电力自动化设备
Electric Power Automation Equipment
ISSN 1006-6047,CN 32-1318/TM

《电力自动化设备》网络首发论文

题目：基于马尔可夫决策过程的电动汽车充电站能量管理策略
作者：黄帅博，陈蓓，高降宇
DOI：10.16081/j.epae.202208031
收稿日期：2022-05-09
网络首发日期：2022-08-31
引用格式：黄帅博，陈蓓，高降宇．基于马尔可夫决策过程的电动汽车充电站能量管理策略[J/OL]．电力自动化设备．https://doi.org/10.16081/j.epae.202208031

网络首发时间：2022-08-31 14:15:26
网络首发地址：https://kns.cnki.net/kcms/detail/32.1318.TM.20220831.0947.001.html

基于马尔可夫决策过程的电动汽车充电站能量管理策略

黄帅博，陈　蓓，高降宇
(上海工程技术大学 电子电气工程学院，上海 201620)

摘要：电动汽车充电站作为并网分布式储能装置，是实现电动汽车与未来能源互联网深度融合的重要组成部分。考虑分时电价和电动汽车用户行为的不确定性，提出了以电动汽车充电站日运营成本最小化为目标的能量管理策略。为了减少对先验信息的依赖和约束，将优化问题建模为一个新的有限回合马尔可夫决策过程模型；基于传统成本模型提出奖惩回报函数，通过主动学习调度决策，得到每辆电动汽车的实时充放电行为；针对模型的高维状态空间问题，设计相应的状态空间和动作空间，采用一种卷积神经网络结构结合强化学习的方法，通过从原始数据观测中提取高质量的经验，获取最优调度策略以达到优化目标。仿真结果表明，与传统的充电策略相比，所提策略可以有效地降低充电站的日运营成本，保护电动汽车的电池，同时能满足电动汽车用户的充电需求。

关键词：电动汽车充电站；充电规划；马尔可夫决策过程；能量管理；深度强化学习
中图分类号：U469.72；TM73　　**文献标志码：**A　　DOI：10.16081/j.epae.202208031

0　引言

全球工业化的加速和人类物质需求的提高，导致化石原料不断减少及其带来的环境问题日益凸显。有数据显示，2021 年全球能源的消耗量达到 1.3865×10^{10} t 油当量，与此同时，化石燃料直接或间接地产生了 3.368×10^{10} t 的碳排放[1]，且仍有增长趋势。作为典型工业化产物的汽车工业，其发展迅速，私家车的数量显著增加，对化石原料的消耗不可忽视。目前，我国汽车油耗占全国油耗总量的 25%，对国外石油的依赖度已达到 60%[2]，长年累月的汽车燃料消耗将进一步加剧能源短缺问题。另一方面，传统的燃料汽车在消耗不可再生能源的过程中，会不可避免地排放一定量的有害气体，从而加剧环境恶化，不符合我国目前所倡导的碳达峰、碳中和新发展理念[3-4]。

新能源电动汽车 EV（Electric Vehicle）有望成为解决上述问题的有效措施之一。相较于传统的燃料汽车，EV 的动力主要来源于电能，其具有低/无污染、高能效等优点，因此 EV 的大规模使用对于改善环境、增强对可再生能源的消纳能力、提升电网的供电质量有积极的促进作用。其中，支持车网互动 V2G（Vehicle to Grid）技术[5-6]的 EV 能够作为柔性负荷，连接到电网中进行充/放电，此类新型负荷具有“时空”属性[7]，可以视作移动储能设备。但受个体用户行为的影响，其充电位置和充电时间分散且无序。因此，EV 的充电管理面临着诸多挑战：①随着 EV 数量的不断增加，充电需求也增加，且充电负荷会与电网其他负荷的用电高峰重合，导致充电成本过高[8]和供需不平衡问题；②EV 用户的停车和充电行为具有不确定性，EV 无序充电会导致电网电压波动，易引起电网的稳定性问题[9]。

针对上述问题，文献[10]提出了一种实时二进制优化模型，将线性规划方法和两阶段凸松弛方案相结合，实时计算接近最优的 EV 充电计划。然而，此类方法依赖模型预测估计 EV 的充电需求、到达时刻、离开时刻，但是在实际中很难得到精确模型。为了减少模型的不精确性对性能的影响，同时考虑到现实中存在的不确定性，近年来以马尔可夫决策过程 MDP（Markov Decision Process）为严格数学基础的强化学习方法被用于解决 EV 充电相关的优化调度问题。例如：文献[11]建立了离线的换电站调度模型，并设计了一种带基线的蒙特卡罗策略梯度强化学习算法求解近似最优解；文献[12]建立了基于博弈论的实时电力交互模型，并设计了一种迁移强化学习算法对模型进行求解。

需要指出的是，上述研究工作采用的是基于数据驱动的强化学习方法，所提模型的训练存在维数灾难或迭代次数过多的问题。为了解决随机环境中高维状态空间表征的问题，许多学者通过引入神经网络来提高强化学习模型对数据的拟合能力，例如：文献[13]提出了经验存储的深度强化学习方法，用于克服风电、光伏和负荷的不确定性变化，并以最大化微电网的经济利益和居民满意度为目标，但未考虑 EV 接入微电网所带来的影响；文献[14]提出了一种基于最大熵值的深度强化学习的充换电负荷实时优化调度策略，考虑了用户因素、

收稿日期：2022-05-09；**修回日期：**2022-08-09
基金项目：国家自然科学基金资助项目(62173222，62073139)；国家科技攻关计划重大项目(2020AAA0109301)
Project supported by the National Natural Science Foundation of China(62173222,62073139) and the National Key Technologies R&D Program of China(2020AAA0109301)

系统因素和市场因素，制定了不同的应用场景，但未考虑大量电池老化带来的经济成本问题。目前，关于电动汽车充电站 EVCS（Electric Vehicle Charging Station）参与“车-路-网”[15]的能量交互，并考虑其经济性和实用性的研究较少。大规模的 EVCS 作为 EV 与电网的“中间商”，是实现 EV 与未来能源互联网深度融合的重要组成部分。

基于上述分析，本文从 EVCS 的角度出发，考虑分时电价和 EV 用户行为的不确定性，将深度 Q 网络 DQN(Deep Q-Network)应用于并网 EVCS，进行 EV 充放电行为的在线优化调度，实现 EVCS 日运营成本最小化。首先，建立了由充电成本、老化成本、惩罚成本组成的传统成本模型，且考虑到传统 MDP 模型无法处理约束的缺点以及用户的行为存在不确定性，构建了一个新的有限回合 MDP 模型，并基于传统成本模型提出了 MDP 的奖惩回报函数；然后，针对随机环境下模型训练遇到的高维状态空间问题，设计了相应的状态空间和动作空间，并采用一种卷积神经网络结构结合强化学习的方法，通过从原始观测数据中提取高质量经验来趋近最优调度以达到优化目标；最后，基于某典型的公共社区停车场数据进行算例分析，验证本文所提基于 MDP 模型的能量管理策略在解决 EV 充放电调度问题方面的有效性和优越性。

1 EVCS 的系统架构

本文研究的 EV 充放电调度策略是由并网的 EVCS 决策执行，目的是通过区域内 EV 与电网进行电能交互，实现 EVCS 日运营成本最小化。EVCS 的结构示意图如图 1 所示。EVCS 可视为并网的分布式储能装置，其中双向直流充电桩用于 V2G 服务，双向交直流电力转换器在本地电网和 EVCS 之间传输电力，以保持直流母线的稳定性。EV 充电装置和电网侧的电力转换器共享 1 条直流母线，减少了基础设施投资，提高了能源转换效率。

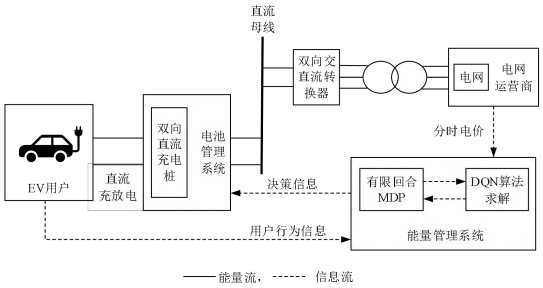


图 1 EVCS 的结构示意图

Fig.1 Structure diagram of EVCS

EVCS 通过协调 EV 用户的充电需求、荷电状态 SOC（State Of Charge）、电网的分时电价进行优化调度，使系统日运营成本最小化。调度过程涉及电网运营商、EV 用户、EVCS 运营商这 3 个角色，其中：电网运营商负责维护电网，并提供基础电力服务，且为了减轻电力需求负担、降低输电维护成本，电网倾向于采用分时电价，以鼓励终端用户参与需求侧能源管理；EVCS 运营商在运营过程中被认为是电网的价格接受者，这意味着其不影响电力市场的清算价格，可以通过电价差、提供停车服务获取利润。EVCS 运营商考虑分时电价和 EV 用户行为的不确定性，以日运营成本最小化和延缓电池老化为目标，制定 EV 的充放电优化调度策略。

2 数学模型

2.1 传统成本模型

EVCS 运营商通过优化调度 EV 的充放电行为以达到日运营成本最小化的目标，目标函数可以表示为：

$$\min C_{\text{tot}} = C_1 + C_2 + C_3 \tag{1}$$

式中：C_{tot} 为 EVCS 日运营成本；C_1 为 EV 充电成本；C_2 为 EVCS 惩罚成本；C_3 为 EV 电池老化成本。

2.1.1 EV 充电成本

EV 充电成本主要是指运行过程中 EV 充放电行为在分时电价差的作用下产生的费用，可以表示为：

$$C_1 = \sum_{t=1}^{T} C_{1,t} \tag{2}$$

$$C_{1,t} = \sum_{i=1}^{N} (e_i^{\text{g2v}} a_t^{\text{buy}} - e_i^{\text{v2g}} a_t^{\text{sell}}) \tag{3}$$

式中：T 为调度周期的时段数量（本文将 1 d 分为 24 个时段，即 $T=24$）；N 为调度周期内接入 EVCS 的 EV 总数量；$C_{1,t}$ 为 t 时段 EV_i 的充放电成本；e_i^{g2v}、e_i^{v2g} 分别为 EV_i 向电网购买、出售的电量；a_t^{buy}、a_t^{sell} 分别为 t 时段向电网购电、售电的电价。

2.1.2 EVCS 惩罚成本

EVCS 惩罚成本主要是指在运行过程中，若 EV 在离开 EVCS 时电池电量没有达到目标电量，则 EVCS 需向用户支付的罚款。若 EV 离开 EVCS 时电池电量大于等于目标电量，则不会产生罚款；若 EV 离开 EVCS 时电池电量小于目标电量，则未满足的电量将以单价 a^{pe} 进行罚款。则惩罚成本可表示为：

$$C_2 = \sum_{i=1}^{N} e_i^{\text{pe}} a^{\text{pe}} \tag{4}$$

$$e_i^{\text{pe}} = E_i^{\text{OFF}} - E_i^{\text{off}} \tag{5}$$

式中：e_i^{pe} 为 EV$_i$ 离开 EVCS 时时未满足目标的电量；E_i^{OFF} 为 EV$_i$ 用户预设的目标电量；E_i^{off} 为 EV$_i$ 离开 EVCS 时的电池电量。

2.1.3 EV 电池老化成本

长时间充放电调度会导致 EV 电池逐渐老化，可用容量不断衰减，性能下降。因此，EVCS 需要承担一部分充放电导致的 EV 电池老化成本，其主要受充放电功率、功率波动等不同因素影响，可表示为：

$$C_3=\sum_{i=1}^{N}(C_{1,i}+C_{2,i}) \tag{6}$$

$$C_1^i=\sum_{t=1}^{T}\delta(P_{t,i}\Delta t)^2 \tag{7}$$

$$C_{2,i}=\sum_{t=1}^{T-1}\beta(P_{t,i}\Delta t-P_{t+1,i}\Delta t)^2 \tag{8}$$

式中：$C_{1,i}$、$C_{2,i}$ 分别为 EV$_i$ 的自然充电损耗成本、充放电状态变化造成的老化成本；Δt 为 EVCS 进行优化调度的时间步长；$P_{t,i}$ 为 t 时段 EV$_i$ 的充放电功率；δ 为电池自然老化系数，是很小的正数；β 为充放电状态切换导致功率变化的老化系数。

电池损耗程度是老化成本的一个关键参数，充电功率会导致电池自然老化，但其损耗较小；充放电状态切换对电池造成的损耗较大，切换状态相邻时段的充放电功率波动越大，则对电池造成的损耗越大。虽然电力电子元器件减少了部分损耗，但充放电过程对电池造成的损耗仍不可忽视。

2.1.4 确定性约束条件

EVCS 在 EV 的可调度时段内将其充电至目标电量，在充电过程中需满足如下确定性约束条件：

$$-P_{\max}\le P_{t,i}\le P_{\max}\quad i=1,2,\cdots,N \tag{9}$$

$$z_t=L_t^{\mathrm{b}}+L_t^{\mathrm{EV}}=L_t^{\mathrm{b}}+\sum_{i=1}^{J}P_{t,i} \tag{10}$$

$$P_t^{\mathrm{grid}}=\sum_{i=1}^{J}P_{t,i} \tag{11}$$

式中：$P_{\max}$ 为 EV 的最大充放电功率，受充放电设备和电池容量限制，其值大于 0 表示充电，值小于 0 表示放电；P_t^{grid} 为 t 时段 EVCS 与电网交互的充放电功率；z_t 为 t 时段 EV 接入后的总负荷；L_t^{b} 为 t 时段电网的基础负荷；L_t^{EV} 为 t 时段 EVCS 内 EV 的综合负荷；J 为 t 时段进行实时充放电的 EV 数量。

2.1.5 不确定性约束条件

本文主要考虑了 EV 用户行为的不确定性，包括 EV 的到达时刻、离开时刻、初始 SOC，通常将这些不确定性因素理想化为服从某种概率分布进行数学模型。本文考虑 EV 充放电更为实际的情况，所提模型不依赖于概率分布，而是对用户数据集进行随机采样，并主动学习得到每辆 EV 的初始信息。则 EV 需满足的不确定性约束条件如下：

$$E_i^{\mathrm{off}}=E_i^{\mathrm{in}}+\sum_{t=t_i^{\mathrm{arr}}}^{t_i^{\mathrm{dep}}}\eta_i P_{t,i} \tag{12}$$

$$h_i=t_i^{\mathrm{dep}}-t_i^{\mathrm{arr}} \tag{13}$$

$$E_i^{\mathrm{in}}+\sum_{t=t_i^{\mathrm{arr}}}^{t_i^{\mathrm{dep}}}\eta_i P_{t,i}\ge E_i^{\mathrm{OFF}} \tag{14}$$

$$E_i^{\mathrm{in}}=E_i^{\mathrm{U}}e_{\mathrm{soc},i} \tag{15}$$

$$0\le E_i^{\mathrm{in}}+\sum_{t=t_i^{\mathrm{arr}}}^{t_i^{\mathrm{dep}}}\eta_i P_{t,i}\le E_i^{\mathrm{U}}\quad i=1,2,\cdots,N \tag{16}$$

式中：E_i^{in} 为 EV$_i$ 到达 EVCS 时的电池电量；h_i 为 EV$_i$ 的可调度时间；t_i^{arr}、t_i^{dep} 分别为 EV$_i$ 到达、离开 EVCS 的时刻；E_i^{U} 为 EV$_i$ 的电池最大限制容量；$e_{\mathrm{soc},i}$ 为 EV$_i$ 到达 EVCS 时电池的初始 SOC；η_i 为 EV$_i$ 充放电过程中的转换效率，可以从本地 EV 的充放电设备获得。

2.2 MDP 模型

传统成本模型存在如下问题：①式（7）、（8）、（12）、（13）假定了电池老化系数、到达 EVCS 的时刻、离开 EVCS 的时刻等参数；②当考虑耦合约束条件式（10）和式（11）时，在转移概率未知的情况下难以在有限的调度周期内获得最优解。传统 MDP 需要根据约束假定转移概率矩阵，往往无法应对 EV 充放电调度任务的广泛性和复杂性。为了解决上述问题，本文设计了有限回合 MDP 和相应的状态空间 $\boldsymbol{S}$、动作空间 $\boldsymbol{A}$，并基于传统成本模型设计决策过程中的奖惩回报函数 R，无需依赖具体的物理模型，可求解得到 EV 的实时充放电策略和 EVCS 的最优日运营成本。

2.2.1 状态空间

在单个时段 t（$t=1,2,\cdots,T$）内，EVCS 通过观察环境的信息特征积累经验，基于此选择充放电动作以达到优化目标。本文中的 EVCS 日运营成本取决于时间步长 Δt 内电价、充放电动作和 EV 到达/离开充电站时刻的变化，因此可以给出 MDP 和智能体的序贯模型，如附录 A 图 A1 所示。

针对后续求解过程中的状态空间维数问题，设计一个有限的状态空间 $\boldsymbol{S}_t$，如式（17）所示。

$$\boldsymbol{S}_t=[E_{t,i},E_i^{\mathrm{in}},E_i^{\mathrm{off}},x_{t,i},O_{\mathrm{leave},t,i},h_i,t,f_{\mathrm{price}}] \tag{17}$$

$$O_{\text{leave},t,i}=\begin{cases}1 & \text{离开}\\ 0 & \text{停留}\end{cases} \tag{18}$$

式中：$E_{t,i}$、E_i^{in}、E_i^{off}为电池状态位，$E_{t,i}$为t时段EV$_i$的电池电量；$x_{t,i}$为充放电状态位，表示t时段EV$_i$的充放电状态；$O_{\text{leave},t,i}$为停车状态位，表示t时段EV$_i$是否停留在EVCS，若停留则取值为0，若离开则取值为1；f_{price}为电网的分时电价。

MDP示意图如附录A图A1所示，智能体为接入EVCS的EV，当EV数量增多时，系统会出现高维状态空间，可根据智能体的数量进行空间划分，将其解耦[16]为多个单独状态子空间。因此，每个解耦子模型的状态空间维数为24×8阶，分别对应24个决策节点（各调度时段的开始时刻）的EV状态信息。

2.2.2 动作空间

在每个决策节点，智能体EV有充电、放电和不充不放这3种可能的动作状态，因此解耦子模型的动作空间为3元组，在不解耦的情况下t时段的空间大小为3^J。显然，解耦模型显著减小了优化问题的规模，提高了搜索速度，增强了实用性。

在解耦子模型i(对应于EV$_i$)中，用$x_{t,i}$表征智能体的充放电行为，具体取值为：

$$x_{t,i}=\begin{cases}-1 & \text{放电}\\ 0 & \text{不充不放}\\ 1 & \text{充电}\end{cases} \tag{19}$$

2.2.3 状态转移

在随机环境中，定义状态转移函数为$\boldsymbol{S}_{t+1}=f(\boldsymbol{S}_t,\boldsymbol{A}_t)$，其中$f$为MDP随机转移概率函数，其过程较难预测，类似为暗箱模型；下一个状态$\boldsymbol{S}_{t+1}$由当前状态$\boldsymbol{S}_t$和当前状态下采取的动作$\boldsymbol{A}_t$决定，如式（20）所示。

$$\boldsymbol{S}_{t+1}=[E_{t+1,i},E_i^{\text{in}},E_i^{\text{off}},x_{t+1,i},O_{\text{leave},t+1,i},h_t,t+1,f_{\text{price}}] \tag{20}$$

本文所设计有限回合MDP的状态转移的开始和结束由式（18）决定，充放电状态转移逻辑如附录A图A2所示。由于充放电状态的转移很难用准确的概率分布数学模型进行合理的描述，本文采用DQN算法进行求解，利用训练模型在经验样本中隐式地学习充放电状态转移的概率分布。

2.2.4 奖惩回报函数

EVCS日运营成本的传统模型以EV充电成本、EVCS惩罚成本、EV电池老化成本为优化目标，本节在此基础上，设计了MDP的奖惩回报函数。模型最终寻优决策使EVCS日运营成本最小化，所得智能体EV的策略由MDP中的奖惩回报函数进行评价，奖惩回报与智能体在当前状态的动作空间(搜索过程)中选择的动作是一一对应的。因此，t时段解耦子模型i的奖惩回报$r_{t,i}$与当前状态的电量有关，如式（21）所示。下一时段的电池电量$E_{t+1,i}$与当前状态选择的充放电行为有关，如式（22）所示。

$$r_{t,i}=\begin{cases}(a_t^{\text{buy}}-a_t^{\text{sell}})\times\min\{x_{t,i}\Delta t,E_{t,i}\} & \text{放电}\\ a_t^{\text{buy}}\times\max\{\Delta E_{t,i},0\}-a_t^{\text{sell}}\times\max\{-\Delta E_{t,i},0\} & \text{充电}\end{cases} \tag{21}$$

$$E_{t+1,i}=\begin{cases}E_{t,i}-\min\{x_{t,i}\Delta t,E_{t,i}\} & \text{放电}\\ E_{t,i} & \text{不充不放}\\ E_{t,i}+\min\{x_{t,i}\Delta t,E_i^{\text{U}}-E_{t,i}\} & \text{充电}\end{cases} \tag{22}$$

式中：$\Delta E_{t,i}=\min\{x_{t,i}\Delta t,E_i^{\text{U}}-E_{t,i}\}$。

3 DQN算法求解模型

3.1 DQN算法

针对传统MDP面临的维数灾难问题，即在环境交互过程中产生的状态空间很大且连续，无法用普通的查表法来求解每一个状态-动作价值Q的问题，本文采用DQN算法，使用深度神经网络来表示状态-动作Q值函数，通过与环境交互学习积累经验以训练求解模型。

MDP的常规求解方法包括数值迭代和策略迭代，实时动态规划算法[17]是改进的启发式搜索算法，但需要预先设定环境的动力学模型。在本文研究中采用的DQN算法无需具体的模型处理数据的不确定性。EV的到达时刻、离开时刻、初始SOC等信息是难以完美预测的，而本文所提方法不依赖于任何先验信息的假设，随机抽取EV接入EVCS。MDP模型能同时获得到达时刻t_i^{arr}、用户设定的充电需求电量E_i^{OFF}、离开时刻t_i^{dep}以及初始SOC $e_{\text{soc},i}$作为状态空间$\boldsymbol{S}$的初始状态信息，并将到达时刻、离开时刻分别作为有限回合MDP的开始和结束标志开始训练模型，通过与环境交互生成经验样本$(\boldsymbol{S}_t,\boldsymbol{A}_t,R,\boldsymbol{S}_{t+1},\text{end})$得到最优策略。DQN算法结构如附录A图A3所示。

3.2 DQN算法的实现

DQN算法不需要先验数据进行训练，而是通过智能体和环境交互记录相关的数据$(\boldsymbol{S}_t,\boldsymbol{A}_t,R,\boldsymbol{S}_{t+1},\text{end})$并将其存储为经验样本池，利用深度神经网络来表示$Q$值函数，且考虑到数据关联会导致网络参数不稳定，通过模型随机更新经验样本池。智能体只需要知道当前状态和动作列表，每个状态-动作组合都有一个与之相关的值，将其称为

状态-动作价值 Q。Q 值函数[18]可表示为：

$$Q(s_t,a_t)=r_t+\gamma\sum_t P\left(s_{t+1}|s_t\right)\pi(s)Q_\pi(s_{t+1}) \quad (23)$$

式中：r_t 为当前状态 s_t 执行动作 a_t 的奖惩回报；$\gamma\in(0,1)$ 为当前状态预期未来奖励的衰减因子；策略 $\pi(s)$ 为状态到动作的映射，表示当前状态 s_t 选择的动作 a_t 转移到下一个状态 s_{t+1}；$P\left(s_{t+1}|s_t\right)$ 为当前状态 s_t 到下一状态 s_{t+1} 的状态转移概率；$Q_\pi(s_{t+1})$ 为执行策略 $\pi(s)$ 后下一状态 s_{t+1} 的状态-动作价值。

由于 MDP 模型中 Q 值、转移概率矩阵是未知的，在训练过程中 DQN 引入了 2 个网络：①固定参数的目标 Q 值网络，用于固定步长更新参数；②根据评价策略更新参数的动作值函数逼近网络，在每一个时段内进行更新逼近，直至完成神经网络的训练。更新策略如下：

$$Q_{t+1}(s_t,a_t)=Q_t(s_t,a_t)+\alpha\left[\left(r_t+\gamma\max_{a_{t+1}}\{Q_t(s_{t+1},a_{t+1})\}\right)-Q_t(s_t,a_t)\right] \quad (24)$$

式中：$Q_t(s_t,a_t)$、$Q_{t+1}(s_t,a_t)$ 分别为当前时段、下一时段的状态-动作价值；$\alpha\in(0,1)$ 为学习率。式（24）表示状态-动作组合 (s_t,a_t) 的下一个时段 Q 值为当前时段的 Q 值加上学习率和下一次估计的误差乘积。新的估计值是当前时段的 Q 值与下一个状态下可能的最大 Q 值之和。

在本文设计的有限回合 MDP 模型中，如果这个回合有终止标识符，则在这个过程中就不再有未来的状态。因此，式（24）中含 γ 的项在更新过程中会衰减至 0，算法的伪代码见附录 A 表 A1。

4 仿真算例分析

4.1 算例设置

本文以某典型公共社区停车场结合某届“电工杯”的 EV 用户真实数据[19]为算例进行仿真分析，部分数据见附录 B 表 B1。算例的仿真规模主要受 J 和 N 这 2 个参数影响，本文设定 J 的取值范围为[0，40]辆，$N=200$ 辆，即每天接入 EVCS 的 EV 总数量为 200 辆，且各时段 EVCS 内的 EV 数量不超过 40 辆。每天随机抽取 EV 用户数据，通过仿真验证所提 MDP 充放电优化调度策略的可行性和有效性。其中 EV 离开充电站时的目标电量 E_i^{OFF} 由用户设置，EV 电池的容量为 20 kW·h，充放电功率为 3 kW（充电时功率为正值，放电时功率为负值），充放电效率 η=0.95。进行调度决策的周期为 24 h，时段间隔为 1 h，期间功率不变，模型参数设置如附录 B 表 B2 所示。

在 EVCS 的运行过程中，采用我国局部地区的分时电价作为购电电价，具体如附录 B 表 B3 所示。且考虑到与电网交互功率会产生相关的维护费用，则 EVCS 实时地将售电电价 a_t^{sell} 设定为购电电价 a_t^{buy} 的 95%，如式（25）所示。

$$a_t^{\mathrm{sell}}=a_t^{\mathrm{buy}}\times 95\% \quad (25)$$

采用第 3 节所提方法训练 MDP 模型，在运行过程中以 EV 的到达时刻、离开时刻分别作为有限回合 MDP 模型的开始和结束标志。模型中 $\boldsymbol{S}$ 含有 8 个变量，即输入的初始状态信息为 8 维向量；神经网络结构采用 2 层全连接层，每层的神经元个数分别为 38、16；输出变量维数与动作空间维数一致为 3。模型的学习率 α=0.002，衰减因子 γ=0.8；设置 EVCS 未满足充电需求的惩罚单价 a^{pe}=1.2 元/(kW·h)。训练过程共进行 12000 个回合，每和回合随机得到 EV 的到达时刻、离开时刻、初始 SOC。

4.2 成本结果分析

4.2.1 不同成本函数的优化结果

为了验证所提 EVCS 能量管理策略的可行性和有效性，基于真实的用户数据进行 MDP 模型训练，并以日运营成本最小化为目标进行优化调度，结果如图 2 所示。对模型训练 12000 个回合，行为策略选用 ε-贪心探索，将前 4000 个回合作为经验样本池，此时 ε=0.8，进行数据初始化后随机选择动作，该过程不学习动作的选择而仅积累经验；训练 4000 个回合之后，模型开始学习搜索最优的动作，ε 在该过程中逐渐减少至 0.0001 并保持不变。ε 的衰减过程表示智能体 EV 从随机选择逐渐转变为“聪明”地选择最优动作。

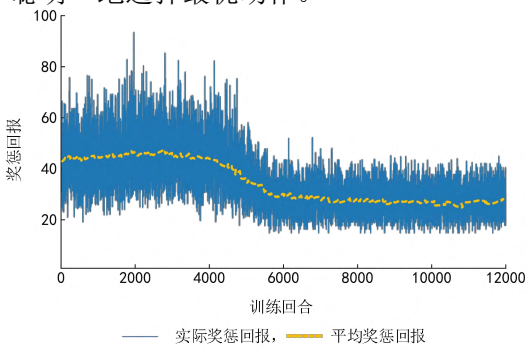


图 2 MDP 模型的训练结果

Fig.2 Training results of MDP model

从图 2 中的曲线可看出，局部区域存在波动，这是因为各训练回合开始随机抽取 EV，且用户的行为存在不确定性，即 EV 的离开会使电量和功率状态突然发生改变，该回合的结束环境也需进行初始化，导致各训练回合存在可控的状态差异，使得奖惩回报曲线产生了一定的波动。

设置相同的超参数，MDP 模型考虑 EV 电池

老化成本对求解过程产生的影响结果见附录 B 图 B1。由图可看出：模型初期训练的过程大致相同，均在进行随机探索和经验积累，是不断试错的过程；训练 4000 个回合左右时，考虑、不考虑 EV 电池老化成本的方案基本寻得一致的收敛方向，由于 EVCS 需要求解充放电切换造成的电池容量损耗，考虑老化成本会使模型的收敛速度更慢，且收敛过程波动更大，同时前期的经验存储也更复杂，这增加了模型的训练难度；最终考虑、不考虑 EV 电池老化成本的方案都能收敛，且考虑电池损耗确实增加了少量的成本，但延缓了电池老化，这更符合实际情况。

总体而言，在超参数相同的情况下，考虑、不考虑 EV 电池老化成本的方案都能稳定收敛，虽然考虑老化成本的方案在模型训练前期的难度增大，但随着训练回合的进行，考虑老化成本带来的影响逐渐减小，2 种方案基本在相同的训练时间内稳定收敛，进行实时调度。

4.2.2　不同策略的成本结果

为了评估本文所提基于 MDP 模型的能量管理策略（本文策略）的有效性，将其与随机延迟充电 RND（Randomly Delayed Charging）[20]策略进行对比分析。2 种策略下的日运营成本比较如图 3 所示（左侧、右侧条形分别对应本文策略、RND 策略）。由图可知：相较于 RND 策略，本文策略下 EVCS 的日运营成本明显减少，下降了 33.6%左右；RND 策略未满足用户需求产生的 EVCS 惩罚成本普遍高于本文策略，且第二天的惩罚成本最大；考虑了 EV 电池老化成本的本文策略利用分时电价差，减少了部分充电成本，但也产生了少量的电池老化成本。

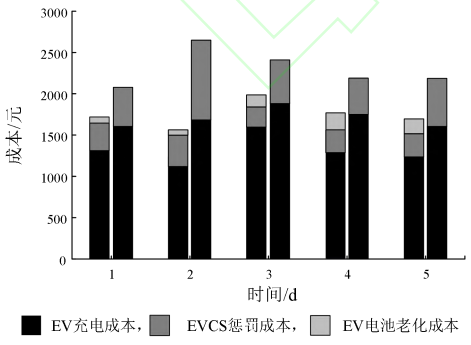


图 3　本文策略和 RND 策略下的日运营成本比较

Fig.3 Comparison of daily operation cost between proposed strategy and RND strategy

2 种策略下 EVCS 的具体成本（5 d）比较如表 1 所示。由表可知：相较于 RND 策略，EVCS 在本文策略下运营 5 d，考虑了 EV 电池老化成本的总运营成本为 8661 元，减少了 33.6%左右，其中充电成本在电价差的作用下为 6618 元，减少了 16.6%左右，EV 电池老化成本和 EVCS 惩罚成本之和为 2043 元，减少了约 43.2%左右；EVCS 经过 5 d 的运营，本文策略、RND 策略下平均每辆 EV 的日运营成本分别为 8.66、11.56 元。基于前文的分析，本文策略通过奖惩回报和优化调度充放电行为以适应不同的用户需求，达到了日运营成本最优。

表 1　本文策略和 RND 策略下 EVCS 的成本比较（5 d）

Table 1 Comparison of EVCS cost between proposed strategy and RND strategy(5 days)

单位：元

策略	EV 充电成本	售电收入	EVCS 惩罚成本	EV 电池老化成本	总运营成本
RND 策略	7937	0	3616	0	11553
本文策略	11654	5036	1699	344	8661

4.3　充放电行为分析

本文策略以 EVCS 日运营成本最小化为目标实时调度 EV 的充放电行为。为了更直观地说明 EV 充放电状态的变化，选取 20 辆 EV 的充放电过程进行分析，并验证考虑电池老化成本的本文策略的有效性。

本文策略下 20 辆 EV 的 SOC 变化曲线如图 4 所示。由图可知：当 EV 到达 EVCS 的时刻处于峰时段（10:00—14:00、17:00—20：00）内时，若 EV 的初始 SOC 较高，则采取放电策略，若 EV 的初始 SOC 较低，则在满足充电需求的前提下，采取不充不放策略；当 EV 到达 EVCS 的时刻处于平时段（07:00—10:00、15:00—17:00）内时，不论 EV 的初始 SOC 是高还是低，都会采取充电策略。可见，在电网峰平谷分时电价的作用下，EVCS 倾向于在峰时段提供 V2G 服务，在其他时段为 EV 充电，在降低充电成本的同时，减少电网峰时段的用电压力。

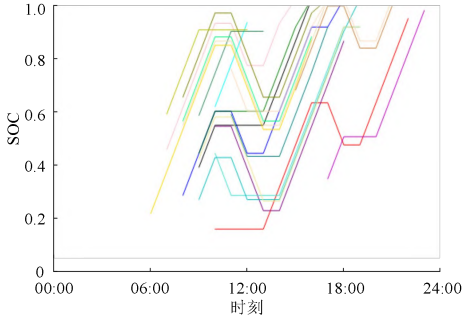


图 4　本文策略下 20 辆 EV 的 SOC 变化曲线

Fig.4 SOC curves of twenty EVs under proposed strategy

此外，图 4 中 EV 充电时 SOC 呈上升趋势，放电时 SOC 呈下降趋势，当 EV 的充放电状态发生改变时，SOC 会保持一段时间不变，这是因为本文策略综合考虑电池老化和用户需求，延长了充放电状态的切换时间，减小了充放电功率的波动。

为了验证考虑老化成本的本文策略的有效性，同样选取图 4 中对应的 20 辆 EV，比较了考虑电池老化成本的基于 MDP 模型的能量管理策略（本文策略）、不考虑电池老化成本的基于 MDP 模型的能量管理策略（MDP 对比策略）、RND 策略下 EV 充放电功率和功率波动，结果如图 5 所示。

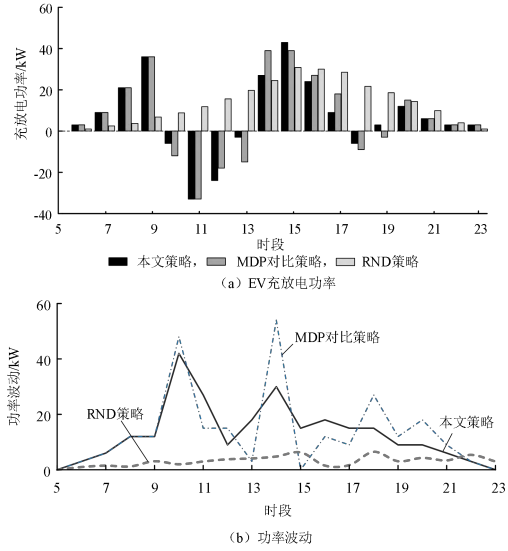


图 5 不同的策略下 20 辆 EV 的充放电功率比较

Fig.5 Comparison of charging and discharging of twenty EVs among different strategies

由图 5 可知：RND 策略下 EV 的充放电功率平缓，波动很小，但由表 1 可知该策略下的惩罚成本较高；MDP 对比策略能够充分利用峰谷电价差，降低运营成本，但是没有考虑电池损耗，导致在切换充放电状态前、后的功率波动较大（例如在时段 10、14、18）；本文策略能够明显减少充放电状态切换前、后的功率波动，而功率波动的减小有利于延长电池的寿命，更符合实际应用需求，且提高了 EV 接入电网时的安全性和稳定性。

总体而言，当 EVCS 面对相同的 EV 用户时：由于用户行为具有不确定性，RND 策略难以满足部分用户的充电需求，虽然充放电功率平缓，不产生电池老化成本，但增加了产生惩罚成本的可能性；本文策略针对充电成本，调度 EV 在峰时段放电，在平、谷时段充电，针对电池老化问题，延长了充放电状态切换时间，以减小相邻时段的功率波动，且全程考虑了用户的充电需求，使日运营成本比 RND 策略降低了 33.6%左右。

5 结论

本文利用 EV 可作为移动储能设备的优点，考虑充电成本、电池老化成本和惩罚成本，降低 EV 到达充电站时刻、离开充电站时刻等不确定性因素对优化目标的影响，将有限回合 MDP 模型应用于 EVCS 的能量管理策略。

为了延缓 EV 电池老化，本文考虑了 EVCS 切换充放电功率造成的电池损耗，适当延长切换时间并采用电力电子元器件进行充放电控制，在一定程度上延缓了电池老化。另外，考虑用户需求设置了相应的惩罚成本，算例结果表明基于 MDP 模型的能量管理策略能基本实时满足 EV 用户的充电需求，具有很强的实用性和扩展性。未来将基于更多的 EV 真实数据进行研究，针对不同用户的特征建立多时间尺度调度模型以进一步完善调度策略，增强鲁棒性的同时，更好地满足用户的需求。

参考文献：

[1] REVEL D. BP statistical review of world energy 2014[EB/OL]. [2022-05-09]. https://www.semanticscholar.org/paper/BP-Statistical-review-of-world-energy-2014-Revel/f266ec52951523fd111d501de0ab3a737c02a702.

[2] 刘振亚. 全球能源互联网跨国跨洲互联研究及展望[J]. 中国电机工程学报，2016，36(19)：5103-5110,5391.
LIU Zhenya. Research of global clean energy resource and power grid interconnection[J]. Proceedings of the CSEE,2016,36(19):5103-5110,5391.

[3] 彭光博，向月，陈文淑乐，等. “双碳”目标下电力系统风电装机与投资发展动力学推演及分析[J/OL]. 电力自动化设备. [2022-04-28]. https://doi.org/10.16081/j.epae.202205013.

[4] 余苏敏,杜洋,史一炜,等. 考虑 V2B 智慧充电桩群的低碳楼宇优化调度[J]. 电力自动化设备,2021,41(9):95-101.
YU Sumin,DU Yang,SHI Yiwei,et al. Optimal scheduling of low-carbon building considering V2B smart charging pile groups[J]. Electric Power Automation Equipment,2021,41(9):95-101.

[5] 吕耀棠,管霖,赵琦,等. 充电式电动汽车停车场的 V2G 模型及接入配电网方案优化研究[J]. 电力自动化设备,2018,38(11):1-7,14.
LÜ Yaotang,GUAN Lin,ZHAO Qi,et al. Research on V2G model of EPVV and its optimal scheme accessing to distribution network[J]. Electric Power Automation Equipment,2018,38(11):1-7,14.

[6] 苏舒，林湘宁，张宏志，等. 电动汽车充电需求时空分布动态演化模型[J]. 中国电机工程学报，2017，37(16)：4618-4629，4887
SU Shu,LIN Xiangning,ZHANG Hongzhi,et al. Spatial and temporal distribution model of electric vehicle charging demand[J]. Proceedings of the CSEE,2017,37(16):4618-4629,4887.

[7] 万雄，彭忆强，邓鹏毅，等. 电动汽车 V2G 关键技术研究综述[J]. 汽车实用技术，2020(2)：9-12.
WAN Xiong,PENG Yiqiang,DENG Pengyi,et al. Summary of research on key technologies of electric vehicle V2G[J]. Automobile Applied Technology,2020(2):9-12.

[8] KANDIL S M,FARAG H E Z,SHAABAN M F,et al. A combined resource allocation framework for PEVs charging stations,renewable energy resources and distributed energy storage systems[J]. Energy,2018,143:961-972.

[9] 原凯,宋毅,李敬如,等. 分布式电源与电动汽车接入的谐波特征研究[J]. 中国电机工程学报,2018,38(增刊 1)：53-57.
YUAN Kai,SONG Yi,LI Jingru,et al. Harmonic characteristics of distributed generation and electric vehicle

supplying access to the grid[J]. Proceedings of the CSEE, 2018,38(Supplement 1):53-57.
[10] YAO L,LIM W H,TSAI T S. A real-time charging scheme for demand response in electric vehicle parking station[J]. IEEE Transactions on Smart Grid,2017,8(1):52-62.
[11] 张文昕，栗然，臧向迪，等. 基于强化学习的电动汽车换电站实时调度策略优化[J/OL]. 电力自动化设备. [2022-05-09]. https://doi.org/10.16081/j.epae.202203003.
[12] ZHANG X S,BAO T,YU T,et al. Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid[J]. Energy,2017,133:348-365.
[13] 赵鹏杰，吴俊勇，王燚，等. 基于深度强化学习的微电网优化运行策略[J/OL]. 电力自动化设备. [2022-05-09]. https://doi.org/10.16081/j.epae.202205032.
[14] 刘敦楠,王玲湘,汪伟业,等. 基于深度强化学习的大规模电动汽车充换电负荷优化调度[J]. 电力系统自动化,2022,46(4):36-46.
LIU Dunnan,WANG Lingxiang,WANG Weiye,et al. Optimal scheduling of electric vehicle load for large-scale battery charging and swapping based on deep reinforcement learning[J]. Automation of Electric Power Systems,2022,46(4):36-46.
[15] XIANG Y,YANG J P,LI X C,et al. Routing optimization of electric vehicles for charging with event-driven pricing strategy[J]. IEEE Transactions on Automation Science and Engineering,2022,19(1):7-20.
[16] WU Y,ZHANG J,RAVEY A,et al. Real-time energy management of photovoltaic-assisted electric vehicle charging station by Markov decision process[J]. Journal of Power Sources,2020,476:228504.
[17] 傅质馨，李潇逸，朱俊澎，等. 基于马尔科夫决策过程的家庭能量管理智能优化策略[J]. 电力自动化设备，2020，40(7)：141-152.
FU Zhixin,LI Xiaoyi,ZHU Junpeng,et al. Intelligent optimization strategy of home energy management based on Markov decision process[J]. Electric Power Automation Equipment,2020,40(7):141-152.
[18] 沈国辉,赵荣生,董晓,等. 基于多信息交互与深度强化学习的电动汽车充电导航策略[J]. 南方电网技术,2022,16(1):108-116.
SHEN Guohui,ZHAO Rongsheng,DONG Xiao,et al. Electric vehicle charging navigation strategy based on multi-information interaction and deep reinforcement learning[J]. Southern Power System Technology,2022,16(1):108-116.
[19] 张谦，蔡家佳，刘超，等. 基于优先权的电动汽车集群充放电优化控制策略[J]. 电工技术学报，2015，30(17)：117-125.
ZHANG Qian,CAI Jiajia,LIU Chao,et al. Optimal control strategy of cluster charging and discharging of electric vehicles based on the priority[J]. Transactions of China Electrotechnical Society,2015,30(17):117-125.
[20] CHANDRA MOULI G R,KEFAYATI M,BALDICK R,et al. Integrated PV charging of EV fleet based on energy prices, V2G, and offer of reserves[J]. IEEE Transactions on Smart Grid,2019,10(2):1313-1325.

作者简介：


黄帅博

黄帅博(1997—)，男，硕士研究生，研究方向为智能电网优化调度（**E-mail:** 18015938513@163.com）；

陈　蓓(1985—)，女，副教授，博士，通信作者，研究方向为滑模控制、微电网系统（**E-mail:** chenbei1631@163.com）；

高降宇(1995—)，男，硕士研究生，研究方向为微电网的合作博弈（**E-mail:** 15755336825@163.com）。

(编辑　陆丹)

Energy management strategy of electric vehicle charging station based on Markov decision process

HUANG Shuaibo,CHEN Bei,GAO Jiangyu

(School of Electrical and Electronic Engineering,Shanghai University of Engineering and Technology,Shanghai 201620,China)

Abstract: As a grid-connected distributed energy storage device, the EVCS(electric vehicle charging station) is an important part of realizing the deep integration of EV(electric vehicles) and the future energy Internet. Considering the time-of-use electricity price and the behavioral uncertainty of EV users, energy management strategy is proposed with the optimal daily operating cost of EVCS as the goal. In order to reduce the dependence on prior information and constraints. First, the charge and discharge scheduling problem are modeled as a new finite round MDP(Markov decision process), and the quantified traditional cost model is used as a model reward and punishment function. By actively learning the scheduling decisions, the real-time charging and discharging action of each EV is solved. Aiming at the high-dimensional state space problem encountered in model, the corresponding state space and action space are designed. By using a convolutional neural network structure combined with reinforcement learning method, the high-quality experience in the original data is extracted to obtain an optimal scheduling strategy, and achieve the optimization goal. The simulation results show that compared with the traditional charging strategy, the proposed strategy can effectively reduce the daily operating cost of EVCS and protect EV batteries, while meeting the needs of EV users.

Key words:electric vehicle charging station;charging planning;Markov decision process;energy management; deep reinforcement learning

附录 A

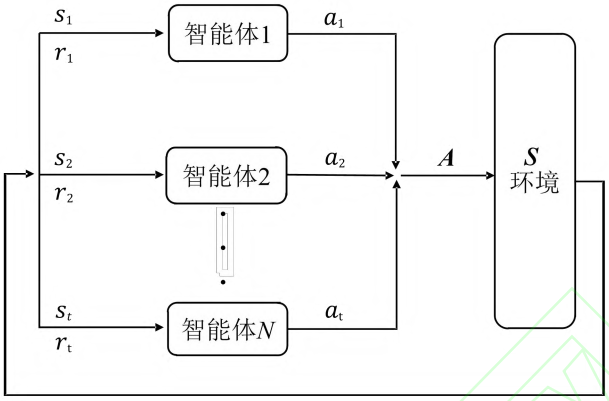


图 A1　MDP 示意图

Fig.A1　Schematic diagram of MDP

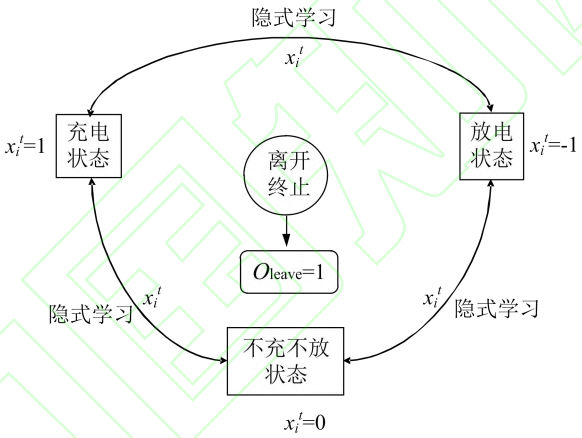


图 A2　充放电状态转移逻辑

Fig.A2　Transfer logic of charging and discharging state

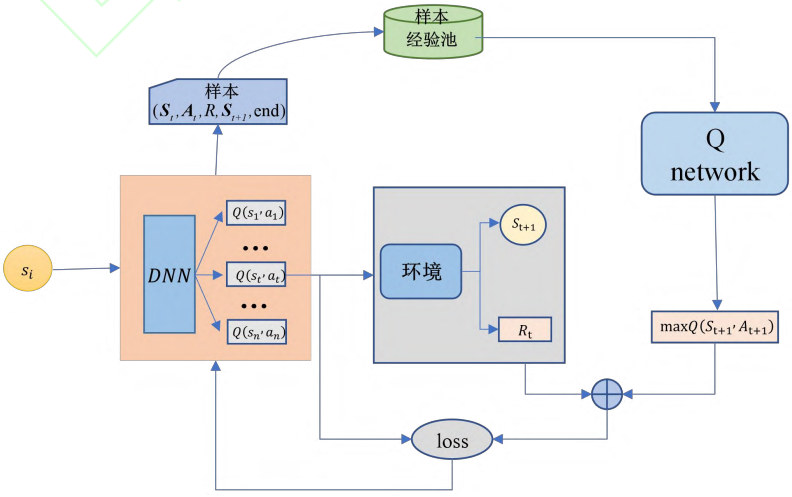


图 A3　DQN 算法结构

Fig.A3　Structure of DQN algorithm

表 A1　DQN 算法伪代码

Table A1　Pseudo-code of DQN algorithm

Algorithm:	DQN Algorithm for the finite-horizon MDP mode
Input：	N; $\boldsymbol{A}$; γ; ε; Q structure; samples; t
output：	Q-network parameters; The optimal strategy $\boldsymbol{\pi}^* / Q^*$
1：	for episode =1…N do
2：	Randomly initialize Q network parameters
3:	for i = 1…T do
4:	Initialize S_1 in St to get its feature $\phi(s_t)$ enter the Q-network to get the Q table
5:	With probability ε select a random action a_t
6:	Otherwise select a_t = max Q^*
7:	Execute a_t and get S_{t+1} corresponding $\phi(s_{t+1})$ and r; whether to terminate the symbol end
8：	Store $(\boldsymbol{S}_t, \boldsymbol{A}_t, R, \boldsymbol{S}_{t+1}, \text{end})$ in memory
9:	$\boldsymbol{S}_{t+1} = \boldsymbol{S}_t$
10:	Sample random minibatch of $(\boldsymbol{S}_t, \boldsymbol{A}_t, R, \boldsymbol{S}_{t+1}, \text{end})$ from D, get Q reality value y_j
11:	if end = True : Set $y_j = r_j$
12:	Else set $y_j = r_j + \gamma \max Q(s_{t+1}, a_{i+1}, \omega)$
13:	Perform a gradient descent step on $\left(y_j - Q(s_j, a_j, w)\right)^2$ (loss)Q Reality-Q estimate
14:	if end = True :
15:	break
16:	Else Repeat from step 5 and Return Q network

附录 B

表 B1　EV 用户参数

Table B1　Parameters of EV users

EV 编号	充电开始时刻	可调度时长/h	充电电量/(kW·h)	电池老化程度/%
1	9:03	5.08	10.45	12
2	8:24	6.29	8.35	5
3	10:04	6.88	12.5	8
4	10:04	4.64	13.38	10
5	9:00	0.37	1.72	7
6	8:07	10.09	7.64	13
7	10:15	10.18	17.31	20
8	12:23	17.11	12.28	13
9	9:15	11.17	1.84	9
10	15:04	5.69	15.37	6
⋮	⋮	⋮	⋮	⋮
1000	17:15	14.98	12.52	3

表 B2　仿真参数设置

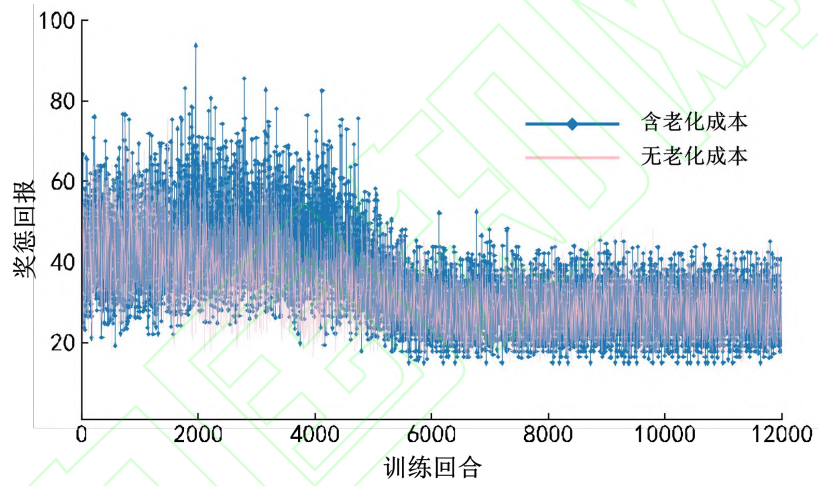
Table B2　Simulation parameter setting

参数	数值	参数	数值	参数	数值
γ	0.8	Δt	1 h	T	24 h
α	0.002	E_i^{U}	20	β	0.001
δ	0.0001	p_i^t	3 kW	η_i	0.95
J	[0,40]	N	[0,200]	ε	(0,1)

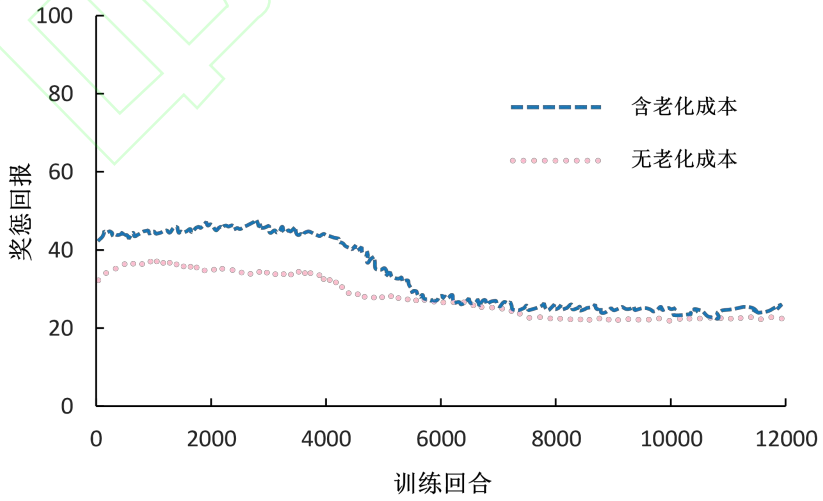
表 B3　分时电价

Table B3　Time-of-use electricity price

	时段	分时电价/[元·(kW·h)$^{-1}$]
峰时段	10:00—14:00，17:00—20:00	0.675
平时段	07:00—10:00，14:00—17:00，20:00—23:00	0.425
谷时段	00:00—07:00，23:00—24:00	0.235



(a) 实际奖惩回报曲线



(b) 平均奖惩回报曲线

图 B1　成本训练结果对比

Fig.B1　Comparison of cost training results