基于深度强化学习的微电网优化运行策略

赵鹏杰，吴俊勇，王 燚，张和生
（北京交通大学 电气工程学院，北京 100044）

摘要：风电、光伏、负荷的不确定性给含有高比例可再生能源的微电网制定运行策略带来了挑战，人工智能技术的发展为解决微电网运行优化问题提供了新思路。基于强化学习框架，将微电网运行问题转化为马尔可夫决策过程，以最大化微电网经济利益和居民满意度为目标，提出一种基于深度强化学习的微电网在线调度方法。为了在深度强化学习训练的过程中高效利用经验，设计一种优先经验存储的深度确定性策略梯度（PES-DDPG）算法，学习各类环境下不同时段的微电网最优调度策略。算例结果表明，PES-DDPG 算法能够为微电网提供有效的调度策略，并实现微电网的实时优化。

关键词：深度强化学习；微电网；马尔可夫模型；优化运行
中图分类号：TM 73　　**文献标志码**：A　　DOI：10.16081/j.epae.202205032

0 引言

微电网是由分布式发电、负荷、储能装置等组成的小规模电网，可以有效提高电网中分布式电源渗透率，是实现“双碳”目标的有效途径[1]。然而，分布式发电的间歇性和不稳定性使得微电网能量管理变得更加困难[2]。

微电网能量管理问题的求解算法包括经典优化算法和启发式算法。经典优化算法主要包括线性规划、混合整数规划等，面对复杂的高维非线性、不连续的目标函数和约束时，该算法存在求解困难的问题[3]。启发式算法对数学模型的依赖性较小，更容易处理非线性问题，但该算法的寻优结果具有随机性，通常为次优[4]。上述2类算法能够解决确定性微电网优化问题，但实际微电网中的可再生能源出力、负荷等因素均是不确定因素，不确定性问题主要采用随机规划和鲁棒优化方法[5-7]进行求解。随机优化方法的难点是如何保证概率分布准确刻画实际不确定性因素的变化规律[8]。鲁棒优化方法采用不确定性集描述不确定性因素的变化范围，但所得优化结果较为保守[9]。

近年来，人工智能技术获得了极大的发展，其中深度强化学习 DRL（Deep Reinforcement Learning）算法因具有解决序贯决策问题的能力而受到电力系统研究人员的关注[10]，目前主要包括基于值函数和基于策略梯度2类算法。

在基于值函数的 DRL 算法研究中：文献[11]建立包含蓄电池和储氢装置的微电网复合储能模型，采用基于值函数的深度 Q 网络 DQN（Deep Q Network）算法解决储能装置的协调控制问题；文献[12]针对社区微电网储能系统，提出 Q-learning 的能源管理策略，提高了储能的效率和可靠性；文献[13]提出采用 DQN 和深度双 Q 网络 DDQN（Double Deep Q Network）算法解决家庭家电优化调度问题，并证明 DDQN 算法比 DQN 算法更适合求解最小化成本问题；文献[14]提出一种基于 DQN 算法的微能源网能量管理模型，通过与遗传算法进行对比，验证了 DQN 算法在解决能量管理问题时的有效性。基于值函数的 DRL 算法将微电网连续变量离散化，导致寻优结果不精确。

在基于策略梯度的 DRL 算法研究中：文献[15]建立含有空调系统和储能系统的智能家居能源成本最小化模型，证明了深度确定性策略梯度 DDPG（Deep Deterministic Policy Gradient）算法可有效处理模型的不确定性；文献[16]考虑居民行为、实时电价和室外温度的不确定性，建立实时需求响应模型，提出一种基于信赖域策略优化的需求响应算法，实现了不同类型设备的优化调度；文献[17]采用双延迟深度确定性梯度算法解决了光储电站储能运行问题。

本文针对含有风机、光伏、燃气轮机、储能设备和负荷的典型微电网架构，首先详细描述将微电网运行优化问题转化为马尔可夫决策过程 MDP（Markov Decision Process）的方法及步骤，然后采用 DDPG 算法求解微电网连续变量优化问题，为提高 DDPG 算法的收敛性，设计一种优先经验存储的深度确定性策略梯度 PES-DDPG（Priority Experience Storage Deep Deterministic Policy Gradient）算法，最后基于某微电网 2018 年的风机、光伏及负荷数据[18]进行算例分析，验证了 PES-DDPG 算法能够提高 DDPG 算

收稿日期：2021-07-27；**修回日期**：2022-03-30
在线出版日期：2022-04-12
基金项目：中央高校基本科研业务费专项资金资助项目（2020YJS162）
Project supported by the Fundamental Research Funds for the Central Universities（2020YJS162）

法的收敛稳定性以及该算法处理微电网能量优化问题时的有效性和优越性。

1 微电网系统及优化模型

1.1 微电网结构

本文微电网模型如图1所示，包括风机、光伏、燃气轮机、储能设备、负荷以及与外部电网接口。微电网运营商的职责是调节各节点的功率流动，实现微电网经济运行。

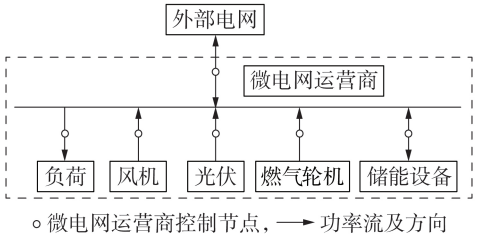

图1 微电网模型

Fig.1 Model of microgrid

1.2 微电网设备模型

1)燃气轮机。

燃气轮机通过燃烧天然气为微电网提供可调节的电力供应，有效降低微电网对外部电网的依赖，其燃料成本以二次函数表示，如式(1)所示。

$$C_t^{\mathrm{MT}}=a^{\mathrm{MT}}(P_t^{\mathrm{MT}})^2+b^{\mathrm{MT}}P_t^{\mathrm{MT}}+c^{\mathrm{MT}} \tag{1}$$

式中：C_t^{MT}为t时段燃气轮机的燃料成本；P_t^{MT}为t时段燃气轮机的实际输出电功率；a^{MT}、b^{MT}、c^{MT}为成本系数。

燃气轮机的输出功率满足如下约束条件：

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

$$-R_{\mathrm{down}}^{\mathrm{MT}}\leqslant P_t^{\mathrm{MT}}-P_{t-1}^{\mathrm{MT}}\leqslant R_{\mathrm{up}}^{\mathrm{MT}} \tag{3}$$

式中：$P_{\max}^{\mathrm{MT}}$、$P_{\min}^{\mathrm{MT}}$分别为燃气轮机的最大、最小输出功率；$R_{\mathrm{up}}^{\mathrm{MT}}$、$R_{\mathrm{down}}^{\mathrm{MT}}$分别为上、下爬坡约束。

2)储能设备。

储能设备由蓄电池组成，其可与具有随机性和波动性的可再生能源协调运行，发挥“削峰填谷”的作用，保证微电网的可靠性和经济性。考虑蓄电池的充放电功率和储能荷电状态SOC(State Of Charge)，储能设备充放电表达式为：

$$S_{t+1}^{\mathrm{soc}}=S_t^{\mathrm{soc}}-\frac{s_t^{\mathrm{ch}}\eta_{\mathrm{ch}}P_t^{\mathrm{ch}}\Delta t}{S_{\mathrm{es}}}-\frac{s_t^{\mathrm{dis}}P_t^{\mathrm{dis}}\Delta t}{\eta_{\mathrm{dis}}S_{\mathrm{es}}} \tag{4}$$

式中：S_t^{soc}为t时段储能设备的SOC；s_t^{ch}、s_t^{dis}分别为t时段储能设备充电和放电状态值，其值为0表示关闭，为1表示开启；P_t^{ch}、P_t^{dis}分别为t时段储能设备的充电和放电功率，设充电功率为负，放电功率为正；η_{ch}、η_{dis}分别为储能设备的充电和放电效率；S_{es}为储能设备的额定容量；Δt为时间间隔。

储能设备充放电满足如下约束条件：

$$-P_{\max}^{\mathrm{ch}}\leqslant P_t^{\mathrm{ch}}\leqslant 0 \tag{5}$$

$$0\leqslant P_t^{\mathrm{dis}}\leqslant P_{\max}^{\mathrm{dis}} \tag{6}$$

$$S_{\min}^{\mathrm{soc}}\leqslant S_t^{\mathrm{soc}}\leqslant S_{\max}^{\mathrm{soc}} \tag{7}$$

$$s_t^{\mathrm{ch}}+s_t^{\mathrm{dis}}\in\{0,1\} \tag{8}$$

式中：$P_{\max}^{\mathrm{ch}}$、$P_{\max}^{\mathrm{dis}}$分别为储能设备的最大充电和放电功率，均为正值；$S_{\min}^{\mathrm{soc}}$、$S_{\max}^{\mathrm{soc}}$分别为储能设备SOC的最小值和最大值。式(8)表示同一时段内储能设备不能同时进行充电和放电操作。

3)需求响应负荷。

微电网运营商可通过电价或其他需求响应手段调节居民负荷的消费特性，为微电网调度运行提供辅助服务，但需求响应不能一味改变用户的用电行为，造成用户体验下降。微电网运营商在执行调度指令时应考虑用户满意度因素。用户满意度有如下特性[19]：用户倾向于消耗更多的能量直至达到目标用电计划；当用户消耗的能量接近目标计划时，用户满意度将逐渐饱和。用户满意度U_t^{L}表达式为：

$$U_t^{\mathrm{L}}=\begin{cases}v_t\dfrac{P_t^{\mathrm{L}}}{L_t}-\dfrac{z_t(P_t^{\mathrm{L}})^2}{2L_t^2} & \dfrac{P_t^{\mathrm{L}}}{L_t}\leqslant\dfrac{v_t}{z_t}\\[2ex] \dfrac{z_t}{2v_t} & \dfrac{P_t^{\mathrm{L}}}{L_t}>\dfrac{v_t}{z_t}\end{cases} \tag{9}$$

$$0\leqslant P_t^{\mathrm{L}}\leqslant L_t \tag{10}$$

式中：v_t、z_t为t时段用户满意度系数；P_t^{L}、L_t分别为t时段用户的实际用电功率和计划用电功率。

4)微电网母线。

微电网中含有大量以风机、光伏为代表的分布式电源，为实现可再生能源的完全消纳，默认风机、光伏功率全部并网。微电网中母线要保持功率平衡，可建模为：

$$P_t^{\mathrm{grid}}+P_t^{\mathrm{MT}}+P_t^{\mathrm{PV}}+P_t^{\mathrm{WT}}+P_t^{\mathrm{ES}}=P_t^{\mathrm{L}} \tag{11}$$

$$P_t^{\mathrm{ES}}=P_t^{\mathrm{dis}}+P_t^{\mathrm{ch}} \tag{12}$$

$$P_t^{\mathrm{grid}}=P_t^{\mathrm{b}}-P_t^{\mathrm{s}} \tag{13}$$

式中：P_t^{grid}为t时段微电网同外部电网的交互功率；P_t^{PV}、P_t^{WT}分别为t时段光伏、风机并网功率；P_t^{ES}为t时段储能功率；P_t^{b}、P_t^{s}分别为t时段微电网向外部电网购电的功率和售电的功率。

1.3 微电网优化模型

微电网运营商通过求解以下优化问题来确定运行方案：

$$\max\left[\sum_{t=1}^{T}(\sigma_t^{\mathrm{s}}P_t^{\mathrm{s}}-\sigma_t^{\mathrm{b}}P_t^{\mathrm{b}}+\sigma_tP_t^{\mathrm{L}}-C_t^{\mathrm{MT}})+\alpha\sum_{t=1}^{T}U_t^{\mathrm{L}}\right] \tag{14}$$

式中：σ_t^{b}、σ_t^{s}分别为t时段微电网向外部电网购电的价格和售电的价格；σ_t为t时段微电网向负荷售电的价格；α为用户满意度的等效价格因子；T为总优化时段数。

以式(14)为目标函数，式(1)—(13)为约束条件，形成混合整数二次规划MIQP(Mixed Integer Quadratic Programming)问题。

2　微电网 MDP

强化学习的本质是使智能体和环境交互，智能体基于观察到的环境状态选择动作，环境对该动作作出响应，智能体获得环境的反馈后调整下一步动作，最终实现智能体对环境的最优响应。

在强化学习框架中，如果智能体的环境及其状态满足马尔可夫属性，则该强化学习任务称为MDP。MDP通常由5元组$(\boldsymbol{S},\boldsymbol{A},\boldsymbol{P}_r,\boldsymbol{r},\mu)$表示，其中：$\boldsymbol{S}$为环境状态空间，设$s_t$为$t$时段的环境状态；$\boldsymbol{A}$为动作空间，设$a_t$为$t$时段智能体的动作；$\boldsymbol{P}_r$为概率，设$P_r\left(s_{t+1}\middle|s_t,a_t\right)$为在状态$s_t$采取动作$a_t$后状态转移到$s_{t+1}$的概率；$\boldsymbol{r}$为即时奖励，设$r_t$为在状态$s_t$采取动作$a_t$后获得的即时奖励；$\mu$为策略，是状态到动作概率分布的映射，状态、动作和奖励序列构成策略的轨迹τ。

假设未来奖励在每个时段的折扣因子为γ，在T时段终止，累积奖励定义如式(15)所示。强化学习的目标是找到最优策略μ^*，使得所有状态的预期回报最大，如式(16)所示。

$$R_t=\sum_{i=0}^{T-1}-\gamma^i r_{t+1} \tag{15}$$

$$\mu^*=\mathrm{argmax}\ E_\mu\left[R_t\middle|\mu\right] \tag{16}$$

式中：R_t为t时段的累积奖励；argmax表示求解最大化问题时对应的参数；$E_\mu[\cdot]$表示变量的期望值。

定义策略μ的动作值函数为：

$$Q_\mu(s,a)=E_\mu\left[R_t\middle|s_t=s,a_t=a\right] \tag{17}$$

式中：$Q_\mu(s,a)$为从状态s开始采用动作a后遵循策略μ的预期回报。

利用强化学习寻找最优策略μ^*的问题可等价转化为寻找最大的动作价值函数问题，最大的动作价值函数$Q^*(s,a)$对应的策略就是最优策略μ^*，如式(18)所示。

$$Q^*(s,a)=\max_\mu Q_\mu(s,a) \tag{18}$$

微电网MIQP问题转化为MDP进行求解的关键是：环境状态$\boldsymbol{S}$、动作$\boldsymbol{A}$、奖励$\boldsymbol{r}$的表达式；策略μ的获取。

1)状态空间。

微电网运行优化过程中所需的环境信息共同组成智能体状态空间。环境信息可分为时变信息和时不变信息。为简化状态空间，将时不变信息设定为智能体自身的已知信息，将时变信息作为本文模型的状态信息，时变信息包括用户预测负荷、风机预测出力、光伏预测出力、储能状态、分时电价，因此，状态空间可以描述为：

$$\boldsymbol{S}=[t,L_t,P_t^{\mathrm{PV}},P_t^{\mathrm{WT}},S_t^{\mathrm{soc}},\sigma_t^{\mathrm{b}},\sigma_t^{\mathrm{s}}] \tag{19}$$

2)动作空间。

在微电网中，控制动作包括实际负荷功率、燃气轮机输出功率和储能充放电功率，智能体动作空间可定义为：

$$\boldsymbol{A}=[P_t^{\mathrm{L}},P_t^{\mathrm{MT}},P_t^{\mathrm{ES}}] \tag{20}$$

3)奖励函数。

智能体选择任一动作后，环境会给予奖励，一般为正，而惩罚成本为较大的负数，智能体为了获得最大奖励，会逐渐约束动作满足动作变化率。惩罚成本表达式为：

$$\begin{aligned}C_t^{\mathrm{c}}=&-\zeta\max\left\{P_t^{\mathrm{MT}}-P_{t-1}^{\mathrm{MT}}-R_{\mathrm{up}}^{\mathrm{MT}},0\right\}+\\&\zeta\min\left\{P_t^{\mathrm{MT}}-P_{t-1}^{\mathrm{MT}}+R_{\mathrm{down}}^{\mathrm{MT}},0\right\}\end{aligned} \tag{21}$$

式中：ζ为较大的正数。

奖励函数由微电网运行收益、用户满意度和惩罚成本组成：

$$r_t=\sigma_t^{\mathrm{s}}P_t^{\mathrm{s}}-\sigma_t^{\mathrm{b}}P_t^{\mathrm{b}}+\sigma_t P_t^{\mathrm{L}}-C_t^{\mathrm{MT}}+\alpha U_t^{\mathrm{L}}+C_t^{\mathrm{c}} \tag{22}$$

3　PES-DDPG 算法

本节介绍策略μ的求解算法，即PES-DDPG算法。

3.1　DDPG 算法

DDPG算法的基本思想是，给定状态和参数，则只输出1个确定的动作。显然，对于微电网优化问题，针对确定的运行状态，只有唯一的最优调度策略，因此，本文选择DDPG算法作为求解微电网连续变量优化问题的基础算法。

DDPG算法的网络架构如附录A图A1所示。定义确定性动作策略为μ，每步动作a_t可通过$a_t=\mu(s_t)$计算得到。采用神经网络对μ函数以及Q函数进行模拟，分别称为Actor策略网络(θ^μ)和Critic价值网络(θ^Q)。定义函数$J_\pi(\mu)$衡量策略的好坏，$J_\pi(\mu)$的表达式为：

$$J_\pi(\mu)=E_{s\sim p^\pi}\left[Q_\mu(s,\mu(s))\right] \tag{23}$$

式中：p^π为概率分布函数；$Q_\mu(s,\mu(s))$为智能体按照策略μ选择动作产生的Q值。$J_\pi(\mu)$等价为状态s服从p^π分布时，按照策略μ得到的$Q_\mu(s,\mu(s))$的期望值。

智能体通过与微电网环境的交互积累丰富经验后进入网络更新阶段。在更新阶段，首先得到式(24)所示目标奖励y_i、Critic价值网络实际Q值$Q_{\theta^Q}(s_{t,i},a_{t,i})$，根据式(25)所示误差方程得到Critic价值网络误差L，并最小化误差实现Critic价值网络的更新，然后通过求取式(26)所示策略梯度$\nabla_{\theta^\mu}J(\mu_{\theta^\mu})$确定Actor策略网络的更新方向，实现Actor策略网络的更新。其中，下标“i”表示样本经验编号。

$$y_i=r_{t,i}+\gamma Q_{\theta^{Q'}}(s_{t+1,i},\mu_{\theta^{\mu'}}(s_{t+1,i})) \tag{24}$$

$$L=\frac{1}{n}\sum_i\left(y_i-Q_{\theta^Q}(s_{t,i},a_{t,i})\right)^2 \tag{25}$$

$$\nabla_{\theta^\mu}J(\mu_{\theta^\mu})\approx\frac{1}{n}\sum_i\nabla_a Q_{\theta^Q}(s_{t,i},\mu_{\theta^\mu}(s_{t,i}))\nabla_{\theta^\mu}\mu_\theta(s_{t,i}) \quad (26)$$

式中：$Q_{\theta^{Q'}}(s_{t+1,i},\mu_{\theta^{\mu'}}(s_{t+1,i}))$为目标Critic价值网络得到的$Q$值，$\mu_{\theta^{\mu'}}$为目标Actor策略网络得到的策略；$n$为样本经验数；$\nabla_a Q_{\theta^Q}(s_{t,i},\mu_{\theta^\mu}(s_{t,i}))$、$\nabla_{\theta^\mu}\mu_\theta(s_{t,i})$分别为Critic价值网络和Actor策略网络的梯度。

DDPG算法的特点主要有：采取随机策略进行动作的探索，采取确定性梯度策略进行策略的更新；采用Actor-Critic结构，将其分为Actor策略网络和Critic价值网络，并为其创建备份网络，称为目标网络，解决更新不稳定的问题；利用DQN算法的经验回放对网络进行训练，最小化样本的相关性。本文分别对DDPG算法的动作探索机制和经验回放机制进行改进，使该算法更适用于优化问题的求解。

3.2 动态噪声搜索策略

搜索策略是为了搜索到完整的动作状态空间，本文在DDPG算法的训练过程中引入高斯噪声，将动作的选择从确定性过程变为随机性过程，并在随机过程采样得到动作。

动态噪声策略是指，在每次训练过程中，智能体通过策略网络$a_t=\mu(s_t)$生成动作并叠加随机噪声时，噪声幅值随着训练的进行而逐渐变小，使智能体动作完全符合策略μ。噪声幅值ε_k的表达式如式(27)所示。

$$\varepsilon_k=\max\left\{\varepsilon_{\max}-\lambda k,\varepsilon_{\min}\right\} \quad (27)$$

式中：$\varepsilon_{\max}$、$\varepsilon_{\min}$分别为随机选择动作概率的最大值和最小值；λ为衰减系数；k为训练时的迭代轮数。

3.3 经验优先存储策略

在DDPG算法中，将智能体探索到的数据组合(s_t,a_t,r_t,s_{t+1})作为经验并将其存在经验复用池中，随机抽取一批经验对Actor策略网络和Critic价值网络进行训练，由于经验池大小固定，随着训练的进行，新获取的经验不断覆盖旧经验，只有新获取的经验逐渐一致时，才能得到算法收敛结果。

实际应用过程中发现，在探索初期，智能体得到高奖励经验的概率远低于得到低奖励经验的概率，该时期即使获得了奖励极高的经验，随着智能体趋向探索行为，低奖励经验大量进入经验复用池，也会导致高奖励经验的丢失，训练过程中波动性较大，收敛结果不理想。

考虑微电网能量优化问题的特点，一天24个不同时段之间的调度结果和收益存在较大差异，而同一时段的调度结果和收益具有相似性。在同一时段内，如果智能体搜索到的经验比之前的经验奖励高，则将其视为优秀经验，并存储到经验复用池中，否则将其视为普通经验，并按照一定概率存储到经验复用池中。

定义经验池中t时段的经验平均奖励r_t^{mean}为：

$$r_t^{\text{mean}}=\sum_{k=1}^{m}r_{t,k} \quad (28)$$

式中：m为训练时的最大迭代轮数；$r_{t,k}$为t时段第k轮训练时的奖励值。

经验优先存储策略流程图如图2所示。

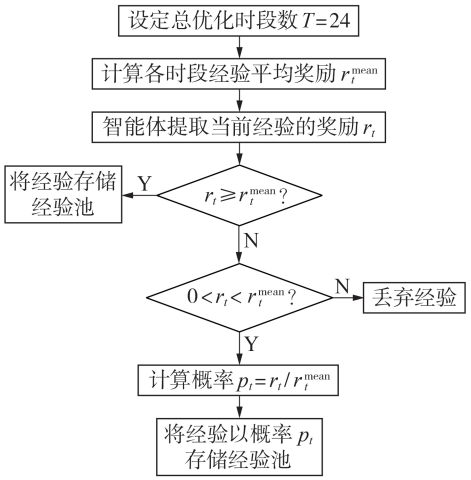

图2 经验优先存储策略流程图

Fig.2 Flowchart of priority experience storage strategy

结合动态噪声搜索策略和经验优先存储策略，本文提出的PES-DDPG算法伪代码如附录A表A1所示。

4 算例分析

4.1 微电网场景设置

本文建立的微电网各设备主要参数如附录A表A2和表A3所示。微电网中负荷最大值为400 kW，风机安装容量为250 kW，光伏安装容量为150 kW。用户满意度系数v_t、z_t分别为2.2、2.5。

4.2 算法及算例设置

本文分别采用PES-DDPG算法、传统DDPG算法、DDQN算法和MIQP算法求解微电网优化运行方案并对结果进行对比。基于微电网马尔可夫模型，PES-DDPG算法和传统DDPG算法的状态空间大小为7，动作空间大小为3。设计的Actor策略网络结构如附录A图A2所示。输入层为7个神经元，对应输入7个环境状态；隐藏层为3层全连接层，各层分别有128个神经元，采用线性修正函数ReLU(Rectified Linear Unit)作为激活函数；输出层为3个神经元，对应模型中的3类动作，采用tanh作为激活函数。对DDQN算法的动作空间进行离散化，将燃气轮机、储能设备、负荷动作离散为9×4×2种动作。强化学习算法的超参数设置如下：Actor策略网络和Critic价值网络学习率均为0.001，折扣因子γ=0.95，最大、最小噪声幅值分别为1和0.004，每批抽样数为128。

本文设置如下2个算例：算例1，微电网环境变量均为确定值，即风机、光伏功率及负荷需求均可准确预测，在此基础上，采用4种算法进行微电网日前运行优化；算例2，选取某地2018年1月1日至3月30日的实际数据，将其进行标幺化处理后作为训练集，采用本文提出的PES-DDPG算法训练智能体，假设微电网环境中不确定性参数分别存在5%、10%、30%的预测误差，面对不确定的环境变量，智能体进行在线优化。

4.3　算例1结果及分析

4.3.1　不同算法优化结果分析

微电网某日数据如附录A图A3所示。针对确定性环境，采用4种算法进行微电网日前运行优化，各方案所得微电网运行收益如表1所示。

表1　微电网运行收益

Table 1　Operation profit of microgrid

算法	收益/元	算法	收益/元
PES-DDPG	709.6	DDQN	665.6
传统DDPG	725.8	MIQP	760.8

由表1可知：由于算例中调用Yalmip工具箱求解器进行求解，因此，MIQP算法获得了理论上的最高收益，为最优的结果，但由于实际中未来24 h内的风机、光伏功率及负荷需求不可能准确预测，因此，MIQP算法的求解结果难以直接应用；PES-DDPG算法和传统DDPG算法的结果较接近最优的结果；由于动作离散，DDQN算法的效果不及传统DDPG算法。

图3展示了PES-DDPG算法、传统DDPG算法和MIQP算法的优化结果。PES-DDPG算法和传统DDPG算法的寻优结果基本一致，体现在：在低电价时，燃气轮机运行成本比电价高，因此燃气轮机仅保持最低出力；储能设备在夜间时段1—5进行充电，在中午高峰期进行放电，通过低买高卖进行套利，在下午时段15—18中等电价时继续进行充电，并在夜间高电价时进行放电套利；在中午时段12—14，光伏出力较多，燃气轮机也保持高出力，微电网向电网反送功率获利。传统DDPG算法与MIQP算法理论最优结果的显著区别是：在传统DDPG算法的结果中，微电网在时段12—14向电网反送功率，而在MIQP算法的结果中，微电网仅在时段12向电网反送功率；在传统DDPG算法的结果中，储能设备在高电价时（夜晚时段19、20）放电，而在MIQP算法的结果中，储能设备不仅在时段19、20放电，还在时段21—24放电，满足夜间微电网内的功率缺额需求，减少向电网购电。

4.3.2　优先经验存储策略分析

针对算例1场景，PES-DDPG算法和传统DDPG算法的收敛结果对比如图4所示。由图可知，在传统DDPG算法的训练后期，随机探索和经验池中的低奖励经验会导致算法收敛的不确定性，如果算法恰好在抽取到低奖励经验训练时停止，则可能无法得到满意的结果。

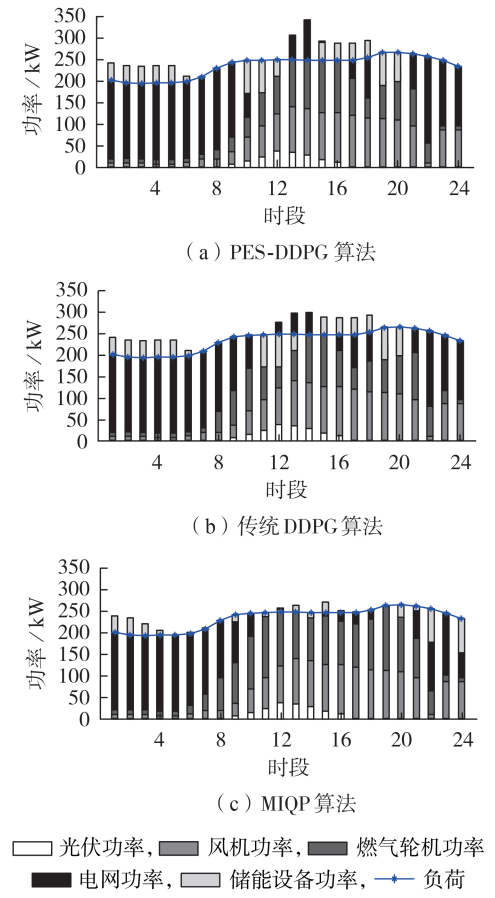

图3　3种算法优化结果

Fig.3　Optimization results of three algorithms

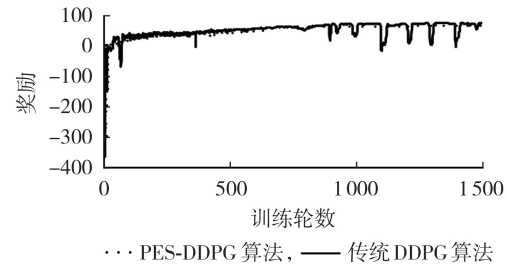

图4　PES-DDPG算法和DDPG算法的收敛结果对比

Fig.4　Comparison of convergence results between PES-DDPG algorithm and traditional DDPG algorithm

在不同的初始条件下，采用传统DDPG算法进行训练，其收敛结果如图5所示。由图可知，在3次训练中，算法均在训练1 000轮左右取得了基本一致的奖励，但是随着训练次数的增加，经验池中随机抽取的普通经验改变了算法的收敛方向，当训练到1 500轮时，训练结果变得更加糟糕，经验质量和训练终止的次数均会对算法的收敛结果产生随机性的影响。

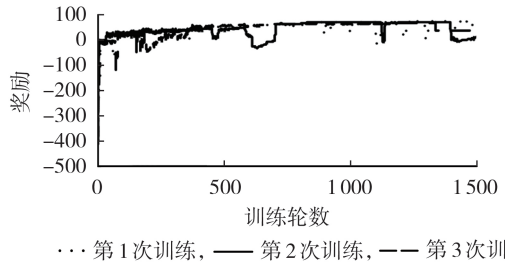

图5 不同初始条件下传统DDPG算法的收敛结果对比

Fig.5 Comparison of convergence results of traditional DDPG algorithm among different initial conditions

为进一步评估PES-DDPG算法的收敛性，定义如下指标：

$$C_{MV}=\frac{\sum_{i=1}^{N} r^i}{N} \tag{29}$$

$$C_V=\frac{\sum_{i=1}^{N}(r^i-C_{MV})^2}{N} \tag{30}$$

$$C_{MAX}=\max(r^i) \tag{31}$$

$$C_{MIN}=\min(r^i) \tag{32}$$

式中：C_{MV}为收敛均值；r^i为完成第i次训练时的奖励，在实际中可对奖励进行缩放；N为训练总次数，每次训练的初始条件不同；C_V为收敛方差；C_{MAX}为收敛最大值；C_{MIN}为收敛最小值。

基于算例1的场景，分别采用PES-DDPG算法和传统DDPG算法进行100次训练，收敛性评价指标如表2所示。

表2 收敛性评价指标对比

Table 2 Comparison of evaluation criteria for convergence

算法	C_{MV}	C_V	C_{MAX}	C_{MIN}
PES-DDPG	57.42	307.51	76.01	25.16
传统DDPG	54.98	438.04	75.98	25.16

由表2可知：在多次训练中，采用PES-DDPG算法和传统DDPG算法得到的C_{MAX}较接近，C_{MIN}指标相同，这说明2种算法的寻优能力相当，均能够找到最优解，同时也会陷入局部最优；采用PES-DDPG算法得到的C_{MV}大于传统DDPG算法，这说明在多次寻优过程中采用PES-DDPG算法取得较优结果的次数多于传统DDPG算法；采用PES-DDPG算法得到的C_V约为传统DDPG算法的70.2%，由于C_V越大，寻优结果波动越大，因此，PES-DDPG算法的稳定性更高。

4.3.3 动态噪声搜索策略分析

本文应用动态噪声搜索策略设计对比策略，以验证搜索策略对PES-DDPG算法收敛性的影响：静态策略，设置静态噪声，在前1 000轮训练中噪声幅值保持不变，在后500轮训练中噪声幅值为0；动态策略，设置线性动态噪声，参数$\lambda=4\times10^{-5}$。图6为静态策略和动态策略下PES-DDPG算法的收敛结果对比。由图可知：当在PES-DDPG算法中采用静态策略时，训练至1 000轮时噪声突然消失，改变了算法的收敛方向，有可能导致最终收敛到的策略比当前的策略差；而当在PES-DDPG算法中采用动态策略时，在训练过程中噪声的影响逐步减弱，对算法收敛性产生的影响较小。

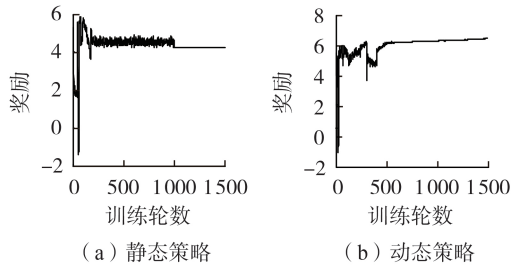

图6 静态策略和动态策略下PES-DDPG算法的收敛结果对比

Fig.6 Comparison of convergence results of PES-DDPG algorithm between static strategy and dynamic strategy

此外，DDQN算法作为另一类基于值函数的DRL算法，本文探讨静态策略和动态策略对DDQN算法收敛性的影响。该算法的静态策略是指在前1 000轮训练中智能体以0.5的概率随机选择动作，1 000轮后智能体选择当前最大Q值所对应的动作；该算法的动态策略是指在训练开始时智能体按照动态衰减的概率随机选择动作，随着训练轮数的增加，随机选择动作的概率最终衰减为0.004，智能体不再随机选择动作，而是选择最大Q值所对应的动作。静态策略和动态策略下DDQN算法的收敛结果对比如图7所示。由图可知：当在DDQN算法中采用静态策略时，在前1 000轮训练中奖励趋势几乎保持不变，每轮奖励在[−100,50]范围内波动；当在DDQN算法中采用动态策略时，在前1 000轮训练中获得的奖励逐渐提高，这说明算法逐渐寻找到了更好的策略，因此，在DDQN算法中采用动态策略的搜索效果优于静态策略。

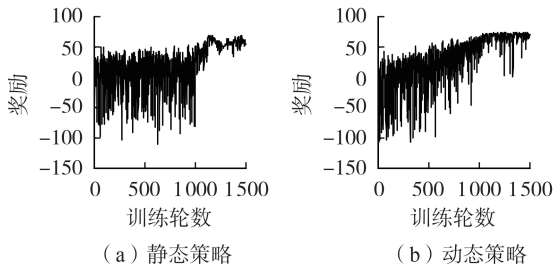

图7 静态策略和动态策略下DDQN算法的收敛结果对比

Fig.7 Comparison of convergence results of DDQN algorithm between static strategy and dynamic strategy

分别在PES-DDPG算法和DDQN算法中采用静态策略和动态策略进行100次训练，收敛性评价指

标如表3所示。

表3　静态策略和动态策略下的收敛性评价指标对比

Table 3　Comparison of evaluation criteria for convergence between static strategy and dynamic strategy

算法	C_{MV}	C_{V}	C_{MAX}	C_{MIN}
PES-DDPG（动态策略）	57.42	307.51	76.01	25.16
PES-DDPG（静态策略）	54.43	348.53	75.81	24.84
DDQN（动态策略）	72.51	3.74	73.90	69.33
DDQN（静态策略）	68.93	17.95	73.54	62.39

由表3可知，在2种算法中采用静态策略时，各项指标均比采用动态策略时的指标差，这说明动态策略提高了2种算法的收敛性和寻优能力。值得注意的是，PES-DDPG算法的收敛性更好，多次训练得到的奖励方差很小，而由于离散化动作空间，DDQN算法的寻优能力弱于PES-DDPG算法。

4.4　算例2结果及分析

鲁棒优化假设不确定性存在于不确定性集中，本文构造一种针对不确定性参数的最坏情况来实现最优求解。本算例中不确定集采用盒式形式，风、光、荷不确定集U的表达式为：

$$U=\{(1-\delta)(P_t^{PV}+P_t^{WT}-P_t^{L})\leqslant P_t^{PV}+P_t^{WT}-P_t^{L}\leqslant (1+\delta)(P_t^{PV}+P_t^{WT}-P_t^{L})\} \tag{33}$$

式中：δ为最大预测误差。

考虑风、光、荷不确定集下的系统能量优化，建立最恶劣的交互收益目标为：

$$\min\sum_{t=1}^{T}(\sigma_t^{s}P_t^{s}-\sigma_t^{b}P_t^{b}+\sigma_t P_t^{L}) \tag{34}$$

微电网能量优化鲁棒模型目标函数等效转化为：

$$\max\left[\min_{P_t^{PV},P_t^{WT},P_t^{L}\in U}\sum_{t=1}^{T}(\sigma_t^{s}P_t^{s}-\sigma_t^{b}P_t^{b}+\sigma_t P_t^{L})-\sum_{t=1}^{T}C_t^{MT}+\alpha\sum_{t=1}^{T}U_t^{L}\right] \tag{35}$$

当微电网内不确定性参数的预测误差分别为5%、10%、30%时，PES-DDPG算法和鲁棒优化模型求解算法所得微电网收益如表4所示，其中，鲁棒优化模型求解算法参考文献[20]。随着预测误差的增大，微电网收益均逐渐降低。采用PES-DDPG算法比鲁棒优化模型求解算法得到的收益更高，且随着预测误差的增大，PES-DDPG算法的优势更明显，其原因在于：预测误差越大，场景越极端，由于鲁棒优化最优解具有高度保守性，因此，极端场景降低了鲁棒优化性能，而PES-DDPG算法是一种无模型和数据驱动的DRL算法，其通过数据的训练学习最优控制策略，且可学习到封装在数据中的不确定性。在30%的预测误差下，采用PES-DDPG算法可以比鲁棒优化模型求解算法提高约8.32%的收益。

2种算法的平均计算时间如表5所示。对于微电网设备优化问题，由于规模小以及复杂度低，求解器可以快速求解，但面对大规模变量及约束问题时，求解器的求解效率难以保证，而DRL智能体具备实时优化的潜力。

表4　不同预测误差下的优化算法结果比较

Table 4　Result comparison between optimization algorithms under different prediction errors

预测误差 / %	收益 / 元	
	PES-DDPG算法	鲁棒优化模型求解算法
5	722.9	719.2
10	684.5	678.6
30	480.3	443.4

表5　优化算法平均计算时间

Table 5　Average calculation time of optimization algorithms

算法	平均计算时间 / s
鲁棒优化模型求解算法	2.48
PES-DDPG算法	0.40

5　结论

本文采用DRL算法求解微电网的优化运行问题，在算法层面，对传统DDPG算法进行改进，提出PES-DDPG算法，通过经验优先存储方法和动态噪声搜索策略提高了算法在训练过程中的收敛稳定性。训练完成的DDPG智能体表现出了DDPG算法处理连续变量的能力以及应对不确定问题时的优越性。通过算例验证了本文所提算法能够实现微电网的优化调度。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] 杨新法，苏剑，吕志鹏，等. 微电网技术综述[J]. 中国电机工程学报，2014，34(1)：57-70.
YANG Xinfa，SU Jian，LÜ Zhipeng，et al. Overview on microgrid technology[J]. Proceedings of the CSEE，2014，34(1)：57-70.

[2] AREFIFAR S A，ORDONEZ M，MOHAMED Y A R I. Energy management in multi-microgrid systems-development and assessment[J]. IEEE Transactions on Power Systems，2017，32(2)：910-922.

[3] 吴雄，王秀丽，刘世民，等. 微电网能量管理系统研究综述[J]. 电力自动化设备，2014，34(10)：7-14.
WU Xiong，WANG Xiuli，LIU Shimin，et al. Summary of research on microgrid energy management system[J]. Electric Power Automation Equipment，2014，34(10)：7-14.

[4] ZIA M F，ELBOUCHIKHI E，BENBOUZID M. Microgrids energy management systems：a critical review on methods，solutions，and prospects[J]. Applied Energy，2018，222：1033-1055.

[5] 李驰宇，高红均，刘友波，等. 多园区微网优化共享运行策略[J]. 电力自动化设备，2020，40(3)：29-36.
LI Chiyu，GAO Hongjun，LIU Youbo，et al. Optimal sharing operation strategy for multi park-level microgrid[J]. Electric Power Automation Equipment，2020，40(3)：29-36.

[6] 鲁卓欣,徐潇源,严正,等. 不确定性环境下数据驱动的电力系统优化调度方法综述[J]. 电力系统自动化,2020,44(21):172-183.
LU Zhuoxin, XU Xiaoyuan, YAN Zheng, et al. Overview on data-driven optimal scheduling methods of power system in uncertain environment[J]. Automation of Electric Power Systems,2020,44(21):172-183.
[7] 程杉,倪凯旋,赵孟雨. 基于Stackelberg博弈的充换储一体化电站微电网双层协调优化调度[J]. 电力自动化设备,2020,40(6):49-55,69.
CHENG Shan, NI Kaixuan, ZHAO Mengyu. Stackelberg game based bi-level coordinated optimal scheduling of microgrid accessed with charging-swapping-storage integrated station[J]. Electric Power Automation Equipment,2020,40(6):49-55,69.
[8] NING C,YOU F Q. Data-driven stochastic robust optimization: general computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era[J]. Computers & Chemical Engineering, 2018, 111:115-133.
[9] SHANG C, YOU F Q. Distributionally robust optimization for planning and scheduling under uncertainty[J]. Computers & Chemical Engineering,2018,110:53-68.
[10] 杨挺,赵黎媛,王成山. 人工智能在电力系统及综合能源系统中的应用综述[J]. 电力系统自动化,2019,43(1):2-14.
YANG Ting,ZHAO Liyuan,WANG Chengshan. Review on application of artificial intelligence in power system and integrated energy system[J]. Automation of Electric Power Systems,2019,43(1):2-14.
[11] 张自东,邱才明,张东霞,等. 基于深度强化学习的微电网复合储能协调控制方法[J]. 电网技术,2019,43(6):1914-1921.
ZHANG Zidong,QIU Caiming,ZHANG Dongxia,et al. A coordinated control method for hybrid energy storage system in microgrid based on deep reinforcement learning[J]. Power System Technology,2019,43(6):1914-1921.
[12] BUI V H, HUSSAIN A, KIM H M. Q-learning-based operation strategy for Community Battery Energy Storage System (CBESS) in microgrid system[J]. Energies,2019,12(9):1789.
[13] LIU Y K, ZHANG D X, GOOI H B. Optimization strategy based on deep reinforcement learning for home energy management[J]. CSEE Journal of Power and Energy Systems,2020,6(3):572-582.
[14] 刘俊峰,陈剑龙,王晓生,等. 基于深度强化学习的微能源网能量管理与优化策略研究[J]. 电网技术,2020,44(10):3794-3803.
LIU Junfeng,CHEN Jianlong,WANG Xiaosheng,et al. Energy management and optimization of multi-energy grid based on deep reinforcement learning[J]. Power System Technology, 2020,44(10):3794-3803.
[15] YU L,XIE W W,XIE D,et al. Deep reinforcement learning for smart home energy management[J]. IEEE Internet of Things Journal,2020,7(4):2751-2762.
[16] LI H P,WAN Z Q,HE H B. Real-time residential demand response[J]. IEEE Transactions on Smart Grid,2020,11(5):4144-4154.
[17] 陈亭轩,徐潇源,严正,等. 基于深度强化学习的光储充电站储能系统优化运行[J]. 电力自动化设备,2021,41(10):90-98.
CHEN Tingxuan,XU Xiaoyuan,YAN Zheng,et al. Optimal operation based on deep reinforcement learning for energy storage system in photovoltaic-storage charging station[J]. Electric Power Automation Equipment,2021,41(10):90-98.
[18] OpenEI DOE open data[DB/OL]. [2021-07-20]. https://openei.org/doe-opendata/dataset.
[19] MAHARJAN S,ZHU Q Y,ZHANG Y,et al. Demand response management in the smart grid in a large population regime[J]. IEEE Transactions on Smart Grid,2016,7(1):189-199.
[20] LÖFBERG J. Automatic robust convex programming[J]. Optimization Methods and Software,2012,27(1):115-129.

作者简介:

赵鹏杰

赵鹏杰(1995—),男,博士研究生,研究方向为微电网、博弈论、深度强化学习(**E-mail**:305491303@qq.com);

吴俊勇(1966—),男,教授,博士研究生导师,博士,通信作者,研究方向为人工智能、智能电网、配电网优化运行等(**E-mail**:wujy@bjtu.edu.cn)。

(编辑 王锦秀)

Optimal operation strategy of microgrid based on deep reinforcement learning

ZHAO Pengjie, WU Junyong, WANG Yi, ZHANG Hesheng

(School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: The uncertainty of wind power, photovoltaic and load brings challenges to the formulation of operation strategy for microgrid with high proportion of renewable energy, and the development of artificial intelligence technology provides a new idea for solving the operation optimization problem of microgrid. Based on the reinforcement learning framework, the operation problem of microgrid is transformed into a Markov decision process, and an online scheduling method of microgrid based on deep reinforcement learning is proposed, which takes the maximum economic benefit of microgrid and residents' satisfaction as its object. In order to effectively use the experience in the training process of deep reinforcement learning, a PES-DDPG (Priority Experience Storage Deep Deterministic Policy Gradient) algorithm is designed to learn the optimal scheduling strategy of microgrid for different periods under each type of environment. Case results show that PES-DDPG algorithm can provide effective scheduling strategy for microgrid and realize real-time optimization of microgrid.

Key words: deep reinforcement learning; microgrid; Markov model; optimal operation

附录 A

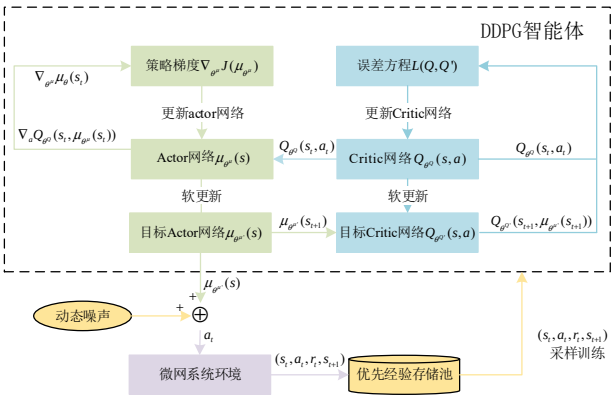

图 A1 DDPG 智能体结构

Fig.A1 Structure of DDPG agent

表 A1 PES-DDPG 算法伪代码

Table A1Pseudo code of PES-DDPG algorithm

初始化经验复用池的容量为 D，抽取样本数 n，最大训练轮数 m,

初始化优化周期 T，目标网络更次频率 T_N ，迭代轮数 k=0

随机初始化策略网络参数 θ^μ 和价值网络参数 θ^Q

初始化目标策略网络参数 $\theta^{\mu'}$ 和目标价值网络参数 $\theta^{Q'}$

For episode = 1， m do

初始化微电网状态，初始化动作探索噪声 ε

For t = 1， T do

根据在线策略网络和探索噪声选择动作 $a_t=\mu(s_t|\theta^\mu)+\varepsilon$

动态更新噪声幅值

智能体执行动作 a_t ，观察即时奖励 r_t

微电网环境对动作 a_t 作出响应，并转换至下一个状态 s_{t+1}

(s_t，a_t，r_t，s_{t+1})作为经验，按照 PES 策略保存在存储池

经验存储池中随机抽取 n 个经验(s_j，a_j，r_j，s_{j+1})

得到目标奖励 y_t (式 24)

最小化损失函数(式 25)更新 Q 网络：

采样得到策略梯度(式 26)更新策略网络：

采用软更新，更新目标网络

$$\theta^{Q'} \longleftarrow \tau\theta^Q+(1-\tau)\theta^{Q'}$$

$$\theta\mu_{\theta^{\mu'}}(s)^{\mu'} \longleftarrow \tau\theta^\mu+(1-\tau)\theta^{\mu'}$$

End for

End for

表 A2 设备工作参数

Table A2 Working parameters of devices

设备	额定容量/kW	参数	数值
燃气轮机	250	a^{MT}	0.0013
		b^{MT}	0.553
		c^{MT}	14.07
储能	200	$\eta_{ch}=\eta_{dis}$	0.95
		S_{min}^{soc}	0.1
		S_{max}^{soc}	1

表 A3 系统峰谷平电价

Table A3 Electricity price in system

时段	购电电价/[元·(kW·h)$^{-1}$]	售电电价/[元·(kW·h)$^{-1}$]
低谷 1—7、23、24	0.37	0.28
平段 8—10、15—18、21、22	0.82	0.65
高峰 11—14、19、20	1.36	0.78

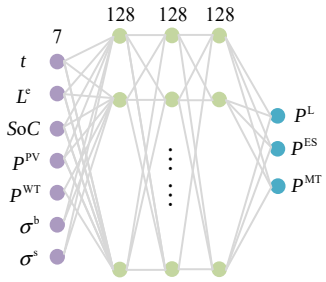

图 A2 神经网络结构

Fig.A2 Structure of neural network

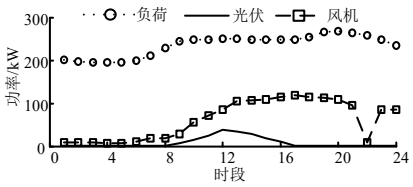

图 A3 可再生能源出力及负荷预测

Fig.A3 Renewable energy output and load forecasting