基于柔性策略-评价网络的微电网源储协同优化调度策略

刘林鹏，朱建全，陈嘉俊，叶汉芳
（华南理工大学 电力学院，广东 广州 510640）

摘要：近年来，微电网中的可再生能源与储能占比不断增大，给其优化调度带来了新的挑战。针对微电网源储协同调度问题中非凸非线性约束带来的求解困难，利用深度强化学习算法构建基于数据的策略函数，通过不断地与环境进行交互学习寻找最优策略，避免了对原非凸非线性问题的直接求解。考虑到训练过程中策略函数可能不满足安全约束，进一步提出了一种利用部分模型信息的微电网源储协同优化调度安全策略学习方法，得到了满足网络安全约束的优化策略。此外，针对强化学习的智能体在训练过程中与环境的交互耗时较长的问题，采用神经网络对环境进行建模以提高学习效率。

关键词：微电网；可再生能源；储能；柔性策略-评价网络；强化学习；深度学习；安全约束
中图分类号：TM 727；TM 73　　**文献标志码**：A　　DOI：10.16081/j.epae.202110036

0 引言

近年来，为了实现可再生能源的就地消纳，微电网中可再生能源的占比日益提高[1-2]。为抑制可再生能源的间歇性和随机性，维持微电网的稳定运行，有必要装设一定比例的储能，实现源储协同运行[3]。在这种背景下，如何充分地考虑可再生能源与储能系统的特点，对微电网进行源储协同优化调度成为一个热点问题。

目前，微电网的优化调度问题已经得到了大量的研究。已有的方法可以分为基于模型的数学优化算法和无模型的强化学习算法2类。基于模型的数学优化算法通常是通过直接求解集中式的数学优化问题以获取最优策略。例如：文献[4]将微电网调度问题转化成二次型最优控制问题，并利用黎卡提方程解的特性对其进行求解；文献[5]将微电网调度问题转化为二阶鲁棒优化模型，利用列约束生成和强对偶原理将原问题分解后交替求解；文献[6]使用KKT（Karush-Kuhn-Tucker）条件及二阶锥松弛技术将微电网调度模型转换为单层的混合整数线性规划问题，并调用CPLEX求解器对其进行求解；文献[7]构建了微电网双层调度模型，并利用交替方向乘子法对其进行求解。上述文献为求解微电网优化调度问题，对原问题中的非凸非线性约束进行了一定简化处理。这些简化处理方法通常建立在一定假设的基础上，它们求得的最优策略与原问题的最优策略在某些情况下并不等价。无模型的强化学习算法将智能体不断与环境进行交互，通过观察交互后得到的结果改进策略。例如：文献[8]使用基于值的深度Q网络DQN（Deep Q-Network）算法得到了微电网的在线调度策略；文献[9]使用基于随机性策略的策略-评价网络AC（Actor-Critic）算法求解微电网的最优调度策略；文献[10]使用基于确定性策略的深度确定性策略梯度DDPG（Deep Deterministic Policy Gradient）算法求解微电网中共享储能的最优控制问题。上述强化学习算法相较于基于模型的数学优化算法的优势在于其不需要模型的信息，可通过观察到的数据寻找最优策略。此外，其得到的策略泛化能力强，在强随机性环境下有较好的表现[8-9]。尽管强化学习方法在微电网优化调度问题的求解过程中有较好的表现，但由于它在训练过程中为保证智能体的探索性能，往往需要在策略探寻过程中加入一定的随机性，这可能导致所搜寻的策略不满足约束条件。为解决这个问题，已有的文献主要采取了以下措施：文献[11]结合了壁垒函数的特性以保证智能体在满足约束的条件下进行策略学习；文献[12]通过在奖励函数中设置惩罚因子，使智能体在学习过程中避开不满足约束条件的策略；文献[13]使用元学习的方式使得策略更新过程满足约束条件。上述方法本质上都是通过无模型学习的方式使得智能体朝着满足约束条件的方向对策略进行更新，但这类方法并不能保证所得策略严格满足约束条件。

针对以上问题，本文结合有模型的数学优化与无模型的强化学习的思想，提出了一种基于柔性策略-评价网络SAC（Soft Actor-Critic）的微电网源储协同优化调度方法。一方面，所提方法在不对原问

收稿日期：2021-06-30；**修回日期**：2021-09-07
基金项目：国家自然科学基金资助项目（51977081）；电力系统国家重点实验室资助课题（SKLD20M15）；广东省自然科学基金资助项目（2019A1515010877）；广东省普通高校青年创新类人才项目（2019GKQNCX040）
Project supported by the National Natural Science Foundation of China（51977081），the Foundation of State Key Laboratory of Power System（SKLD20M15），the Natural Science Foundation of Guangdong Province（2019A1515010877） and the Young Innovative Talents Project of Colleges and Universities in Guangdong Province（2019GKQNCX040）

题进行简化处理的前提下，利用强化学习算法将原问题分解为多个子问题进行求解，并通过贝尔曼最优定理保证了所得策略与原问题最优策略的等价性；另一方面，所提方法利用部分模型信息使得策略严格满足约束条件。此外，为减少智能体在训练过程中与环境的交互时长，本文提出了一种基于深层长短期记忆LSTM(Long Short-Term Memory)网络的环境建模方法。

1 微电网源储协同调度模型

1.1 目标函数

以微电网的运行成本最小化为目标，则有：

$$F=\min_{P_{g,t},P_{s,t},P_{l,t},o_{g,t}}\sum_{t=1}^{24}c_t \tag{1}$$

式中：$P_{g,t}$和$P_{s,t}$分别为t时段机组g和储能s的有功出力，$P_{s,t}$取值为正时表示储能放电，取值为负时表示储能充电，其最大值为$P_s^{\max}$；$P_{l,t}$为t时段联络线l传输的有功功率，其取值为正时表示从主网购电，取值为负时表示向主网售电；$o_{g,t}$为t时段机组g状态，其取值为0时表示处于离线状态，取值为1时表示处于工作状态；c_t为t时段即时成本。

1.2 马尔可夫决策过程

在利用强化学习求解优化问题时，需要先将原问题构建为一个马尔可夫决策过程[14]。本文从时间维度对原问题进行解耦，构建了以下的马尔可夫决策过程。

1)状态。

$$s_t=\{E_{\mathrm{S},t},p_t,P_{\mathrm{wt},t},P_{\mathrm{pv},t},L_t,o_{g,t},t\} \tag{2}$$

式中：s_t为t时段的状态变量集合；$E_{\mathrm{S},t}$为t时段储能电量；p_t为t时段电价；$P_{\mathrm{wt},t}$为t时段风力发电功率；$P_{\mathrm{pv},t}$为t时段光伏发电功率；L_t为t时段总负荷；$o_{g,t}$为决策前机组g的状态。

2)动作。

$$a_t=\{P_{s,t},o_{g,t+1}\} \tag{3}$$

式中：a_t为t时段的动作变量集合；$o_{g,t+1}$为决策后机组g的状态。

3)状态转移。

状态转移用以表示状态变量的改变与动作、环境之间的函数关系。储能电量转移函数如式(4)所示；负荷、风电出力、光伏出力和电价的状态转移函数分别如式(5)—(8)所示，它们分别服从未知的分布D_{L}、D_{wt}、D_{pv}和D_{p}。

$$E_{\mathrm{S},t+1}=E_{\mathrm{S},t}-\left(\frac{P_{\mathrm{dis},t}}{\eta}+P_{\mathrm{cha},t}\eta\right) \tag{4}$$

$$L_{t+1}\sim D_{\mathrm{L}}(\mu_{\mathrm{L},t+1},\sigma^2_{\mathrm{L},t+1}) \tag{5}$$

$$P_{\mathrm{wt},t+1}\sim D_{\mathrm{wt}}(\mu_{\mathrm{wt},t+1},\sigma^2_{\mathrm{wt},t+1}) \tag{6}$$

$$P_{\mathrm{pv},t+1}\sim D_{\mathrm{pv}}(\mu_{\mathrm{pv},t+1},\sigma^2_{\mathrm{pv},t+1}) \tag{7}$$

$$p_{t+1}\sim D_{\mathrm{p}}(\mu_{\mathrm{p},t+1},\sigma^2_{\mathrm{p},t+1}) \tag{8}$$

式中：$P_{\mathrm{cha},t}$和$P_{\mathrm{dis},t}$分别为t时段储能的充电和放电功率；η为储能的充放电效率系数；$\mu_{\mathrm{L},t+1}$、$\mu_{\mathrm{wt},t+1}$、$\mu_{\mathrm{pv},t+1}$和$\mu_{\mathrm{p},t+1}$分别为分布D_{L}、D_{wt}、D_{pv}和D_{p}的均值；$\sigma_{\mathrm{L},t+1}$、$\sigma_{\mathrm{wt},t+1}$、$\sigma_{\mathrm{pv},t+1}$和$\sigma_{\mathrm{p},t+1}$分别为分布D_{L}、D_{wt}、D_{pv}和D_{p}的标准差。

4)奖励。

奖励是智能体每次与环境进行交互时收到的反馈信号，可用于指导策略的更新方向。为了实现微电网的运行成本最小化，本文将奖励设置为即时成本的负值：

$$r_t(s_t,a_t)=-c_t(s_t,a_t) \tag{9}$$

式中：r_t为t时段智能体在状态s_t下做出动作a_t获得的奖励。

5)环境。

在本文的微电网源储协同优化调度模型问题中，智能体所处的环境为原问题在时间维度解耦后的单时段优化问题：

$$\min_{P_{g,t},P_{l,t}} c_t\left(P_{g,t},P_{l,t}\middle|s_t,a_t\right)=\sum_g\Big[o_{g,t}\left(a_gP_{g,t}^2+b_gP_g+c_g\right)+o_{g,t}c_{\mathrm{o}}(o_{g,t}-o_{g,t-1})\Big]+p_tP_{l,t}+c_{\mathrm{s}}\left(\frac{P_{\mathrm{dis},t}}{\eta}-P_{\mathrm{cha},t}\eta\right) \tag{10}$$

$$\begin{cases}P_{i,t}^{\mathrm{in}}=\sum_{j=1}^{N}V_{i,t}V_{j,t}\left[G_{ij}\cos(\delta_{i,t}-\delta_{j,t})+B_{ij}\sin(\delta_{i,t}-\delta_{j,t})\right]\\ Q_{i,t}^{\mathrm{in}}=\sum_{j=1}^{N}V_{i,t}V_{j,t}\left[G_{ij}\sin(\delta_{i,t}-\delta_{j,t})-B_{ij}\cos(\delta_{i,t}-\delta_{j,t})\right]\end{cases} \tag{11}$$

$$V_{\min}\leqslant V_{i,t}\leqslant V_{\max} \tag{12}$$

$$\left|P_{l,t}\right|\leqslant P_l^{\max} \tag{13}$$

$$E_{\mathrm{S},\min}\leqslant E_{\mathrm{S},t}\leqslant E_{\mathrm{S},\max} \tag{14}$$

$$P_g^{\min}\leqslant P_{g,t}\leqslant P_g^{\max} \tag{15}$$

$$P_{\mathrm{cha},t}P_{\mathrm{dis},t}=0 \tag{16}$$

式中：a_g、b_g和c_g为机组g的发电成本系数；c_{o}为停启成本系数；c_{s}为储能充放电成本系数；N为节点总数；$P_{i,t}^{\mathrm{in}}$为t时段节点i的注入有功功率，即节点i处的所有电源出力之和与负荷的差；$Q_{i,t}^{\mathrm{in}}$为t时段节点i的注入无功功率；G_{ij}和B_{ij}分别为节点i和节点j之间的线路导纳的实部和虚部；$V_{i,t}$和$\delta_{i,t}$分别为t时段节点i的电压幅值和电压相角；$V_{\min}$和$V_{\max}$分别为电压幅值的最小值和最大值；$P_l^{\max}$为联络线功率上限；$E_{\mathrm{S},\min}$和$E_{\mathrm{S},\max}$分别为储能电量的下限和上限；$P_g^{\min}$和$P_g^{\max}$分别为机组g出力的下限和上限。

在微电网源储协同调度问题中，决策变量包含机组出力、储能充放电功率、机组的启停状态以及联络线功率。若直接用无模型的强化学习算法搜寻这4个变量对应的策略，将无法保证其搜寻的策略严

格满足约束条件。为解决这一问题，将这4个变量分成了两部分：一部分为储能充放电功率和机组的启停状态，这部分变量通过强化学习的策略网络输出得到；另一部分为机组的出力和联络线功率，这部分变量由策略网络输出储能充放电功率和机组的启停状态后通过CPLEX商业求解器求解式(10)—(15)组成的单时段的优化问题得到。通过这种方式求解这4个决策变量可以保证它们严格满足约束条件。

2　基于SAC的源储协同优化调度

2.1　SAC优化策略

定义微电网源储协同优化调度策略π为：

$$\pi\left(a_t\middle|s_t\right)=\rho\left(a_t\middle|s_t\right) \tag{17}$$

式中：$\rho\left(a_t\middle|s_t\right)$为状态$s_t$下动作$a_t$的概率分布。为使微电网日内运行成本最小，在探寻最优策略π^*过程中，应使得智能体在任意状态s_t下都能输出使得日内成本最小化的动作分布$\rho^*\left(a_t\middle|s_t\right)$。

2.1.1　智能体的目标函数

SAC算法作为无模型的强化学习算法之一，能够有效地在模型未知的情况下，通过不断地与环境进行交互以搜寻最优策略[15]。本文将利用SAC算法学习最优策略的智能体称为SAC智能体。在微电网源储协同优化调度问题中，SAC智能体的目标可定义为最大化智能体调度周期内的总奖励与策略熵的期望值[16]：

$$\max_{\pi} E_{v\sim D_t}\left(r_t(s_t,a_t)+\alpha H_t\left(\pi\left(a_t\middle|s_t\right)\right)\right) \tag{18}$$

式中：$E(\cdot)$表示求期望；v为分布未知的随机量，$v\in\{L,P_{\mathrm{wt}},P_{\mathrm{pv}},p\}$；$\alpha$为策略熵权重系数；$H(\pi)$为策略熵，具体定义如式(19)所示。

$$H(\pi)=E_{a\sim\pi}(-\ln\pi(a)) \tag{19}$$

通过求解式(18)所示的目标函数，所得策略便可实现总奖励的最大化(即运行成本最小化)。另一方面，由于目标函数考虑了将策略熵最大化，所得策略具有更强的探索性能以及更好的鲁棒性。

2.1.2　智能体结构

SAC智能体一般由2个神经网络组成，它们分别为评价网络和策略网络，具体结构如附录A图A1所示。图中，评价网络用于输出反映当前策略好坏程度的状态-动作值函数$Q(s_t,a_t)$；策略网络用于根据当前智能体所处状态输出相应的策略$\pi\left(a_t\middle|s_t\right)$。

SAC智能体的状态-动作值函数$Q^\pi(s_t,a_t)$是智能体从状态s_t和动作a_t开始执行策略π，直到结束状态时所得总奖励与策略熵的期望值：

$$Q^\pi(s_t,a_t)=E\left[\sum_{\tau=t}^{24}\gamma^{\tau-1}\left(r_\tau+\alpha H_\tau\left(\pi\left(a_\tau\middle|s_\tau\right)\right)\right)\middle|s_\tau,a_\tau\right] \tag{20}$$

式中：γ为奖励折扣系数。

根据贝尔曼方程，可以推导出状态-动作值函数$Q^\pi(s_t,a_t)$的递归方程为[14]：

$$Q^\pi(s_t,a_t)=r_t+\alpha H_t(\pi)+\gamma Q(s_{t+1},a_{t+1}) \tag{21}$$

2.1.3　评价网络的参数更新

对于评价网络，其参数是朝着真实状态-动作值函数的方向更新的。因此，基于式(21)以及时序差分算法可得SAC智能体评价网络的参数更新公式为[17]：

$$\min_{\theta_Q}\frac{1}{M}\sum_{i=1}^{M}\left(y_i-Q\left(s_i,a_i\middle|\theta_Q\right)\right)^2 \tag{22}$$

$$y_i=r_i+\gamma Q(s_i',a_i')+\alpha H\left(\pi\left(s_i'\middle|\theta_\pi\right)\right) \tag{23}$$

$$\min_{\alpha}\frac{1}{M}\sum_{i=1}^{M}\left(\alpha H'-\alpha H\left(\pi\left(s_i\middle|\theta_\pi\right)\right)\right)^2 \tag{24}$$

式中：θ_Q和θ_π分别为评价网络和策略网络的参数，可利用文献[18]所提的小批量梯度下降法分别求解式(22)和式(24)以获得θ_Q和α的更新值；H'为目标策略熵；M为小批量更新的样本数量；i表示样本编号，每个样本由(s_i,a_i,r_i,s_i')构成，其中s_i'为转移后状态；a_i'为智能体在s_i'下根据当前策略所得动作。智能体每次与环境进行交互时均会产生一个样本，并将其存入经验回放池中[19]。

2.1.4　策略网络的参数更新

对于策略网络，其参数是朝着最大化总奖励和策略熵的方向进行更新的。因此，可利用梯度上升法求解式(25)对其参数θ_π进行更新。

$$\max_{\theta_\pi}\frac{1}{M}\sum_{i=1}^{M}\left(Q\left(s_i,\pi\left(s_i\middle|\theta_\pi\right)\right)+\alpha H\left(\pi\left(s_i\middle|\theta_\pi\right)\right)\right) \tag{25}$$

SAC智能体不断地与环境进行交互产生新的样本并存入经验回放池中，且每次与环境进行交互后都根据经验回放池中的样本对评价网络和策略网络进行一次参数更新。在超参数设置合理的前提下，通过一定次数的交互训练后，SAC智能体的策略最终可收敛到最优策略[20]。

通过这种方式，可以将原问题分解为多个子问题求解。根据贝尔曼最优定理，所得策略与原问题最优策略具有等价性，相关证明见附录B。

2.2　基于深层LSTM网络的环境建模

由于SAC智能体每次与环境进行交互时，都需要求解一个由式(10)—(16)组成的单时段优化问题，这将导致训练的时间大幅增加。为加快SAC智能体的训练速度，本文利用深层LSTM网络对环境进行建模。

深层LSTM神经网络是循环神经网络RNN(Recurrent Neural Network)的一种类型，其基本结构如附录C图C1所示。从图中可以看出，RNN的隐藏层包含了当前时刻的输入信息以及上一时刻的输入信

息，因此它具有记忆功能。为解决RNN的梯度爆炸和消失问题，LSTM对RNN进行了改进，其结果如附录C图C2所示，图中σ表示Logistic函数，输出区间为(0,1)。LSTM在RNN的基础上引入内部状态$\boldsymbol{c}_t$，用于传递循环信息，引入外部状态$\boldsymbol{h}_t$用于接收内部状态传递的信息，具体如下：

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t \tag{26}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh\boldsymbol{c}_t \tag{27}$$

$$\tilde{\boldsymbol{c}}_t=\tanh\left(\boldsymbol{W}_c\boldsymbol{x}_t+\boldsymbol{U}_c\boldsymbol{h}_{t-1}+\boldsymbol{b}_c\right) \tag{28}$$

式中：$\odot$表示向量元素相乘；$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$分别为遗忘门、输入门和输出门，它们控制其对应的信息通过比例，且$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$中各元素取值范围为[0,1]；$\boldsymbol{W}_c$、$\boldsymbol{U}_c$和$\boldsymbol{b}_c$为可学习的神经网络参数。

与传统的前馈神经网络类似，使用小批量梯度下降法更新LSTM网络参数θ_n：

$$\nabla\theta_n=\frac{1}{K}\sum_{j=1}^{K}\frac{\partial\left\|y_j-\hat{y}_j\left(x_j\middle|\theta_n\right)\right\|}{\partial\theta_n} \tag{29}$$

$$\theta_n^{t+1}=\theta_n^t-\beta\nabla\theta_n \tag{30}$$

式中：K为小批量样本数目；x_j、y_j分别为样本j的特征与标签；$\hat{y}_j$为样本j的LSTM网络输出量；β为学习率。

3 算例分析

3.1 参数设置

以图1所示的微电网为例对所提方法进行测试，相关参数见附录D。评价网络与策略网络结构参数以及用于环境建模的深层LSTM网络超参数见附录E。所有算例均基于MATLAB R2021a实现，并在64位Windows系统、Intel Core i7-6700K@3.7 GHz的环境下运行。

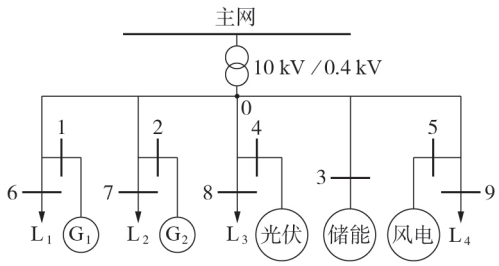


图1 微电网结构

Fig.1 Structure of microgrid

由于深层LSTM网络的训练是一种“端到端”的有监督学习方法，因此在训练前，首先需要准备一定数量的样本。本文通过CPLEX商业求解器求解1 000个不同场景下由式(10)—(16)组成的优化问题，得到了1 000个样本，并将90%的样本作为训练集，用于训练深层LSTM网络；将其余10%的样本作为测试集，用于测试模型的准确性。每个样本包含了用于训练的标签和特征，其中标签为$\boldsymbol{c}_t$，特征为$\{P_{s,t},o_{g,t+1},E_{S,t},p_t,P_{wt,t},P_{pv,t},L_t,o_{g,t}\}$。

3.2 智能体的离线训练过程

为验证SAC智能体在随机环境下的学习能力，假设负荷、风电出力、光伏出力和电价分别服从式(31)—(34)中均值和标准差的高斯分布。

$$\begin{cases}\mu_{L,t}=-0.02t^3+0.53t^2-1.89t+61.81\\ \sigma_{L,t}=0.05\mu_{L,t}\end{cases} \tag{31}$$

$$\begin{cases}\mu_{wt,t}=-0.001t^3+0.04t^2-0.51t+13.2\\ \sigma_{wt,t}=0.15\mu_{wt,t}\end{cases} \tag{32}$$

$$\begin{cases}\mu_{pv,t}=-0.16t^2+4.11t-10.74\\ \sigma_{pv,t}=0.10\mu_{pv,t}\end{cases} \tag{33}$$

$$\begin{cases}\mu_{p,t}=-0.001t^2+0.04t+0.26\\ \sigma_{p,t}=0.05\mu_{p,t}\end{cases} \tag{34}$$

图2展示了SAC智能体在设置的随机环境训练时，微电网的运行成本期望随训练次数增加而变化的过程，其中该期望值通过最近100次训练结果的平均值近似表示。从图2中可以看出：在训练前期，微电网运行成本的期望值随着训练次数的增加而降低；在完成2 400次训练之后，微电网运行成本的期望值基本保持不变，因此可以认为此时SAC智能体找到了近似最优策略。

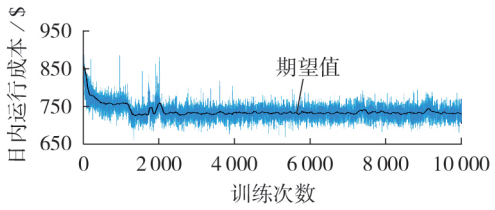


图2 SAC智能体训练过程

Fig.2 Training process of SAC agent

为验证本文所提方法的优势，图3展示了无模型的SAC智能体在设置的随机环境训练时的运行成本变化情况。其中，无模型的SAC智能体采用了文献[12]中的方法，在奖励函数中对于不满足约束条件的策略设置了惩罚因子。在本算例中，对不满足式(12)的策略增加一个值为$200的惩罚成本。从图3中可以看出，这种在奖励函数中增加惩罚因子的无模型强化学习方法无法保证策略严格满足约束条件，造成其运行成本产生较大波动。

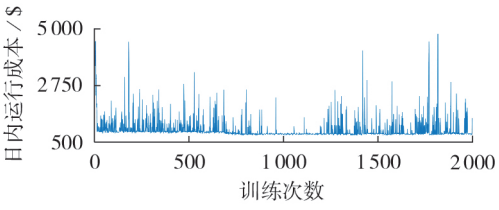


图3 无模型的SAC智能体训练过程

Fig.3 Training process of model-free SAC agent

3.3　智能体在线决策分析

将离线训练后的SAC智能体用于微电网源储协同优化调度的在线决策，并与短视（myopic）策略进行对比。其中，短视策略通过求解式（35）中的单时段优化问题得到。

$$\{P_{g,t}, P_{s,t}, P_{l,t}, o_{g,t}\} = \underset{P_{g,t}, P_{s,t}, P_{l,t}, o_{g,t}}{\operatorname{argmin}} \; c_t \tag{35}$$

图4展示了2种策略连续进行1个月的在线决策的情况。从图中可以看出，所提方法的优化效果明显优于短视策略。采用短视策略时，微电网在该月运行成本均值为$766.90；而采用本文策略后，微电网在该月运行成本均值为$726.36（比短视策略所得运行成本降低了5.29%），这主要得益于本文所提的方法具有远视能力，能全局考虑调度周期内的情况以获得更优的结果。

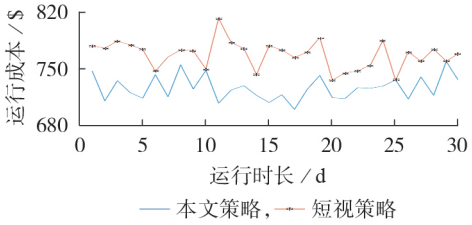


图4　运行1个月的结果对比

Fig.4　Comparison of results in a month

进一步地，图5以第一天的在线决策结果为例，详细展示了采用本文所提方法进行在线决策时各时段的状态变量以及动作变量情况。可以发现，在电价较低时，微电网需要从主网购电以满足负荷需求。由于此时微电网自备机组的运行成本比购电成本高，所以发电机处于停机状态。另一方面，储能选择在电价较低时尽可能充电，随后在电价较高时放电以获取更高的利益。

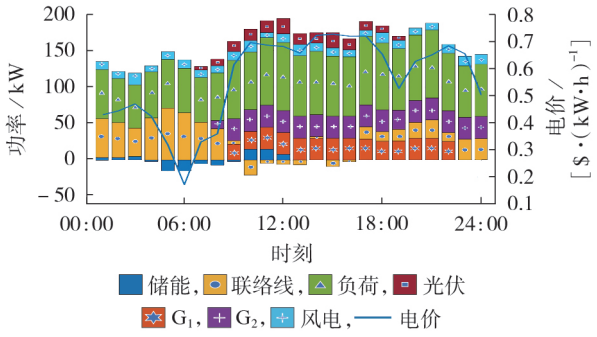


图5　日内在线决策结果

Fig.5　Intra-day online decision results

3.4　LSTM网络环境建模分析

为测试本文所提的LSTM网络环境建模方法的有效性，将基于原环境和深层LSTM网络模型得到的微电网的源储协同优化调度策略进行对比分析。

图6展示了不同测试场景下基于原环境和深层LSTM网络模型得到的成本对比情况。从图中可以看出，基于深层LSTM模型的输出成本曲线与基于原环境的成本曲线基本重合，均方根误差仅为0.3153，这说明深层LSTM模型所建的环境与原环境近似等效。

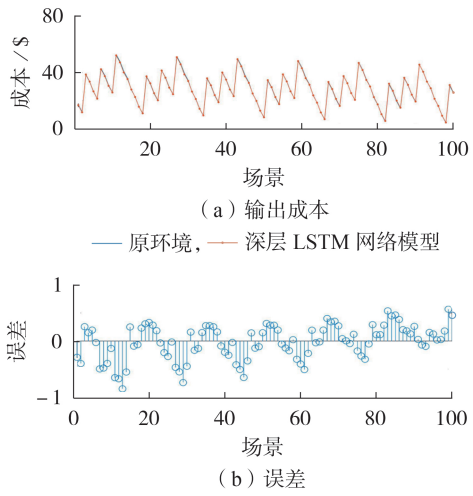


图6　深层LSTM网络误差分析

Fig.6　Error analysis of deep LSTM network

表1进一步对比了SAC智能体在原环境与深层LSTM网络所建环境下的离线训练时长以及在线决策的平均成本。从表中可见，深层LSTM网络所构建的环境减少了80.03%的离线训练时长，而在线决策平均成本仅与原环境相差0.01%。这表明所提深层LSTM网络环境建模在不影响在线决策精度的前提下，显著减少了智能体的离线训练时长。需要说明的是，尽管智能体的离线训练时间较长，但在在线决策阶段，由于可以直接利用离线训练好的策略网络进行决策，其耗时仅为0.41 s，因而可以满足在线决策的需求。

表1　2种环境模型效果对比

Table 1　Comparison of effects between two environment models

环境	离线训练时长 / s	在线决策平均成本 / $
原环境	128 080	726.36
深层LSTM网络模型所建的环境	25 580	726.45

4　结论

本文提出了一种基于SAC的微电网源储协同调度策略，得到的主要结论如下：

1）所提方法能够通过不断地与环境进行交互的方式获得最优策略，并基于部分模型信息进行策略搜寻，确保所得策略满足安全约束；

2）所提环境建模方法在不影响策略准确性的前提下，减少了SAC智能体的训练时长，提高了SAC智

能体的学习效率；

3)所提方法对模型信息的依赖程度较低，仅用时0.41 s便可获得显著优于短视策略的解，可以较好地满足微电网源储协同调度的在线决策要求。

附录见本刊网络版(http://www.epae.cn)。

参考文献：

[1] 骆钊，卢涛，马瑞，等. 可再生能源配额制下多园区综合能源系统优化调度[J]. 电力自动化设备，2021，41(4)：8-14.
LUO Zhao，LU Tao，MA Rui，et al. Optimal scheduling of multi-park integrated energy system under renewable portfolio standard[J]. Electric Power Automation Equipment，2021，41(4)：8-14.

[2] 杨健，唐飞，廖清芬，等. 考虑可再生能源随机性的微电网经济性与稳定性协调优化策略[J]. 电力自动化设备，2017，37(8)：179-184，200.
YANG Jian，TANG Fei，LIAO Qingfen，et al. Optimal strategy for coordinating economy and stability of microgrid considering the randomness of renewable energy[J]. Electric Power Automation Equipment，2017，37(8)：179-184，200.

[3] 陈丽丽，牟龙华，许旭锋，等. 储能装置运行策略及运行特性对微电网可靠性的影响[J]. 电力自动化设备，2017，37(7)：70-76.
CHEN Lili，MOU Longhua，XU Xufeng，et al. Influence of operation strategy and operation characteristics of energy storage device on reliability of microgrid[J]. Electric Power Automation Equipment，2017，37(7)：70-76.

[4] 夏超英，苗海丽. 基于二次型最优控制的微电网实时源储协同调度策略[J]. 中国电机工程学报，2019，39(3)：721-730，951.
XIA Chaoying，MIAO Haili. A real-time source storage cooperative scheduling strategy for microgrid based on quadratic optimal control[J]. Proceedings of the CSEE，2019，39(3)：721-730，951.

[5] 刘一欣，郭力，王成山. 微电网两阶段鲁棒优化经济调度方法[J]. 中国电机工程学报，2018，38(14)：4013-4022，4307.
LIU Yixin，GUO Li，WANG Chengshan. Two-stage robust optimal economic scheduling method for microgrid[J]. Proceedings of the CSEE，2018，38(14)：4013-4022，4307.

[6] 黄张浩，张亚超，郑峰，等. 基于不同利益主体协调优化的主动配电网日前-实时能量管理方法[J]. 电网技术，2021，45(6)：2299-2308.
HUANG Zhanghao，ZHANG Yachao，ZHENG Feng，et al. Day-ahead and real-time energy management for active distribution network based on coordinated optimization of different stakeholders[J]. Power System Technology，2021，45(6)：2299-2308.

[7] 王皓，艾芊，吴俊宏，等. 基于交替方向乘子法的微电网群双层分布式调度方法[J]. 电网技术，2018，42(6)：1718-1727.
WANG Hao，AI Qian，WU Junhong，et al. Multi-level distributed scheduling method for micro-grid group based on alternate direction multiplier[J]. Power System Technology，2018，42(6)：1718-1727.

[8] CAO T，SHEN Z，ZHANG G. LSTM-aided reinforcement learning for energy management in microgrid with energy storage and EV charging[C]//15th International Conference on Mobile Ad-Hoc and Sensor Networks(MSN). Hong Kong，China：IEEE，2020：13-18.

[9] WEI Y，ZHANG Z，YU F R，et al. Power allocation in hetnets with hybrid energy supply using actor-critic reinforcement learning[C]//Globecom IEEE Global Communications Conference. Singapore：IEEE，2017：1-5.

[10] ODONKOR P，LEWIS K. Control of shared energy storage assets within building clusters using reinforcement learning[C]//International Design Engineering Technical Conferences. Quebec，Canada：ASME，2018：86-94.

[11] CHENG R，OROSZ G，MURRAY R M，et al. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu，Hawaii，USA：AAAI，2019：3-6.

[12] ELFWING S，SEYMOUR B. Parallel reward and punishment control in humans and robots：safe reinforcement learning using the MaxPain algorithm[C]//2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics. New Orleans，LA，USA：IEEE，2017：140-147.

[13] DJORDJE G，SEBASTIAN R. Safe reinforcement learning through meta-learned instincts[C]//Proceedings of the ALIFE 2020：the 2020 Conference on Artificial Life. [S.l.]：MIT Press，2020：283-291.

[14] SUTTON R，BARTO A. Reinforcement learning：an introduction[M]. Cambridge，USA：MIT Press，2017：156.

[15] SUGIYAMA M. Statistical reinforcement learning. Modern machine learning approaches[J]. Chapman & Hall / CRC，2015：11(4)：1330-1340.

[16] WANG W，YU N，GAO Y，et al. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems[J]. IEEE Transactions on Smart Grid，2020，11(4)：3008-3018.

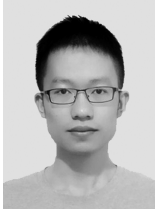
[17] BAIRD L. Residual algorithms：reinforcement learning with function approximation[M]. Tahoe City，California，USA：Elsevier，1995：67.

[18] 李彬，彭曙蓉，彭君哲，等. 基于深度学习分位数回归模型的风电功率概率密度预测[J]. 电力自动化设备，2018，38(9)：15-20.
LI Bin，PENG Shurong，PENG Junzhe，et al. Probability density prediction of wind power based on deep learning quantile regression model[J]. Electric Power Automation Equipment，2018，38(9)：15-20.

[19] 陈希亮，曹雷，李晨溪，等. 基于重抽样优选缓存经验回放机制的深度强化学习方法[J]. 控制与决策，2018，33(4)：600-606.
CHEN Xiliang，CAO Lei，LI Chenxi，et al. A deep reinforcement learning method based on re-sampling optimal cache experience replay mechanism[J]. Control and Decision，2018，33(4)：600-606.

[20] 祁文凯，桑国明. 基于延迟策略的最大熵优势演员评论家算法[J]. 小型微型计算机系统，2020，41(8)：90-98.
QI Wenkai，SANG Guoming. The maximum entropy dominant actor critic algorithm based on delay strategy[J]. Journal of Small and Micro Computer Systems，2020，41(8)：90-98.

作者简介：



刘林鹏

刘林鹏(1996—)，男，江西赣州人，硕士研究生，主要研究方向为人工智能在电力系统中的应用(**E-mail**：1105918406@qq.com)；

朱建全(1982—)，男，广西玉林人，副教授，博士研究生导师，通信作者，主要研究方向为电力系统建模、分析与优化及电力市场等(**E-mail**：zhujianquan@scut.edu.cn)；

陈嘉俊(1996—)，男，广东江门人，硕士研究生，主要研究方向为人工智能在电力系统中的应用及电力市场(**E-mail**：550990770@qq.com)。

(编辑 李玮)

Cooperative optimal scheduling strategy of source and storage in microgrid based on soft actor-critic

LIU Linpeng,ZHU Jianquan,CHEN Jiajun,YE Hanfang
(School of Electric Power,South China University of Technology,Guangzhou 510640,China)

Abstract: In recent years,the proportion of renewable energy and energy storage in microgrid is increasing, which brings new challenges to its optimal scheduling. Aiming at the difficulty in solving the cooperative optimal scheduling problem of source and storage in microgrid due to the non-convex nonlinear constraints, the deep reinforcement learning algorithm is used to construct the data-based strategy function,and the optimal strategy is found out through continuous interactive learning with the environment,so that avoiding the direct solution of the original non-convex nonlinear problem. Considering the strategy function may not meet the security constraints in the training process,furthermore,a learning method of cooperative optimal scheduling secure strategy of source and storage in microgrid based on partial model information is proposed,and the optimal strategy meeting the network security constraints is obtained. In addition,aiming at the problem of long time-consuming due to the interaction between agents and environment in the training process for reinforcement learning,the neural network is used to model the environment,so as to improve the learning efficiency.

Key words: microgrid;renewable energy;energy storage;soft actor-critic;reinforcement learning;deep learning; security constraint

(上接第 19 页 continued from page 19)

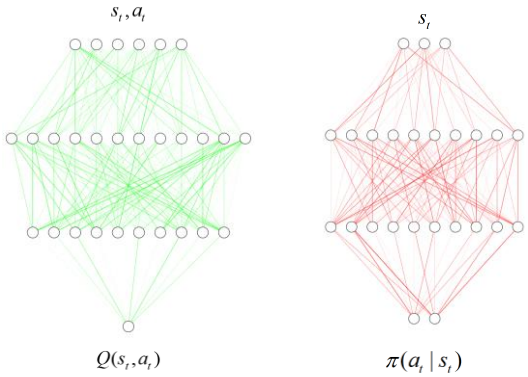
Review and prospect of robust optimization and planning research on generation and transmission system

YUAN Yang[1],ZHANG Heng[1],CHENG Haozhong[1],LIU Lu[1],ZHANG Xiaohu[2],LI Gang[2],ZHANG Jianping[2]
(1. Key Laboratory Control of Power Transmission and Conversion,Ministry of Education,
Shanghai Jiao Tong University,Shanghai 200240,China;
2. East China Branch of State Grid Corporation of China,Shanghai 200120,China)

Abstract: With the uncertainty of power system increasing gradually,the application of robust optimization and planning research on generation and transmission system to resist the uncertainty of extremely scenes has become a significant research method. Firstly,the robust optimization is divided into classical robust optimization and distributionally robust optimization from the perspective of whether the probability distribution characteristics of uncertain factors are considered,the mathematical models and uncertain set characteristics of these two kinds of robust optimization are sorted out. Secondly,the existing classical robust optimization and distributionally robust optimization research on generation and transmission system are divided into three aspects:considering the uncertainty of node injection power,considering the uncertainty of power capacity growth and cost,and considering the uncertainty of transmission network state,and the research framework and limitations of robust optimization planning research on generation and transmission system are refined. Finally,the problems worthy of further study in robust optimization planning on power generation and transmission system are prospected,which provides ideas and directions for the robust optimization planning follow-up research on power generation and transmission system.

Key words: generation and transmission system;optimization and planning;uncertainty;robust optimization; distributionally robust optimization

附录 A



(a)评价网络　　(b)策略网络

图 A1 评价网络与策略网络结构
Fig.A1 Structure of critic and actor

附录 B：所得策略与原问题最优策略的等价性说明

本文基于马尔科夫决策理论，将多时段优化问题解耦为单时段优化问题，所得单时段最优之和在理论上等价于全时段最优，具体分析如下。

1）在本文所构建的马尔科夫决策框架下，根据文献[14]中的贝尔曼最优方程以及状态动作值函数的定义，可将多时段优化问题的目标函数写作由当前时段的奖励以及下一时段的状态动作值函数组成的表达式：

$$\max_{\pi} E(\sum_{t=\tau}^{\tau+24} r_t) = \max_{a_\tau} Q(s_\tau, a_\tau) = \max_{a_\tau} E(r_\tau(s_\tau, a_\tau) + \max_{a_{\tau+1}} Q(s_{\tau+1}, a_{\tau+1}))$$

对上式中的状态动作值函数进行展开，得：

$$\max_{a_\tau} E(r_\tau(s_\tau, a_\tau) + \max_{a_{\tau+1}} Q(s_{\tau+1}, a_{\tau+1})) =$$
$$\max_{a_\tau} E(r_\tau(s_\tau, a_\tau) + \max_{a_{\tau+1}} E(r_{\tau+1}(s_{\tau+1}, a_{\tau+1}) + \max_{a_{\tau+2}} Q(s_{\tau+2}, a_{\tau+2})))$$

根据最优子结构的定义可知，一个问题的最优解可由其最优子结构组成，而通过上式可知该多时段优化问题的最优子结构为：

$$\max_{a_\tau} E(r_\tau(s_\tau, a_\tau) + \max_{a_{\tau+1}} Q(s_{\tau+1}, a_{\tau+1}))$$

由于上述最优子结构中的状态动作值函数 Q 在本文方法中是通过神经网络近似得到的，在求解单时段的奖励后即可得到子结构的最优解。

2）在本文方法中，状态动作值函数 Q 由深层前馈神经网络输出得到，根据神经网络的通用近似定理，所提方法在理论上可以通过神经网络逼近真实的状态动作值函数。因此，使用该神经网络得到的策略与使用真实值函数得到的策略是等价的，即将原问题解耦为单时段优化问题所得最优策略与原问题的全时段最优策略理论上是等价的。需要说明的是，在实际应用中难以满足这个条件，得到的是近似最优解。

附录 C

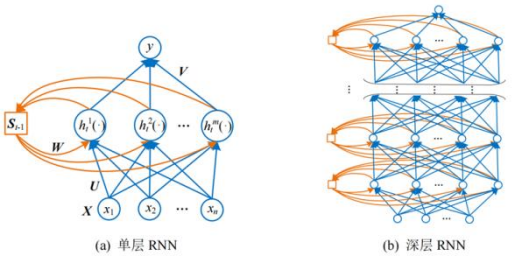


图 C1 RNN 基本结构

Fig.C1 Basic structure of RNN

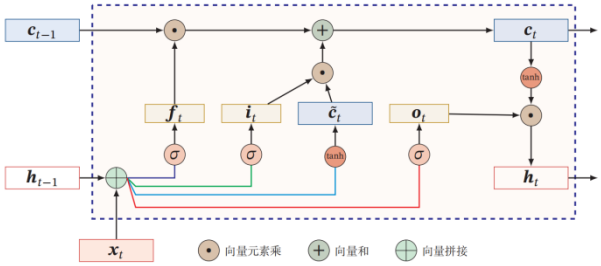


图 C2 LSTM 循环单元结构

Fig.C2 Cycle unit structure of LSTM

附录 D：微电网相关参数

表 D1 微电网线路参数

Table D1 Parameters of microgrid lines

线路编号	始节点	末节点	电阻/p.u.	电抗/p.u.
1	0	1	0.0320	0.0050
2	0	2	0.0640	0.0100
3	0	3	0.0192	0.0030
4	0	4	0.0640	0.0100
5	0	5	0.1280	0.0200
6	1	6	0.0320	0.0050
7	2	7	0.0512	0.0080
8	4	8	0.0640	0.0100
9	5	9	0.0960	0.0150

表 D2 储能参数

Table D2 Parameters of energy storage

参数	数值
η	0.9
$E_{S,min}$ /(kW・h)	18
P_s^{max} /kW	15
C_e/$	0.02

表 D3 发电单元参数

Table D3 Parameters of generators

发电单元	a/(\$ • kW^{-1})	b/(\$ • kW^{-1})	c/(\$ • kW^{-1})	C_0/\$
G_1	0.0036	0.2779	2.8	12
G_2	0.0037	0.2128	5.5	5

附录 E：算法超参数设置

表 E1 SAC 智能体参数

Table E1 Parameters of SAC agent

参数	数值	参数	数值
评价/策略网络隐藏层数	3	γ	0.99
评价网络隐藏层神经元数量	250	α 初始值	1
评价网络隐藏层激活函数	ReLU	H'	-3
评价网络输出层激活函数	tanh	M	128
策略网络隐藏层神经元数量	200	β_Q	10^{-3}
策略网络隐藏层激活函数	ReLU	β_π	10^{-2}

表 E2 深层 LSTM 网络超参数

Table E2 Hyper parameters of deep LSTM network

超参数	数值
隐藏层层数	3
隐藏层神经元数量	250
LSTM 状态激活函数	tanh
LSTM 门激活函数	sigmoid
小批量样本数目	128
学习率	0.0025