DOI:10.19651/j.cnki.emt.2210388

基于改进深度确定性策略梯度算法的微电网能量优化调度*

李 瑜[1] 张占强[1] 孟克其劳[2] 魏皓天[1]
(1.内蒙古工业大学信息工程学院 呼和浩特 010080;
2.内蒙古工业大学能源与动力工程学院 呼和浩特 010051)

摘 要: 针对微电网中分布式发电设备存在输出不确定性和间歇性问题,以及传统的深度确定性策略梯度算法存在收敛速度慢、鲁棒性差、容易陷入局部最优的缺点。本文提出了一种基于优先经验回放的深度确定性策略梯度算法,以微电网系统运行成本最低为目标,实现微电网的能量优化调度。首先,采用马尔可夫决策过程对微电网优化问题进行建模;其次,采用 Sumtree 结构的优先经验回放池提升样本利用效率,并且应用重要性采样来改善状态分布对收敛结果的影响。最后,本文利用真实的电力数据进行仿真验证,结果表明,提出的优化调度算法可以有效地学习到使微电网系统经济成本最低的运行策略,所提出的算法总运行时间比传统算法缩短了 7.25%,运行成本降低了 31.5%。

关键词: 优先经验回放;微电网能量优化调度;深度确定性策略梯度算法
中图分类号: TM734 **文献标识码:** A **国家标准学科分类代码:** 470.4054

Energy optimal dispatch of microgrid based on improved depth deterministic strategy gradient algorithm

Li Yu[1] Zhang Zhanqiang[1] Meng Keqilao[2] Wei Haotian[1]
(1. The College of Information Engineering, Inner Mongolia University of Technology, Hohhot 010080, China;
2. The College of Energy and Power Engineering, Inner Mongolia University of Technology, Hohhot 010051, China)

Abstract: In view of the output uncertainty and intermittent problems of distributed power generation equipment in microgrid, and the shortcomings of traditional deep deterministic policy gradient algorithm, such as slow convergence speed, poor robustness, and easy to fall into local optimum. In this paper, a deep deterministic policy gradient algorithm based on prioritized experience replay is proposed, aiming at the lowest operating cost of the microgrid system, to realize the energy optimal scheduling of the microgrid. First, the Markov decision process is used to model the microgrid optimization problem; secondly, the prioritized experience replay pool with Sumtree structure is used to improve the efficiency of sample utilization, and importance sampling is applied to improve the influence of state distribution on the convergence results. Finally, this paper uses real power data for simulation verification. The results show that the proposed optimal scheduling algorithm can effectively learn the operation strategy that minimizes the economic cost of the microgrid system. At the same time, the introduction of prioritized experience replay and importance sampling improves the performance of the algorithm.

Keywords: prioritized experience replay; microgrid energy optimization scheduling; deep deterministic policy gradient algorithm

0 引 言

近几年,各个国家都在大力发展新能源技术,能源的转型是解决世界能源危机、环境问题和实现社会经济可持续发展的必经之路[1-2]。中国提出了在 2030 年前实现“碳达峰”、2060 年之前实现“碳中和”的双碳目标[2-3]。随着人工智能被广泛应用于各个行业,智能电网和能源互联网也成为研究热点[4-6]。但由于可再生能源输出存在不确定性和

收稿日期:2022-06-20
* 基金项目:内蒙古自治区科技重大专项计划项目(2020ZD0016,2021ZD003)、内蒙古自治区科技计划项目(2020GG0281)资助

间歇性问题[7-8]，传统深度强化学习算法存在经验采样效率低和难以处理高维状态空间表征的问题，所以如何提高微电网运行的灵活性和稳定性，仍然存在巨大的挑战。

为了克服可再生能源在输出不确定情况下微电网可以稳定运行，近年来许多学者提出基于数据驱动的强化学习方法来解决微电网的优化调度问题[9]。文献[10]中提出了一种基于 Q 网络算法的多光储虚拟同步机频率协调控制策略，应用强化学习计算微电网的功率缺额。文献[11]中提出了一种基于 Q 网络算法的微电网经济调度算法，考虑了可再生能源、电动汽车和储能的相互协调运作，系统的负荷波动和碳排放问题。文献[12]提出了采用增强型学习率自适应算法改进传统 BP 神经网络学习方式，合理调整神经元之间的权重值，解决了微电网蓄电池难以准确估计其荷电状态的问题。文献[13]提出了基于深度 Q 网络(deep Q network，DQN)算法的多智能体强化学习方法，用于在消费者主导的微电网中实现实时价格的用户需求响应。文献[14]提出了一种将 Actor-Critic 算法与双延迟深度确定性策略梯度算法相结合的多智能体深度强化学习方法，解决了点对点能源交易问题和内部能源转换问题。文献[15]提出了一种基于博弈论和强化学习的多智能体结构，降低了微电网的成本和计算复杂度，解决了强化学习中的维数灾问题。文献[16]中采用蒙特卡洛法训练深度神经网络，把蓄电池的动作离散化作为神经网络的输出，采用非线性规划求解剩余决策变量，通过 Q 算法输出最优策略。文献[17]利用双深度 Q 网络(double deep Q network，DDQN)算法实现了配电网电压控制优化，DDQN 算法避免了 DQN 算法中 Q 值被高估了的问题。但这些研究中，一方面，没有考虑强化学习的样本经验值之间时间相关性问题，以及根据决策主体与环境的交互获得单一固定的状态转移过程，忽略了学习环境中不确定因素引起的状态转移随机性。另一方面，在经验回放时利用的是采样效率低，学习效果差的均匀采样和随机采样。此外，应用强化学习输出微电网的决策时，多数的奖励机制为稀疏奖励，而稀疏奖励是深度强化学习在解决任务时面临的核心问题[18]。

1992 年 Lin 等[19]提出“经验回放(experience replay，ER)”的概念，在训练神经网络期间，存储在经验回放池中的经验可以混合最近更新的经验，打破了时间相关性。2016 年 Schaul 等[20]改变传统采样方式，提出了优先经验回放(prioritized experience replay，PER)，更频繁地采样高期望值的经验，并引入 TD-error 来衡量被采样经验值的优先级。文献[21]采用优先经验回放机制来改进 DDPG 算法，通过在 OpenAI Gym 中进行实验，表明该算法可以减少训练时间，提高训练过程的稳定性。文献[22]中提出了一种经验奖励值 reward 与 TD-error 相结合的优先经验回放机制。文献[23]中提出了一种单步多周期预测方法，采用优先经验回放解决了现有多步预测方法存在的误差积累问题，并通过误差分布对预测值进行修正，提高了预测精度，以修正后的预测值为调度依据，实现了微电网系统运行成本。文献[24]采用了深度 Q 网络和优先经验回放的算法，在 DQN 算法中引入 PER 解决了经验值之间的相关性，增加了微电网在能源交易时的收益。但是，上述算法中的经验池存储结构为一维或多维向量结构，经验采样速度慢，不能最大程度上减少训练时间，也并未考虑 PER 的加入会改变状态分布，从而影响收敛结果。

针对上述问题，本文提出了一种基于优先经验回放的深度确定性策略梯度(deep deterministic p-olicy gradient of prioritized experience replay，DD-PG-PER)算法。所提算法中，优先经验回放存储结构采用 Sumtree 数据结构，提升学习过程中样本的获取效率，加速学习过程，减少智能体与环境的交互次数，引入重要性采样(importance sampling，IS)来进行偏差退火处理，改善 PER 产生的偏差对状态分布的改变。

1　微电网模型

1.1　微电网系统模型

本文考虑了一个具有独立供需基础设施的微电网。微电网系统的模型如图 1 所示，该系统由储能系统(energy storage system，ESS)、分布式发电设备(distributed generation equipment，DGE)、主电网(main grid，MG)和负荷构成，其中 DGE 由风力发电机(wind turbines，WT)和光伏发电系统(photovoltaic，PV)组成[25]。微电网通过与 MG 的连接，可以不断地与电力市场进行交易。

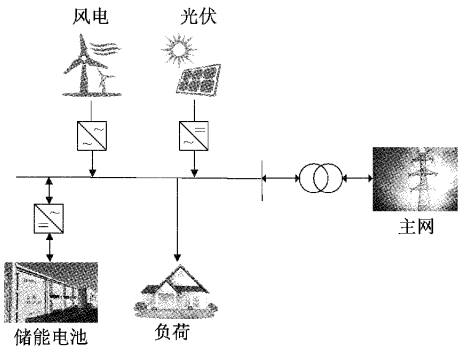


图 1　微电网系统结构

1.2　储能系统模型

考虑到经济性和安全性，文中采用集中式储能装置，所以 ESS 至少可以为该区域提供 1 h 的电量，在微电网系统中 ESS 供电成本是最低的，且微电网在运行时可以将剩余的能量存储，在每一个时间 t 内，将 ESS 进行建模，如式(1)～(3)所示。

$$\begin{cases} SOC_{\min} \leqslant SOC_t \leqslant SOC_{\max} \\ SOC_t = SOC_{t-1} + \dfrac{\delta_t \cdot \mu_c \cdot P_t^E}{CAP_{\max}^{bat}} - \dfrac{(1-\delta_t) \cdot P_t^E}{\rho_{disc} \cdot CAP_{\max}^{bat}} \end{cases} \tag{1}$$

$$0 \leqslant P_t^E \leqslant P_{MAX}^E \tag{2}$$

$$C_{ess}^t = P_t^E \cdot c_l \tag{3}$$

式(1)中，$SOC_{\min}$ 和 $SOC_{\max}$ 分别为储能电池的最低和最高荷电状态，μ_c 和 ρ_{disc} 分别为充电效率和放电效率，$CAP^{bat}_{\max}$ 为储能电池的最大容量。式(2)中，P^E_t 为 t 时刻的充、放电功率，P^E_{MAX} 为最大充、放电功率，采样二进制数 δ_t 表示电池在 t 时刻的状态，当 $\delta_t=1$ 时为充电状态，$\delta_t=0$ 时为放电状态。式(3)中，C^t_{ess} 为 t 时刻储能运行成本，c_l 为储能损耗成本。

1.3　主电网模型

在考虑了 DGE 的输出存在不确定性和间歇性问题，让微电网系统单独工作在孤岛模式下，无法安全稳定地满足微电网的供需平衡。当微电网系统能量不足时，可以向 MG 购电满足系统负荷，同时将地方实时电价波动的情况考虑在内，将灵活多样的供电方式融入系统，进而实现运行成本最低化。具体如式(4)和(5)所示。

$$\begin{cases} P_{G\_\min} \leqslant P^t_G \leqslant P_{G\_\max} \\ (P^t_G)^2+(Q^t_G)^2 \leqslant (S^t_G)^2 \end{cases} \tag{4}$$

$$\begin{cases} C^t_{\text{grid}} = \varsigma_t \cdot P^t_G \cdot c^t_{buy} + (1-\varsigma_t)\cdot P^t_G \cdot c^t_{sell} \\ C^{grid}_{\min} \leqslant c^t_{buy}、c^t_{sell} \leqslant C^{grid}_{\max} \end{cases} \tag{5}$$

式(4)中，P^t_G 和 Q^t_G 分别为 t 时刻微电网与 MG 交易的有功功率和无功功率，S^t_G 为视在功率，$P_{G\_\min}$、$P_{G\_\max}$ 分别为 MG 有功输出下限及上限。式(5)中，C^t_{grid} 为 t 时刻 MG 的运行成本，其中采用二进制 ς_t 表示 MG 当前的状态，当 $\varsigma_t=1$ 时，微电网从 MG 购电，当 $\varsigma_t=0$ 时，微电网向 MG 卖电，c^t_{buy} 和 c^t_{sell} 为 t 时刻购电电价和售电电价，$C^{grid}_{\min}$、$C^{grid}_{\max}$ 分别为 MG 最低和最高电价。

2　深度强化学习算法

2.1　马尔可夫决策过程

微电网的能量优化策略通过强化学习方法输出。在每个学习回合中 Agent 都会根据环境的当前反馈状态执行下一个动作，作为回报 Agent 会收到一个奖励和下一个状态的信息[26]，如图 2 所示。本文通过 MDP 方式对微电网能量优化问题进行建模，用一个四元组 (S,A,R,P) 来表示，其中 S 是状态集，A 是动作集，R 是奖励函数，P 是状态转移概率。

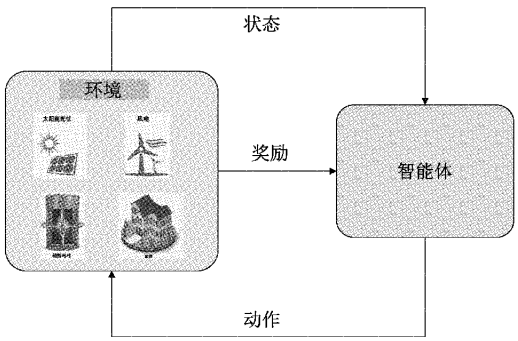


图 2　强化学习交互示意图

1)状态集 S

微电网在 t 时刻的系统状态 $s_t \in S$ 可以被描述为：

$$s_t = \{P^{load}_t, P^{generation}_t, SOC^k_t, C^t_g\} \tag{6}$$

式(6)中，P^{load}_t 表示为 t 时刻负荷功率，$P^{generation}_t$ 表示为 t 时刻 DGE 的发电功率，SOC^k_t 表示为 t 时刻第 k 个 ESS 的荷电状态，$k \in [0,m]$，C^t_g 为 t 时刻 MG 的实时电价。

2)动作集 A

根据 t 时刻微电网系统状态 s_t，可以调度电网和储能电池。本文将微电网系统的动作变量 $a_t \in A$ 描述为如式(7)：

$$a_t = \{P^{grid}_t, P^{E\_k}_t\} \tag{7}$$

式中：P^{grid}_t 表示 t 时刻电网的输出功率、$P^{E\_k}_t$ 表示 t 时刻第 k 个储能的充、放电功率。

3)奖励函数 R

从微电网运行的角度看，以其运行成本最低化为目标，所以将 t 时段内奖励函数设为系统运行成本并取其负值。具体如式(8)所示。

$$r_t(s_t,a_t) = -(C^t_{grid} + C^t_{\text{css}} + C^t_{\text{gen}}) \tag{8}$$

式(8)中，微电网系统的运行成本包括 t 时刻 ESS 的运行成本 C^t_{ess} 和 MG 运行成本 C^t_{grid}，具体如式(3)和(5)，C^t_{gen} 为可再生能源发电成本。

4)状态转移概率 P

状态转移概率 $P:S_t \times A_t \times S_{t+1} \to prob(\cdot)$ 表示为系统从 t 时刻的状态到 $t+1$ 时刻状态的转换概率，其中 $prob(\cdot) \in [0,1]$，可以被描述如式(9)：

$$prob^a_{s,s'} = P\{s_{t+1}=s' \mid s_t=s, a_t=a\} \tag{9}$$

5)优化目标函数

微电网接入设备在一个时间周期内 $t \in T=\{0,1,2,\cdots,23\}$ 的运行成本费用是提高用户经济效益的重要指标，优化目标函数设为与奖励函数一致，优化目标函数如式(10)：

$$\min_T C = \sum_{t=0}^{23}[\gamma \cdot r_t(s_t,a_t)] \tag{10}$$

式中：$\gamma \in [0,1]$ 为折扣因子。

2.2　求解算法

DDPG 算法作为 Actor-Critic 算法的改进和升级，Actor-Critic 算法可以在连续动作空间中根据学习到的策略 π 筛选出随机动作，但是随机性策略具有网络收敛速度慢的问题，需要大量的训练数据，因此将随机性策略梯度算法改为确定性策略梯度算法，解决了算法网络收敛速度慢的问题，同时为了解决无法探索环境的问题，引入了 Off-Policy 采样。DDPG 算法分为策略网络(Actor)和价值网络(Critic)两部分，策略网络中包括 Main-PolicyNet 和 Target-PolicyNet，价值网络中包括 Main-QNet 和 Target-QNet。具体结构如图 3 所示。

定义 $\mu(s \mid \theta_\mu)$ 和 $Q(s,a \mid \theta_Q)$ 分别代表策略网络函数

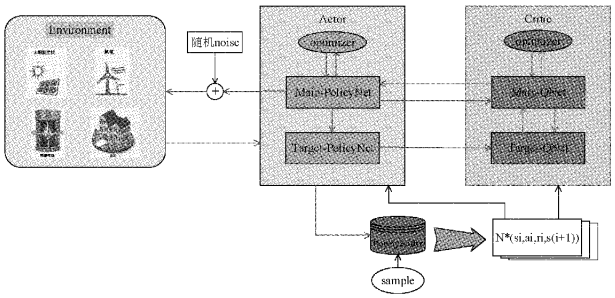


图 3　DDPG 算法结构

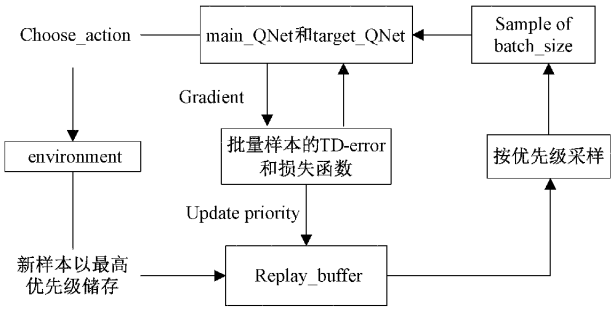


图 4　PER 机制示意图

和价值网络函数，其中用 θ_μ 和 θ'_μ 分别表示 Main-PolicyNet 和 Target-PolicyNet 的神经网络参数，用 θ_Q 和 θ'_Q 分别表示 Main-QNet 和 Target-Qnet 的神经网络参数，θ_μ 通过 gradient 方式更新，如式(11)，θ_Q 通过最小化损失函数进行更新，如式(12)，θ'_μ 和 θ'_Q 通过 soft 更新方式如式(13)所示。

$$\nabla_{\theta_\mu} J = \frac{1}{N}\sum_{j=1}^{N} \nabla_{\mu(s_j|\theta_\mu)} Q[(s_j,\mu(s_j \mid \theta_\mu) \mid \theta_Q] \cdot \nabla_{\theta_\mu}\mu(s_j \mid \theta_\mu) \tag{11}$$

$$\begin{cases} y_j = r_j + \gamma \cdot Q'_\mu[s_{j-1},\mu'(s_{j+1} \mid \theta'_\mu) \mid \theta'_Q] \\ L = \dfrac{1}{N}\sum_{j-1}^{N}[y_j - Q(s_j,a_j \mid \theta_Q)]^2 \end{cases} \tag{12}$$

$$soft\_updata: \begin{cases} \theta'_Q \leftarrow \tau \cdot \theta_Q + (1-\tau)\theta'_Q \\ \theta'_\mu \leftarrow \tau \cdot \theta_\mu + (1-\tau)\theta'_\mu \end{cases}, \tau \in (0,1) \tag{13}$$

式(11)、(12)中，N 表示为抽样数量。式(12)中，y_j 为 Target-QNet 的输出值，$Q(s_j,a_j \mid \theta_Q)$ 为 Main-QNet 的输出值。r_j 为 j 时刻的回报期望值。有关 DDPG 算法原理的推导可以参考文献[27]。

2.3　优先经验回放

不同类型的神经网络在强化学习中的策略优化是近似拟合过程。然而，作为一种监督学习模型，深度神经网络要求数据是独立的、同分布的，为了缓解传统策略梯度法与神经网络相结合时出现的不稳定性，在 DDPG 中使用经验回放机制来消除训练数据之间的相关性。

经验回放包括两部分：存储样本和经验采样。传统经验采样是 Agent 直接使用传入的交互数据进行训练，经验回放是当 Agent 获得一组交互数据，例如 (S,A,R,S')，这组数据将不直接用于神经网络训练，而是存储在经验池中，然后从经验池中提取批量经验进行训练，从缓存器中对经验数据批次采样的过程称为经验回放。但是传统经验采样和经验回放采用的分别是均匀采样和随机采样，都不是高效利用数据的方法，而提出的 PER 用的采样策略，为经验池中的每个样本计算优先级，增大有价值的训练样本在采样时的概率，PER 原理如图 4 所示。

由于在强化学习中通常利用 TD-error 去更新动作值函数 $Q(s,a)$，TD-error 可以隐含的反映出 Agent 从经验中学习的效果，在本文中选取 TD-error 的绝对值 $|\delta|$ 作为评价经验价值的指标，如式(14)：

$$|\delta_j| = y_j - Q(s_j,a_j \mid \theta_Q) \tag{14}$$

式中：y_j 如式(12)。传统 DDPG 算法使用的是随机均匀采样，导致 δ 的值波动剧烈，使得训练结果效果较差，所以在 Agent 训练时经验数据的采样顺序是提升算法性能的重点。于是定义经验抽样概率和经验优先级概率，如式(15)、(16)：

$$P(j) = \frac{p_j^\alpha}{\sum_k p_k^\alpha} \tag{15}$$

$$p_j = |\delta_j| + \varepsilon \tag{16}$$

式(15)中，当 $\alpha = 0$ 时，即为均匀采样。式(16)中，p_j 为第 j 个经验值的优先级概率，$\varepsilon \in (0,1)$ 是为防止未经采样的 p_j 为 0。$|\delta|$ 越大，对预期的动作值校正就更为积极，对应的优先概率越高。

如果每次抽样时都需要根据经验概率进行排序，这样会加大计算过程，所以采用 Sumtree 的数据结构形式存储经验优先概率值，具体数据结构如图 5 所示，叶子节点数为 capacity，所以 Sumtree 总容量为 2×capacity−1，树形结构的最后一层叶节点中存储样本优先级概率，每一个叶子节点的父节点概率值 p_{root_k} 等于所有子节点样本的优先级概率和，每个叶节点对应一个索引值，利用索引值，可以实现对应样本的存取，样本存储结构如图 6 所示。

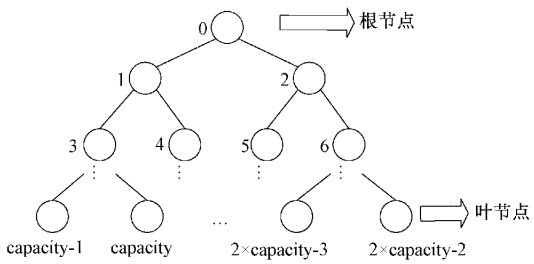


图 5　Sumtree 数据结构

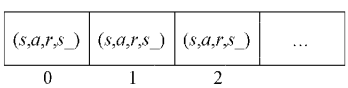


图 6　样本数据存储结构

在进行抽样时，将优先级按顺序从 0 到 $\sum_k p_k^\alpha$ 分成 n 个区间，如式(17)，在划分的每个区间内随机抽取一个值

$p_x=(p_1,p_2,\cdots,p_{batch\_size})$。之后,从根节点处向下搜寻对应的叶子节点,抽取对应叶子节点存储的样本数据 $data_j=[s_j,a_j,r_j,s_j']$,搜索规则:假设随机抽取的数为 p_1,从根节点开始比较,如果 $p_1<p_{root\_0}$,则走右边子节点;若 $p_1>p_{root\_0}$,则走左边子节点,但 $p_1'=p_{root\_0}-p1$,直到找到最后一层叶子节点。

$$n=\frac{\sum_k p_k^{\sigma}}{batch\_size} \tag{17}$$

2.4　重要性采样

在 PER 中使用 TD-error 的绝对值 $|\delta|$ 作为评价经验值是否值得学习的指标,根据 $|\delta|$ 的大小将经验进行优先级赋值,那些高 TD-error 的经验会被经常用来训练,不可避免地改变了状态访问频率,导致神经网络的训练过程容易振荡甚至发散,所以引入了重要性采样 (importance sampli-ng,IS)。这样既保证样本被选到的概率不同,保证了梯度下降时具有相同的收敛结果,同时抑制了神经网络训练过程中振荡情况的产生。定义重要性采样权重(Importance sampling weigh,ISW)为 w_i,如式(18):

$$w_i=(\frac{1}{N}\cdot\frac{1}{P(i)})^{\sigma} \tag{18}$$

式中:N 为经验池容量,$P(i)$ 如式(15),σ 为权重系数,控制修正程度,随着训练逐渐收敛权重系数线性增加到 1,当 $\sigma=1$ 时,表示 PER 对收敛结果的影响完全抵消。为了提高收敛时的稳定性,于是将式(18)经过归一化处理得到式(19):

$$w_j=\frac{(N\cdot P(j))^{-\sigma}}{\max\limits_i(w_i)}=\frac{(N\cdot P(j))^{-\sigma}}{\max\limits_i(N\cdot P(i))^{-\sigma}}=(\frac{P(j)}{\min\limits_i P(i)})^{-\sigma} \tag{19}$$

由于 DDPG 算法中,策略网络的参数依赖价值网络的选取,而价值网络中的参数由价值网络的损失函数来更新,于是得到了新的损失函数,如式(20),于是将式(12)中的损失函数 L 替换为 L_P。

$$L_P=\frac{1}{N}\sum_{j=1}^{N}w_j[y_j-Q(s_j,a_j\mid\theta_Q)]^2 \tag{20}$$

2.5　算法流程

DDPG-PER 算法如下:

1:初始化价值网络 $Q(s,a\mid\theta_Q)$、策略网络 $\mu(s\mid\theta_Q)$ 的神经网络参数 θ_Q 和 θ_μ,replay_buffer 容量
2:初始化目标网络的神经网络参数 θ_Q' 和 θ_μ'
3:初始化抽样概率 $P(j)$ 和 IS 权重的参数 α、σ,target_Qnet 神经网络参数更新频率 updata_step,采样批量 bath_size
4:for $i=1$ to episodes
5:初始化状态 S 作为当前状态集的第一个状态 s_0
6:　for $t=1$ to MAX_STEPS
7:　根据 new policy 选取 new action a_t
8:　得到 r_t 和 s_{t+1},存储经验 (s_t,a_t,r_t,s_{t+1}) 在 replay_buffer,设置 $p_t=\max\limits_{i<t}p_i$
9:　if $i\geqslant$updata_step and t%updata_every==0
10:　for j to updata_every
11:　样本服从经验抽样概率式(15) $P(j)$
12:　根据式(19)计算 ISW:w_j,根据式(16)计算样本优先概率 p_j
13:　end for
14:　更新 Critic 网络,更新 Actor 网络
15:　end if
16:end for
17:end for

3　实验结果分析

3.1　参数设置

实验的验证采用如图 1 所示的微电网结构。神经网络的训练使用的是某地 2015 年全年的实测数据来模拟风电和光伏,风电最大输出为 2.5 MW,光伏最大输出为 0.3 MW,负荷采用文献[28]中的数据模拟,实时电价采用的是某地一般工商业用电电价。模型训练之后,使用 CASIO 2019 年的数据进行测试。定义主电网售电价格为当前购电价格的 80%,可再生能源发电成本设置为主电网平均电价的 10%,售电与购电电价如表 1 所示,该微电网中配备了 2 个容量不同的储能装置,参数如表 2 所示。

表 1　购电与售电价格

	价格(元/kW·h)		
购电	谷时段	平时段	峰时段
	0.32	0.81	1.33
售电	0.256	0.648	1.064

本文设定 DDPG-PER 算法中 Sumtree 数据结构的经验池容量为 2^{19},权重系数 $\sigma=0.4$。算法中学习率为 1×10^{-4},折扣因子 $\gamma=0.99$,引入动作噪声为 0.1,每训练 50 次更新一次 target-QNet 的网络参数。

3.2　算法性能分析

根据图 1 建立了 OpenAIGym 类的环境接口,在每一回合在数据集中随机抽取 24 h 的数据进行训练,设置了 3 组不同的种子数 $seed=0,10,20$,采用多层感知(multi-layers perception,MLP)神经网络结构进行训练,对 DDPG 和 DDPG-PER 算法进行了比较,如图 7 所示,从曲线可以看出 DPPG-PER 的性能曲线大于 DDPG,在 50 回合之前,因为智能体需要对动作空间进行探索,导致曲线抖动较大,并且在 50 回合时,DDPG 算法曲线有较大的波动,说明 DDPG 算法抗鲁棒性能较差。在 50 回合后,智能体通过初期经验积累,并且随着神经网络的第一次更新,曲线上升到

表 2　储能装置参数

	参数							
	$SOC_{\min}$	$SOC_{\max}$	最大充、放电功率 P_{MAX}^{E} /kW	最大容量 $CAP_{\max}^{bat}$ /kW	初始容量/kW	充电效率 μ_c	放电效率 ρ_{disc}	损耗成本 c_l /(kW/元)
储能 1	0.1	1	120	2 000	1 000	0.98	0.98	0.01
储能 2	0.1	1	80	1 500	800	0.98	0.98	0.01

−5 000 附近。在 100 回合后，就表现出了两个算法的优劣性，经过第二次神经网络的更新，DDPG-PER 奖励曲线上升到−2 000 附近，并且后续波动范围也逐渐减小，说明此时智能体已经学习到了一个优秀的策略。而 DDPG 算法在更新后未学习到一个优秀的策略，导致曲线未上升，与第一次更新后的趋势保持一致。

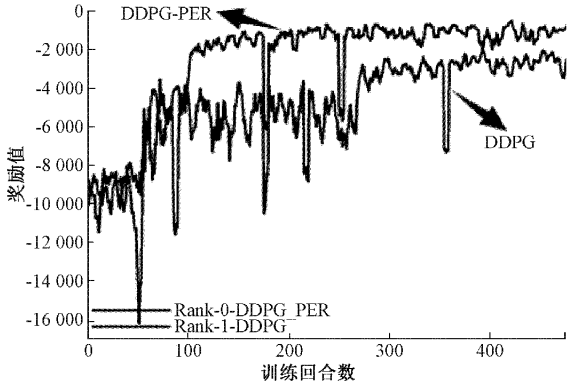


图 7　DDPG 和 DDPG-PER 在不同 seed 下性能曲线

提出的算法主要目的是微电网的运行成本最低，成本函数的标准差是反映回报稳定性的重要指标，如图 8 所示，并且为进一步验证 DDPG-PER 算法的收敛稳定性，通过计算收敛均值、收敛方差来验证，得到的结果如表 3 所示。

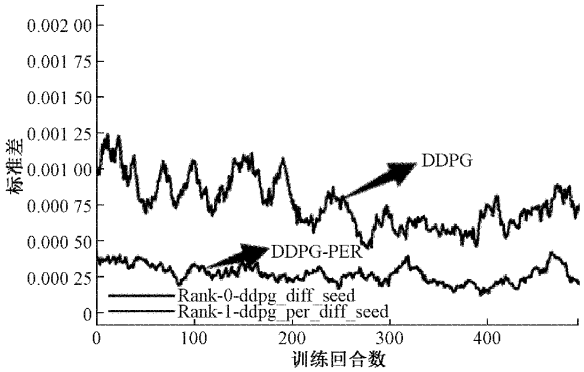


图 8　DDPG 和 DDPG-PER 优化目标函数标准差曲线

表 3　算法收敛性能对比

	运行时间/s	收敛均值	收敛方差
DDPG	2 565.373	−5 473.63	81.03
DDPG-PER	2 379.164	−3 464.393	77.63

在选取 TD-error 作为 PER 的衡量指标，Agent 会抽取高 TD-error 的样本进行学习，根据式(16)将进行区间划分为：高 TD-error 和低 TD-error。batch_size 大小的选取，会导致区间划分的不同，于是，我们设置 batch_size＝32，64，128 对 DDPG 和 DDPG-PER 的算法性能进行比较，结果如图 9 所示，将算法在不同 batch_size 下的性能排序结果在图例中表示，即 Rank-0 表示对比曲线中性能最好的，Rank-5 最差，图例中排序从高到低依次对应图中从上到下的曲线。例如 Rank-0 表示 batch_size＝64 时 DDPG-PER 算法性能曲线。从图中可以看出，不同 batch_size 的选取，两个算法的性能表现均不同，但 DDPG-PER 算法的表现整体优于 DDPG 算法。当 batch_size＝64 时，DDPG-PER 的性能最好，平均 Reward 收敛于−7 000 附近，而 batch_size＝128 时两种方法的性能均较差，并且在加入噪声后曲线波动较大，因为根据式(16)，batch_size 越大，其划分的区间个数越多，而每个区间包含的值越小，这就对于噪声的加入变得更加敏感，可知不同的 batch_size 的选取对算法的性能影响较大。

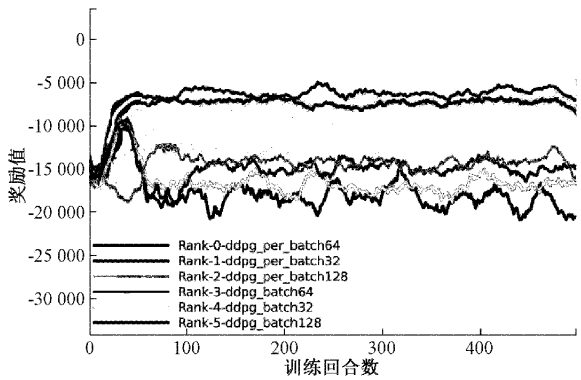
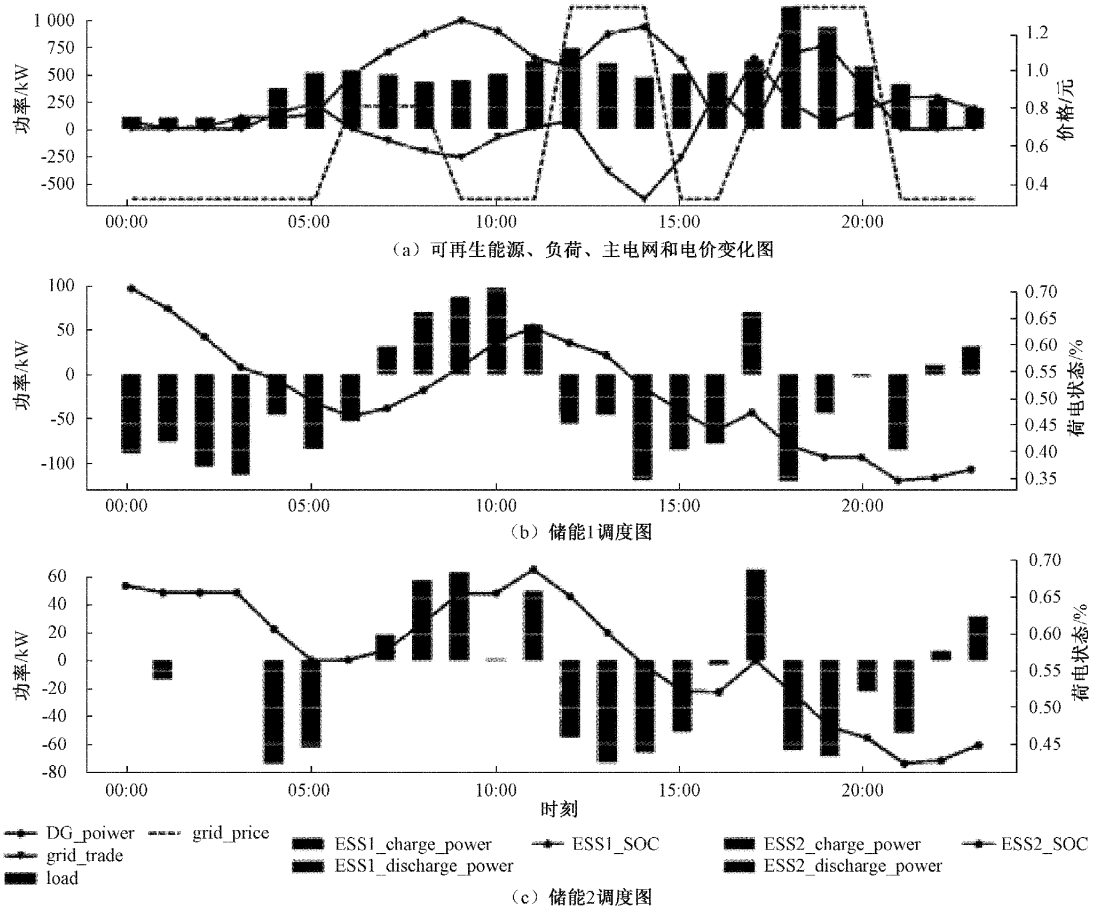


图 9　DDPG 和 DDPG-PER 在不同 batch_size 下性能曲线

3.3　调度决策

根据 3.2 节算法性能的分析，以及进一步验证提出方法的有效性，我们选取某一天的数据进行测试，调度结果如图 10 所示。图 10(a)表示某一天内负荷、可再生能源的功率变化，微电网主电网能量交易功率以及电价变化；图 10(b)表示储能 1 的充、放电功率和荷电状态；图 10(c)表示储能 2 的充、放电功率和荷电状态。

从图 10(a)可以看出，当可再生能源不足以维持负荷需求时，如 00:00～05:00、18:00～21:00，图 10(b)和 10(c)展示了储能 1 和储能 2 均处在放电状态，以满足负荷需求，但为了协调调度储能工作，当储能 1 足以提供负荷需求，储能 2 则不进行动作，如 02:00～03:00 时，减少了微电网



（a）可再生能源、负荷、主电网和电价变化图

（b）储能1调度图

（c）储能2调度图

图 10　微电网出力情况及储能和主电网调度结果

从主电网的购电量;由图 10(a)看出,当可再生能源足以维持负荷的需求,如 07:00～11:00 时,图 10(b)和图 10(c)展示了储能 1 和储能 2 进行充电,并且随着可再生能源输出功率的升高,储能 1 和 2 的充电功率也增加,当储能充电达到最大充电功率时,多余的电量售卖到主电网,从而使经济性最大化;由图 10(a)看出,当可再生能源足以维持负荷,并且此时电价较高,如 13:00～15:00 时,多余的电量售卖到主电网,同时,由图 10(b)和(c)看出储能 1 和储能 2 也进行放电售卖到主电网,这说明提出的调度策略可以根据电价的变化来调整储能装置的输出功率,从相应的增加或减少从主电网的购电量,以降低微电网的运行成本;由图 10(a)看出,当可再生能源足以维持负荷并且电价较低时,储能 1 和储能 2 进行充电。

值得指出的是,提出的调度策略仅根据微电网当前的运行状态得到了实时的调度决策,不需要对可再生能源及负荷进行数据的预测。这些结果说明了提出的调度策略可以在可再生能源输出不确定的情况下及时调度分布式储能设备,实现微电网的稳定运行,可以有效地降低微电网的运行成本,提高微电网的鲁棒性。文献[12]中基于预测对储能装置进行调度的局限性在于,神经网络预测需要大量的数据进行训练,并且对数据的变化较为敏感,如电池的温度,日常充放电损耗等引起的变化都可能导致预测出现失误,而在线实时调度根据的是当前微电网的状态进行调度,所以提出的在线实时调度方法具有更高的可靠性。

4　结　　论

本文提出了一种基于改进深度确定性策略梯度算法的微电网优化调度策略,该方法中考虑了可再生能源输出的不确定性以及电网的阶梯电价。使用 MDP 将微电网的优化调度问题进行建模,为了加速训练神经网络得到最优调度策略,提出了优先经验回放的经验抽取改进传统 DDPG 算法的经验回放方式,该方法的提出可以在训练过程中更加频繁地学习有价值的经验,从而提升学习效率。与传统 DDPG 算法相比,DDPG-PER 算法缩短了总训练时间,同时,在算法中引入了 Sumtree 的存储结构加快数据获取,同时采用重要性采样改善因 PER 的引入对算法收敛结果的影响,结果表明 DDPG-PER 在抗鲁棒性方面具有很好的表现。为了验证所提方法的有效性,本文采用真实的电力数据进行仿真验证,仿真结果表明,该方法可以有效地输出微电网的调度决策,使得微电网运行成本最低。

参考文献

[1] 赖香霖,徐阳. 世界能源发展趋势与中国能源安全研究[J]. 内蒙古煤炭经济,2021(8):63-64.

[2] 邹才能,何东博,贾成也,等. 世界能源转型内涵、路径及其对碳中和的意义[J]. 石油学报,2021,42(2):233-247.

[3] 杨宇,于宏源,鲁刚,等. 世界能源百年变局与国家能源安全[J]. 自然资源学报,2020,35(11):2803-2820.

[4] FAN L P, LI J D, PAN Y, et al. Research and application of smart grid early warning decision platform based on big data analysis[C]. 2019 4th International Conference on Intelligent Green Building and Smart Grid(IGBSG),2019:645-648.

[5] 郭方洪,徐博文,张文安,等. 基于学习优化的智能电网能量管理研究综述[J]. 控制与决策,2022,37(5):1089-1101.

[6] 张瑶,王傲寒,张宏. 中国智能电网发展综述[J]. 电力系统保护与控制,2021,49(5):180-187.

[7] 程逸帆,乔飞,侯珂,等. 区域微电网群两级能量调度策略优化研究[J]. 仪器仪表学报,2019,40(5):68-77.

[8] 於跃,高文根,何飞帆,等. 基于三频段分解的混合储能功率分配策略研究[J]. 电子测量与仪器学报,2021,35(9):27-33.

[9] ARWA E O, FOLLY K A. Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review[J]. IEEE Access,2020,8(12):208992-209007.

[10] 张华强,牟晨东,赵玫,等. 基于强化学习的多光储虚拟同步机频率协调控制策略[J]. 电气传动,2021,51(19):36-42.

[11] 刘金华,柯钟鸣,周文辉. 基于强化学习的微电网能源调度策略及优化[J]. 北京邮电大学学报,2020,43(1):28-34.

[12] 朱晓青,马定寰,李圣清,等. 基于 BP 神经网络的微电网蓄电池荷电状态估计[J]. 电子测量与仪器学报,2017,31(12):2042-2048.

[13] SHOJAEIGHADIKOLAEI A,GHASEMI A,JONES K R, et al. Demand responsive dynamic pricing framework for prosumer dominated microgrids using multiagent reinforcement learning[C]. 2020 52nd North American Power Symposium(NAPS),2021:1-6.

[14] CHEN T Y,BU S R,LIU X,et al. Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning[J]. in IEEE Transactions on Smart Grid,2022,13(1):715-727.

[15] HAO J, GAO D W, ZHANG J J. Reinforcement learning for building energy optimization through controlling of central HVAC system[J]. in IEEE Open Access Journal of Power and Energy, 2020, 7(9): 320-328.

[16] 余宏晖,林声宏,朱建全,等. 基于深度强化学习的微电网在线优化[J/OL]. 电测与仪表:1-7[2023-03-12]. http://kns.cnki.net/kcms/detail/23.1202.TH.20211021.1651.007.html.

[17] YANG Q L, WANG G, SADEGHI A, et al. Two timescale voltage control in distribution grids using deep reinforcement learning[J]. IEEE Transactions on Smart Grid,2020,11(3):2313-2323.

[18] 杨惟轶,白辰甲,蔡超,等. 深度强化学习中稀疏奖励问题研究综述[J]. 计算机科学,2020,47(3):182-191.

[19] LIN L J. Self-improving reactive agents based on reinforcement learning planning and teaching[J]. Machine learning,1992,8(3):293-321.

[20] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. ArXiv Preprint,2015, ArXiv:1511.05952.

[21] HOU Y,LIU L,WEI Q,et al. A novel DDPG method with prioritized experience replay[C]. 2017 IEEE International Conference on Systems, Man, and Cybernetics(SMC),2017:316-321.

[22] GAO J, LI X, LIU W, et al. Prioritized experience replay method based on experience reward[C]. 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), 2021: 214-219.

[23] JOGUNOLA O,TSADO Y. Trading strategy in a local energy market, a deep reinforcement learning approach[C]. 2021 IEEE Electrical Power and Energy Conference(EPEC),2021:347-352.

[24] LI Y, WANG R, YANG Z. Optimal scheduling of isolated microgrids using automated reinforcement learning based multi period forecasting[J]. IEEE Transactions on Sustainable Energy, 2022, 13(1): 159-169.

[25] 刘文洲,胡治辉. 微电网智能负荷计量控制装置研究与设计[J]. 国外电子测量技术,2020,39(10):72-77.

[26] 许杨子,强文,刘俊,等. 基于改进深度强化学习算法的电力市场监测模型研究[J]. 国外电子测量技术,2020,39(1):82-87.

[27] 万里鹏,兰旭光,张翰博,等. 深度强化学习理论及其应用综述[J]. 模式识别与人工智能,2019,32(1):67-81.

[28] MANZOUL M A, ELHASSAN F, HASSAN M. Comparison of deep reinforcement learning algorithms in enhancing energy trading in microgrids[C]. 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering(ICCCEEE),2021:1-6.

作者简介

李瑜,硕士研究生,主要研究方向为微电网优化与控制等。

E-mail:ly1120850221@163.com