格与用户的最佳响应负荷), 有助于算法更快地收敛到更优的策略, 进一步验证了主从博弈框架对求解 BIES 优化问题的有效性。结合表 6, DDPG_normal 的性能与 PPO_Stackelberg 和 TD3_Stackelberg 相当, 而 DDPG_AE 表现更优, 说明所提自适应动作探索机制能够有效平衡算法对动作的探索和利用, 从而提升了算法的收敛精度。从训练时间角度, 不同于 DDPG 和 TD3 的离线策略方法, PPO 作为在线策略方法只使用当前策略生成的数据进行训练, 使得它的训练速度更快, 但其样本利用率更低, 难以找到最优策略。TD3 虽然也是离线策略方法, 但它使用了两套网络表示不同的 Q 值, 因此需要用到 6 个网络, 训练时间比 DDPG 更长。训练好后的 TD3 智能体计算最优方案时间略快。

3.5 最优调度方案展示

图 7 展示了电能优化调度结果, 折线图给出了每个时段需求侧对应的用户负荷, 柱状图则给出了每个时段供给侧的电能调度方案。以 5 点—6 点为例, 需求侧电负荷为 395 kWh, BEO 向上层电网出售了 70 kWh 的电力并通过 EB 机组把 35 kWh 的电能转换为了热能。同时, DG 机组、CHP 机组、WT 机组和蓄电池分别向需求侧贡献 180 kWh、38 kWh、230 kWh 和 52 kWh 的电能。由图 7 可知, 在夜间时需求侧负荷处于低谷期, 在午间和晚间的用电高峰期需求侧电负荷增大, 18 点后需求侧电负荷呈递减趋势。夜间时储电系统 ESS 处于充电状态, 而到了白天工作时段用电高峰期才进行放电。从上级电网购入的电力在大多时段是被充分利用的, 只在少数时段将剩余电力卖回给上级电网。正午时段的太阳能 PV 发电量充足, 从而减小了其余发电机组的发电功率。风能 WT 全天发电且发电功率处于较高水平, 热电联产 CHP 则是在夜间为用户提供更多的电能。

类似地, 图 8 展示了热能优化调度结果, 折线图 and 柱状图分别给出了需求侧热负荷和热能调度方案。需求侧热负荷的高峰期在夜间, 基本维持在 300 kWh 左右, 燃气锅炉 GB 在夜间给用户提供了

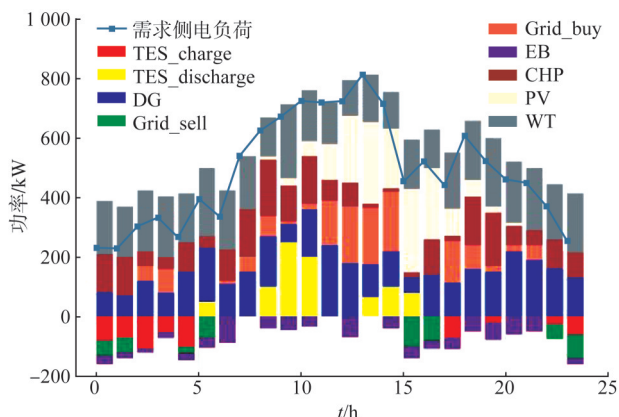


图 7 电能优化调度结果

Fig. 7 Electric energy optimization scheduling results

更多的热量。由于白天阳光照射的缘故, 需求侧热负荷较低, 且在 17 点达到最低值。

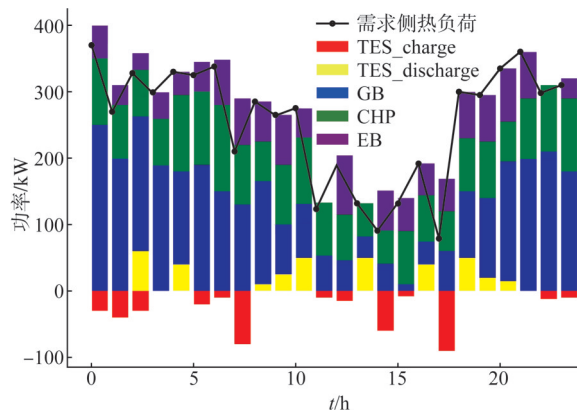


图 8 热能优化调度结果

Fig. 8 Heat energy optimization scheduling results

图 9 给出了需求响应前后 BEO 制定的电价对比图。由图 9 可知, 用户对用电高峰期的电价更加敏感, 他们通过需求响应将可转移电负荷移至夜间, 所以用电高峰期的电价在主从博弈后显著降低, 而非高峰期的电价随着负荷的增加相应增高。类似地, 图 10 显示了需求响应前后 BEO 制定的热价对比图。

4 结语

本文结合主从博弈框架和阶梯式奖惩碳交易机制, 建立建筑综合能源系统优化模型, 设计引入自适应动作探索机制的 DDPG 算法求解所建模型。实验结果表明, 模型和算法中的改进机制是有效的, 与代表性算法相比, 所提算法能够获得更好的收敛精度、稳定性和总收益。本文也存在不足之处: 1)

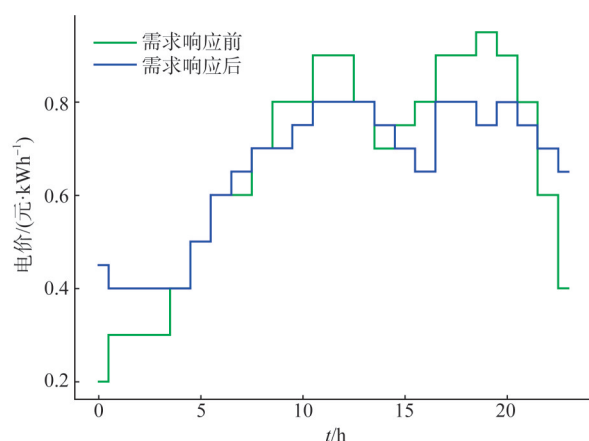


图9 BEO制定的零售电价

Fig. 9 Retail electricity prices of BEO

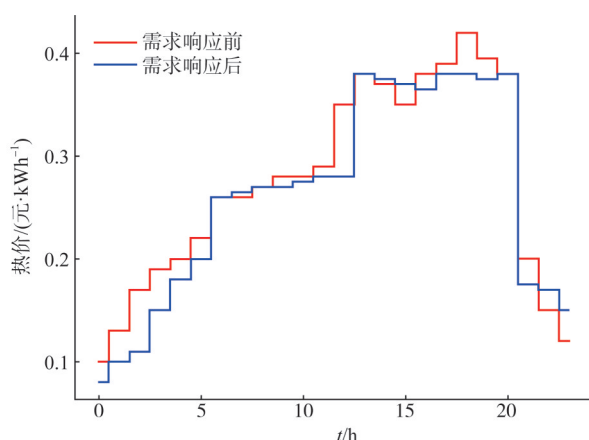


图10 BEO制定的零售热价

Fig. 10 Retail heat prices of BEO

所提算法未考虑能源设备分布的拓扑结构,可利用图神经网络提取系统的拓扑特征,进一步提高系统的收益并降低碳排放量;2)所提算法的运行时间较长,可结合迁移学习加快算法收敛速度。

参考文献

- [1] JIANG Q, WANG H, KONG Q, et al. On-orbit remote sensing image processing complex task scheduling model based on heterogeneous multiprocessor [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023 (61): 1001718.1 – 1001718.11.
- [2] YU L, QIN S, ZHANG M, et al. A review of deep reinforcement learning for smart building energy management[J]. IEEE Internet of Things Journal, 2021, 8(15): 12046 – 12063.
- [3] NTONTIN K, BOULOGGEORGOS A, BJÖRNSON E, et al. Wireless energy harvesting for autonomous reconfigurable intelligent surfaces [J]. IEEE Transactions on Green Communications and Networking, 2023, 7(1): 114 – 129.
- [4] BARBATO A, CAPONE A. Optimization models and methods for demand-side management of residential users: A survey[J]. Energies, 2014, 7(9): 5787 – 5824.
- [5] 任炬光, 张力, 金立, 等. 考虑可再生能源消纳的建筑综合能源系统日前经济调度模型[J]. 工程科学与技术, 2023, 55(2): 160 – 170.
REN Juguang, ZHANG Li, JIN Li, et al. Day-ahead economic dispatch model of building integrated energy systems considering the renewable energy consumption[J]. Advanced Engineering Sciences, 2023, 55(2): 160 – 170.
- [6] 石文喆, 李冰洁, 尤培培, 等. 基于深度强化学习的建筑能源系统优化策略[J]. 中国电力, 2023, 56(6): 114 – 122.
SHI Wenzhe, LI Bingjie, YOU Peipei, et al. Optimization strategy of building energy system based on deep reinforcement learning[J]. Electric Power, 2023, 56(6): 114 – 122.
- [7] 高力强, 费若雯, 刘敏, 等. 含有氢能汽车和储能的新型光伏建筑能源系统建模[J]. 西安建筑科技大学学报(自然科学版), 2022, 54(3): 414 – 422.
GAO Liqiang, FEI Ruowen, LIU Min, et al. Modelling of a novel photovoltaics building energy system considering hydrogen vehicles and energy storage [J]. Journal of Xi'an University of Architecture & Technology (Natural Science Edition), 2022, 54(3): 414 – 422.
- [8] 张大海, 负毓韵, 王小君, 等. 计及光热电站及建筑热平衡的冷热电综合能源系统优化运行[J]. 高电压技术, 2022, 48(7): 2505 – 2514.
ZHANG Dahai, YUAN Wenyun, WANG Xiaojun, et al. Operational optimization of integrated cooling, heating and power energy system considering concentrating solar power plant and heat balance of building [J]. High Voltage Engineering, 2022, 48(7): 2505 – 2514.
- [9] 刘晓华, 张涛, 刘效辰, 等. 面向双碳目标的建筑能源系统再认识[J]. 力学学报, 2023, 55(3): 699 – 709.
LIU Xiaohua, ZHANG Tao, LIU Xiaochen, et al. Rethinking of the building energy systems towards the carbon neutral target [J]. Chinese Journal of Theoretical and Applied Mechanics, 2023, 55(3): 699 – 709.
- [10] 陈淑琴, 陆敏艳, 谭洪卫, 等. 基于多目标优化的办公建筑可再生能源系统集成优化配置方案研究[J]. 太阳能学报, 2018, 39(11): 3147 – 3154.
CHEN Shuqin, LU Minyan, TAN Hongwei, et al. Research on integration renewable energy systems in office building based on multi-objective optimization[J]. Acta Energetica Sinica, 2018, 39(11): 3147 – 3154.
- [11] 李香龙, 赵乐, 王瀚秋, 等. 基于柔性负荷和建筑多能源系统的电网峰谷调节算法[J]. 沈阳工业大学学报, 2023, 45(4): 361 – 365.
Li Xianglong, ZHAO Le, WANG Hanqiu, et al. Power grid peak and valley adjustment algorithm based on flexible load and building multi-energy system [J]. Journal of Shenyang University of Technology, 2023, 45(4): 361 – 365.
- [12] 徐占伯, 周春翔, 吴江, 等. 基于边云协同的建筑能源系统分布式供需协同优化[J]. 中国科学: 信息科学, 2023, 53(3): 517 – 534.
XU Zhanbo, ZHOU Chunxiang, WU Jiang, et al. Edge-cloud framework-based distributed optimization of a building energy system with supply-demand coordination [J]. Scientia Sinica (Informationis), 2023, 53(3): 517 – 534.
- [13] 安佳坤, 贺春光, 刘洪, 等. 基于强化学习的建筑集群需求侧能量管理方法[J]. 电力建设, 2021, 42(5): 16 – 26.
AN Jiakun, HE Chunguang, LIU Hong, et al. Demand-side

- energy management method for building clusters applying reinforcement learning[J]. Electric Power Construction, 2021, 42(5): 16–26.
- [14] WANG, D, LIU B, JIA H, et al. Peer-to-peer electricity transaction decisions of the user-side smart energy system based on the SARSA reinforcement learning[J]. CSEE Journal of Power and Energy Systems, 2022, 8(3): 826–837.
- [15] 李媛, 迟昆, 王洲, 等. 基于强化学习的电-气-热多微网系统定价策略[J]. 南方电网技术, 2024, 18(1): 94–101.
LI Yuan, CHI Kun, WANG Zhou, et al. Pricing strategy for electric-gas-heat multi-microgrid system based on reinforcement learning[J]. Southern Power System Technology, 2024, 18(1): 94–101.
- [16] LIANG Z, HUANG C, SU W, et al. Safe reinforcement learning-based resilient proactive scheduling for a commercial building considering correlated demand response [J]. IEEE Open Access Journal of Power and Energy, 2021, 8(3): 85–96.
- [17] PENG Y, SHEN H, TANG X, et al. Energy consumption optimization for heating, ventilation and air conditioning systems based on deep reinforcement learning [J]. IEEE Access, 2023(11): 88265–88277.
- [18] PLAPPERT M, HOUTHOOFT R, DHARIWAL P, et al. Parameter space noise for exploration[DB/OL]. (2017–6–6) [2018–1–31]. <https://arxiv.org/abs/1706.01905>.
- [19] LI M, HUANG T, ZHU W. Adaptive exploration policy for exploration–exploitation tradeoff in continuous action control optimization[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(12): 3491–3501.
- [20] YANG P, JIANG H, LIU C, et al. Coordinated optimization scheduling operation of integrated energy system considering demand response and carbon trading mechanism [J]. International Journal of Electrical Power & Energy Systems, 2023(147): 108902.1–108902.10.
- [21] 叶宇静, 邢海军, 米阳, 等. 考虑低碳需求响应及主从博弈的综合能源系统低碳优化调度[J]. 电力系统自动化, 2024, 48(9): 34–43.
YE Yujing, XING Haijun, MI Yang, et al. Low-carbon optimal scheduling of integrated energy system considering low-carbon demand response and stackelberg game [J]. Automation of Electric Power Systems, 2024, 48(9): 34–43.
- [22] FENG C, WANG Y, ZHENG K. Smart meter data-driven customizing price design for retailers[J]. IEEE Transactions on Smart Grid, 2020, 11(3): 2043–2054.
- [23] LI K, YE N, LI S, et al. Distributed collaborative operation strategies in multi-agent integrated energy system considering integrated demand response based on game theory[J]. Energy, 2023(273): 127137.1–127137.11.
- [24] 王俐英, 林嘉琳, 董厚琦, 等. 计及阶梯式碳交易的综合能源系统优化调度[J]. 系统仿真学报, 2022, 34(7): 1393–1404.
- WANG Liying, LIN Jialin, DONG Houqi, et al. Optimal dispatch of integrated energy system considering ladder-type carbon trading [J]. Journal of System Simulation, 2022, 34(7): 1393–1404.
- [25] SHENGREN H, SALAZAR E M, VERGARA P P, et al. Performance comparison of deep RL algorithms for energy systems optimal scheduling [C]//2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), October 10–12, 2022, Novi Sad, Serbia. New York: IEEE, 2022: 1–6.
- [26] 王桂兰, 张海晓, 刘宏, 等. 基于近端策略优化算法含碳捕集的综合能源系统低碳经济调度[J]. 计算机应用研究, 2024, 41(5): 1508–1514.
WANG Guilan, ZHANG Haixiao, LIU Hong, et al. Low carbon economic scheduling of integrated energy systems based on proximal policy optimization algorithm with carbon capture [J]. Application Research of Computers, 2024, 41(5): 1508–1514.
- [27] 彭寒梅, 赵长桥, 谭貌, 等. 基于多智能体深度强化学习的大容量电池储能电站功率分配策略[J]. 南方电网技术, 2025, 19(9): 82–93.
PENG Hanmei, ZHAO Changqiao, TAN Mao, et al. Power allocation strategy for large-capacity battery energy storage power station based on multi-agent deep reinforcement learning [J]. Southern Power System Technology, 2025, 19(9): 82–93.
- [28] 唐文虎, 韦慈航, 柳洲, 等. 考虑负荷削减公平性的配电网两阶段韧性提升策略[J]. 广东电力, 2025, 38(9): 52–67.
TANG Wenhui, WEI Cihang, LIU Zhou, et al. Two-stage resilience enhancement strategy for distribution networks considering fairness of load shedding [J]. Guangdong Electric Power, 2025, 38(9): 52–67.
- [29] 李嘉文, 余涛, 张孝顺, 等. 基于改进深度确定性梯度算法的 AGC 发电功率指令分配方法[J]. 中国电机工程学报, 2021, 41(21): 7198–7212.
LI Jiawen, YU Tao, ZHANG Xiaoshun, et al. AGC power generation command allocation method based on improved deep deterministic policy gradient algorithm[J]. Proceedings of the CSEE, 2021, 41(21): 7198–7212.

收稿日期: 2024-05-16; 网络首发日期: 2024-09-29

作者简介:

申晓宁(1981), 女, 通信作者, 教授, 博士, 研究方向为计算智能、调度技术, sxnysyt@sina.com;

陈星晖(1998), 男, 硕士研究生, 研究方向为深度强化学习、综合能源系统, 362104236@qq.com;

陈文言(2000), 女, 硕士研究生, 研究方向为起启发式算法、软件项目调度。