基于软脉冲双延迟深度确定性策略梯度算法的自动发电控制

席 磊[1,2],赵俊苗[1],黄浩超[1],施 宇[1],王文涛[1],李宗泽[1]
(1. 三峡大学 电气与新能源学院,湖北 宜昌 443002;
2. 三峡大学 梯级水电站运行与控制湖北省重点实验室,湖北 宜昌 443002)

摘要:强化学习能够有效获取随机问题的最优解,但进行价值函数估计时存在低估和高估偏差,严重影响自动发电控制的性能。为此,提出一种基于软脉冲双延迟深度确定性策略梯度的自动发电控制算法,通过引入玻尔兹曼softmax算子,利用其平滑性质和误差控制能力,平衡Q值的高估和低估偏差。为解决引入softmax算子带来的计算负担增加问题,融入脉冲神经网络,通过模拟生物神经元的时序活动降低计算负担。通过对改进的IEEE标准两区域自动发电控制模型以及基于西南电网的风光水火储一体化多区域自动发电控制模型进行仿真,验证了所提算法的有效性,且与其他算法相比,所提算法具有更小的频率偏差和更优的控制性能。

关键词:自动发电控制;强化学习;脉冲神经网络;策略梯度;控制性能
中图分类号:TM73 **文献标志码**:A DOI:10.16081/j.epae.202602015

0 引言

“双碳”目标加速了新型电力系统的发展。具有波动性和随机性特性的新能源大规模接入[1-4]所带来的强随机扰动,严重影响电网的频率稳定性及自动发电控制(automatic generation control,AGC)性能,如控制性能指标(control performance standard,CPS)和区域控制偏差(area control error,ACE)。在实际电力系统运行中,集中式AGC受到广域通信网络的时延和带宽限制,并面临地理分散性带来的信息同步挑战,这使其在实现跨区域协同控制方面受到约束。

基于马尔可夫随机过程的强化学习[5-6]能够有效获取随机问题的最优解,被广泛应用于AGC,本文将其命名为RL-AGC。Q学习及其衍生算法是目前在RL-AGC领域应用最为广泛的强化学习算法[7-8]。文献[9]将Q学习作为AGC算法,依赖奖励Q函数来选择动作,进而形成闭环控制以提高整个系统的灵活性和适应性。文献[10]采用具有先验知识的Q学习作为AGC算法,以提高强化学习的收敛速度和学习效率。文献[11]在传统的深度Q学习的基础上,采用置信区间上界策略优先级采样机制代替均匀随机采样机制,促使智能体快速收敛到最优策略。文献[12]提出一种基于虚拟狼群策略的分层Q学习控制策略的RL-AGC算法,实现了AGC的多区域协同控制。

收稿日期:2025-08-17;**修回日期**:2026-01-28
在线出版日期:2026-02-11
基金项目:国家自然科学基金资助项目(52277108,52477104)
Project supported by the National Natural Science Foundation of China(52277108,52477104)

然而,上述RL-AGC算法是受限于预定义动作集的离散强化学习,本文称之为离散RL-AGC算法,该算法在面对复杂且连续变化的状态空间[13]时,往往无法输出精确的动作,这使强化学习的学习过程变得缓慢,甚至无法收敛到最优解。为此,基于连续强化学习的RL-AGC被引入AGC系统,本文称之为连续RL-AGC算法,该算法无须预定义动作集,在面对高维连续状态空间时,能够输出精确的动作。文献[14]提出一种基于专家经验回放的贪婪Actor-Critic算法,该算法利用高斯分布策略生成连续的动作值,使得强化学习能够在高维连续状态空间中寻找到最优解,有效应对了AGC系统中新能源出力的强随机性。文献[15]提出一种面向新型电力系统多能协同的深度强化学习算法,解决了新能源大规模接入电网带来的随机扰动使AGC性能变差的问题。文献[16]提出一种将智能体执行策略建模为连续动作空间的AGC控制方法,以探索高维状态-动作的协同最优解。然而,上述连续RL-AGC算法在Q值估计时,由于最大值操作对Q值近似误差的放大作用,容易产生Q值的高估,使得强化学习在决策过程中对某些动作的价值产生过高估计,从而导致AGC的性能下降。

文献[17]提出一种多经验池概率回放的双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient,TD3)的连续RL-AGC算法,通过使用2个Q网络的最小值进行延迟策略的更新,在实现AGC连续控制的同时,解决了强化学习中Q值高估的问题。文献[18]提出一种基于知识嵌入型深度强化学习的电力系统频率紧急控制方法,能够显著提高智能体的策略学习效率与决策质量。文献

[19]提出一种面向综合能源系统的多智能体协同RL-AGC算法，不仅解决了强化学习中Q值高估的问题，还快速获取到强化学习的最优策略，使电网频率的稳定性明显提高。然而，上述连续RL-AGC算法在解决Q值高估问题的同时，却忽略了Q值的低估偏差，导致AGC机组输出的调节功率不足以完全抵消ACE，进而影响电网频率的稳定性和AGC性能。

文献[20]通过引入玻尔兹曼softmax算子来改进Q值估计方法，采用加权平均操作平衡Q值的高估与低估偏差。然而，该方法在每次Q值更新时都需要计算2个Q网络的输出并通过softmax算子进行融合，导致算法计算负担增加，使算法收敛速度变慢，进而影响AGC性能。文献[21]引入脉冲神经网络(spiking neural network，SNN)，利用脉冲神经元的时序活动来精细建模强化学习中的状态转移和奖励信号，更高效地更新Q值网络，显著加快了算法的收敛速度，进而提高了AGC的性能。

因此，本文提出一种融入玻尔兹曼softmax算子及SNN的连续RL-AGC算法，即软脉冲双延迟深度确定性策略梯度(soft spike twin delayed deep deterministic policy gradient，SoftSpike-TD3)算法。通过在SNN的输出层引入softmax机制，将离散的脉冲发放率转化为连续的概率分布，有效解决传统TD3算法中Q值高估的问题，同时保持SNN的低计算特性，通过减轻算法计算负担来加快算法收敛速度，进而提高AGC的性能。最后，通过改进的IEEE标准两区域AGC模型以及基于西南电网的风光水火储一体化多区域AGC模型进行仿真，验证所提算法的有效性。

1 风光水火储系统数学模型

相较于传统的AGC控制框架，新型电力系统的AGC模型通过融合多元异构电源及可调度柔性负荷进行优化调整，从而构建出兼具高灵活性与高能效性的频率调节机制。

1.1 风电机组及光伏机组

风能和太阳能发电[22]单元作为新型电力系统中的关键可再生能源组成部分，是AGC模型中的重要部分。风机的输出功率P_W可描述为：

$$P_W=\frac{1}{2}\rho A V_W^3 C \tag{1}$$

式中：ρ为空气的密度；A为转子的截面积；V_W为风速额定值；C为转子叶片的功率系数。设T_W为风机的时间常数，ΔP_W为风电场的功率变化，风电机组的模型结构如附录A图A1所示。

设P_{PV}为光伏发电的输出功率，ΔP_{PV}为光伏功率偏差，T_P为光伏机组的时间常数，ΔP_{solar}为太阳能发电的功率偏差量，P_{solar}为光伏发电系统的额定功率。光伏发电机组模型如附录A图A2所示，其输出功率相对于稳态值的偏离量为：

$$\Delta P_{solar}=0.6\sqrt{P_{solar}} \tag{2}$$

1.2 火电机组及水电机组

基荷与调峰任务主要由火电机组承担，其输出平稳且持续；而水电机组则凭借快速爬坡能力提供瞬时灵活性，用于平抑负荷侧的突发扰动。为便于分析频率响应，2类机组均可降阶为“调速环节+原动机”结构：火电机组对应汽轮机驱动，水电机组对应水轮机驱动。调速器、汽轮机以及水轮机经线性化后的传递函数模型[23]为：

$$G_g(s)=\frac{1}{1+sT_g} \tag{3}$$

$$G_t(s)=\frac{1}{1+sT_t} \tag{4}$$

$$G_s(s)=\frac{1+sT_{w1}}{1+sT_{w2}}\frac{1-sT_{wh}}{1+s\times 0.5T_{wh}} \tag{5}$$

式中：$G_g(s)$为调速器的数学模型；$G_t(s)$为汽轮机的数学模型；$G_s(s)$为水轮机的数学模型；T_g为调速器自身的响应迟延；T_t为汽轮机的时间常数；T_{wh}为水轮机流体惯性的等效时间常数；T_{w1}、T_{w2}分别为水电机组调节阀的时间常数和调速器的时间常数。

2 SoftSpike-TD3

SoftSpike-TD3是在TD3的基础上引入softmax算子，用于更新Actor-Critic的值函数，从而解决Q值高估和低估的问题。然而，softmax算子的引入增加了算法的计算负担，导致算法的收敛速度较慢。为了克服这一缺点，本文利用脉冲神经元的时序活动来精细建模强化学习中的状态转移和奖励信号，更高效地更新Q值网络，显著加快收敛速度，从而有效提升AGC性能。

2.1 TD3

TD3作为深度确定性策略梯度(deep deterministic policy gradient，DDPG)的衍生算法，通过Actor-Critic架构与深度神经网络的结合，实现对连续控制方法的优化。TD3主要通过如下3项关键技术的引入，有效解决Q值函数[24]的高估问题。

1)采用双Q学习机制。在这一机制下，算法部署了2个独立的Q网络，分别对动作价值函数进行评估。在更新过程中，算法选取2个Q网络评估结果中的较小值作为目标Q值，以此降低单一Q网络估计偏差所带来的影响。

2)引入策略延迟更新策略。通过降低策略网络的更新频率，即在固定训练步数后才对策略网络进行1次更新，有效避免了因频繁更新策略网络而导致的算法不稳定性问题。这种策略更新延迟机制有

助于减少算法在训练过程中的过度估计偏差，从而提高算法的稳定性。

3）通过在目标策略中引入噪声，实现了目标策略的平滑化。这种噪声的引入增加了动作选择的随机性，有助于避免算法陷入局部最优解，从而提高算法的探索能力和性能。

尽管TD3通过上述改进有效解决了Q值的高估问题，但目标策略与行为策略之间的差异仍然会导致策略优化过程中的低估偏差，进而影响算法的稳定性控制性能。

2.2　玻尔兹曼softmax算子

softmax算子为平衡Q值的高估和低估偏差提供了一种更平滑的价值估计方法。本文通过计算2个Q网络评估结果中最小值的softmax，来估计目标价值，从而有效地降低高估和低估偏差的影响。

在计算目标Q值时，首先利用下一时刻的状态s'和动作a'，计算2个Critic网络的softmax权重，softmax函数定义为：

$$\mathrm{softmax}_{\beta}\left(Q(s,\cdot)\right)=\int_{a\in\mathcal{A}}\frac{\exp\left(\beta Q(s,a)\right)}{\int_{a'\in\mathcal{A}}\exp\left(\beta Q(s,a')\right)\mathrm{d}a'}Q(s,a)\mathrm{d}a \tag{6}$$

式中：β为softmax算子的温度参数；a为动作；$\mathcal{A}$为连续且有界的动作空间；$Q(s,a)$为在状态s下采取动作a所获得的预期回报；$\mathrm{softmax}_{\beta}\left(Q(s,\cdot)\right)$表示对$Q$值函数$Q(s,a)$应用softmax操作。

利用双估计器的softmax算子的方法，通过$y_i=r+\gamma\mathcal{T}_{\mathrm{SD3}}(s')$来估计Critic网络$Q_i$的目标值，其中$y_i$为$t+1$时刻的目标值，$r$为奖励函数，$\gamma$为折现因子，$\mathcal{T}_{\mathrm{SD3}}(s')$为在状态$s'$下通过softmax计算的目标值函数。$\mathcal{T}_{\mathrm{SD3}}(s')$表达式为：

$$\mathcal{T}_{\mathrm{SD3}}(s')=\mathrm{softmax}_{\beta}\left(\hat{Q}_i(s',\cdot)\right) \tag{7}$$

$$\hat{Q}_i(s',a')=\min\left(Q_i(s',a';\theta_i^-),Q_{-i}(s',a';\theta_{-i}^-)\right) \tag{8}$$

式中：θ_{-i}^-为目标网路的参数。$\hat{Q}_i$为2个Critic网络Q值的最小值，Q_{-i}为第i个Critic网络配对的另一个网络。

引入玻尔兹曼softmax算子对连续动作空间中的所有可能动作进行加权平均，平衡了Q值高估和低估的问题，进而加快了算法的收敛速度，然而，同时也增加了算法的计算负担。

2.3　SNN

SNN[25]的核心思想是模拟生物神经元的动作电位变化，通过脉冲序列来传递信息。与传统的人工神经网络相比，SNN在处理时间序列数据和执行异步更新方面具有优势，这使其在处理动态和时间依赖性任务时表现出更高的效率和适应性。

AGC系统的状态信号（如频率偏差Δf、联络线功率偏差ΔP_{tie}）是典型的时序信号。SNN的神经元模型（leukemia inhibitory factor，LIF）具有内生的膜电位积分特性，适合处理和记忆时序信息，能够更有效地捕捉状态信号的变化率和历史趋势，这对于预测系统动态至关重要。

编码器模块首先将从环境中观测到的连续状态向量进行归一化处理，观察值转换为群体中每个神经元的刺激强度A_{E}，即：

$$A_{\mathrm{E}}=\exp\left\{-\frac{1}{2}\left[(s-\mu)/\sigma\right]^2\right\} \tag{9}$$

式中：μ、σ为群体编码的参数。

在预设的周期内为归一化的状态向量生成脉冲序列，A_{E}充当神经元的突触前输入。考虑突触电流衰减的LIF神经元动态特性为：

$$I(t)=I_{\mathrm{R}}+I_{\mathrm{C}} \tag{10}$$

式中：$I(t)$为神经元的总电流输入；I_{R}为神经元的脉冲电流，符合线性电阻器特性；I_{C}为电容器电流。I_{R}和I_{C}的特性直接影响脉冲的生成和传播，进而影响动作的生成。

对各个维度的动作a_i，有：

$$\mu_a=\frac{1}{H}\sum_{i=1}^{N}a_i \tag{11}$$

$$\sigma_a=\sqrt{\frac{1}{H}\sum_{i=1}^{N}\left(a_i-\mu_a\right)^2} \tag{12}$$

式中：μ_a为动作a的均值；H为样本数量；σ_a为动作a的标准差。

2.4　SoftSpike-TD3

为解决强化学习中的高估和低估以及因计算负担大而导致的收敛速度慢的问题，本文在TD3的基础上引入softmax算子和SNN，提出SoftSpike-TD3算法。该算法通过在双估计器上构建玻尔兹曼softmax算子改善Q值的低估偏差，同时融合SNN来减轻计算负担，加快算法的收敛速度。

Actor网络通过群体编码将输入状态转换为脉冲信号，并通过时空反向传播进行训练，生成动作概率函数。Critic网络评估动作的优劣，并通过奖惩值指导优化Actor网络的动作策略。

使用扩展的时空反向传播更新SNN参数。通过函数$z(v)$近似脉冲梯度[16]，如式（13）所示。

$$z(v)=\begin{cases}1 & \left|v-V_{\mathrm{th}}\right|<h\\ 0 & \text{其他}\end{cases} \tag{13}$$

式中：V_{th}为发放阈值；h为渐变的阈值窗。

SoftSpike-TD3中Q值的更新公式为：

$$Q_{t+1}\left(s,a;\theta_{t+1}\right)\leftarrow\mu_0Q_t\left(s,a;\theta_t\right)+\alpha_t\left(r+\eta_{t+1}C_a+\gamma\max_{a'}Q_t^-(s',a';\theta_t^-)-\mu_0Q_t\left(s,a;\theta_t\right)\right) \tag{14}$$

式中：Q_{t+1}为第t+1次迭代时的动作值函数；θ_i^-为第i个Critic的目标网络参数；μ_0为平滑参数；α_t为学习率；η_{t+1}为第$t+1$次迭代时的动作索引；C_a为动作成本；θ_t为第t次迭代时的网络参数。迭代过程为：

$$Q_{t+1}(s,a)=r_t(s,a)+\gamma E_{s'\sim p(\cdot|s,a)}\left(V_t(s')\right) \tag{15}$$

式中：$r_t(\cdot)$为即时奖励函数；$E_{s'\sim p(\cdot|s,a)}$为下一个状态s'的期望，$p(\cdot|s,a)$为状态转移概率；$V_t(\cdot)$为第t次迭代的状态值函数。

引入SNN可以降低计算负担，适合实时控制任务。鉴于存在的低估问题，本文采用softmax算子平滑优化过程，解决局部最优的问题，提高算法的收敛速度和稳定性。

最小Q值计算公式为：

$$\hat{Q}(s',\hat{a}')=\min_{i=1,2}Q_i(s',\hat{a}';\theta_i^-) \tag{16}$$

式中：$\hat{a}'$为目标策略加噪声后的候选动作。

Critic网络更新公式为：

$$L_{\text{critic}}=\frac{1}{N}\sum_s\left(Q_i(s,a;\theta_i)-y_i\right)^2 \tag{17}$$

式中：L_{critic}为Critic网络的损失函数。

策略梯度更新公式为：

$$\nabla_{\phi_i}J=\frac{1}{N}\sum_s\nabla_{\phi_i}\pi(s;\phi_i)\nabla_a Q_i(s,a;\theta_i)\big|_{a=\pi(s;\phi_i)} \tag{18}$$

式中：$\nabla_{\phi_i}J$为策略目标函数J关于参数ϕ_i的梯度，ϕ_i为第i个Actor网络的参数；$\pi(s;\phi_i)$为第i个确定性策略网络。

目标网络更新公式为：

$$\theta_i^-\leftarrow\tau\theta_i+(1-\tau)\theta_i^- \tag{19}$$

$$\phi_i^-\leftarrow\tau\phi_i+(1-\tau)\phi_i^- \tag{20}$$

式中：τ为目标网络软更新系数；θ_i为第i个Critic在线网络参数；ϕ_i^-为第i个Actor在线网络参数。

SoftSpike-TD3通过结合softmax算子和SNN，在解决高估和低估偏差问题的同时，优化了计算效率，增强了训练过程的稳定性和高效性。SoftSpike-TD3的伪代码如附录B表B1所示。

2.5 参数设置

1）softmax算子温度参数$\beta(0<\beta<1)$，用于控制Q值估计的平滑程度和探索性。β越大，softmax越接近均匀分布，融合结果越平滑；β越小，融合结果越接近最大Q值。本文β取为0.5。

2）折现因子$\gamma(0<\gamma<1)$，指明了偏重于当前奖励，还是长期奖励。γ值越趋近于1，越偏重于长期奖励；γ值越趋近于0，则越偏重于当前奖励。考虑到智能体追求长期回报，本文γ取为0.95。

3）学习因子$\alpha(0<\alpha<1)$，决定了赋给算法更新部分的信任度。α较大时，SoftSpike-TD3的收敛速度加快，但很可能收敛到局部最优值；α较小时，SoftSpike-TD3的搜索空间会得到保证，收敛的稳定性会随之提高。本文α取为0.1。

4）目标网络软更新系数$\tau(0<\tau<1)$，用于控制主网络参数向目标网络参数的融合速度。τ越小，更新越慢，目标网络更稳定，训练更平滑；τ越大，更新越快，目标网络更敏感，可能引入振荡。本文取0.001。

5）经验池容量B越大，样本多样性越高，训练更稳定，但占用内存线性增加；B越小，样本重复快，易过拟合，训练波动大。

6）发放阈值$V_{\text{th}}(0.8<V_{\text{th}}<1.5)$，决定了神经元发放脉冲所需的膜电位积累程度，直接影响SNN的脉冲发放率和信息编码密度。

7）渐变阈值窗h的调优范围设定为$0.8<h<0.99$。

参数设置如附录B表B2所示。

3 基于SoftSpike-TD3的AGC系统设计

本文提出的SoftSpike-TD3能够解决传统RL-AGC算法存在的学习过程缓慢、高估、低估等问题，因此本文设计基于SoftSpike-TD3的AGC系统。

在AGC系统中，各区域的智能体为SoftSpike-TD3算法，同时将其作为本区域的AGC控制器，通过观测AGC系统当前的输出动作，并根据环境返回的奖励信号动态调整控制策略，进而输出最优功率调节指令。

分布式AGC各区域智能体负责实时采集所在区域电网的运行数据，对每个区域的ACE、互联电网频率偏差Δf、CPS的数据进行长期记录、实时监控和计算，并将其作为AGC控制器的可观测状态s_i。本文采用北美电力可靠性委员提出的CPS标准[26]和Δf（合格范围为±0.2 Hz）对AGC性能进行评估。具体指标如下：若CPS1的值大于等于200%，则无须考虑CPS2的值，CPS指标是合格的；若CPS1的值大于等于100%且小于200%，CPS2大于等于90%，则CPS的指标是合格的；若CPS1的值小于100%，则CPS指标是不合格的。

3.1 奖励函数

ACE作为关键指标，直接反映了区域控制的精准度以及实时动作的准确性和有效性。CPS可评估AGC系统在较长时间内的频率调节效果，从而为电网的长期稳定运行提供保障。为确保ACE输出的稳定性以及CPS指标的长期稳定，本文的目标奖励函数为k时刻ACE瞬时值和CPS瞬时值的加权，如式（21）所示。

$$R(k)=-\eta e_{\text{ACE}}^2(k)-(1-\eta)\left(e_{\text{CPS1}}(k)-200\right)^2 \tag{21}$$

式中：$R(k)$为第k次迭代的奖励值；$e_{\mathrm{ACE}}(k)$为k时刻ACE的值；$e_{\mathrm{CPS1}}(k)$为k时刻CPS1的值；η为权重因子，当其值为0.5时，AGC系统能够实现控制性能最优。e_{ACE}及e_{CPS1}的计算公式分别为：

$$e_{\mathrm{ACE}}=\Delta P_{\mathrm{tie}}+T\Delta f \tag{22}$$

$$e_{\mathrm{CPS1}}=\left(2-\frac{\sum e_{\mathrm{ACE\text{-}1min}}\Delta F_{\mathrm{ave\text{-}1min}}}{-10B\varepsilon_0^2}\right)\times 100\,\% \tag{23}$$

式中：ΔP_{tie}为联络线功率；T为频率响应系数；$e_{\mathrm{ACE\text{-}1min}}$为1 min内ACE的平均值；$\Delta F_{\mathrm{ave\text{-}1min}}$为1 min内频率偏差的绝对值；$\varepsilon_0$为全年1 min内频率平均值偏差的均方根。

3.2　状态和动作空间

ACE的动态调整能力使得AGC能够更好地应对负荷波动和新能源发电的不确定性，并有助于减小频率偏差，确保电力系统的持续高效运行。本文中智能体的状态即为当前时刻ACE的状态向量$\boldsymbol{s}_k$，即：

$$\boldsymbol{s}_k=\left[e_{\mathrm{ACE}}(k)\right] \tag{24}$$

智能体依据状态信息输出动作a_k，且传统机组按照等比例容量进行功率分配，从而得到各传统机组的发电功率指令，具体公式为：

$$\Delta P_{k,g}=a_k\frac{P_{\mathrm{max},g}}{\sum_{j=1}^{N_{\mathrm{total}}}P_{\mathrm{max},j}} \tag{25}$$

式中：$\Delta P_{k,g}$为传统机组g的输出功率指令；$P_{\mathrm{max},g}$为机组g的最大调频容量；N_{total}为总机组数。

AGC系统中包括可控性柔性负荷，柔性负荷的显著特征是其输出功率和消纳能力具有不确定性，可控性柔性负荷的发电功率指令为：

$$\Delta P_{\mathrm{f},g}=\begin{cases}\dfrac{E_{\mathrm{c},g}-E_{\mathrm{cmin},g}}{E_{\mathrm{cmax},g}}a_k & a_k>0\\ \dfrac{E_{\mathrm{cmax},g}-\left(E_{\mathrm{c},g}-E_{\mathrm{cmin},g}\right)}{E_{\mathrm{cmax},g}}a_t & a_k<0\\ 0 & a_k=0\end{cases} \tag{26}$$

$$\Delta P=n\Delta P_{k,g}+m\Delta P_{\mathrm{f},g} \tag{27}$$

式中：$\Delta P_{\mathrm{f},g}$为柔性负荷g的输出功率指令；$E_{\mathrm{c},g}$为柔性负荷g当前的总储能量；$E_{\mathrm{cmax},g}$、$E_{\mathrm{cmin},g}$分别为柔性负荷g总储能量的上、下限；n为系统中的传统机组数量；m为系统中的可控性柔性负荷数量；ΔP为总调节功率指令。

若按照传统机组的等比例容量进行功率分配，则分配值过低会导致柔性负荷的优势没有被充分利用，而分配值过高会导致功率指令与柔性负荷实际输出之间出现显著偏差，进而对电力系统的安全构成威胁。因此，可控性柔性负荷在接收到智能体输出的总调节功率指令后，将执行1次协同功率分配。

3.3　控制流程

SoftSpike-TD3在完成预学习阶段的离线策略训练之后，即可进入正式运行阶段。具体的操作流程如下。

在预学习阶段，系统引入电网负荷扰动，收集动态响应数据，并以4 s为周期进行数据实时更新。智能体在每个周期内采集AGC系统的实时状态变量$\boldsymbol{s}_k$，计算相应的动作a_k，根据当前状态和动作获得即时奖励r_k，以评估动作的效果并指导策略的优化。通过与AGC系统的交互，智能体根据当前策略和探索选择并执行动作a，观察奖励r和下一个状态s'，将经验元组(s,a,r,s')储存到经验缓冲区中。再采样小容量的N个经验元组和K个噪声，计算噪声扰动后的动作a以及目标Q值。在此基础上，引入玻尔兹曼softmax算子来计算目标值，以更新Critic网络、Actor网络和目标网络。经过大量的离线训练，智能体的策略网络逐步被优化，最终能学习到适应于随机干扰的最优控制策略，并获得稳定收敛的策略函数$\boldsymbol{\pi}_\theta\left(\boldsymbol{s}_k\middle|\theta_\pi\right)$。

在在线运行阶段，将预学习时获得的策略网络集成至AGC系统中。在在线运行过程中，AGC系统持续监测电网状态，收集实时数据$\boldsymbol{s}_k$。策略网络$\boldsymbol{\pi}_\theta\left(\boldsymbol{s}_k\middle|\theta_\pi\right)$利用这些状态信息，迅速确定当前最优的总调节指令，从而生成每台机组相应的调节指令。

各区域SoftSpike-TD3智能体与本区域AGC系统的交互流程如图1所示。

4　算例分析

4.1　改进的IEEE标准两区域AGC模型

本文在IEEE标准两区域负荷频率控制模型的基础上，融入水力发电、风电和飞轮储能，如附录C图C1所示。

4.1.1　预学习

在正式投入在线学习阶段之前，要先基于SoftSpike-TD3算法的AGC系统完成充分随机探索的过程，即预学习。为确保训练目标与工程实际相符合，在两区域模型中引入周期为1 200 s、幅值为1 000 MW的连续正弦信号，以训练智能体获取最优策略，共训练15轮。在训练结束后，对各算法最后一轮的频率、CPS1、ACE进行收集与对比分析，以评估所提算法的控制性能。图2展示了DDPG[27]、深度Q网络(deep Q-network，DQN)算法[28]、软演员-评论家(soft Actor-Critic，SAC)算法[29]、软深度确定性策略梯度(soft deep deterministic policy gradient，SD3)[30]、TD3、SoftSpike-TD3算法预学习过程的性能指标曲线(以A区域为例)。

预学习性能指标如图2所示。由图2(a)的预学

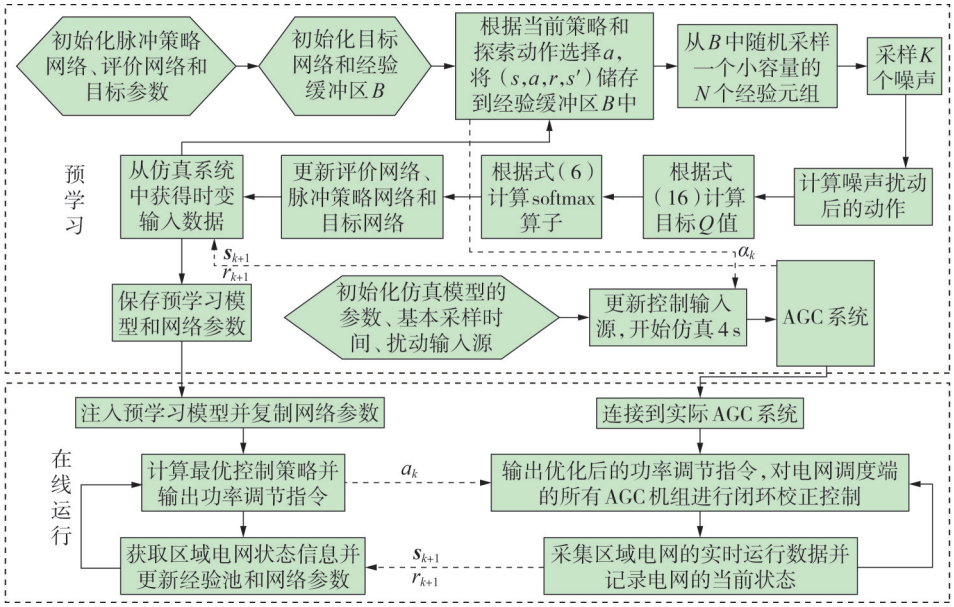

图1 SoftSpike-TD3智能体与AGC系统交互流程

Fig.1 Interaction flow between SoftSpike-TD3 agent and AGC system

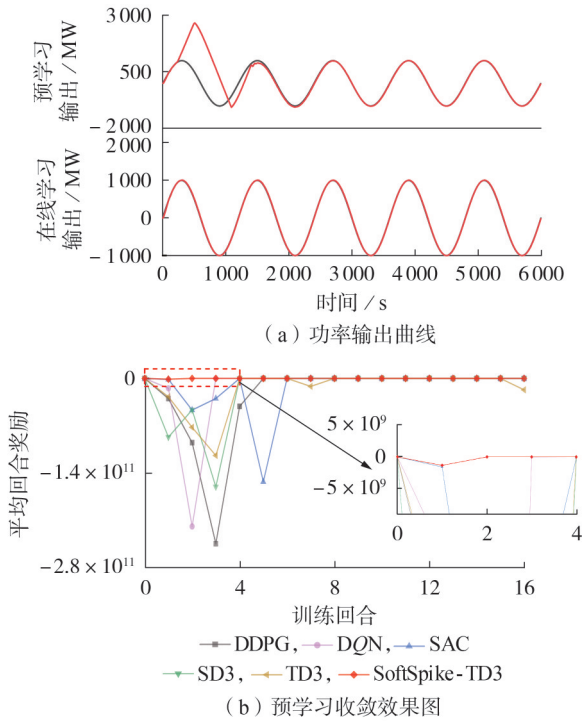

图2 预学习性能指标

Fig.2 Pre-learning performance indicators

习曲线可知，SoftSpike-TD3约在1 700 s收敛于最优策略。将这段时间内的最优策略应用于在线运行阶段，由2(a)在线运行曲线可知，SoftSpike-TD3从0 s开始就能完全跟踪正弦负荷扰动。由图2(b)可知，SoftSpike-TD3在训练过程中展现出显著的收敛特性，其奖励值在2个训练周期后即达到稳定状态，这一收敛速度明显优于其他算法。各算法的CPS1平均值、ACE平均值和频率偏差如附录D图D1所示。在频率控制精度方面，SoftSpike-TD3的最大频率偏差仅为0.01 Hz，处于预设的±0.2 Hz阈值内，特别是在1 000 s后的稳态运行阶段，其频率波动范围控制在±0.003 Hz以内，优于其他算法24.4%~85.5%。在ACE指标方面，SoftSpike-TD3将平均值稳定维持在±2 MW范围内，优于其他算法32.5%~87.2%。在CPS1指标方面，SoftSpike-TD3的平均值波动范围明显小于其他算法，该算法展现出良好的控制稳定性，且其CPS2值为100%，符合标准。

4.1.2 阶跃扰动

为验证算法在AGC动态过程中的控制性能，本文通过引入阶跃负荷扰动模拟实际运行中的负荷突变场景。对DDPG、DQN、SAC、SD3、TD3、SoftSpike-TD3等算法进行对比分析，各算法均采用统一的奖励函数进行训练，实验结果以区域A为例进行说明。

阶跃扰动性能指标如图3所示。由图3(a)可知，在负荷突变工况下，相较于其他算法，SoftSpike-TD3不仅能够将ACE绝对值维持在较低水平，且在负荷稳定后ACE能快速收敛至零值附近，优于其他算法7.5%~99.5%。由图3(b)可知，SoftSpike-TD3的CPS1值稳定趋近于200%的理想值，比其他算法更加稳定且CPS2符合标准。由图3(c)可知，传统DDPG算法因动作Actor函数的高估问题而导致频率波动较大，而SD3通过构建softmax算子方法有效解决了该问题，使频率稳定性提升24.4%~99.1%。综合各项指标分析可知，SoftSpike-TD3在负荷突变场景下表现出最优的综合控制性能，其次为SD3。

4.1.3 随机方波扰动

为验证SoftSpike-TD3在复杂负荷波动场景下的

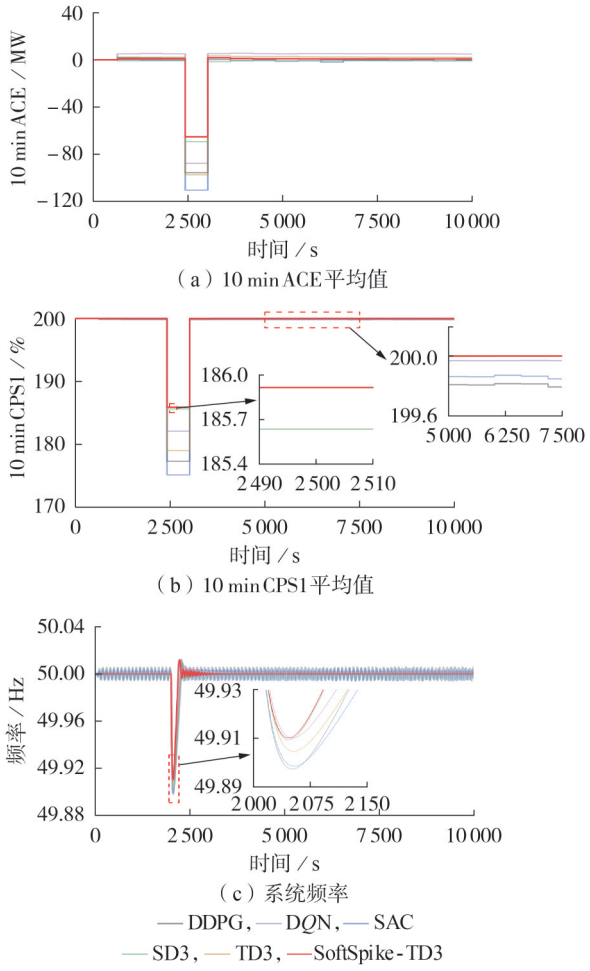

图 3　阶跃扰动性能指标

Fig.3　Step disturbance performance indicators

控制性能，本文引入随机方波信号模拟系统负荷的连续突变情况，设置 30 000 s 的仿真时长进行实时性能测试。图 4 展示了区域 A 在不同算法下的功率输出曲线，其中 SoftSpike-TD3 表现出显著的稳定性优势，其输出曲线在负荷突变时呈现最小的超调量。

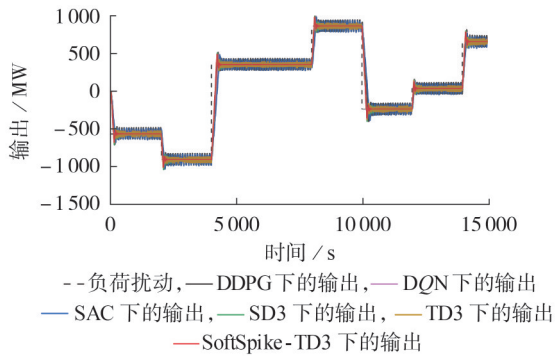

图 4　随机方波扰动下的机组出力曲线

Fig.4　Unit output curves under random square wave disturbance

进一步通过图 5 中的量化指标对比分析，可以更直观地评估各算法的控制性能。实验选取了 4 个关键性能指标，即 $|\Delta f|$、ACE 绝对值以及 CPS1 和 CPS2 值（图中仅展示了 $|\Delta f|$、ACE 绝对值）。与其他算法相比，SoftSpike-TD3 能够使 ACE 绝对值降低 6.1 %～53.9 %，使 $|\Delta f|$ 减少 12.2 %～55.0 %，同时维持更高的 CPS1 和 CPS2 指标值。

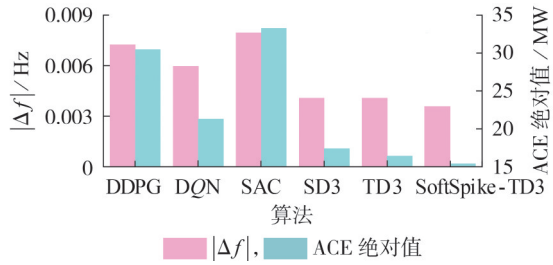

图 5　随机方波扰动下各算法的控制性能

Fig.5　Control performance of each algorithm under random square wave disturbance

4.2　风光水火储一体化多区域 AGC 模型

为充分考虑多种电源的大规模协调运行，本文构建一个风光水火储一体化的三区域 AGC 模型，其发电单元由传统火电、水电、风电、光伏、飞轮蓄能和电动汽车储能组成，利用飞轮蓄能和电动汽车储能的瞬时响应能力，可以有效参与电网调频，如附录 C 图 C2 所示。系统中光伏和风电的功率输出曲线如附录 D 图 D2 所示。其中传统发电机组与新能源机组的相关运行数据均来源于西南电网的实际运行数据。

4.2.1　变幅值正弦扰动

本文引入时长为 7 540 s 的变幅值正弦负荷扰动。6 种算法的 ACE 和频率响应特性如附录 D 图 D3 所示。以重庆地区为例，相较于其他算法，SoftSpike-TD3 的 $|\Delta f|$ 提升了 42.4 %～95.4 %，ACE 实现了 42.1 %～94.8 % 的降幅，CPS1 优于其他算法 27.7 %～96.6 %，同时 CPS2 符合标准。

4.2.2　白噪声扰动

为模拟电力系统实际运行中持续存在的随机负荷波动，本文采用白噪声信号作为扰动输入，构建 30 000 s 的仿真时长，以全面评估 SoftSpike-TD3 的控制性能。以重庆地区为例，SoftSpike-TD3 的实际负荷跟踪性能如附录 D 图 D4 所示。由图可知，曲线具有很好的平滑度。白噪声扰动系统频率的响应曲线如附录 D 图 D5 所示。相较于其他算法，SoftSpike-TD3 能够使 $|\Delta f|$ 减少 39.6 %～94.7 %。图 6 为 3 个区域 ACE 性能指标。以重庆地区为例，相较于其他算法，所提算法能够使 ACE 绝对值降低 56.9 %～95.0 %，使 CPS1 优于其他算法 22.2 %～93.4 %，使 CPS2 的值均为 100 %，符合标准。

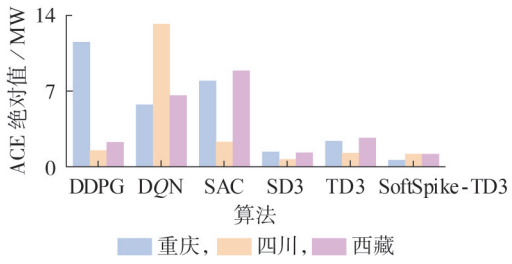

图 6 白噪声扰动下算法的控制性能

Fig.6 Control performance of algorithms under white noise

4.2.3 随机负荷扰动

引入无规则随机负荷扰动，以模拟真实电网环境，构建 24 h 实时仿真场景。图 7 为重庆地区 SoftSpike-TD3 的负荷跟踪曲线。由图可知，负荷扰动存在较大波动时，该算法仍能较好地跟随负荷变化。各算法在不同区域的 CPS 如附录 B 表 B3 所示。以重庆地区为例，相较于其他算法，SoftSpike-TD3 能够使 ACE 绝对值降低 23.9 %～86.6 %，使 $|\Delta f|$ 减少 21.2 %～86.8 %，使 CPS1 提高 0.6 %～12.5 %，且 CPS2 符合标准。

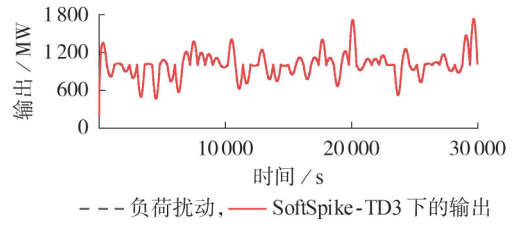

图 7 SoftSpike-TD3 控制器输出功率曲线

Fig.7 Output power curve of SoftSpike-TD3 controller

仿真结果表明，在各种不同的运行工况下，所提算法均能够满足 AGC 性能的要求。特别是在强随机扰动条件下，SoftSpike-TD3 展现出优异的频率稳定性和综合控制性能。

5 结论

为解决高比例新能源接入带来的随机扰动导致电力系统频率调节能力下降和控制性能恶化的问题，本文提出一种连续 RL-AGC 算法，即 SoftSpike-TD3。主要结论如下。

1）所提算法通过在连续控制 Actor-Critic 的值函数估计中引入玻尔兹曼 softmax 算子，有效地平衡了 Q 值高估和低估的问题，提高了系统的控制性能。

2）该算法采用 SNN 模拟生物神经元的时序活动，优化了 Q 值网络的更新速度，减轻了计算负担，进一步增强了 AGC 系统的稳定性。

3）通过对改进的 IEEE 标准两区域 AGC 模型以及基于西南电网的三区域 AGC 模型进行仿真，验证了所提算法的有效性。与其他算法相比，所提算法具有更优的 AGC 性能。

4）SoftSpike-TD3 在在线运行阶段的实时性满足 AGC 的秒级控制要求。

所提算法采用 softmax 算子实现了更准确的 Q 值估计，但在分布式 AGC 的增益参数共享方面具有更高的 Q 值要求。为此，下一步工作的重点将是研究增益参数共享机制，以实现更好的控制性能。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] 谢桦，胡一菡，张思宇，等. 考虑新能源分钟级功率波动的源网荷储协同频率优化运行方法[J]. 电力自动化设备，2025，45(6)：141-147.
XIE Hua，HU Yihan，ZHANG Siyu，et al. Source-grid-load-energy storage cooperative frequency optimization operation method considering minute-level power fluctuation of new energy[J]. Electric Power Automation Equipment，2025，45(6)：141-147.

[2] 龙传煜，张靖，何宇，等. 基于绿证-碳交易等价交互的综合能源系统低碳优化调度[J/OL]. 电力自动化设备.[2026-01-10]. https://doi.org/10.16081/j.epae.202512021.

[3] 张珺，时斌，崔晨雨，等. 分布式能源参与北美批发市场的聚合节点模型及对我国的启示[J]. 电力自动化设备，2026，46(2)：156-168.
ZHANG Jun，SHI Bin，CUI Chenyu，et al. Aggregated nodal model for distributed energy participating in North American wholesale market and implication for China[J]. Electric Power Automation Equipment，2026，46(2)：156-168.

[4] 余光正，崔长付，陈甜甜，等. 电-碳市场下可再生能源发电商参与多市场交易策略[J/OL]. 电力自动化设备.[2026-01-10]. https://doi.org/10.16081/j.epae.202512011.

[5] 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]. 电力自动化设备，2022，42(10)：92-99.
HUANG Shuaibo，CHEN Bei，GAO Jiangyu. Energy management strategy of electric vehicle charging station based on Markov decision process[J]. Electric Power Automation Equipment，2022，42(10)：92-99.

[6] TIAN M，LI X X，ZHU Z Y，et al. Robust voltage control for active distribution networks via safe deep reinforcement learning against state perturbations[J]. Protection and Control of Modern Power Systems，2026，11(1)：192-207.

[7] 梁涛，柴露露，谭建鑫，等. 基于深度强化学习算法的氢耦合电-热综合能源系统优化调度[J]. 电力自动化设备，2025，45(1)：59-66.
LIANG Tao，CHAI Lulu，TAN Jianxin，et al. Optimal scheduling of hydrogen coupled electrothermal integrated energy system based on deep reinforcement learning algorithm[J]. Electric Power Automation Equipment，2025，45(1)：59-66.

[8] HU Z J，MA R J. Adaptive event-triggered tracking control via switching functions[J]. Automatica，2026，185：112813.

[9] 余涛，周斌. 基于强化学习的互联电网CPS自校正控制[J]. 电力系统保护与控制，2009，37(10)：33-38.
YU Tao，ZHOU Bin. Reinforcement learning based CPS self-tuning control methodology for interconnected power systems[J]. Power System Protection and Control，2009，37(10)：33-38.

[10] 李红梅，严正. 具有先验知识的 Q 学习算法在AGC中的应用[J]. 电力系统自动化，2008，32(23)：36-40，99.
LI Hongmei，YAN Zheng. Application of Q-learning approach

with prior knowledge to non-linear AGC system[J]. Automation of Electric Power Systems,2008,32(23):36-40,99.

[11] 席磊,王昱昊,陈宋宋,等. 面向综合能源系统的多智能体协同AGC策略[J]. 电机与控制学报,2022,26(4):77-88.
XI Lei,WANG Yuhao,CHEN Songsong,et al. Multi-agent collaborative AGC strategy for integrated energy system[J]. Electric Machines and Control,2022,26(4):77-88.

[12] 席磊,柳浪,黄悦华,等. 基于虚拟狼群策略的分层分布式自动发电控制[J]. 电力系统自动化,2018,42(16):65-72.
XI Lei,LIU Lang,HUANG Yuehua,et al. Hierarchical and distributed control method for automatic generation control based on virtual wolf pack strategy[J]. Automation of Electric Power Systems,2018,42(16):65-72.

[13] 杨鹏,赵子珩,王中冠,等. 基于状态空间线性变换的主动配电网分布式电压控制[J]. 电力自动化设备,2023,43(1):64-72.
YANG Peng,ZHAO Ziheng,WANG Zhongguan,et al. Distributed voltage control of active distribution network based on state space linear transformation[J]. Electric Power Automation Equipment,2023,43(1):64-72.

[14] XI L,SHI Y,QUAN Y,et al. Research on the multi-area cooperative control method for novel power systems[J]. Energy,2024,313:133912.

[15] 席磊,李亚楠,刘治洪,等. 基于贝叶斯软演员评论家回溯损失的自动发电控制[J/OL]. 南方电网技术. [2025-06-04]. https://link.cnki.net/urlid/44.1643.TK.20250422.1759.034.

[16] 柳丹,任建宇,席磊,等. 基于高维协同软演员-评论家的多智能体自动发电控制[J]. 南方电网技术,2025,19(4):93-106.
LIU Dan,REN Jianyu,XI Lei,et al. Multi-agent automatic power generation control based on high dimensional collaborative soft actor-critic[J]. Southern Power System Technology,2025,19(4):93-106.

[17] 李嘉文,余涛,张孝顺,等. 基于改进深度确定性梯度算法的AGC发电功率指令分配方法[J]. 中国电机工程学报,2021,41(21):7198-7212.
LI Jiawen,YU Tao,ZHANG Xiaoshun,et al. AGC power generation command allocation method based on improved deep deterministic policy gradient algorithm[J]. Proceedings of the CSEE,2021,41(21):7198-7212.

[18] 李佳旭,吴俊勇,史法顺,等. 基于知识嵌入型深度强化学习的电力系统频率紧急控制方法[J]. 电力系统自动化,2026,50(1):97-107.
LI Jiaxu,WU Junyong,SHI Fashun,et al. Emergency frequency control method for power systems based on knowledge-embedded deep reinforcement learning[J]. Automation of Electric Power Systems,2026,50(1):97-107.

[19] 席磊,黄浩超,刘治洪,等. 基于深度确定性策略梯度算法的综合能源系统自动发电控制[J]. 高电压技术,2025,51(12):5941-5953.
XI Lei,HUANG Haochao,LIU Zhihong,et al. Automatic generation control of integrated energy system based on deep deterministic policy gradient algorithm[J]. High Voltage Engineering,2025,51(12):5941-5953.

[20] LIU C H,ZHAO Y B. Swap softmax twin delayed deep deterministic policy gradient[C]//2023 6th International Symposium on Autonomous Systems(ISAS). Nanjing,China:IEEE,2023:1-6.

[21] 刘晓德,郭宇飞,陈元培,等. 基于脉冲强化学习的连续运动控制仿真与优化[J]. 系统仿真学报,2025,37(10):2662-2671.
LIU Xiaode,GUO Yufei,CHEN Yuanpei,et al. Simulation and optimization of continuous motion control based on spiking reinforcement learning[J]. Journal of System Simulation,2025,37(10):2662-2671.

[22] 杨博,谢蕊,武少聪,等. 基于指数分布优化器的混合光伏-温差系统最大功率点跟踪[J]. 电力系统保护与控制,2024,52(16):12-25.
YANG Bo,XIE Rui,WU Shaocong,et al. Hybrid PV-TEG system maximum power point tracking based on an exponential distribution optimizer[J]. Power System Protection and Control,2024,52(16):12-25.

[23] 姚宇,陈武晖,张庚午,等. 直流电网稳定性分析的改进节点阻抗方法[J]. 电力系统保护与控制,2025,53(23):113-126.
YAO Yu,CHEN Wuhui,ZHANG Gengwu,et al. Enhanced nodal impedance method for stability assessment in DC power grid[J]. Power System Protection and Control,2025,53(23):113-126.

[24] RAMASWAMY A,HÜLLERMEIER E. Deep Q-learning:theoretical insights from an asymptotic analysis[J]. IEEE Transactions on Artificial Intelligence,2022,3(2):139-151.

[25] ESHRAGHIAN J K,WARD M,NEFTCI E O,et al. Training spiking neural networks using lessons from deep learning[J]. Proceedings of the IEEE,2023,111(9):1016-1054.

[26] 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.

[27] JALEELI N,VANSLYCK L S. NERC's new control performance standards[J]. IEEE Transactions on Power Systems,1999,14(3):1092-1099.

[28] 詹华,江昌旭,苏庆列. 基于分层强化学习的电动汽车充电引导方法[J]. 电力自动化设备,2022,42(10):264-272.
ZHAN Hua,JIANG Changxu,SU Qinglie. Electric vehicle charging navigation method based on hierarchical reinforcement learning[J]. Electric Power Automation Equipment,2022,42(10):264-272.

[29] 贺庆辰,秦斌. 基于改进SAC算法的城轨列车混合储能系统动态功率分配策略[J]. 湖南电力,2024,44(4):11-19.
HE Qingchen,QIN Bin. Dynamic power allocation strategy for hybrid energy storage system of urban rail trains based on improved SAC algorithm[J]. Hunan Electric Power,2024,44(4):11-19.

[30] MENG F P,WANG H R,LI R T,et al. Maritime long-distance communication system assisted by UAV-RIS based on AE-SD3 algorithm[C]//2024 4th International Conference on Electronic Information Engineering and Computer Communication(EIECC). Wuhan,China:IEEE,2025:422-425.

作者简介:

席　磊(1982—),男,教授,博士研究生导师,博士,主要研究方向为电力系统运行与控制、自动发电控制(**E-mail**:xilei2014@163.com);

赵俊苗(2002—),女,硕士研究生,主要研究方向为自动发电控制(**E-mail**:17530865652@163.com);

李宗泽(2001—),男,博士研究生,通信作者,研究方向为信息物理系统网络攻击与防御(**E-mail**:lizongze0608@163.com)。

(编辑　王锦秀)

Automatic generation control based on soft pulse twin delayed deep deterministic policy gradient algorithm

XI Lei[1,2], ZHAO Junmiao[1], HUANG Haochao[1], SHI Yu[1], WANG Wentao[1], LI Zongze[1]

(1. College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China;
2. Hubei Provincial Key Laboratory for Operation and Control of Cascaded Hydropower Station, China Three Gorges University, Yichang 443002, China)

Abstract: The reinforcement learning can effectively obtain the optimal solution of stochastic problem, but it has underestimation and overestimation biases when estimating the value function, seriously affecting the performance of automatic generation control. Therefore, an automatic generation control algorithm based on the soft pulse twin delayed deep deterministic policy gradient is proposed, which introduces Boltzmann softmax operator, and balances the overestimation and underestimation deviations of Q value by using its smoothing property and error control ability. In order to solve the increased computational burden problem brought by the introduction of softmax operator, a spiking neural network is integrated, and the computational burden is reduced by simulating the temporal activity of biological neurons. The effectiveness of the proposed algorithm is validated through the simulation on the modified IEEE standard two-area automatic generation control model and the integrated wind-solar-water-thermal-storage multi-area automatic generation control model based on the Southwest Power Grid. Compared with other algorithms, the proposed algorithm has lower frequency deviation and better control performance.

Key words: automatic generation control; reinforcement learning; spiking neural network; policy gradient; control performance