计算机科学
COMPUTER SCIENCE

基于多智能体深度强化学习的光储充电站动态定价及能源调度策略

陈锦韬，林兵，林崧，陈静，陈星

引用本文

陈锦韬，林兵，林崧，陈静，陈星. 基于多智能体深度强化学习的光储充电站动态定价及能源调度策略[J]. 计算机科学，2025，52(9)：337-345.

CHEN Jintao, LIN Bing, LIN Song, CHEN Jing, CHEN Xing. Dynamic Pricing and Energy Scheduling Strategy for Photovoltaic Storage Charging Stations Based on Multi-agent Deep Reinforcement Learning [J]. Computer Science, 2025, 52(9): 337-345.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700197

基于多智能体深度强化学习的光储充电站动态定价及能源调度策略

陈锦韬[1,3]　林　兵[2,3]　林　崧[1]　陈　静[3]　陈　星[3]
1 福建师范大学计算机与网络空间安全学院　福州 350117
2 福建师范大学物理与能源学院　福州 350117
3 福建省网络计算与智能信息处理重点实验室　福州 350116
(963200594@qq.com)

摘　要　光储充电站运营收益的提升,能够使充电站运营商加大对光储充电站基础设施的投资和部署,从而缓解日益增长的电动汽车渗透到电网时带来的负荷压力。针对光储充电站的运营收益提升问题,提出了一种基于多智能体深度强化学习的动态定价及能源调度策略,旨在提高完全合作关系下光储充电站的整体运营收益。首先,以最大化所有光储充电站的总运营收益为目标,将在单个光储充电站运营商下的多个光储充电站和电动汽车建模成马尔可夫博弈模型;其次,采用多智能体双延迟确定性策略梯度算法进行模型求解,通过制定充电服务价格和储能系统的充放电策略,以达到总运营收益最大化的目标,并通过余弦退火方法对算法学习率进行调整,提升该算法的收敛速率和收敛阈值;最后,为防止完全合作关系下多站可能出现的价格垄断问题,引入反需求函数对充电服务价格进行约束。实验结果表明,所提策略和对比方法相比,提高了4.17%～66.67%的充电站运营收益,且所用的反需求函数能够有效预防多站的价格垄断问题。

关键词:多智能体深度强化学习;光储充电站;能源调度;动态定价;反需求函数
中图分类号　TP391

Dynamic Pricing and Energy Scheduling Strategy for Photovoltaic Storage Charging Stations Based on Multi-agent Deep Reinforcement Learning

CHEN Jintao[1,3],LIN Bing[2,3],LIN Song[1],CHEN Jing[3] and CHEN Xing[3]
1 College of Computer and Cyber Security,Fujian Normal University,Fuzhou 350117,China
2 College of Physics and Energy,Fujian Normal University,Fuzhou 350117,China
3 Fujian Key Laboratory of Network Computing and Intelligent Information Processing,Fuzhou 350116,China

Abstract　The improvement in the operational profits of photovoltaic storage charging stations(PSCSs) can enable charging station operators to increase their investment and deployment of PSCSs infrastructure,thereby alleviating the load pressure on the grid caused by the growing penetration of electric vehicles(EVs). To address the issue of improving PSCSs operational profits,a dynamic pricing and energy scheduling strategy based on multi-agent deep reinforcement learning(MADRL) is proposed to enhance the overall operational profits of PSCSs under a fully cooperative relationship. Firstly,aiming to maximize the total operational profits of all PSCSs,multiple PSCSs and EVs under a single PSCS operator are modeled as a Markov game. Secondly,the multi-agent twin delayed deep deterministic policy gradient(MATD3) algorithm is used to solve the model,setting the selling price of charging services and the charging and discharging strategies of the energy storage system(ESS) to achieve profit maximization. The cosine annealing method is employed to adjust the learning rate of the algorithm,improving its convergence rate and threshold. Finally,to prevent potential price monopolies under a fully cooperative relationship among multiple stations,an inverse demand function is introduced to constrain the selling price of charging services. Experimental results show that the proposed strategy improves the operational profits of charging stations by 4.17% to 66.67% compared to benchmark methods,and using

到稿日期:2024-07-30　返修日期:2024-11-29
基金项目:国家自然科学基金(62072108);福建省高校产学合作项目(2022H6024,2021H6026);福建省高校物理学学科联盟教学改革项目(FJPHYS-2022-B02);福建省促进海洋与渔业产业高质量发展专项资金(FJHYF-ZH-2023-02);福建省技术创新重点攻关及产业化项目(2024XQ004)
This work was supported by the Natural Science Foundation of China(62072108), University-Industry Cooperation of Fujian Province(2022H6024,2021H6026),Fujian University Physics Union(FJPHYS-2022-B02),Special Funds for Promoting High-quality Development of Marine and Fishery Industries in Fujian Province(FJHYF-ZH-2023-02) and Fujian Key Technological Innovation and Industrialization Projects(2024XQ004).
通信作者:林兵(WheelLX@163.com)

the inverse demand function effectively prevents price monopolies among multiple stations.

Keywords Multi-agent deep reinforcement learning, Photovoltaic storage charging station, Energy scheduling, Dynamic pricing, Inverse demand function

1 引言

随着全球环境与能源问题的日益凸显，电动汽车(Electric Vehicles，EVs)作为一种环境友好型的交通工具，受到了广泛关注，并得到了快速发展。在我国，新能源汽车的保有量近年来显著增长，截至2022年已超过1310万辆，其中纯EVs的占比高达79.79%[1]。这一增长也带来了新的挑战：如此数量的EVs会产生庞大的充电需求，且其充电行为具有随机性和突发性，当其大规模渗透到电网中时，会给电网带来巨大的负担和调节压力。而结合光伏(Photovoltaic，PV)系统和储能系统(Energy Storage System，ESS)的光储充电站(Photovoltaic Storage Charging Station，PSCS)不仅能促进PV发电的就近消纳，通过“削峰填谷”平抑站内负荷波动，还可利用储能的灵活调节能力获取经济利益，对于缓解EV大规模入网带来的压力具有重要意义[2]。

随着PSCS基础设施的发展和电力零售市场的自由化，EVs用户可以通过多种方法来优化其充电成本，例如前往充电服务价格较低的PSCS充电、选择充电服务价格较低的时段充电，而充电站运营商可以通过能源调度策略控制ESS充放电以降低运营成本，同时利用售价策略来获得盈利[3]。因此制定合理的PSCS充电服务价格和能源调度策略，可以极大地提升PSCS的运营收益，从而加速充电站运营商对PSCS基础设施的投资和部署，对于缓解EVs大规模入网的压力和推动EVs行业的可持续发展具有重要意义。

PSCS能源调度策略能通过控制ESS的充放电行为，使其在电网电价较低时储能，在用电高峰时段供电，做到“低储高发”，还能实现PV发电与负荷需求的动态匹配，减小弃光率，从而降低运营成本以提升充电站的盈利能力。而制定一个合理的EVs充电服务价格可以确保充电站在运营过程中实现成本回收和盈利，同时也能够维持行业的健康竞争环境。目前关于PSCS同时制定能源调度策略和EVs充电服务价格的研究较少，且大多研究针对单一站点或只考虑其中一点[4-6]，而对于单个充电站运营商而言，其对充电站管理是多个同时进行的，站与站之间是完全合作关系，但现今对于多站的研究大多是建立在非同充电站运营商下的非合作关系[7-8]，且没有相关研究提出解决多站完全合作关系下可能出现的价格垄断问题。

针对上述问题，本文采用反需求函数解决多PSCS在完全合作关系下可能出现的价格垄断问题，利用多智能体深度强化学习(Multi-Agent Deep Reinforcement Learning，MADRL)方法动态制定多PSCS对EVs充电服务价格和能源调度策略，以实现整体运营收益的最大化。本文的主要贡献如下。

1)本文以最大化充电站运营收益为目标，将多个PSCS和EVs建模成马尔可夫博弈模型，用模型模拟环境与各主体之间的信息交互，并采用反需求函数建立起充电站充电服务价格与EVs用户充电需求之间的联系，该函数确保了随着充电服务价格的提高，充电需求相应减少。

2)本文采用MADRL中的多智能体双延迟确定性策略梯度(Multi-Agent Twin Delayed Deep Deterministic Policy Gradient，MATD3)[9]算法对模型求解，且为了提高收敛速率和阈值，引入了余弦退火方法来对学习率进行调整。该方法通过优化学习率的衰减过程，成功提升了算法的训练效率和寻优能力。

3)仿真实验结果表明，本文改进的MATD3算法在多个PSCS的动态定价和能源调度方面表现良好，具有较高的实用参考价值。同时，实验还验证了所采用的反需求函数能有效防止完全合作关系下充电站之间出现价格垄断现象。

2 相关工作

目前，对PSCS收益优化的研究主要分为规划和运营两大方向。规划主要通过分析EVs的特性，合理选择PSCS的位置和规模，以降低建设和运营成本，提高利用率[10-11]；运营则主要从动态定价[12-13]、能源调度[14-15]和EVs充电调度[16-17]3个方面提高PSCS的整体收益。针对运营方面，现有研究大多基于分时定价或动态定价、EVs充电需求等因素，利用需求响应、储能出力调节、调整电网购电量等方式，通过优化方法最大化PSCS收益。而根据所用优化方法，主要可分为启发式方法和强化学习(Reinforcement Learning，RL)方法。

对于启发式方法，Yu等[18]开发了一个PV微电网优化模型，针对分时电价和需求响应，考虑了功率平衡和储能等多重约束。通过非支配排序遗传算法，他们优化了系统运行成本和电网交换电量，提升了经济效益并优化了电网负荷平衡。Hao[19]考虑功率平衡和EVs充放电功率、PV出力、电网交互功率、峰谷电价差，以PSCS单日利润最大为目标，采用遗传算法求解，不仅能提高PSCS的经济收益，还有助于电网的稳定运行。Su等[20]基于EVs电池的荷电状态，提出了一个综合考虑PV发电、EVs充放电、电价波动和储能状态的能源管理策略。他们采用混合遗传算法和粒子群优化(GAPSO)来优化EVs充放电计划，以减小微电网运营成本的方式来提高利润。Lyu等[21]构建了一个基于能源成本和收益的经济函数，且引入PV利用率和电池退化作为可持续性目标，利用预测性全局优化算法(PGO)来求解这个经济函数，并采用自适应策略来协调经济目标与可持续性目标之间的平衡，有效地提升了PSCS的经济性。

上述运营优化方法都是利用已有数据进行数学建模，这种启发式方法或是博弈论方法，其通常假设参与者的策略是固定的，当环境动态变化时，固定策略可能无法适应快速变化的环境。而现实场景下，PSCS的运营环境具有非线性、动态特性，且状态空间较大，如大量EVs用户的行为决策、ESS的充放电状态，并不适合上述方法。RL方法作为一种无模型方法，其能通过学习机制适应环境的非线性和动态特性，逐步

优化策略，在复杂的多智能体环境中实现实时的决策优化，从而提供更具鲁棒性和灵活性的解决方案。

在 RL 方面，Muriithi 等[22]利用 RL 解决了可再生能源和微电网之间的不稳定性，为了确保实际的应用，将能源管理系统制定成马尔可夫决策过程（Markov Decision Process，MDP），并考虑分时电价和固定电池的退化成本，通过实现能源成本最小化来提升利润。Cui 等[23]通过建立一个车辆-道路模拟模型，利用深度 RL 求解，提高了快速充电站的利润，缓解了高峰期的充电需求压力和路网拥堵，提高了用户对充电的满意度。Lee 等[24]提出三阶段隐私和安全感知深度 RL 框架，通过对 ESS 的充放电、数据加密、配电系统运营商的充放电 3 个阶段阶梯式递进，最大限度地降低功率损耗并提高充电站利润。Qian 等[25]提出了 MADRL 方法来学习多个充电站的充电定价策略，并使用不完全信息来近似定价博弈的纳什均衡，旨在确定单个充电站的最优充电价格，有效降低充电站之间的竞争并提高利润。尽管上述研究中的模型比启发式算法更能灵活应对充电需求和电价的波动，但其研究范围限定在单个 PSCS 或非同充电站运营商下的多个 PSCS。这些模型尚未涵盖单个充电站运营商下多个 PSCS 在合作模式下的互动，也未考虑如何避免在这种合作环境中可能出现的价格垄断现象。

综上所述，本文对单个运营商下的多个 PSCS 以完全合作关系建模，利用反需求函数防止价格垄断问题，以最大化所有 PSCS 运营收益为目标，通过 MADRL 方法求解对应的最佳的充电服务价格和能源调度策略。

3 系统模型

本文系统模型如图 1 所示，主要包含 4 部分。

1）主电网 P：主电网以分时电价向 PSCS 进行售电。

2）单个充电站运营商下的多个光储充电站 $PSCS=\{g_1, g_2, \cdots, g_n\}$：PSCS 由 PV 系统、ESS 和充电桩组成。PV 系统负责将太阳能转换为电能；ESS 用于存储 PV 系统和电网传输的电能；充电桩仅作为能源传输的媒介。

3）电动汽车 $EVs=\{e_1, e_2, \cdots, e_m\}$：EVs 包含当前提出充电需求的未充电 EVs 与正在站内充电的 EVs。

4）能源调度中心 SC：负责记录和监控每个 PSCS 产生的所有数据，并决策其购电量以及 ESS 的充放电量。

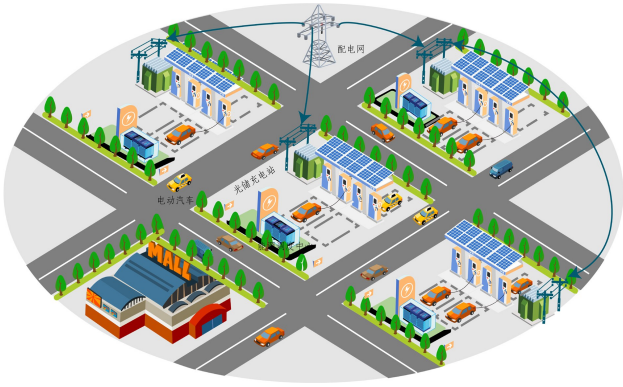


图 1 系统模型

Fig. 1 System model

本文将一天分成 T 个时段，在每个时段中模型各组件的交互流程如下。

1）能源调度中心会在每个时段前生成每个 PSCS 当前时段的充电服务价格，并将其发布给 EVs 车主的终端设备。

2）EVs 车主会根据每个 PSCS 发布的价格信息进行选择，未充电的车主会考虑前往哪个 PSCS，正在充电的车主会选择是否继续充电，两者会将选择信息（包括到达时间、充电功率、离站时间、是否继续充电等）发送给能源调度中心。

3）能源调度中心会将收集的信息整合，计算每个 PSCS 当前时段的总需求消耗（含上一时段未满足量），再根据供需情况和主电网的分时电价向主电网进行适当购电。

4）在时段结束后，能源调度中心会根据每个 PSCS 售电收益和购电成本以及损耗成本，计算总收益。

综上，可得本文模型的能量流和信息流流向，如图 2 所示。

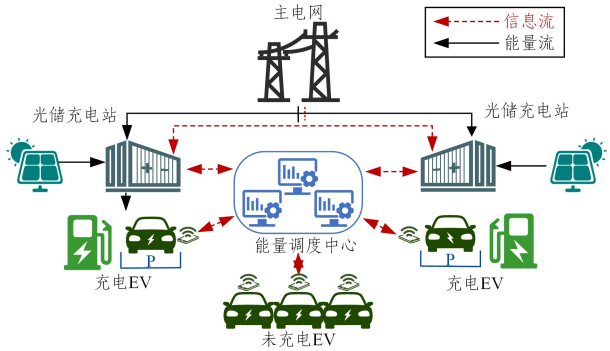


图 2 模型能量流与信息流流向

Fig. 2 Modeling energy flow and information flow direction

3.1 目标函数

本文的优化问题目标为最大化所有 PSCS 的总收益，其定义如下：

$$\max\ r=\sum_{t=0}^{T}\sum_{i=1}^{N}u_{i,t}-c_{i,t}^{\mathrm{grid}}-c_{i,t}^{\mathrm{ess}} \tag{1}$$

$$u_{i,t}=q_{i,t}^{\mathrm{ev}}\cdot\pi_{i,t}^{\mathrm{ev}} \tag{2}$$

$$c_{i,t}^{\mathrm{grid}}=q_{i,t}^{\mathrm{grid}}\cdot\pi_{P,t}^{\mathrm{grid}} \tag{3}$$

$$c_{i,t}^{\mathrm{ess}}=\alpha\cdot\beta\cdot q_{i,t}^{\mathrm{ess}} \tag{4}$$

其中，t 为当前时段，$u_{i,t}$ 是第 i 个 PSCS g_i 在 t 时段售给 EVs 的总收益；$c_{i,t}^{\mathrm{grid}}$ 是 g_i 在 t 时段向主电网的购电成本；$c_{i,t}^{\mathrm{ess}}$ 是 g_i 的 ESS 在 t 时段的损耗成本；$q_{i,t}^{\mathrm{ev}}$ 为 g_i 在 t 时段内售给 EVs 的总电量；$\pi_{i,t}^{\mathrm{ev}}$ 为 g_i 在 t 时段的充电服务价格；$q_{i,t}^{\mathrm{grid}}$ 为 g_i 在 t 时段向主电网的购电量；$\pi_{P,t}^{\mathrm{grid}}$ 是主电网 P 在 t 时段的分时电价；α 是储能充放电效率，β 是储能单位电量交换的损耗成本。

3.2 约束条件

本文的约束条件如下。

能量平衡约束：由于主电网、PV 系统存在功率上限，ESS 存在容量上限，同时根据图 2 的能量流流向和供需平衡，可得能量平衡约束式如下：

$$q_{i,t}^{\mathrm{pv}}+q_{i,t}^{\mathrm{ess}}+q_{i,t}^{\mathrm{grid}}=q_{i,t}^{\mathrm{ev}} \tag{5}$$

其中，$q_{i,t}^{\mathrm{pv}}$ 表示 t 时段中 g_i 的 PV 系统的发电量。

ESS 约束：将 ESS 的充放电量 $q_{i,t}^{\mathrm{ess}}$ 约束至可供给范围内（$q_{i,t}^{\mathrm{ess}}$ 值为正代表放电，值为负代表充电），其约束式如下：

$$q_{i,t}^{\mathrm{ess}}=\begin{cases}\min(q_{i,t}^{\mathrm{ess}},q_{\max}^{\mathrm{ess}}), & q_{i,t}^{\mathrm{ess}}\geqslant 0\\ \max(q_{i,t}^{\mathrm{ess}},q_{\min}^{\mathrm{ess}}), & q_{i,t}^{\mathrm{ess}}<0\end{cases} \tag{6}$$

其中，$q_{\max}^{\mathrm{ess}}$ 和 $q_{\min}^{\mathrm{ess}}$ 分别表示 g_i 的 ESS 充放电量的上、下限。

PV 约束：为了尽量减少 PV 系统的弃光现象，本文优先消耗 PV 系统的发电量，其约束式如下：

$$q_{i,t}^{\mathrm{pv}}=\begin{cases}\min(q_{i,t}^{\mathrm{pv,ori}},q_{i,t}^{\mathrm{ev}}-q_{i,t}^{\mathrm{ess}}), & q_{i,t}^{\mathrm{ess}}<0\\ \min(q_{i,t}^{\mathrm{pv,ori}},q_{i,t}^{\mathrm{ev}}), & q_{i,t}^{\mathrm{ess}}\geqslant 0\end{cases} \tag{7}$$

其中，$q_{i,t}^{\mathrm{pv,ori}}$ 表示 g_i 的 PV 系统在 t 时段初始发电量，$q_{i,t}^{\mathrm{ev}}-q_{i,t}^{\mathrm{ess}}$ 表示 t 时段 PSCS 和 EVs 所需求的总电量。

主电网约束：主电网 P 是补充 PSCS 电能的主要来源之一，当 PSCS 的 ESS 充放电量与 PV 系统提供的电量少于 EVs 总需求电量时，PSCS 就会向主电网进行购电以保证供给，因此主电网的约束式如下：

$$q_{i,t}^{\mathrm{grid}}=\max(0,\min(q_{i,t}^{\mathrm{ev}}-q_{i,t}^{\mathrm{ess}}-q_{i,t}^{\mathrm{pv}},q_{\max}^{\mathrm{grid}})) \tag{8}$$

其中，$q_{i,t}^{\mathrm{ev}}-q_{i,t}^{\mathrm{ess}}-q_{i,t}^{\mathrm{pv}}$ 值为 PSCS 提供的电量与 EVs 总需求电量的差值，$q_{\max}^{\mathrm{grid}}$ 为主电网最大可供给量。

EVs 约束：EVs 总需求电量需在 PSCS 最大可供给范围之内，其约束式如下：

$$q_{i,t}^{\mathrm{ev}}=\min(q_{i,t}^{\mathrm{pv}}+q_{i,t}^{\mathrm{ess}}+q_{i,t}^{\mathrm{grid}},q_{i,t}^{\mathrm{ev,ori}}) \tag{9}$$

其中，$q_{i,t}^{\mathrm{ev,ori}}$ 为 EVs 的初始总需求电量。

3.3 EVs 选择计算

对于第 j 辆 EVs e_j，其选择为到各个 PSCS 的路径损耗和充电成本总和最小的 PSCS 进行充电。

$$w_{e_j}=\min(f(x_1^{e_j}),f(x_2^{e_j}),\cdots,f(x_i^{e_j}),\cdots,f(x_n^{e_j})) \tag{10}$$

$$f(x_i^{e_j})=(v_{j,t}^{\mathrm{energy,dep}}-(v_{j,t}^{\mathrm{energy}}-(l_{e_j}\cdot d_i^{e_j})))\cdot\pi_{i,t}^{\mathrm{ev}} \tag{11}$$

其中，w_{e_j} 表示 e_j 选择前往充电的 PSCS，$f(x_i^{e_j})$ 表示 e_j 到 g_i 的最小路径损耗和充电成本总和的函数；$v_{j,t}^{\mathrm{energy,dep}}$ 表示 e_j 充电结束时的电量百分比(由式(12)给出)；$v_{j,t}^{\mathrm{energy}}$ 表示 e_j 的当前电量百分比；l_{e_j} 表示 e_j 的每公里耗电百分比；$d_i^{e_j}$ 表示 e_j 和 g_i 之间的距离。

3.4 反需求函数

由于本文定价环境为单个充电站运营商下的多个 PSCS，因此为了防止 PSCS 之间出现价格垄断，扰乱充电市场环境与 EVs 可能前往其他充电站运营商充电而产生收益损失的现象，本文参考了经济学中反需求函数的概念[26]，设置了一个反需求函数，即把充电服务价格视作充电量的函数关系，具体如下：

$$v_{j,t}^{\mathrm{energy,dep}}=\mathrm{e}^{\frac{-(\pi_{i,t}^{\mathrm{ev}})^2}{2k^2}} \tag{12}$$

其中，函数的输入是充电服务价格 $\pi_{i,t}^{\mathrm{ev}}$，输出是 EVs 充电结束时的电量百分比 $v_{j,t}^{\mathrm{energy,dep}}$。该函数可限制 PSCS 的充电服务价格，使其越高时，EVs 车主的意愿充电量越低，从而防止出现完全合作关系下多站间的价格垄断现象。k 是该函数的系数，其值大小由实验过程对比而确定。

4 基于 MADRL 的动态定价和能源调度策略

4.1 马尔可夫博弈模型

在 RL 中，智能体通过与环境不断交互，利用奖励函数反映环境的变化与动作之间的关联，从而更新自身的策略，最终使奖励达到最大化。因此 RL 中通常把问题建模成一个包含 4 元素的 MDP：1)状态空间 $\boldsymbol{S}$；2)动作空间 $\boldsymbol{A}$；3)状态转移函数 f；4)奖励函数 R。

而 MARL 是将 RL 拓展至多智能体系统(Multi Agent System，MAS)中，同时也把 MDP 模型拓展成马尔可夫博弈模型。其马尔可夫博弈通常定义为 $\langle N,\boldsymbol{S},\{\boldsymbol{O}^i\}_{i\in N},\{\boldsymbol{A}^i\}_{i\in N},f,\{r^i\}_{i\in N}\rangle$，其中 N 为智能体的数量；$\boldsymbol{S}$ 代表所有智能体可能的状态空间，这是一个全局信息；$\{\boldsymbol{O}^i\}_{i\in N}$ 则为智能体从环境中得到自身的部分可观测信息；$\{\boldsymbol{A}^i\}_{i\in N}$ 是智能体 i 的动作集合；f 是环境的状态转移函数；$\{r^i\}_{i\in N}$ 是智能体 i 的奖励函数。而 MADRL 是在 MARL 中引入神经网络进行训练的一种方法，本文采用的是 MADRL 方法，因此对应的马尔可夫博弈模型的元素定义如下。

定义 1(智能体 N) 本文将 PSCS 作为与环境交互的智能体，EVs 则作为环境的一部分，其中单个智能体用 i 表示，智能体总数为 n。

定义 2(状态空间 $\boldsymbol{S}$) 环境的状态表示为 $\boldsymbol{s}_t=\{t,\boldsymbol{H}_t,\boldsymbol{\Pi}_t^{\mathrm{grid}},\boldsymbol{V}_t^{\mathrm{soc}},\boldsymbol{Q}_t^{\mathrm{pv}}\}$，其中 t 表示当前时段；$\boldsymbol{H}_t=\{h_{1,t},h_{2,t},\cdots,h_{i,t},\cdots,h_{n,t}\}$，$h_{i,t}$ 表示智能体 i 在 t 时段的 EVs 总需求电量；$\boldsymbol{\Pi}_t^{\mathrm{grid}}=\{\pi_{P,t}^{grid}\}$，$\pi_{P,t}^{\mathrm{grid}}$ 表示主电网 P 在 t 时段的主电网的分时电价；$\boldsymbol{V}_t^{\mathrm{soc}}=\{v_{1,t}^{\mathrm{soc}},v_{2,t}^{\mathrm{soc}},\cdots,v_{i,t}^{\mathrm{soc}},\cdots,v_{n,t}^{\mathrm{soc}}\}$，$v_{i,t}^{\mathrm{soc}}$ 表示智能体 i 在 t 时段 ESS 的 SOC(State of Charge)情况；$\boldsymbol{Q}_t^{\mathrm{pv}}=\{q_{1,t}^{\mathrm{pv}},q_{2,t}^{\mathrm{pv}},\cdots,q_{i,t}^{\mathrm{pv}},\cdots,q_{n,t}^{\mathrm{pv}}\}$，$q_{i,t}^{\mathrm{pv}}$ 表示智能体 i 在 t 时段 PV 系统的发电量。

定义 3(观测空间 $\{\boldsymbol{O}^i\}_{i\in N}$) 本文采用常见的环境假设，即假设环境对每个智能体是部分可观测的，因此每个智能体 i 只能观测自身的状态，故观测空间为 $\boldsymbol{o}_t^i=\{t,h_{i,t},\pi_{P,t}^{\mathrm{grid}},v_{i,t}^{\mathrm{soc}},q_{i,t}^{\mathrm{pv}}\}$。

定义 4(动作空间 $\{\boldsymbol{A}^i\}_{i\in N}$) 每个智能体的动作表示为 $\boldsymbol{a}_t^i=\{\lambda_{i,t},q_{i,t}^{\mathrm{ess}}\}$，其中 $q_{i,t}^{\mathrm{ess}}$ 是智能体 i 在 t 时段的 ESS 充放电量；$\lambda_{i,t}$ 是智能体 i 在 t 时段的动态自适应价格调整系数，而智能体 i 在 t 时段的充电服务价格则为 $(\pi_{i,t}^{\mathrm{ev}}+\lambda_{i,t})$，此时充电服务价格的取值范围为 $[\pi_{i,t}^{\mathrm{ev}}+\lambda_{\min},\pi_{i,t}^{\mathrm{ev}}+\lambda_{\max}]$。

定义 5(状态转移函数 f) 状态转移函数 $f(\boldsymbol{s}_t,\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)\rightarrow\boldsymbol{s}_{t+1}$ 是所有智能体在当前环境下采取各自的动作后，环境转换到下一个状态的概率。

定义 6(奖励函数 $\{r^i\}_{i\in N}$) 本文的奖励函数设计是由上文提到的目标函数。因为 MADRL 本质上也是在试错中不断地搜索最佳策略，为了加速搜索，本文增添一项新的储能超限惩罚 $c_{i,t}^{\mathrm{punish}}$ 来减少其对不好策略的搜索次数。同时为了更好学得充电服务价格对 EVs 总需求电量的影响和主电网分时电价与 ESS 的关联性，本文分别添加 $r_{i,t}^{\mathrm{energy}}$ 和 z_i 两个奖励引导函数，其中 z_i 是通过实验过程不断调整完善设置的一个函数。因此每个智能体 i 的奖励 r_i 表示如下：

$$r_t^i=u_{i,t}-c_{i,t}^{\mathrm{ess}}-c_{i,t}^{\mathrm{punish}}+r_{i,t}^{\mathrm{energy}}+z_i \tag{13}$$

其中，$c_{i,t}^{\mathrm{punish}}$ 的设置如下：

$$c_{i,t}^{\mathrm{punish}}=\begin{cases}\xi\cdot(v_{i,t}^{\mathrm{soc}}-v_{\max}^{\mathrm{soc}})\cdot C, & v_{i,t}^{\mathrm{soc}}>v_{\max}^{\mathrm{soc}}\\ \xi\cdot(v_{\min}^{\mathrm{soc}}-v_{i,t}^{\mathrm{soc}})\cdot C, & v_{i,t}^{\mathrm{soc}}<v_{\min}^{\mathrm{soc}}\end{cases} \tag{14}$$

其中，$v_{\max}^{\mathrm{soc}}$，$v_{\min}^{\mathrm{soc}}$ 分别表示 PSCS 的 ESS 的 SOC 上、下限阈值，C 表示 ESS 的容量，ξ 表示 ESS 充放电超限惩罚系数。而 $r_{i,t}^{\mathrm{energy}}$ 的表示如下：

$$r_{i,t}^{\mathrm{energy}}=\zeta\cdot h_{i,t} \tag{15}$$

其中,ζ 是经过多次实验后取最优值,设置为 8。

在 MADRL 中,其核心目标是通过优化策略,以最大化长期累积折扣奖励,即最大化折扣回报。

$$r_t=\sum_{i=1}^{n}\sum_{t=0}^{T}\gamma^t\cdot r_t^i \tag{16}$$

其中,γ 为折扣因子,是 MADRL 的一个超参数,用于反映其相对于即时奖励的不确定性和延迟性。γ 的值越接近 1,意味着未来奖励的当前价值越高;γ 的值越接近 0,意味着智能体更倾向于获取即时奖励而不是等待未来的奖励。

4.2 MATD3 算法

为了解决深度确定性策略梯度(Deep Deterministic Policy Gradient,DDPG)算法中只用单 Q 网络而存在 Q 值高估问题,提出了一种改进算法 TD3 算法。通过采用双 Q 网络和延迟策略更新等技术,TD3 有效地降低了价值估计的偏差,提高了算法的稳定性和学习效率。但 TD3 只适合单智能体系统,为了将 TD3 算法适配于 MAS,其进一步发展成 MATD3 算法。与单智能体 TD3 算法相比,MATD3 通过智能体之间的信息共享来优化多智能体系统中的策略协作,其属于 CTDE(Centralized Training Decentralized Execution)架构,在训练阶段通过集中化的 Critic 来评估多个智能体的动作,而在执行阶段,智能体则根据各自的局部信息进行独立决策。同时,MATD3 采用所有智能体共享一个经验回放池,相比于智能体独立拥有经验回放池,共享经验回放池更适配合作关系下的 MAS,其能够使智能体访问到整个环境中的所有经验,大大增加有效样本量,提高样本利用效率,并降低策略学习中的不稳定性,减少智能体在学习过程中由于环境变化而产生的波动,使算法缩短训练时间并提升决策效率。

MATD3 算法中每个智能体共有 6 个神经网络,包含一对 Actor 及对应目标网络和两对 Critic 及对应的目标网络。Actor 网络(网络用 μ 表示,网络参数设为 θ^{μ})用来训练策略,其动作生成方式如下:

$$\boldsymbol{a}_t^i=\mu_i(\boldsymbol{o}_t^i)+\mathcal{N}_{i,t} \tag{17}$$

因为在智能体早期的学习阶段,智能体对环境的知识是缺乏的,所以为了更多地探索环境,本文添加了一项高斯噪声 $\mathcal{N}_{i,t}$ 来扩充探索范围。

Critic 网络(网络用 $Q_{i,j}^{\mu}$ 表示,j 为网络的索引,网络参数设为 θ^{Q})用来评估智能体采取动作的好坏程度。Critic 网络的输入是一对状态-动作对($\boldsymbol{s}_t^i$,$\boldsymbol{a}_t^i$),输出是动作价值函数 $Q_{i,j}^{\mu}(\boldsymbol{s}_t^i,\boldsymbol{a}_t^i)$。

MATD3 算法对于单个智能体的 Actor 网络是通过策略梯度进行更新的。

$$\nabla_{\theta^{\mu_i}}J(\mu_i)\approx\frac{1}{B}\sum_{b=1}^{B}\nabla_{\theta^{\mu}}\mu_i(\boldsymbol{a}_b^i)\cdot\nabla_a Q_{i,1}^{\mu_i}(\boldsymbol{s}_b,\boldsymbol{a}_b^1,\cdots,\boldsymbol{a}_b^n)\,|\,\boldsymbol{a}_i=\mu_i(\boldsymbol{o}_i) \tag{18}$$

其中,B 表示从经验回放池采样的样本大小,b 则是采样数据的索引。

对于 Critic 网络,则通过梯度下降更新方式来减少损失以调整网络的参数,其损失函数为:

$$L(\theta_j^Q)=\frac{1}{B}\sum_{b=1}^{B}(y_i-Q_{i,j}^{\mu}(\boldsymbol{s}_b,\boldsymbol{a}_b^1,\boldsymbol{a}_b^2,\cdots,\boldsymbol{a}_b^n))^2,\ \forall\, j\in\{1,2\} \tag{19}$$

$$y_i=r_b^i+\gamma\min_{j=1,2}Q_{i,j}^{\mu'}(\boldsymbol{s}_{b+1},\mu_i'(\boldsymbol{o}_1')+\varepsilon,\cdots,\mu_i'(\boldsymbol{o}_n')+\varepsilon) \tag{20}$$

其中,y_i 为目标函数值;$Q_{i,j}^{\mu'}$ 为第 j 个 Critic 目标网络输出的动作价值函数;μ' 为 Actor 的目标网络;$\varepsilon\sim clip(\varepsilon,-c,c)$ 是 MATD3 算法为了实现动作价值函数沿动作空间的平滑化而在目标动作中添加的噪声,c 为随机噪声参数。

目标网络的更新采取的是一种软更新的方式,即让目标网络缓慢更新,逐渐接近网络,其更新式为:

$$\begin{cases}\theta_i^{\mu'}=\tau\cdot\theta_i^{\mu}+(1-\tau)\cdot\theta_i^{\mu'}\\ \theta_{i,j}^{Q'}=\tau\cdot\theta_{i,j}^{Q}+(1-\tau)\cdot\theta_{i,j}^{Q'},\ \forall\, j\in\{1,2\}\end{cases} \tag{21}$$

其中,$\theta_i^{\mu'}$,$\theta_{i,j}^{Q'}$ 分别为 Actor 目标网络和第 j 个 Critic 目标网络的参数;τ 为软更新参数,其目的是稳定该算法的学习过程,避免目标动作价值函数变化的跨度太大。

4.3 余弦退火方法

在神经网络的训练过程中,学习率的设定对于模型的权重更新至关重要。在训练初期,采用较大的学习率,能够使模型迅速响应梯度信息,快速地向损失函数的低值区域靠近,从而接近全局最优解。随着训练的进展,逐渐减小学习率,使模型对权重调整更细致,确保能够精准地定位到最优解。

模型权重的初始值在训练初期是随机设定的,如果采用较大的学习率可能会导致模型在优化过程中产生波动,影响其收敛稳定性。因此,本文采用余弦退火方法对学习率进行调整(本文的学习率衰减如图 3 所示),使学习率在整个训练周期内从初始的较高值逐渐衰减,使算法在训练初期能够快速探索解空间,在训练后期则能够细致地逼近最优解。

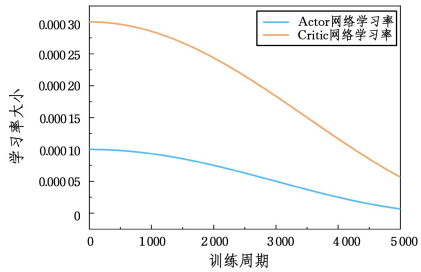


图 3 学习率衰减曲线

Fig. 3 Curves of learning rate decay

余弦退火方法的计算式如下:

$$\eta_i=\eta_{\min}+\frac{1}{2}(\eta_{\max}-\eta_{\min})\cdot\left(1+\cos\left(\frac{T_{\mathrm{cur}}}{T_{\max}}\cdot\pi\right)\right) \tag{22}$$

其中,η_i 表示当前训练回合的学习率;$\eta_{\min}$,$\eta_{\max}$ 分别表示学习率变化的最小值、最大值;T_{cur},$T_{\max}$ 分别表示当前训练回合、最大训练回合。

4.4 算法框架

MATD3 的算法框架如图 4 所示。在 MATD3 算法框架内,智能体通过执行动作与环境进行交互,并捕获环境反馈的奖励及状态信息,随后算法将这些交互经验存储于经验回放池中。该算法利用存储的经验数据,采用双延迟机制来提升学习过程的稳定性,并通过策略梯度方法对智能体的 Actor 网络实施参数更新。同时,智能体还通过 Critic 函数预测基于当前策略的期望累积回报,从而评估并优化策略的长期效能。

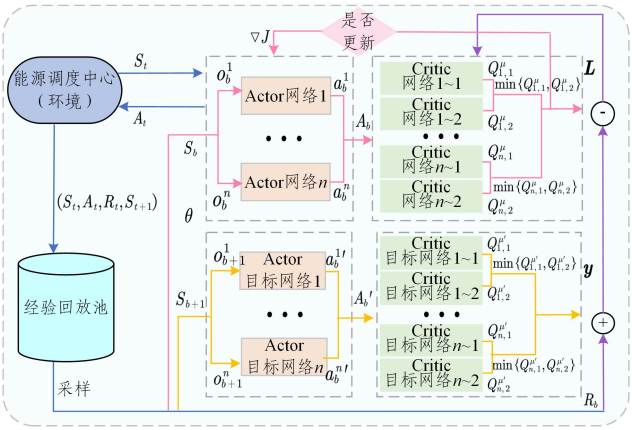


图 4 MATD3 算法框架

Fig. 4 MATD3 algorithmic framework

本文改进的 MATD3 的具体执行步骤如算法 1 所示。

算法 1 改进的 MATD3 算法

/ * 改进的 MATD3 算法计算多个 PSCS 的充电服务价格和 ESS 充放电量 * /

输入：训练回合数 MaxEpisode；最大时间步数 MaxStep

输出：PSCS 充电服务价格和 ESS 充电量策略

1. 初始化：初始化每个智能体的 Actor 网络，Critic 网络的参数 θ^{μ}，θ^{Q_1}，θ^{Q_2} 及其学习率 η^{μ}，η^{Q_1}，η^{Q_2}，以及对应的目标网络参数 $\theta^{\mu'}$，$\theta^{Q_1'}$，$\theta^{Q_2'}$ 及其学习率 $\eta^{\mu'}$，$\eta^{Q_1'}$，$\eta^{Q_2'}$；
2. 初始化经验回放池 D
3. for episode=1 to MaxEpisode do：
4. 初始化环境状态 $\mathbf{s}_0$；
5. for t=0 to MaxStep do：
6. for i=1 to n do：
7. 对于智能体 i，根据式(17)得到动作 $\mathbf{a}_t^i$；
8. end for
9. 执行动作 $\mathbf{a}_t=(\mathbf{a}_t^1,\mathbf{a}_t^2,\cdots,\mathbf{a}_t^n)$，得到奖励 r_t 和下一个状态 $\mathbf{s}_{t+1}$；
10. 将样本数据 $(\mathbf{s}_t,\mathbf{a}_t,r_t,\mathbf{s}_{t+1})$ 存储到经验回放池 D；
11. if 经验回放池达到一定容量及达到训练更新步数：
12. 从经验回放池 D 随机采样 B 个小批量样本 $(\mathbf{s}_b,\mathbf{a}_b,r_b,\mathbf{s}_{b+1})$；
13. for i=1 to n do：
14. 根据式(20)计算出目标函数值 y_i；
15. 根据式(19)计算损失函数以更新两个 Critic 网络；
16. 根据式(22)更新学习率 η^{Q_1}，η^{Q_2}；
17. if 达到网络更新频率：
18. 根据式(18)更新 Actor 网络；
19. 根据式(21)更新 Actor 目标网络和两个 Critic 目标网络；
20. 根据式(22)更新学习率 η^{μ}，$\eta^{\mu'}$，$\eta^{Q_1'}$，$\eta^{Q_1'}$；
21. end if
22. end if
23. end if
24. end if
25. end if

根据算法 1，下面给出算法的具体步骤。

第 1 步 首先初始化算法所有网络的参数以及经验回放池，并初始化训练回合、时间步数以及环境状态。

第 2 步 在每个训练回合的每个时间步中，每个智能体会根据环境状态计算出动作，并执行动作使环境进入下一个状态，同时得到当前动作对应的奖励值，并将当前时间步的样本数据 $(\boldsymbol{s}_t,\boldsymbol{a}_t,\boldsymbol{r}_t,\boldsymbol{s}_{t+1})$ 存入经验回放池 D 中。

第 3 步 当达到可以采样的经验回放池 D 长度和训练更新步数，开始采样进行智能体训练，先计算出目标值，再根据目标值来计算损失函数，通过梯度下降的方式更新两个 Critic 网络，同时使用余弦退火方法更新其学习率，并判断是否达到网络更新频率，若达到，则更新 Actor 网络及所有目标网络，同时更新其学习率。

第 4 步 判断当前时间步是否已达到最大时间步数，若是，则进行下一个训练回合，否则进入下一个时间步。再判断是否达到最大训练回合数，若是，则结束训练，否则进行下一回合训练。

5 实验

5.1 实验设置

在本文的 PSCS-EVs 仿真环境中，交通网络是根据福州市某地区的地图进行仿真所得，车辆信息和 PSCS 运行信息则是通过数据集所得。本文数据集来自福建某公司所采集的实际数据（数据集包括 EVs 的充电时间、离开时间、充电功率，PSCSPV 发电量和充电服务价格、主电网的分时电价等）。将上述数据按每天 24 个时段进行处理得到本文的数据集，EVs 总数量为 120 辆。其中输入的主电网的分时电价、PSCS 充电服务价格和 PV 发电量如图 5 所示，PSCS 的 ESS 参数如表 1 所列。

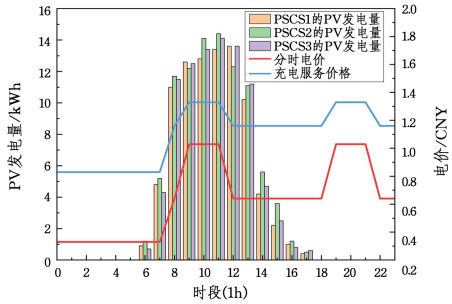


图 5 输入

Fig. 5 Input

表 1 PSCS 参数

Table 1 PSCS parameters

参数	值
时段长度 Δt	1 h
储能容量 C	250 kWh
储能 SOC 初始值 $v_{\text{init}}^{\text{soc}}$	0.5
储能 SOC 上限阈值 $v_{\max}^{\text{soc}}$、下限阈值 $v_{\min}^{\text{soc}}$	0.97，0.03
主电网最大可供给量 $q_{\max}^{\text{grid}}$	150 kWh
储能单位交换电量损耗成本 β	0.08 CNY
ESS 充放电超限惩罚系数 ξ	1 500
储能充放电效率 α	0.95

本文算法的每个智能体的 Actor 网络和 Critic 网络均由 4 个全连接层组成，分别包含 256，256 和 128 个神经元；激活函数为泄漏修正线性单元（Leaky Rectified Linear Unit，LeakyReLU）；使用适应性矩估计（Adaptive Momentum Esti-

mation，Adam）优化器来迭代训练更新神经网络权重，一共训练 5000 个训练回合，算法具体参数如表 2 所列。本文模拟实验是在 64 位 Windows 10 系统环境下，硬件平台为 Intel(R) CoreTM i7-7700HQ CPU @ 2.80 GHz 2.80 GHz，内存为 16 GB，1 块 NVIDIA GeForce GTX 1060 显卡进行运算，改进的 MATD3 采用 Python3.6 和 PyTorch 架构进行编程和训练。

表 2 算法参数

Table 2 Algorithmic parameters

参数	值
折扣因子 γ	0.935
软更新因子 τ	5×10^{-3}
经验回放池大小 D	5×10^{5}
采样样本数 B	2000
智能体个数 n	3
最大训练回合 $MaxEpisode$	5000
每回合最大时间步数 $MaxStep$	24
动态自适应电价调整系数$[\lambda_{\min},\lambda_{\max}]$	$[-0.1,0.1]$
Actor 网络及对应目标网络的初始学习率 $\eta^{\mu},\eta^{\mu'}$	1×10^{-4}
Critic 网络及对应目标网络初始学习率 $\eta^{Q_1},\eta^{Q_2},\eta^{Q_1'},\eta^{Q_2'}$	3×10^{-4}
训练更新步数	70
Actor 网络及对应目标网络的更新频率	2
Actor 目标网络添加的高斯噪声的标准差	0.2
Actor 目标网络添加的高斯噪声截断范围	0.5

5.2 结果评价与分析

为了验证本文改进的 MATD3 算法的有效性，采用 MADDPG[27]，MMADDPG[28]，MATD3 这 3 种 MADRL 方法作为对比方法，3 种算法的 Actor 网络和 Critic 网络的网络层数和参数设置都与本文算法一致。

模型训练回报曲线如图 6 所示。

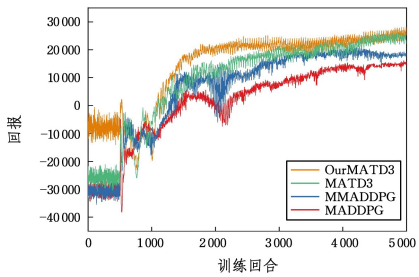
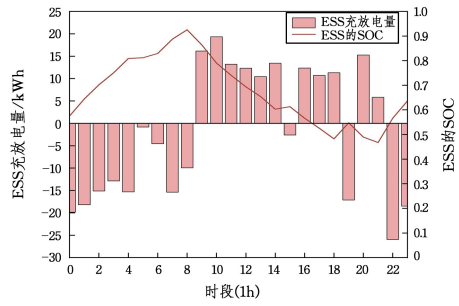


图 6 不同算法收敛结果对比

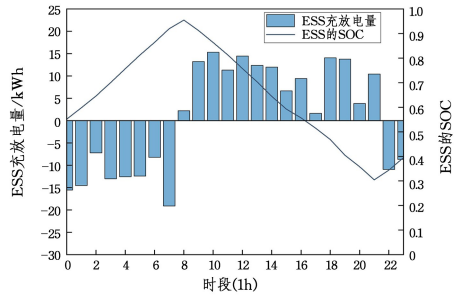
Fig. 6 Comparison of convergence results of different algorithms

由于本文算法是离线学习，因此需要先将经验回放池存到一定容量才能进行学习，故在前 500 个训练回合中，模型只通过随机选择动作对环境进行探索来积累经验。在 500 个训练回合后，模型开始从已有的经验中进行学习，逐步寻找最优策略，回报开始波动并呈上升趋势，约到 3000 个训练回合后，回报开始逐渐趋于稳定，在 5000 个训练回合后回报值在 25000左右收敛。图 6 还展示了对比算法的训练情况，从图中可以看出，本文改进的算法的收敛速度和收敛阈值都优于其余 3 种算法，其中 MADDPG 算法所取得的回报值最低，回报收敛值在 15000 左右，其次是 MMADDPG 算法，回报收敛值在 18000 左右，这两种算法由于采用的都是一个 Q 网络，无法避免 Q 值高估情况，因此结果相比本文改进的算法较差。而 MATD3 算法的回报最终收敛于 24000 左右，与本文算法相差不大，但是本文算法的收敛速度比之快了 1000 个训练回合，这说明本文采用余弦退火衰减学习率的方法十分有效。通过对比这 3 种算法，本文改进的算法的回报值比之分别提升了 66.67%，38.89%和 4.17%。

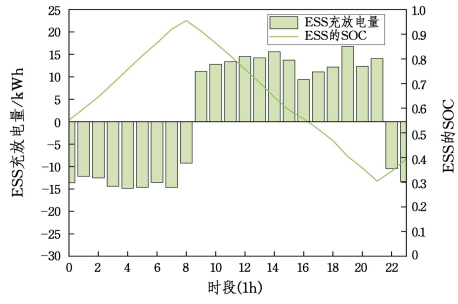
图 7 给出了本文算法学习得到的多个 PSCS 的 ESS 充放电量策略。结合图 5 可以发现，本文算法能够很好地学到在主电网分时电价低时（即 0—7 时）进行购电，在分时电价高时（即 8—22 时）进行售电，从而使 PSCS 的购电成本大幅降低。同时发现所得策略能保证 PSCS 的 ESS 的 SOC 在日终时刻能够稳定回归至约 0.5 的水平，同时在充电过程中能够将 SOC 提升至接近 0.9 的高位，而在放电过程中则能保持 SOC 不低于 0.2 的安全阈值，这种策略保障 ESS 的高效运行，有助于延长储能电池的整体使用寿命。



(a) PSCS1 的 ESS 充放电量



(b) PSCS2 的 ESS 充放电量



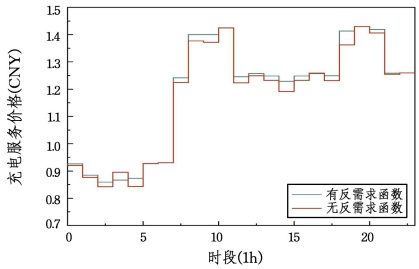
(c) PSCS3 的 ESS 充放电量

图 7 各 PSCS 的 ESS 充放电量

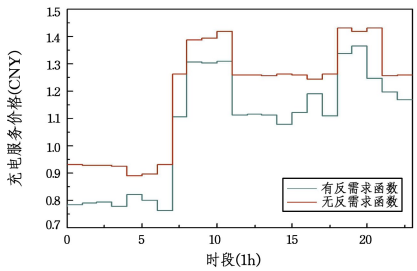
Fig. 7 ESS charging and discharging for each PSCS

本文对比分析了本文算法在考虑和不考虑反需求函数情况下对多个 PSCS 充电服务价格的影响，实验结果如图 8 所示。实验结果表明，在未引入反需求函数的情况下，各 PSCS 的充电服务价格普遍偏高，且趋向于算法所允许的最大价格。这一现象说明了 MADRL 在合作定价环境中存在价格垄断问题。当算法引入反需求函数后，从图中可以看出，PSCS2 的充电服务价格波动明显，而 PSCS1 和 PSCS3 的价格波动较小。原因在于 PSCS2 的充电需求远大于 PSCS1 和 PSCS3，在引入反需求函数后，充电服务价格成为充电需求的函数，而这一关联在充电需求较大的站点（如 PSCS2）表现得更加明显。

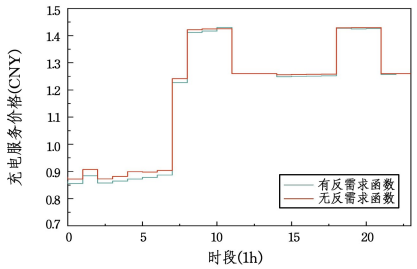
反需求函数的本质是为了防止价格垄断，在其限制下，为了最大化 PSCS 的总收益，调整 PSCS2 的充电服务价格能够比调整 PSCS1 和 PSCS3 带来更高的收益回报，因为 PSCS2 的充电需求基数大，价格的微小变动就会对充电量产生显著影响，从而显著影响整体收益。图中的结果也表明，PSCS 的充电服务价格得到了有效调控，避免了价格垄断，验证了本文提出的反需求函数在合作定价环境中能够有效解决价格垄断问题，确保充电市场健康发展，保障 PSCS 的总收益。



(a)PSCS1 的充电服务价格



(b)PSCS2 的充电服务价格



(c)PSCS3 的充电服务价格

图 8 各 PSCS 的充电服务价格

Fig. 8 Charging service prices for each PSCS

同时，本文反需求函数的系数是通过一系列实验来反推设置的，其中实验增量设置为 0.1，图 9 给出了在不同系数下取得的实验结果。通过对比分析不同系数对算法性能的影响，可以发现当系数设定为 1.5 时，算法能够获得最优的回报值，而当系数调整为 1.4 或 1.6 时，回报值均略有下降。

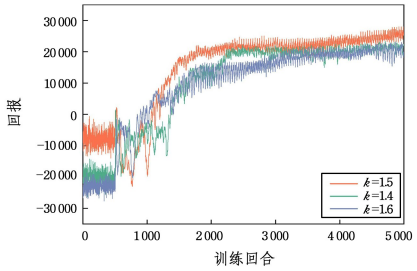
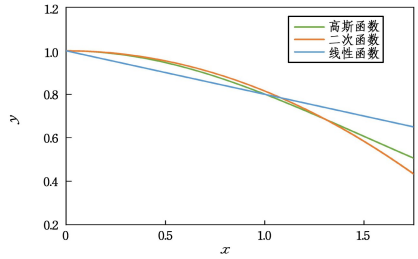


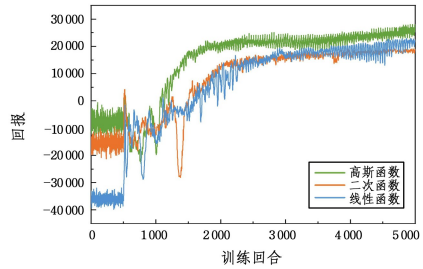
图 9 不同反需求函数系数的训练收敛结果

Fig. 9 Training convergence results for different inverse demand function coefficients

本文还对高斯函数作为反需求函数的表现进行了评估，并将其与线性和二次函数进行了对比分析。实验结果如图 10 所示，其中图 10(a)为 3 种函数的函数图，图 10(b)为实验收敛图。实验结果表明，相较于线性和二次函数，本文采用的高斯函数在模拟 EVs 充电需求与充电服务价格这一关系时表现更佳，能够更准确地反映充电服务价格变化对 EVs 充电需求的影响。



(a)3 种不同的函数图像



(b)不同函数下的算法收敛结果

图 10 不同反需求函数的训练收敛结果

Fig. 10 Training convergence results of different anti demand functions

结束语 本文提出了一种改进的 MATD3，用于解决单个充电站运营商下多个 PSCS 之间的动态定价及能源调度策略，同时利用反需求函数对多站的合作定价进行约束。实验结果证明，改进后的 MATD3 算法有效提升了算法的收敛速度和阈值，且制定的能源调度策略能够做到“低储高发”，在降低购电成本的同时有效提高了 PSCS 的整体运营收益，此外反需求函数的引入也成功避免了完全合作关系下多站间的价格垄断现象。因此，本文所用方法可以为完全合作关系下多站的动态定价和能源调度提供科学有效的参考。

在下一步的工作中，会进一步优化 EVs 充电服务价格需求响应的模型，将主电网的负载需求整合进模型中，并引入 V2G 技术为 PSCS 提供新的收益模式，通过调度 EVs 的充电功率获得电网辅助调峰服务补贴或峰谷电价差的收益，降低 EVs 的使用成本，同时实现电网负载的平衡，增加 PSCS 的运营收入。

参 考 文 献

[1] FANG Q. Energy storage operation strategy of optical storage charging station based on PPO algorithm [J]. Electric Engineering, 2024(2): 97-100.

[2] LIU Y X, ZHANG S, GUO L, et al. A coordinated optimal scheduling method of distribution grid and photovoltaic storage charging station taking into account electric energy-standby coupling [J]. Power System Technology, 2024, 48(8): 3175-3185.

[3] LEE S,CHOI D H. Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: A privacy-preserving deep reinforcement learning approach [J]. Applied Energy,2021,304:117754.

[4] AFFOLABI L,SHAHIDEHPOUR M,GAN W,et al. Optimal transactive energy trading of electric vehicle charging stations with on-site PV generation in constrained power distribution networks [J]. IEEE Transactions on Smart Grid,2021,13(2):1427-1440.

[5] WANG R S,CHEN Z,XING Q,et al. A Modified Rainbow-Based Deep Reinforcement Learning Method for Optimal Scheduling of Charging Station [J]. Sustainability,2022,14(3):1884.

[6] ZHANG S,LIU J J,SU Y T. Research on real-time control strategy of optical storage charging station based on PSO-DDPG algorithm [J]. Mechanical and Electrical Information, 2023(17):5-9.

[7] YANG X Y,CUI T X,WANG H R,et al. Multiagent Deep Reinforcement Learning for Electric Vehicle Fast Charging Station Pricing Game in Electricity-Transportation Nexus [J]. IEEE Transactions on Industrial Informatics,2024 20(4):6345-6355.

[8] QIAN T,SHAO C,LI X,et al. Multi-agent deep reinforcement learning method for EV charging station game [J]. IEEE Transactions on Power Systems,2021,37(3):1682-1694.

[9] ACKERMANN J,GABLER V,OSA T,et al. Reducing overestimation bias in multi-agent domains using double centralized critics[J]. arXiv:1910.01465,2019.

[10] HUANG Z Q,LIN B,LU Y,et al. A charging station siting and capacity determination method for multi-objective optimization [J]. Journal of Fujian Normal University(Natural Science Edition),2024,40(2):23-35.

[11] QUE H K,FENG X F,GUO W C,et al. Charging station layout model based on fuzzy bi-objective planning [J]. Computer Science,2022,39(3):751-757.

[12] WANG S,BI S,ZHANG Y A. Reinforcement learning for real-time pricing and scheduling control in EV charging stations [J]. IEEE Transactions on Industrial Informatics,2019,17(2):849-859.

[13] XU H,WU Q,WEN J,et al. Joint bidding and pricing for electricity retailers based on multi-task deep reinforcement learning [J]. International Journal of Electrical Power & Energy Systems,2022,138:107897.

[14] SHIN M J,CHOI D H,KIM J. Cooperative management for PV/ESS-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach [J]. IEEE Transactions on Industrial Informatics,2019,16(5):3493-3503.

[15] KABIR M E,ASSIC C,TUSHAR M H K,et al. Optimal scheduling of EV charging at a solar power-based charging station [J]. IEEE Systems Journal,2020,14(3):4221-4231.

[16] ZHOU X Y,QIAN L P,HUANG Y P,et al. An Optimization Method for Electric Vehicle Charging Scheduling Based on Ant Colony Algorithm [J]. Computer Science. 2020,47(11):280-285.

[17] WU W T,LIN Y,LIU R H,et al. Online EV charge scheduling based on time-of-use pricing and peak load minimization: Properties and efficient algorithms [J]. IEEE Transactions on Intelligent Transportation Systems,2020,23(1):572-586.

[18] YU W W,LIU S L,CHEN Q G,et al. Multi-objective optimal scheduling of photovoltaic microgrids considering electric vehicle charging and demand-side response [J]. Proceedings of the CSU-EPSA,2018,30(1):88-97.

[19] HAO Y. Energy optimization management of new energy charging station based on genetic algorithm [J]. Telecom Power Technologies,2019,36(11):27-28,31.

[20] SU L,JIANG X C,WANG W,et al. Optimized energy management of microgrids taking into account electric vehicles and photovoltaic energy storage [J]. Automation of Electric Power Systems,2015,39(9):164-171.

[21] LYU C,ZHAN S,ZHANG Y,et al. Synergistic two-stage optimization for multi-objective energy management strategy of integrated photovoltaic-storage charging stations [J]. Journal of Energy Storage,2024,89:111665.

[22] MURIITHI G,CHOWDHURY S. Optimal energy management of a grid-tied solar pv-battery microgrid: A reinforcement learning approach [J]. Energies,2021,14(9):2700.

[23] CUI L,WANG Q,QU H,et al. Dynamic pricing for fast charging stations with deep reinforcement learning [J]. Applied Energy,2023,346:121334.

[24] LEE S,CHOI D H. Three-Stage Deep Reinforcement Learning for Privacy-and Safety-Aware Smart Electric Vehicle Charging Station Scheduling and Volt/VAR Control [J]. IEEE Internet of Things Journal,2023,11(5):8578-8589.

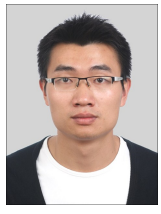
[25] QIAN T,SHAO C,LI X,et al. Multi-agent deep reinforcement learning method for EV charging station game [J]. IEEE Transactions on Power Systems,2021,37(3):1682-1694.

[26] VARIAN H R. Intermediate microeconomics with calculus: a modern approach[M]. New York: W. W. NORTON & Company,2014:114-115.

[27] LOWE R,WU Y I,TAMAR A,et al. Multi-agent actor-critic for mixed cooperative-competitive environments [C]// Advances in Neural Information Processing Systems. 2017.

[28] ZHANG Y,YANG Q,AN D,et al. Multistep multiagent reinforcement learning for optimal energy schedule strategy of charging stations in smart grid [J]. IEEE Transactions on Cybernetics,2022,53(7):4292-4305.

CHEN Jintao, born in 2000, postgraduate. His main research interests include resource scheduling and reinforcement learning.



LIN Bing, born in 1986, Ph.D, associate professor, postgraduate supervisor, is a member of CCF (No. 83773M). His main research interests include cloud computing technology and computational intelligence.

(责任编辑:喻藜)