基于SAC和TD3的含电动汽车虚拟电厂调度策略

陶力[1,2,5]，杨夏喜[3]，顾金辉[4]，魏兵兵[3]，张琳[2,5]，王嘉宁[4]

（1.华北电力大学 经济与管理学院，北京 102206；2.南瑞集团有限公司（国网电力科学研究院有限公司），江苏 南京 210003；3.苏州市产品质量监督检验院，江苏 苏州 215104；4.华北电力大学 电气与电子工程学院，北京 102206；5.北京科东电力控制系统有限责任公司，北京 100194）

摘要：虚拟电厂（VPP）可以聚合分布式电源（DER）参与电力市场和辅助服务市场运行，为配电网和输电网提供管理和辅助服务，其运行和控制得到广泛关注。针对含电动汽车（EV）充电站的虚拟电厂，构建了基于柔性行动器-评判器（SAC）算法和双延迟深度确定性策略梯度（TD3）算法的VPP与EV充电站主从博弈模型。通过训练主从博弈网络参数，计算博弈均衡时的策略和解。算例结果表明，上述模型训练完成后，可以有效地降低EV充电站运行费用以及平缓功率，基于SAC强化学习方法能够整合VPP内部DER，并引导电动汽车有序充电。在VPP作为价格接受者参与日前电力市场时，也能够给出优化的交易策略；当VPP与EV之间存在主从博弈时，EV用确定性策略算法可以降低充电成本，VPP用随机性策略算法则可以提高收益。

关键词：虚拟电厂；SAC算法；TD3算法；电动汽车；主从博弈；实时调度

中图分类号：TM721 **文献标识码**：A **DOI**：10.19457/j.1001-2095.dqcd24441

Scheduling Strategy of Virtual Power Plant with Electric Vehicle Based on SAC and TD3

TAO Li[1,2,5]，YANG Xiaxi[3]，GU Jinhui[4]，WEI Bingbing[3]，ZHANG Lin[2,5]，WANG Jianing[4]

（1.*School of Economics and Management*，*North China Electric Power University*，*Beijing* 102206，*China*；2.*Nanrui Group Co.*，*Ltd.*（*State Grid Electric Power Research Institute Co.*，*Ltd.*），*Nanjing* 210003，*Jiangsu*，*China*；3.*Suzhou Institute of Product Quality Supervision and Inspection*，*Suzhou* 215104，*Jiangsu*，*China*；4.*School of Electrical & Electronic Engineering*，*North China Electric Power University*，*Beijing* 102206，*China*；5.*Beijing Kedong Electric Power Control System Co.*，*Ltd.*，*Beijing* 100194，*China*）

Abstract: Virtual power plant（VPP）can integrate distributed energy resource（DER）to participate in the operation of power market and auxiliary service market，and provide management and auxiliary services for distribution network and transmission network. Its operation and control have been widely concerned. Aiming at the virtual power plant containing electric vehicle（EV）charging stations，the Stackelberg game model of VPP and EV charging stations was constructed based on soft actor-critic（SAC）algorithm and twin delay deep deterministic policy gradient（TD3）algorithm. By training the network parameters of Stackelberg game，the strategy and solution in game equilibrium was calculated. The calculation example results show that the model proposed can effectively reduce the operating cost and smooth power of EV charging stations after the completion of training，and the SAC reinforcement learning method can integrate the internal DER of VPP and guide the orderly charging of EV. When VPP participates in day-ahead power market as price taker，it can also give optimal trading strategy. When there is Stackelberg game between VPP and EV，EV can reduce charging cost by using deterministic strategy algorithm，while VPP can improve revenue by using stochastic strategy algorithm.

Key words: virtual power plant（VPP）；SAC algorithm；TD3 algorithm；electric vehicle（EV）；Stackelberg game；real-time dispatch

由于电动汽车（electric vehicle，EV）具有低能耗、低排放的优势，预计其接入电网的比例持续

基金项目：国家电网科技项目（5100-202040443A-0-0-00）

作者简介：陶力（1981—），男，博士，高级工程师，Email：power_tao@sina.com

增加。电动汽车能够借助充电桩实现与电网之间的互动（vehicle to grid，V2G），在减少用户成本的同时，起到辅助电网安全稳定运行的作用[1]，是一种非常有潜力的分布式电源（distributed energy resource，DER）。然而，由于EV接入电网的时空不确定性，其入网充电时间与充电电量均具有较高的随机性，这也给电网优化控制带来了极大的挑战。

针对EV充电的优化问题，文献[2]对用户取车时的目标电池荷电状态（state of charge，SOC）做出约束，提出采用双延迟深度确定性策略梯度（twin delay deep deterministic policy gradient，TD3）算法连续控制充电桩的充电功率，但是没有考虑到充电管理系统能够提高实际充电时的效率；文献[3]采用深度Q网络算法控制EV的充电行为，能够降低充电费用及平抑网络功率波动，但是这种算法只能用于充电功率的分档调节；文献[4]将电动汽车充、放电调度问题建模为带约束的马尔可夫决策过程，然后采用提高强化学习安全性的约束型策略优化（constrained policy optimization，CPO）算法求解；文献[5]将电动汽车实时电压控制问题转化为EV无功控制和V2G两种模式的马尔可夫博弈，并采用深度确定性策略梯度（deep deterministic policy gradient，DDPG）算法求解，效果较好。上述基于深度强化学习（deep reinforcement learning，DRL）的方法利于EV充电站内部优化，但无法通过与传统运筹学模型相结合来与外部资源联合优化。

与此同时，随着分布式能源、储能、通信、并行计算等技术的发展，含电动汽车的虚拟电厂（VPP）可以将分布式可再生能源发电、储能装置、电动汽车等资源聚合成一个整体，作为一个特殊的电厂参与电力市场竞争[6-8]。文献[9]提出风电商和EV聚合商通过合作博弈组成VPP参与电力市场投标竞争，并采用Shapley值法进行收益分配。文献[10]提出以VPP为售电商的EV主从博弈模型，能够优化自身售电策略，并引导EV有序充电。然而此类电力市场模型往往呈现非凸、非线性、维度高的特点，采用传统运筹学方法求解时难度较大。系统中可再生能源的间歇性以及电动汽车负荷需求的不确定性造成了供需双方的随机波动，传统的调度方法难以准确地适应实际环境的动态变化，也难以对智能体控制的EV充电站优化调度。

随着人工智能技术的发展，DRL方法在电力系统中也越来越受到重视[11-13]，其可以模拟不完全信息的交易或博弈，能从高维、连续的状态空间中提取高阶数据特征，对含不确定性的可再生能源出力、电动汽车有序充电、电力市场交易等模型有较强的表达及特征挖掘能力。文献[14]采用DDPG算法，以解决VPP经济调度问题的最优解，但没有考虑电动汽车充电，也没有考虑VPP在日前市场的交易。文献[15]采用基于优先经验回放的深度确定性策略梯度（deep deterministic policy gradient with prioritized experience replay，DDPG-PER）算法作为电力市场竞价策略，当出清模型非凸时，获得的收益超过数学规划方法。此外，无模型的深度强化学习算法也已被应用于求解Nash博弈[16-17]、Stackelberg博弈[18-19]、平均场博弈（mean field games，MFG）[20]等多种博弈论模型。

从现有文献来看，传统博弈论方法一般局限于求解完全信息静态博弈问题。传统的强化学习（reinforcement learning，RL）算法虽然可以动态模拟不完全信息的重复博弈，但应用范围局限于低维的离散状态/动作空间，且收敛结果不稳定。本文针对含EV充电站、分布式机组、储能、可再生能源等灵活性资源的VPP，提出基于深度强化学习的VPP与EV主从博弈模型，其中VPP采用柔性行动器-评判器（soft actor-critic，SAC）算法，EV聚合商采用TD3算法。VPP整合配电网内分布式能源并制定合理的售电策略来引导EV的有序入网，以实现多种新能源间的协调互补与整体优化。本文的主要贡献如下：

1）对于智能体优化控制的VPP与EV充电站构成的主从博弈模型，提出交替训练行动器-评判器算法网络参数的求解流程与方法。

2）算例从博弈论中策略类型的角度，研究了混合策略与纯策略在模型中的不同效果，并给出了初步的解释。

3）算例对比了博弈与不博弈下模型的结果，表明主从博弈模型能降低EV用电成本，提高社会福利。

1　电力市场交易流程及VPP结构

VPP服务器利用通信技术将可控分布式发电机组、风电、光伏、储能及电动汽车充电站等资源聚合，形成整体参与电网市场交易及电网运行。VPP容量较小，可作为价格接受者参与日前电力

市场（day-ahead market，DAM）和实时平衡市场（real-time balancing market，RBM）的电力交易[21]。

1.1　含VPP电力市场交易流程

电价机制采用市场清算电价（pay-as-bid，PAB），其规则[22]如下：在第n天的能量市场交易结束之前，VPP以系统运营商（independent system operator，ISO）的出清结果为日前电力市场价格曲线，根据对电动汽车负荷量、可再生能源发电量的预测，通过与EV主从博弈产生的均衡解，形成VPP日前优化与实时调度的控制策略。并依据训练好的深度强化学习模型，向ISO独立申报第n+1天24个交易时段的电量交易信息。在随后的实时平衡市场中，VPP根据EV以及新能源机组的实时功率波动，调整内部储能与可控分布式机组（distributed generator，DG）的出力、EV充电电价以降低功率偏差，对于无法平衡的功率偏差则在实时平衡市场中以惩罚电价进行交易。

VPP在日前电力市场中的购电策略可以表示为$P^{\mathrm{D}}=\{P_1^{\mathrm{D}},P_2^{\mathrm{D}},\cdots,P_T^{\mathrm{D}}\}$，其购电成本$C_t^{\mathrm{B}}$为

$$C_t^{\mathrm{B}}=\begin{cases}-P_t^{\mathrm{D}}\lambda_t^{\mathrm{D}}-P_t^{\mathrm{R}}\lambda_{\mathrm{b},t}^{\mathrm{R}} & P_t^{\mathrm{R}}\geqslant 0\\ -P_t^{\mathrm{D}}\lambda_t^{\mathrm{D}}-P_t^{\mathrm{R}}\lambda_{\mathrm{s},t}^{\mathrm{R}} & P_t^{\mathrm{R}}<0\end{cases}\tag{1}$$

$$P_{\min}^{\mathrm{D}}\leqslant P_t^{\mathrm{D}}\leqslant P_{\max}^{\mathrm{D}}\tag{2}$$

式中：P_t^{D}，P_t^{R}为VPP在日前电力市场和实时平衡市场中的购电/售电量；$P_{\min}^{\mathrm{D}}$，$P_{\max}^{\mathrm{D}}$为购售电量上、下限，与联络线功率约束有关；λ_t^{D}为日前电力市场的出清价格，由ISO在日前根据出清结果确定；$\lambda_{\mathrm{b},t}^{\mathrm{R}}$，$\lambda_{\mathrm{s},t}^{\mathrm{R}}$为实时平衡市场中的惩罚性购售电价[10]。

1.2　VPP结构及数学模型

VPP作为一个整体对外参与电力市场，对内实现各DER、储能、EV充电站的协调运行控制，可以提高在电力市场中的竞争力[23]。

1.2.1　EV充电站

本文考虑一个包含K个充电桩、完全由智能体控制的EV充电站，其中第i台EV充、放电的数学模型如下所示：

$$e_{i,t}=\begin{cases}e_{i,t-1}+\dfrac{\eta_{\mathrm{e}}^{\mathrm{ch}}P_{i,t}^{\mathrm{EV}}\Delta t}{Q_{\mathrm{e}}} & P_{i,t}^{\mathrm{EV}}\geqslant 0\\ e_{i,t-1}+\dfrac{P_{i,t}^{\mathrm{EV}}\Delta t}{\eta_{\mathrm{e}}^{\mathrm{dis}}Q_{\mathrm{e}}} & P_{i,t}^{\mathrm{EV}}<0\end{cases}\tag{3}$$

$$e_{i,t,\min}\leqslant e_{i,t}\leqslant e_{\max}\tag{4}$$

$$e_{i,t,\min}=\max\left\{e_{\min},e_{\mathrm{n}}-\frac{\eta_{\mathrm{e}}^{\mathrm{ch}}P_{i,t}^{\mathrm{ch,max}}(t_{\mathrm{l},i}-t)}{Q_{\mathrm{e}}}\right\}\tag{5}$$

$$\begin{cases}-P_{i,t}^{\mathrm{dis,max}}\leqslant P_{i,t}^{\mathrm{EV}}\leqslant P_{i,t}^{\mathrm{ch,max}}\\ t_{\mathrm{a},i}<t\leqslant t_{\mathrm{l},i}\end{cases}\tag{6}$$

式中：$P_{i,t}^{\mathrm{EV}}$为t时刻充、放电功率，$P_{i,t}^{\mathrm{EV}}\geqslant 0$代表EV充电量，$P_{i,t}^{\mathrm{EV}}<0$代表放电量；$t_{\mathrm{a},i}$，$t_{\mathrm{l},i}$分别为EV的到达、离开时刻；$e_{i,t}$，$e_{i,t,\min}$分别为第$i$台EV在时刻$t$的SOC和满足用户要求的最低SOC；$e_{\max}$，$e_{\min}$分别为EV电池容量限制的最大SOC和最小SOC；$e_{\mathrm{n}}$为电动汽车出发时期望的最小SOC；$\eta_{\mathrm{e}}^{\mathrm{ch}}$，$\eta_{\mathrm{e}}^{\mathrm{dis}}$分别为EV电池的充、放电效率；$Q_{\mathrm{e}}$为电池总容量；$\Delta t$为时间间隔；$P_{i,t}^{\mathrm{ch,max}}$，$P_{i,t}^{\mathrm{dis,max}}$分别为充、放电功率$P_{i,t}^{\mathrm{ch}}$和$P_{i,t}^{\mathrm{dis}}$的最大值。

充电站中不同EV在每个时段内的充、放电功率由同一个充电站智能体控制。充电站智能体依次观察每台EV当前的状态，并确定下个时刻EV的动作。第i台EV的状态包括当前时间、充电站内充电桩使用率、VPP制定的EV充电价格、第i台EV的SOC、第i台EV预计剩余的取车时间，即

$$S_i^{\mathrm{EV}}=\left\{t,\eta^{\mathrm{EV}},\lambda_t^{\mathrm{EV}},e_i,t_{\mathrm{l},i}\right\}\tag{7}$$

其动作为每台EV在t时刻充、放电量，即

$$A_i^{\mathrm{EV}}=\left\{P_{i,t}^{\mathrm{EV}}\right\}\tag{8}$$

1.2.2　分布式机组发电

可控分布式机组一般为用户侧的小型燃气机组或柴油机组，运行成本$C_{i,t}^{\mathrm{DG}}$主要考虑发电成本$C_{i,t}^{\mathrm{DG1}}$、启停成本$C_{i,t}^{\mathrm{DG2}}$，其运行特性与约束条件为

$$\begin{cases}C_{i,t}^{\mathrm{DG}}=C_{i,t}^{\mathrm{DG1}}+C_{i,t}^{\mathrm{DG2}}\\ C_{i,t}^{\mathrm{DG1}}=a_i^{\mathrm{DG}}(P_{i,t}^{\mathrm{DG}})^2+b_i^{\mathrm{DG}}P_{i,t}^{\mathrm{DG}}+c_i^{\mathrm{DG}}\\ C_{i,t}^{\mathrm{DG2}}=\displaystyle\sum_{t=1}^{T}\sum_{i=1}^{N_{\mathrm{G}}}[c_i^{\mathrm{on}}h_{i,t}^{\mathrm{g}}(1-h_{i,t-1}^{\mathrm{g}})+\\ \qquad c_i^{\mathrm{off}}h_{i,t-1}^{\mathrm{g}}(1-h_{i,t}^{\mathrm{g}})]\end{cases}\tag{9}$$

$$P_{i,t,\min}^{\mathrm{DG}}\leqslant P_{i,t}^{\mathrm{DG}}\leqslant P_{i,t,\max}^{\mathrm{DG}}\tag{10}$$

$$\Delta P_{i,t}^{\mathrm{DG}}=P_{i,t}^{\mathrm{DG}}-P_{i,t-1}^{\mathrm{DG}}\tag{11}$$

$$P_{i,\mathrm{down}}^{\mathrm{DG}}\leqslant\Delta P_{i,t}^{\mathrm{DG}}\leqslant P_{i,\mathrm{up}}^{\mathrm{DG}}\tag{12}$$

式中：$P_{i,t}^{\mathrm{DG}}$为第i台分布式机组在t时刻的输出功率；$h_{i,t-1}^{\mathrm{g}}$为机组的启停状态，1表示运行，0表示停运；a_i^{DG}，b_i^{DG}，c_i^{DG}分别为第i台分布式机组的耗量参数；c_i^{on}，c_i^{off}分别为第i台常规机组的启动和停机成本；$P_{i,t,\max}^{\mathrm{DG}}$，$P_{i,t,\min}^{\mathrm{DG}}$分别为DG输出功率上、下限；$\Delta P_{i,t}^{\mathrm{DG}}$为功率变化量；$P_{i,\mathrm{up}}^{\mathrm{DG}}$，$P_{i,\mathrm{down}}^{\mathrm{DG}}$分别为分布式机组爬坡速率上、下限。

式(9)为分布式机组成本耗量函数，式(10)为输出功率约束，式(12)约束分布式机组功率调整的爬坡速率。

1.2.3　储能

本文储能单元的运行特性与约束条件为

$$f_{i,t}^{\mathrm{ES}}=\begin{cases}f_{i,t-1}^{\mathrm{ES}}+\eta_{\mathrm{b}}^{\mathrm{ch}}P_{i,t}^{\mathrm{ES}}\Delta t & P_{i,t}^{\mathrm{ES}}\geqslant 0\\ f_{i,t-1}^{\mathrm{ES}}+\dfrac{P_{i,t}^{\mathrm{ES}}\Delta t}{\eta_{\mathrm{b}}^{\mathrm{dis}}} & P_{i,t}^{\mathrm{ES}}<0\end{cases}\tag{13}$$

$$f_{\min}^{\mathrm{ES}}\leqslant f_{i,t}^{\mathrm{ES}}\leqslant f_{\max}^{\mathrm{ES}}\tag{14}$$

$$f_{i,1}^{\mathrm{ES}}=f_{i,T}^{\mathrm{ES}}\tag{15}$$

式中：$P_{i,t}^{\mathrm{ES}}$为储能充放电量，$P_{i,t}^{\mathrm{ES}}\geqslant 0$代表充电量，$P_{i,t}^{\mathrm{ES}}<0$代表放电量；$f_{i,t}^{\mathrm{ES}}$为$t$时刻在储能单元中存储的能量；$f_{\min}^{\mathrm{ES}}$，$f_{\max}^{\mathrm{ES}}$分别为储能单元的最小、最大容量；$f_{i,1}^{\mathrm{ES}}$，$f_{i,T}^{\mathrm{ES}}$分别为一天开始与结束时刻储能单元的能量；$\eta_{\mathrm{b}}^{\mathrm{ch}}$，$\eta_{\mathrm{b}}^{\mathrm{dis}}$为储能单元的充、放电效率。

1.2.4　可再生能源发电

风电聚合商在日前给出预测的24 h风电出力。风力预测相对误差通常大于负荷预测的相对误差，且该误差的标准差随着预测水平的增大而增大。本文保守估计其误差服从均值为0、标准差为δ的正态分布[24]。其出力可表示为

$$P_{\mathrm{wr},i,t}=P_{\mathrm{wf},i,t}+\Delta p_{\mathrm{w},i,t}\tag{16}$$

$$\Delta p_{\mathrm{w},i,t}\sim N(0,\delta_{\mathrm{w},i,t}^{2})\tag{17}$$

$$\delta_{\mathrm{w},i,t}=\frac{1}{50}Q_{\mathrm{w},i}+\frac{1}{5}P_{\mathrm{wf},i,t}\tag{18}$$

式中：$P_{\mathrm{wr},i,t}$为风电在t时刻的功率实际值；$P_{\mathrm{wf},i,t}$为风电在t时刻的功率预测值；$\Delta p_{\mathrm{w},i,t}$为风电功率预测误差；$\delta_{\mathrm{w},i,t}$为风电在$t$时刻的风电出力预测误差标准差；$Q_{\mathrm{w},i}$为风电装机容量。

风电设备的建造成本为一次性投入，本文将其忽略。

2　VPP与EV主从博弈模型

在Stackelberg主从博弈[25]中，假定博弈中的所有参与方都为理性人，以使己方利益最大化为目标。领导者先提出一个策略，然后跟随者根据领导者采取的策略，调整策略使自己的效用最大化。

在本文中，假定VPP为博弈主体，EV为博弈从体，根据DRL算法，得出每个博弈主体的最佳策略。各主体与环境相互作用，以优化其长期奖励为目标进行策略的学习。VPP的控制变量为$\{\lambda_t^{\mathrm{EV}},P_{i,t}^{\mathrm{ES}},P_t^{\mathrm{D}},P_{i,t}^{\mathrm{DG}},\forall i,\forall t\}$，其目标函数和约束条件如下：

$$\begin{cases}\min\displaystyle\sum_{t=1}^{T}\{C_t^{\mathrm{B}}-C_t^{\mathrm{EV}}+\sum_{i=1}^{N}C_{i,t}^{\mathrm{DG}}+L_t^{\mathrm{VPP}}\}\\ \text{s.t. 式(1)~式(2)，式(7)~式(18)}\end{cases}\tag{19}$$

$$C_t^{\mathrm{EV}}=\begin{cases}\displaystyle\sum_{k=1}^{K}\lambda_t^{\mathrm{EV}}P_{k,t}^{\mathrm{EV}} & P_{k,t}^{\mathrm{EV}}\geqslant 0\\ \displaystyle\sum_{k=1}^{K}\lambda_t^{\mathrm{EV}}P_{k,t}^{\mathrm{EV}} & P_{k,t}^{\mathrm{EV}}<0\end{cases}\tag{20}$$

$$P_t^{\mathrm{R}}=\sum_{i=1}^{K}P_{i,t}^{\mathrm{EV}}+\sum_{i=1}^{M}P_{i,t}^{\mathrm{ES}}-\sum_{i=1}^{N}P_{i,t}^{\mathrm{DG}}-P_{\mathrm{wr},i,t}-P_t^{\mathrm{D}}\tag{21}$$

$$\sum_{t=1}^{T}\lambda_t^{\mathrm{D}}=\sum_{t=1}^{T}\lambda_t^{\mathrm{EV}}\tag{22}$$

$$\lambda_{t,\min}^{\mathrm{EV}}\leqslant\lambda_t^{\mathrm{EV}}\leqslant\lambda_{t,\max}^{\mathrm{EV}}\tag{23}$$

式中：C_t^{EV}为整个充电站的用电成本，即VPP从EV获得的收入；λ_t^{EV}为VPP制定的t时刻EV的充、放电价格，满足对应时刻的价格上、下限约束；L_t^{VPP}为VPP各约束的惩罚项。

训练时，为处理模型中的等式约束，本文引入L_t^{VPP}作为VPP各约束的惩罚项，其表达式为

$$L_t^{\mathrm{VPP}}=\alpha^{\mathrm{EV}}|\sum_{t=1}^{T}\lambda_t^{\mathrm{D}}-\sum_{t=1}^{T}\lambda_t^{\mathrm{EV}}|+\alpha^{\mathrm{ES}}\sum_{i=1}^{M}|f_{i,1}^{\mathrm{ES}}-f_{i,T}^{\mathrm{ES}}|\tag{24}$$

式中：α^{EV}，α^{ES}为惩罚项的系数，取值为足够大的正数，以激励智能体满足模型约束。

训练中，为引导智能体满足每辆EV的SOC不等式约束，引入L_t^{EV}作为惩罚项，其计算公式为

$$L_t^{\mathrm{EV}}=\sum_{i=1}^{K}L_{i,t}^{\mathrm{EV}}\tag{25}$$

$$L_{i,t}^{\mathrm{EV}}=\begin{cases}\beta|e_{i,t,\min}-e_{i,t}| & e_{i,t}\leqslant e_{i,t,\min}\\ \beta|e_{i,t}-e_{\max}| & e_{i,t}\geqslant e_{\max}\end{cases}\tag{26}$$

式中：β为惩罚项的系数。

本文中，VPP的状态为时间、微型汽轮机发电量、电动汽车充电站充电桩使用率、储能SOC、DAM电价、电动汽车充电站的电价累计值、风电功率预测值，即

$$S^{\mathrm{VPP}}=\left\{t,P_{1:N,t}^{\mathrm{DG}},\eta_t^{\mathrm{EV}},f_{1:M,t}^{\mathrm{ES}},\lambda_t^{\mathrm{D}},\tau_t^{\mathrm{EV}},P_{\mathrm{wf},1:W,t}\right\}\tag{27}$$

其中　　$\tau_t^{\mathrm{EV}}=\sum_{t}\lambda_t^{\mathrm{EV}}$

VPP的动作为微型汽轮机发电变化量、电动汽车充电站充电价格、储能动作、日前售电量，即

$$A^{\mathrm{VPP}}=\left\{\Delta P_{1:N,t}^{\mathrm{DG}},\lambda_t^{\mathrm{EV}},P_{1:M,t}^{\mathrm{ES}},P_t^{\mathrm{D}}\right\}\tag{28}$$

3　基于深度强化学习的模型求解

强化学习基本框架如图1所示。

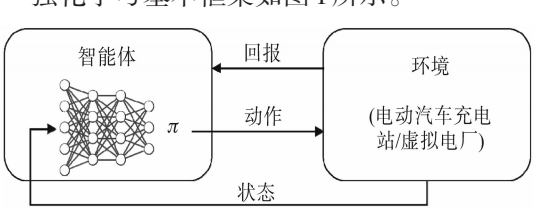

图1　强化学习基本框架

Fig.1　Basic framework for reinforcement learning

3.1　行动器-评判器算法框架

行动器-评判器(actor-critic，AC)框架是强化学习连续动作领域的一类重要算法，包含了

DDPG算法[26]、TD3算法[27]、SAC算法[28]等多种无模型(model-free)的、离轨策略(off-policy)的算法。其中，DDPG与TD3为确定性策略，SAC为随机性策略。

3.1.1　优化目标

强化学习算法的训练目标为通过与环境互动，寻找最优策略π^*，使得智能体在有限马尔科夫决策过程(Markov decision process，MDP)[29]中，累积回报的期望最大，即

$$\pi^* = \arg\max_{\pi} \mathop{\mathbb{E}}_{\tau\sim\pi}[R(\tau)] \tag{29}$$

$$R(\tau) = \sum_{t=0}^{T} r_t \tag{30}$$

式中：τ为策略π在环境中形成的状态-动作轨迹，即$\tau=(s_0,a_0,s_1,a_1,\cdots)$；$R(\tau)$为智能体在每幕的总回报；$r_t$为时刻$t$的回报。

策略则由参数为θ的神经网络表示，本文将确定性策略记为$\mu_\theta(s)$，即$a=\mu_\theta(s)$；将随机性策略记为$\pi_\theta(\cdot|s)$，即$a\sim\pi_\theta(\cdot|s)$。

为提高算法的探索能力，防止过快收敛，SAC算法采用了熵正则化，其目标函数为

$$J(\pi)=\sum_{t=0}^{T}\mathop{\mathbb{E}}_{\tau\sim\pi}\{r(s_t,a_t,s_{t+1})+\alpha H[\pi(\cdot|s_t)]\} \tag{31}$$

$$H[\pi(\cdot|s_t)]=\mathop{\mathbb{E}}_{a_t\sim\pi(\cdot|s_t)}[-\log P(a_t)] \tag{32}$$

式中：α为温度系数，即熵项的权重；H为在策略π、状态s_t下采取动作的熵项。

其贝尔曼方程(Bellman equation)为

$$Q^\pi(s,a)=\mathop{\mathbb{E}}_{\tau\sim\pi}[r(s,a,s')+\gamma V^\pi(s')] \tag{33}$$

$$V^\pi(s)=\mathop{\mathbb{E}}_{a\sim\pi}[Q^\pi(s,a)]+\alpha H[\pi(\cdot|s)] \tag{34}$$

式中：$V^\pi(s)$为状态值函数；γ为奖励折扣因子，表示一个状态的价值由该状态的奖励以及后续状态价值按一定的衰减比例求和组成。

在强化学习中，为得到最优策略π^*，核心思想是用价值函数来对最优策略进行结构化搜索，通过迭代策略评估来寻找满足贝尔曼方程的最优价值函数(optimal value function)V^*和Q^*。

3.1.2　动作的选择

智能体的动作由当前Actor网络的输出决定。对于确定性策略算法，为增加对环境的探索能力，对输出动作加噪处理，即

$$a_t=clip[\mu_\theta(s_t)+\varepsilon,a_{\mathrm{L}},a_{\mathrm{H}}]\quad \varepsilon\sim N(0,\sigma) \tag{35}$$

式中：$clip()$为将动作限制在$[a_{\mathrm{L}},a_{\mathrm{H}}]$范围内。

对于随机性策略算法，智能体的动作由网络输出参数确定的分布决定，即

$$a_t\sim\pi_\theta(\cdot|s_t) \tag{36}$$

3.1.3　网络的更新

AC框架的网络由策略网络(actor)、价值网络(critic)、目标策略网络、目标价值网络组成，其参数分别用θ，ϕ，θ_{targ}和ϕ_{targ}表示。策略网络采用策略梯度方法，进行梯度上升更新，用于建立由状态s_t到动作a_t的映射，对于DDPG和TD3，假如从缓冲记忆库D中抽取一批数据$B=[(s,a,r,s',\delta)]$，其网络参数更新梯度为

$$\nabla_\theta\frac{1}{|B|}\sum_{s\in B}Q_\phi[s,\mu_\theta(s)] \tag{37}$$

对于SAC，其网络参数更新梯度为

$$\nabla_\theta\frac{1}{|B|}\sum_{s\in B}\{\min_{j=1,2}Q_{\phi_j}[s,\tilde{a}_\theta(s)]-\alpha\log\pi_\theta[\tilde{a}_\theta(s)|s]\} \tag{38}$$

其中，为了使得式(38)可微，$\tilde{a}_\theta(s)$为通过重参数化技巧(reparameterization trick)得到的动作，本文采用挤压高斯策略(squashed Gaussian policy)获得，即

$$\tilde{a}_\theta(s,\xi)=\tanh[\mu_\theta(s)+\sigma_\theta(s)\odot\xi]\quad \xi\sim N(0,I) \tag{39}$$

式中：$\odot$为向量间对应元素相乘。

价值网络相当于传统强化学习算法中的状态值函数，即从初始状态出发得到的期望累积回报，采用梯度下降方法更新，目的是对策略网络建立的映射作出评价，即进行Q值估计。对于DDPG，其网络参数更新梯度为

$$\nabla_\phi\frac{1}{|B|}\sum_{(s,a,r,s',d)\in B}[Q_\phi(s,a)-y(r,s',d)]^2 \tag{40}$$

其中

$$y(r,s',d)=r+\gamma(1-d)Q_{\phi_{\mathrm{targ}}}[s',\mu_{\theta_{\mathrm{targ}}}(s')] \tag{41}$$

式中：$y(r,s',d)$为目标。

对于TD3和SAC，为避免出现DDPG中常见的价值高估问题，采用两个结构相同的价值网络估算Q值，并取最小值，其网络参数更新梯度为

$$\nabla_{\phi_j}\frac{1}{|B|}\sum_{(s,a,r,s',d)\in B}[Q_{\phi_j}(s,a)-y(r,s',d)]^2\quad j=1,2 \tag{42}$$

对于TD3，有：

$$y(r,s',d)=r+\gamma(1-d)\min_{j=1,2}Q_{\phi_{\mathrm{targ},j}}[s',a'(s')] \tag{43}$$

$$a'(s')=clip[\mu_{\theta_{\mathrm{targ}}}(s')+\varepsilon,a_{\mathrm{L}},a_{\mathrm{H}}]\quad \varepsilon\sim N(0,\sigma) \tag{44}$$

对于SAC，进一步采用了熵正则化技巧：

$$y(r,s',d)=r+\gamma(1-d)[\min_{j=1,2}Q_{\phi_{\text{targ},j}}(s',\tilde{a}')-\alpha\log\pi_\theta(\tilde{a}'|s')] \tag{45}$$

目标策略网络、目标价值网络的参数分别从策略网络、价值网络软更新，即

$$\theta_{\text{targ}}\leftarrow\tau\theta+(1-\tau)\theta_{\text{targ}} \tag{46}$$

$$\phi_{\text{targ}}\leftarrow\tau\phi+(1-\tau)\phi_{\text{targ}} \tag{47}$$

式中：τ为更新参数，本文取0.005。

3.2 含EV的VPP调度的深度强化学习算法流程

本文建立的模型中，EV充电站采用TD3算法训练智能体，而VPP采用SAC算法训练智能体。这两个智能体之间存在主从博弈关系。

本文采用交替训练的方法来模拟VPP与EV充电站多阶段博弈过程。为提高训练稳定性，VPP向EV售电价格采用软更新方法，计算公式为

$$\lambda_t^{\text{EV}}=\xi\lambda_{\theta,t}^{\text{EV}}+(1-\xi)\lambda_t^{\text{EV}} \tag{48}$$

式中：ξ为更新系数，本文取0.01。

在完成上述计算后，VPP在整个训练过程中，智能体可以记录训练过程中的日前电力市场购售电量的滑动平均值$\bar{P}_t^{\text{D}}$作为在日前向ISO申报的实际购售电量。

4 算例分析

在一个包含可再生能源、储能、分布式发电、电动汽车等资源的VPP中验证本文采用的强化学习方法。风电、电价曲线均取自北欧电力市场瑞典中南部地区2020年6月7日的数据[30]，并按汇率进行调整。

可再生能源平均出力及DAM电价曲线如图2所示。

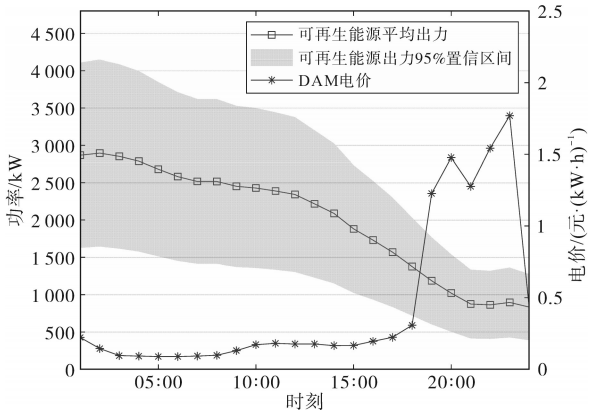

图2 可再生能源平均出力及DAM电价曲线

Fig.2 Renewable energy average output and DAM tariff curve

电动汽车充电负荷曲线来自文献[31]。电动汽车数量曲线如图3所示。

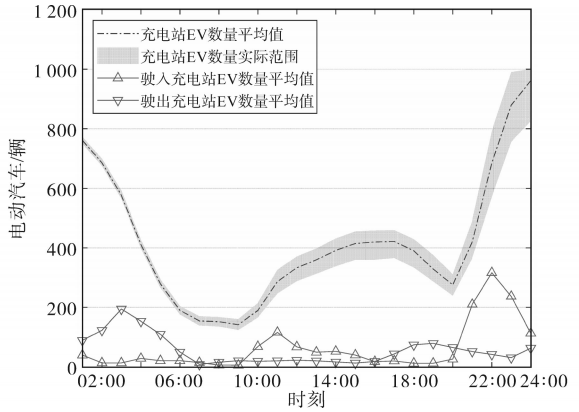

图3 电动汽车数量曲线

Fig.3 Electric vehicle volume curves

设置风电额定总有功出力为3 MW；可控分布式机组发电总有功出力为10 MW，其余参数如表1所示；储能容量1 MW·h，充、放电效率均为0.95；EV充电站容纳电动汽车总数为1 000辆，单个电动汽车的电池容量为50 kW·h，电池最大充、放电功率为10 kW，充、放电效率均为0.95；EV抵达充电站时的起始SOC服从正态分布$N(0.34,0.1)$，充电时间服从正态分布$N(8.5,1)$；风电出力预测误差标准差设为预测值的15%。

表1 虚拟发电厂分布式发电参数

Tab.1 Unit operation data of VPP distributed energy resource

参数	数值	参数	数值
a^{DG}/(¥/(MW²·h))	12.9	c^{off}/¥	0
b^{DG}/(¥/(MW·h))	1.5	$P_{\text{up}}^{\text{DG}}$/kW	1 000
c^{DG}/¥	2.0	$P_{\text{down}}^{\text{DG}}$/kW	1 000
c^{on}/¥	100	—	—

4.1 算法参数设置

为验证本文方法的有效性，分别采用DDPG，TD3，SAC作为EV充电站的强化学习算法，将TD3，SAC作为VPP的强化学习算法，形成6个算例，各算例算法的设置如表2所示。

表2 各算例算法设置

Tab.2 Algorithm settings for different cases

算法设置	VPP采用TD3	VPP采用SAC
EV充电站采用DDPG	Case 1	Case 4
EV充电站采用SAC	Case 2	Case 5
EV充电站采用TD3	Case 3	Case 6

对于EV充电站采用的算法，策略网络和价值网络的隐含层层数均为2层，每层有128个神经元，隐含层的激活函数均为Leaky ReLU（泄露修正线性单元），折扣因子为0.99，mini-batch大小为128，缓冲记忆库大小为20 000，τ为0.001，采用Adam优化器更新网络权重。DDPG算法的价值网络学习率为0.001，策略网络学习率为0.000 1；

TD3算法的价值网络学习率为0.001，策略网络学习率为0.001；SAC算法的价值网络学习率为0.000 5，策略网络学习率为0.000 3。

对于VPP采用的算法，策略网络和价值网络的隐含层层数均为2层，每层有256个神经元，TD3算法的价值网络学习率为0.000 1，策略网络学习率为0.000 2，其余参数与EV充电站的算法相同。

4.2　训练过程与收敛性

本文算法采用Python3.8编写，采用Pytorch 1.6.0作为深度学习框架。网络的参数更新使用CUDA并行计算架构加速，并在NVIDIA GeForce GTX 1660 GPU上执行；智能体与环境互动及其更新部分使用Numba[32]即时编译器技术加速。本文算法在Intel Core i7-8750H CPU @ 2.20GHz和内存为8GB的电脑上运行，每个完整的算例平均需要用69 min。

本文搭建的框架中，每小时计算一次回报，智能体的回报均为经济收益减去惩罚量。智能体根据获得的回报每小时对网络参数进行更新，则每24 h的训练为1幕（eposide）。设置EV充电站智能体的训练幕数为450，VPP智能体的训练幕数为7 350，EV充电站与VPP平均每幕的回报如图4和图5所示。

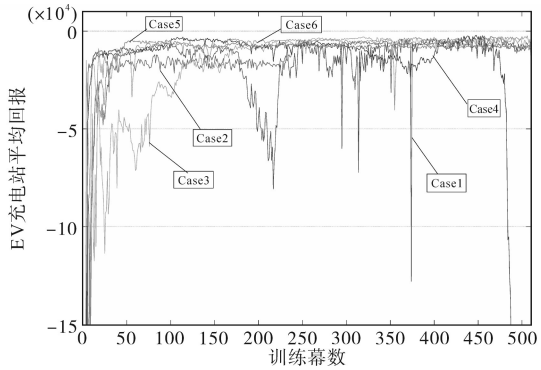

图4　EV充电站智能体训练过程的回报

Fig.4　Rewards of the EV charging station agent training process

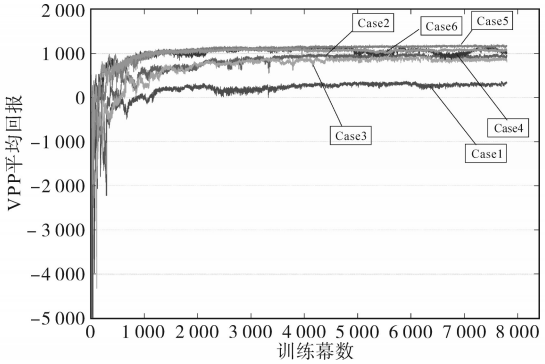

图5　VPP智能体训练过程的回报

Fig.5　Rewards of the VPP agent training process

训练过程中VPP向EV售电价格的迭代收敛情况如图6所示。

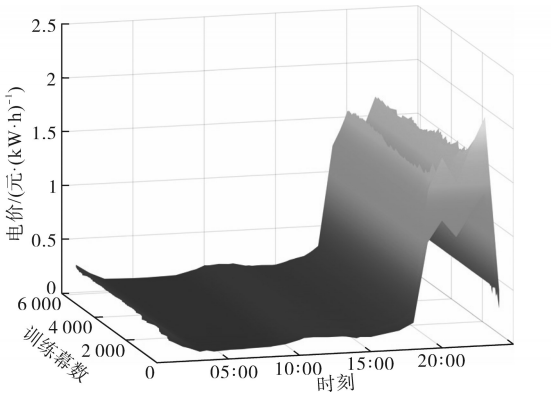

图6　VPP向EV售电价格迭代收敛情况

Fig.6　Iterative convergence of VPP to EV power sales prices

从图4～图6可以看出，EV智能体能在100幕后逐渐收敛，VPP智能体能够在4 000幕左右逐渐收敛，并在之后缓慢优化。VPP向EV售电价格随迭代的进行能够逐渐收敛。训练过程中收益的波动则主要受随机量影响。从训练过程来看，DDPG算法的稳定性和成功率不如TD3和SAC。从奖励数值来看，VPP采用SAC算法时获得的奖励更高。

4.3　强化学习优化调度结果分析

以TD3作为EV充电站的智能体训练算法，以SAC作为VPP的智能体训练算法，利用历史数据对智能体进行离线训练，其计算收敛后得到的调度结果如图7所示。

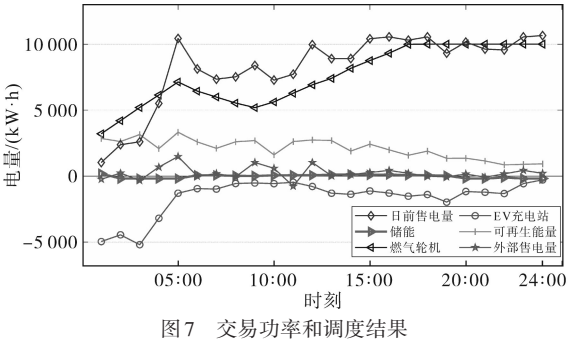

图7　交易功率和调度结果

Fig.7　Trading power and scheduling results

由图7可以看出，储能在电价的引导下充、放电，在00:00—18:00的低电价时段充电，而在18:00—24:00的高电价时段放电。燃气轮机在低电价时段发电少，在高电价时段前提前提高发电功率，并在高电价时段以额定功率发电。VPP制定的EV充电价格结果如图8所示。

由图8可以看出，由于式(23)的约束，VPP制定的EV充电价格始终在上、下限之间波动。EV充电站会根据价格信号改变用电行为，实现负荷

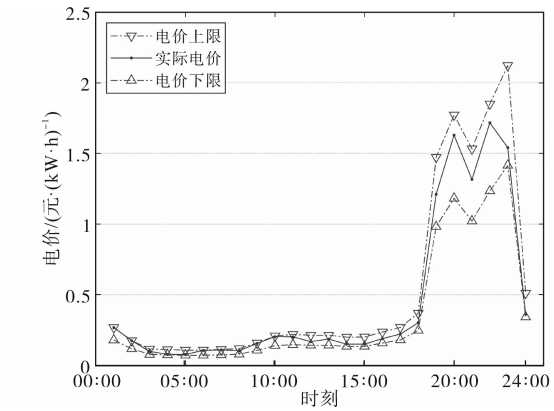

图8　EV充电站分时价格曲线

Fig.8　Price curves of EV time-sharing charging

削峰填谷并提高了供电可靠性，有利于VPP的长期稳定运营。

4.4　算法类型的影响

在线调度阶段，采用蒙特卡洛法计算VPP收益，分别用两种方案确定日前购售电量。

方案1：采用整个训练过程中，日前电力市场购售电量P_t^{D}的滑动平均值作为向ISO申报的实际购售电量。

方案2：采用整个训练过程中，日前电力市场购售电量减去实时平衡市场购售电量，即$P_t^{\mathrm{D}}-P_t^{\mathrm{R}}$的滑动平均值作为向ISO申报的实际购售电量。

在智能体训练完成后，采用蒙特卡洛方法模拟调度100 d，并取平均值作为收益结果。模型采用不同算法求解的结果如表3所示。从结果看出，在本文建立的VPP与EV充电站主从博弈模型中，当VPP采用随机性策略时，其净收益更高，且惩罚更小；当EV充电站采用确定性策略时，其成本更低。采用两种方法确定DAM实际购售电量时，VPP收益没有明显差异，这也能说明本文方法计算出的结果是较优的。

表3　各算例求解结果

Tab.3　The results of solving each case

VPP算法	EV充电站算法	日前购售电量方案	VPP收益/元	EV充电站成本/元	总惩罚
TD3	DDPG	方案1	84 475	8 619	5.90
TD3	SAC	方案1	79 414	14 970	5.97
TD3	TD3	方案1	80 012	15 823	3.58
SAC	DDPG	方案1	84 963	12 544	0.27
SAC	SAC	方案1	81 590	16 601	5.27
SAC	TD3	方案1	80 991	13 204	1.87
TD3	DDPG	方案2	84 167	8 607	6.16
TD3	SAC	方案2	78 541	15 019	5.91
TD3	TD3	方案2	81 203	15 859	3.77
SAC	DDPG	方案2	84 610	12 569	0.26
SAC	SAC	方案2	82 053	16 529	5.49
SAC	TD3	方案2	79 509	13 416	5.77

从实际训练过程来看，由于智能体的价值网络在起始时估值不准，DDPG算法可能存在价值高估的问题并在迭代中不断恶化，导致训练过程崩溃。因此本文建议用TD3作为EV充电站的算法。

从结果来看，VPP更倾向于采用基于随机性策略算法，即博弈论中的混合策略；而EV充电站倾向于采用确定性策略算法，即纯策略。这是由于VPP占据主从博弈中的主体地位，而EV充电站占据从体地位。根据海萨尼的证明[33]，完全信息情况下的混合战略均衡可以解释为不完全信息情况下纯战略均衡的极限。在不了解EV充电站支付矩阵的情况下，VPP不能确定EV充电站将选择什么样的纯策略。根据海萨尼转换，这种不确定性等价于VPP不确定EV充电站的具体类型。然而，对于占据从体地位的EV充电站，其参与的博弈是完全信息的，因此采用纯策略更优。

4.5　博弈的影响

为研究博弈对模型均衡解的影响，将VPP智能体的训练幕数设置为6 150，EV充电站智能体的训练幕数设置为450。在采用博弈模型时，采用3.2节流程交替训练主、从智能体。在不采用博弈模型时，则先训练完成VPP智能体，再训练EV充电站智能体。

采用方案1确定日前购售电量，分别计算在主从博弈和不采用主从博弈模型下的收益，如表4、表5所示。

表4　采用VPP主从博弈模型VPP收益与EV成本

Tab.4　VPP benefits and EV costs using VPP Stackelberg game model

VPP算法	EV充电站算法	VPP收益/元	EV充电站成本/元
TD3	DDPG	80 946	7 092
TD3	SAC	79 925	13 224
TD3	TD3	80 364	13 227
SAC	DDPG	78 227	11 789
SAC	SAC	80 713	17 286
SAC	TD3	79 554	10 201

表5　不采用VPP主从博弈模型VPP收益与EV成本

Tab.5　VPP benefits and EV costs without using VPP Stackelberg game model

VPP算法	EV充电站算法	VPP收益/元	EV充电站成本/元
TD3	DDPG	85 135	22 947
TD3	SAC	87 203	16 806
TD3	TD3	87 897	12 575
SAC	DDPG	89 281	15 453
SAC	SAC	88 749	24 941
SAC	TD3	92 594	16 497

从表中计算结果看出，主从博弈降低了EV充电站的成本，同时降低了VPP的总收益，这是由于VPP的部分收入来自于EV充电站支付的电费。虽然VPP的目标为收益最大化，但由于与EV充电站存在博弈，限制VPP过度榨取收益，使其只能获取相对最优的收益，且总体下降。本文建立的模型不仅能降低EV用户成本，且实现了全社会成本下降。

5　结论

本文针对电动汽车虚拟电厂调度问题，提出以SAC和TD3算法训练得到智能体并以此进行VPP和EV充电站调度，通过算例验证了方法的有效性，所得结论如下：

1）本文提出的虚拟电厂智能体能够学习向EV售电价格策略，对内部资源优化调度，参与电力市场交易；电动汽车聚合代理商智能体能够学习EV充、放电调度策略。

2）基于TD3强化学习方法能够对VPP内EV充电站调度进行优化控制，在电动汽车数量较多的情况下能够有效计算。

3）主从博弈降低了EV充电站的成本，也降低了VPP的总收益，这表明该算法可以限制VPP的过度榨取收益，使全社会成本降低。

参考文献

[1] 董军刚，王丽荣．一种分布式电源的自适应虚拟同步控制策略研究[J]. 电气传动，2019，49(7)：78-81，89.
DONG Jungang，WANG Lirong. Research on an adaptive virtual synchronization control strategy for distributed power sources [J]. Electric Drive，2019，49(7)：78-81，89.

[2] 赵星宇，胡俊杰．集群电动汽车充电行为的深度强化学习优化方法[J]. 电网技术，2021，45(6)：2319-2327.
ZHAO Xingyu，HU Junjie. Deep reinforcement learning based optimization method for charging of aggregated electric vehicles [J]. Power System Technology，2021，45(6)：2319-2327.

[3] 李航，李国杰，汪可友．基于深度强化学习的电动汽车实时调度策略[J]. 电力系统自动化，2020，44(22)：161-167.
LI Hang，LI Guojie，WANG Keyou. Real-time dispatch strategy for electric vehicles based on deep reinforcement learning[J]. Automation of Electric Power Systems，2020，44(22)：161-167.

[4] LI H，WAN Z，HE H. Constrained EV charging scheduling based on safe deep reinforcement learning[J]. IEEE Transactions on Smart Grid，2020，11(3)：2427-2439.

[5] SUN X，QIU J. A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method[J]. IEEE Transactions on Industrial Informatics，2021，17(10)：6852-6863.

[6] 周亦洲，孙国强，黄文进，等．多区域虚拟电厂综合能源协调调度优化模型[J]. 中国电机工程学报，2017，37(23)：6780-6790，7069.
ZHOU Yizhou，SUN Guoqiang，HUANG Wenjin，et al. Optimized multi-regional integrated energy coordinated scheduling of a virtual power plant[J]. Proceedings of the CSEE，2017，37(23)：6780-6790，7069.

[7] 陈新和，裴玮，邓卫，等．数据驱动的虚拟电厂调度特性封装方法[J]. 中国电机工程学报，2021，41(14)：4816-4828.
CHEN Xinhe，PEI Wei，DENG Wei，et al. Data-driven virtual power plant dispatching characteristic packing method[J]. Proceedings of the CSEE，2021，41(14)：4816-4828.

[8] 刘思源，艾芊，郑建平，等．多时间尺度的多虚拟电厂双层协调机制与运行策略[J]. 中国电机工程学报，2018，38(3)：753-761.
LIU Siyuan，AI Qian，ZHEN Jianping，et al. Bi-level coordination mechanism and operation strategy of multi-time scale multiple virtual power plants[J]. Proceedings of the CSEE，2018，38(3)：753-761.

[9] 王晛，张华君，张少华．风电和电动汽车组成虚拟电厂参与电力市场的博弈模型[J]. 电力系统自动化，2019，43(3)：155-162.
WANG Xian，ZHANG Huajun，ZHANG Shaohua. Game model of electricity market involving virtual power plant composed of wind power and electric vehicles[J]. Automation of Electric Power Systems，2019，43(3)：155-162.

[10] 张高，王旭，蒋传文．基于主从博弈的含电动汽车虚拟电厂协调调度[J]. 电力系统自动化，2018，42(11)：48-55.
ZHANG Gao，WANG Xu，JIANG Chuanwen. Stackelberg game based coordinated dispatch of virtual power plant considering electric vehicle management[J]. Automation of Electric Power Systems，2018，42(11)：48-55.

[11] 乔骥，王新迎，张擎，等．基于柔性行动器-评判器深度强化学习的电-气综合能源系统优化调度[J]. 中国电机工程学报，2021，41(3)：819-832.
QIAO Ji，WANG Xinying，ZHANG Qing，et al. Optimal dispatch of integrated electricity-gas system with soft actor-critic deep reinforcement learning[J]. Proceedings of the CSEE，2021，41(3)：819-832.

[12] 于一潇，杨佳峻，杨明，等．基于深度强化学习的风电场储能系统预测决策一体化调度[J]. 电力系统自动化，2021，45(1)：132-140.
YU Yixiao，YANG Jiajun，YANG Ming，et al. Prediction and decision integrated scheduling of energy storage system in wind farm based on deep reinforcement learning[J]. Automation of Electric Power Systems，2021，45(1)：132-140.

[13] 彭刘阳，孙元章，徐箭，等．基于深度强化学习的自适应不确定性经济调度[J]. 电力系统自动化，2020，44(9)：33-42.
PENG Liuyang，SUN Yuanzhang，XU Jian，et al. Self-adaptive uncertainty economic dispatch based on deep reinforcement

learning[J]. Automation of Electric Power Systems, 2020, 44(9):33-42.

[14] LIN L, GUAN X, PENG Y, et al. Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy[J]. IEEE Internet of Things Journal, 2020, 7(7): 6288-6301.

[15] YE Y, QIU D, SUN M, et al. Deep reinforcement learning for strategic bidding in electricity markets[J]. IEEE Transactions on Smart Grid, 2020, 11(2): 1343-1355.

[16] LIANG Y, GUO C, DING Z, et al. Agent-based modeling in electricity market using deep deterministic policy gradient algorithm[J]. IEEE Transactions on Power Systems, 2020, 35(6): 4180-4192.

[17] 李宏仲，王磊，林冬，等. 多主体参与可再生能源消纳的Nash博弈模型及其迁移强化学习求解[J]. 中国电机工程学报，2019，39(14)：4135-4150.
LI Hongzhong, WANG Lei, LIN Dong, et al. A Nash game model of multi-agent participation in renewable energy consumption and the solving method via transfer reinforcement learning[J]. Proceedings of the CSEE, 2019, 39(14): 4135-4150.

[18] ZHAN Y, LIU C H, ZHAO Y, et al. Free market of multi-leader multi-follower mobile crowdsensing: an incentive mechanism design by deep reinforcement learning[J]. IEEE Transactions on Mobile Computing, 2020, 19(10): 2316-2329.

[19] GU B, YANG X, LIN Z, et al. Multiagent actor-critic network-based incentive mechanism for mobile crowdsensing in industrial systems[J]. IEEE Transactions on Industrial Informatics, 2021, 17(9): 6182-6191.

[20] LI L, CHENG Q, XUE K, et al. Downlink transmit power control in ultra-dense UAV network based on mean field game and deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 15594-15605.

[21] PANDZIC H, MORALES J M, CONEJO A J, et al. Offering model for a virtual power plant based on stochastic programming [J]. Applied Energy, 2013, 105(5): 282-292.

[22] 郭红霞，白浩，刘磊，等. 统一电能交易市场下的虚拟电厂优化调度模型[J]. 电工技术学报，2015，30(23)：136-145.
GUO Hongxia, BAI Hao, LIU Lei, et al. Optimal scheduling model of virtual power plant in a unified electricity trading market[J]. Transactions of China Electrotechnical Society, 2015, 30(23): 136-145.

[23] 张高. 含多种分布式能源的虚拟电厂竞价策略与协调调度研究[D]. 上海：上海交通大学，2019.
ZHANG Gao. Bidding strategy and coordinated dispatch of virtual power plant with multiple distributed energy resources[D]. Shanghai: Shanghai Jiao Tong University, 2019.

[24] ORTEGA-VAZQUEZ M A, KIRSCHEN D S. Estimating the spinning reserve requirements in systems with significant wind power generation penetration[J]. IEEE Transactions on Power Systems, 2009, 24(1): 114-124.

[25] FUDENBERG D, TIROLE J. Game theory[M]. Cambridge, Mass: MIT Press, 1992.

[26] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6): A187.

[27] FUJIMOTO S, VAN Hoof H, MEGER D. Addressing function approximation error in actor-critic methods[J]. arXiv Preprint arXiv: 1802.09477, 2018.

[28] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning, PMLR, 2018: 1861-1870.

[29] SUTTON Richard S, BARTO Andrew G. 强化学习[M]. 俞凯，译. 第2版. 北京：电子工业出版社，2019：45-68.
SUTTON Richard S, BARTO Andrew G. Reinforcement learning[M]. Translated by YU Kai. Second Edition. Beijing: Publishing House of Electronics Industry, 2019: 45-68.

[30] NORD Pool. Historical market data[DB/OL]. [2021-01-04]. https://www.nordpoolgroup.com/historical-market-data.

[31] 罗卓伟，胡泽春，宋永华，等. 电动汽车充电负荷计算方法[J]. 电力系统自动化，2011，35(14)：36-42.
LUO Zhouwei, HU Zechun, SONG Yonghua, et al. Study on plug-in electric vehicles charging load calculating[J]. Automation of Electric Power Systems, 2011, 35(14): 36-42.

[32] Numba. Numba[DB/OL]. [2020-12-10]. https://github. com/numba/numba.

[33] 张维迎. 博弈论与信息经济学[M]. 上海：上海人民出版社，1996：267-268.
ZHANG Weiyin. Game theory and information economics[M]. Shanghai: Shanghai People's Publishing House, 1996: 267-268.

收稿日期：2022-06-24
修改稿日期：2022-07-08