基于协同奖励函数多目标强化学习的智能频率控制策略研究

韩保军[1],高 强[2],代 飞[1],杨 宵[1],吕 颖[2],许忠义[3],付希越[4]

(1.国网河南省电力公司,河南 郑州 450052;2.中国电力科学研究院有限公司,北京 100192;3.国网浙江省电力有限公司松阳县供电公司,浙江 松阳 323400;4.武汉大学电气与自动化学院,湖北 武汉 430072)

摘 要:在含大规模风电并网系统的智能频率控制策略中,仅考虑CPS控制准则易造成频率短期集中越限,严重影响智能自动发电控制(AGC)策略的控制效果。提出一种基于协同奖励函数的多目标强化学习(TOPQ-MORL)智能频率控制策略,该策略构建了计及多维度频率控制性能评价标准的协同奖励函数,实现了多维度频率控制性能标准在时间尺度上的配合评价。采用TOPQ学习策略对智能体动作空间进行全局寻优,有效解决了传统贪婪策略下的Q函数线性加权多目标强化学习算法运算效率不佳的问题。标准两区域互联电网AGC控制模型仿真研究结果表明:所提智能AGC控制策略能有效改善频率控制性能,显著提高系统在全时间尺度上的频率质量。

关 键 词:风电并网;智能频率控制策略;多维度频率控制性能标准;TOPQ-MORL算法;协同奖励函数

DOI:10.19781/j.issn.1673-9140.2023.02.003 **中图分类号**:TM933 **文章编号**:1673-9140(2023)02-0018-12

Intelligent frequency control strategy based on multi-objective reinforcement learning of cooperative reward function

HAN Baojun[1], GAO Qiang[2], DAI fei[1], YANG Xiao[2], LÜ Ying[2], XU Zhongyi[3], FU Xiyue[4]

(1.State Grid Henan Electric Power Company, Zhengzhou 450052, China; 2.China Electric Power Research Institute, Beijing 100192, China; 3.Songyang Power Supply Company, State Grid Zhejiang Electric Power Co., Ltd., Songyang 323400, China; 4.School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China)

Abstract: In the intelligent frequency control strategy with large-scale wind power grid-connected system, only considering the CPS control criterion can easily cause the frequency off-limit in a short time, which seriously affects the control effect of the intelligent AGC control strategy. This paper proposes a multi-objective collaborative reward function reinforcement learning algorithm (TOPQ-MORL) intelligent frequency control strategy, which constructs a collaborative reward function that takes into account the multi-dimensional frequency control performance evaluation criteria, and realizes the coordinating evaluation of multi-dimensional frequency control performance standards on the time scale .The TOPQ learning strategy is used to optimize the action space of the agent globally, which effectively solves the problem of poor calculation efficiency of the Q function linear weighted multi-objective reinforcement learning algorithm under the traditional greedy strategy. The simulation results of the AGC control model of the standard two-region interconnected power grid shows that the intelligent AGC control strategy proposed in this paper can effectively improve the frequency

收稿日期:2022-03-11;**修回日期**:2022-09-20

基金项目:国家重点研发计划(2017YFB0902600)

通信作者:付希越(1999—),女,硕士研究生,主要从事电力系统运行与控制研究;E-mail:fuxiyue@whu.edu.cn

control performance and improve the frequency quality of the system on the full-time scale obviously.

Keywords: Grid-connected wind power; Intelligent frequency control strategy; Multi-dimensional frequency control performance standard; TOPQ-MORL algorithm; collaborative reward function

自动发电控制(automatic generation control, AGC)是实现电力系统实时有功－负荷供需平衡的重要手段，其中，AGC频率控制策略的优劣将直接决定AGC的频率控制效果[1]。目前，高比例可再生能源和高比例电力电子设备(即“双高”)正成为电力系统发展的重要趋势和关键特征，系统频率开始表现为长期频率稳定问题和短期频率稳定问题[2]。工程实际中所采用的阈值分区AGC控制策略[3]已经无法满足频率长、短期动态行为重大变化带来的日益复杂的互联电网频率控制需求[4-6]。

近年来，基于强化学习的智能频率控制策略由于其不依赖模型以及不需要精确历史训练样本和系统先验知识的特点得到大量关注[7-12]。文献[13]提出一种半监督群体预学习方法，解决了强化学习Q控制器在预学习试错阶段的系统镇定和快速收敛问题；文献[14]提出多步随机最优松弛算法，解决了非马尔科夫环境下火电机组长延时回滞问题，提高CPS性能的同时兼顾了系统的鲁棒性；文献[15]在传统Q学习选择动作机制中引入人工情感量化器，解决了传统Q学习不能输出连续动作的问题，提高了互联电网频率控制性能；文献[16]提出利用神经网络预测机制替换Q强化学习的Q表动作控制器，提高了控制器的收敛速度；文献[17]提出的多智能体深度强化学习算法能够实时优化动作集合，解决了神经网络训练效果不佳问题并有效提高了新能源利用率。事实上，上述文献通常是把电力系统频率控制器的智能学习问题转化为以CPS1为学习准则的单目标强化学习问题。但是，AGC本质是通过实时的频率控制动作来维持系统频率稳定，CPS1频率控制性能评价标准通过统计长期的历史频率数据来评估系统长期AGC频率控制效果，BAAL频率控制性能评价标准通过短期约束频差在任意30 min中波动的均值是否越限来评估系统短期AGC频率控制效果。现有文献仅用长期CPS1性能评价标准来指导实时AGC控制行为，在传统长期频率稳定问题占主导的电力系统中较为合适[18]。在长、短期频率稳定问题均凸显的“双高”电力系统中，仍然采用长期CPS学习准则来指导实时AGC控制行为，极易造成频率短期集中越限，严重影响智能AGC控制策略的实时控制效果[19]。

随着新能源并网以及智能电网的发展，电网频率控制评价标准正从单一尺度评价考核到多尺度的多维评价综合考核过渡[20]。文献[21-22]证明了引入BAAL与CPS1指标共同对系统频率进行多维度约束，能有效提高系统在不同尺度下的频差分布约束引导能力。文献[23]提出了一种多目标强化学习(multi objective reinforcement learning，MORL)频率控制策略。但是，该策略是以机组调节成本和碳排放量这两类经济性指标作为学习准则。目前，还没有基于BAAL与CPS1学习准则的智能AGC控制策略的相关报道。此外，目前多目标强化学习的多目标协调因子容易遗漏关键动作，造成智能体动作集探索不充分[24-26]。

针对上述问题，本文把电力系统频率控制器的智能学习问题转化为考虑长期CPS1指标和短期BAAL指标协同影响的多目标强化学习问题。采用长时间尺度CPS1指标和短时间尺度BAAL指标来协同引导频率控制器更加合理地评估和量化环境特征，在构建以CPS和BAAL指标协同奖励函数的基础上，提出了基于协同奖励函数多目标强化学习的智能频率控制策略。然后改进多目标强化学习的搜索策略，提出了一种基于TOPQ策略的多目标强化学习算法，用于智能频率控制器的动作集探索。

1　互联电网频率控制性能评价标准

1.1　CPS1频率控制性能评价标准

北美电力可靠性公司采用BAL(BAL-001)扰动控制系列指标来评价互联电网的频率控制质量，

其中又以CPS1(BAL-001-2:R1)指标在中国应用最为广泛,如下:

$$A_{\mathrm{VG},1,T}\left[\left(\frac{A^{m}_{\mathrm{CE},1\,\mathrm{min}}}{-10B_m}\cdot\Delta F_{1\mathrm{min}}\right)\right]\leqslant\varepsilon^2 \tag{1}$$

式中,$\Delta F_{1\,\mathrm{min}}$为控制区域频率偏移在1 min内的平均值;$A^{m}_{\mathrm{CE},1\,\mathrm{min}}$为控制区域$m$区域功率偏差在1 min内的平均值;$B_m$为区域$m$频率偏差系数,代表分配给区域$m$的频率调节责任;$A_{\mathrm{VG},1,T}(\cdot)$为求12个月的平均值;$\varepsilon$为区域$m$的频率偏移目标控制上限。

仅以实际频率高于计划频率为例,将式(1)展开有:

$$\frac{1}{T}\int_0^{\mathrm{T}}\frac{\Delta F}{\varepsilon}\cdot\left[\frac{\Delta P_{\mathrm{tie}}}{-10B_m\varepsilon}+\frac{\Delta F}{\varepsilon}\right]\mathrm{d}t\leqslant 1 \tag{2}$$

式中,T为整个时间周期,$\Delta F/\varepsilon$为本区域自身的频率偏差贡献度,$\Delta P_{\mathrm{tie}}/-10B_m\varepsilon$为其他区域对本区域的频率贡献度,$\Delta P_{\mathrm{tie}}/(-10B_m\varepsilon+\Delta F/\varepsilon)$为综合频率偏差贡献度。为分析方便,定义$(\Delta F/\varepsilon)\cdot(\Delta P_{\mathrm{tie}}/-10B_m\varepsilon+\Delta F/\varepsilon)$为综合频率偏差因子,用$\psi$表示。

CPS1指标松弛了每一时刻秒级和分钟级频率偏差的绝对控制,采用1 min内频率偏移的平均值作为统计基本单元,对评价区域T时段频差时间序列的滚动均方根进行统计评价。根据概率与统计理论,当T时段足够长的时候,式(1)等价于T时段系统频率偏差合格率大于99.99%。因此,CPS1是一个长时间尺度的反映互联电网频率质量的评价指标。

1.2 BAAL频率控制性能评价标准

近年来,随着大规模新能源并网,电网频率安全成为频率控制的重要目标,北美电力可靠性公司于2013年提出了BAAL(BAL-001-2:R2)评价指标,并于2016年开始实施,即

$$T\left[A^{m}_{\mathrm{CE},1\,\mathrm{min}}\geqslant-10B_m\frac{(F_{\mathrm{FIL-high}}-F_{\mathrm{S}})^2}{(F_{\mathrm{A}}-F_{\mathrm{S}})_{1\,\mathrm{min}}}\right]\leqslant T_{\mathrm{v}} \tag{3}$$

$$T\left[A^{m}_{\mathrm{CE},1\,\mathrm{min}}\leqslant 10B_m\frac{(F_{\mathrm{FIL-low}}-F_{\mathrm{S}})^2}{(F_{\mathrm{A}}-F_{\mathrm{S}})_{1\,\mathrm{min}}}\right]\leqslant T_{\mathrm{v}} \tag{4}$$

式中,F_{A}为实际频率值;F_{S}为计划频率值;$F_{\mathrm{FIL\text{-}high}}/F_{\mathrm{FIL\text{-}low}}$为高/低频率触发限制;$T_{\mathrm{v}}$为规定的允许连续越限时长。$T[\cdot]$为持续越限时间。

仅以实际频率高于计划频率为例,式(3)同理可变形为

$$T\left[\frac{1}{T''}\int_{T'}^{T'+T''}\frac{\Delta F}{\varepsilon}\cdot\left[\frac{\Delta P_{\mathrm{tie}}}{-10B_m\varepsilon}+\frac{\Delta F}{\varepsilon}\right]\mathrm{d}t\geqslant 1\right]\leqslant T_{\mathrm{v}} \tag{5}$$

式中,T'为任意开始越限的时间节点;T''为持续越限时间;T_{v}为规定的允许连续越限时长,本文规定连续越限时长为5 min。

BAAL指标主要是通过T_{v}来松弛互联电网区域间空间支援的绝对限制。因此,BAAL指标是一个短时间尺度下的反映互联电网频率控制安全的评价指标。

1.3 BAAL与CPS1标准协调的频率控制性能分析

与传统频率控制性能考核曲线[27]相比,BAAL指标对区域控制偏差(area control error,ACE)的约束会随着允许连续越限时间的变化而动态变化,因此同时计及CPS1和BAAL指标时的考核曲线需要增加时间维度的影响,其多维频率控制性能指标下的频率控制性能考核曲线如图1所示。

由图1可以看出,CPS1指标通过松弛频率偏差在时间尺度上的连续分布以加强对ACE幅值的限制能力。而BAAL指标则是通过牺牲对ACE幅值的限制能力来保证频率偏差在连续越限时间上的短时约束能力。因此,BAAL的作用是在CPS1标准保证系统长期频率质量的基础之上防止短期频率质量恶化。

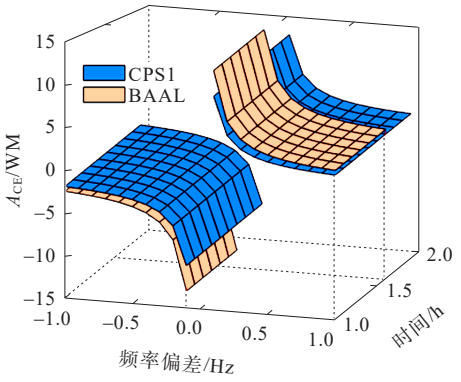
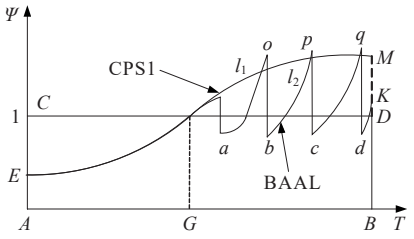


图1 BAAL和CPS1的随时间变化考核曲线

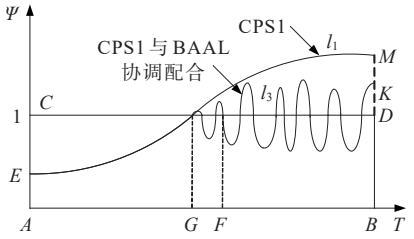
Figure 1 Time-varying assessment-curve of BAAL and CPS1

为进一步分析这两类指标的频率控制行为引导特点,图2给出了在不同性能指标引导下的综合频率偏差因子ψ的变化曲线。如图2所示,G点为频率越限临界点,在T时段内,l_1是仅满足CPS1指标

的φ曲线，l_2是仅满足BAAL指标的φ曲线，l_3是考虑CPS1与BAAL指标协调配合的φ曲线。基于文1.2、1.3可知，φ具有反映本区域频率变化和频率质量优劣的能力，也即若φ曲线在T时段内的面积大于区域S_{ABCD}（图中$ABCD$区域）会导致频率偏差越过规定的频率性能控制阈值ε。



（a）无协调配合时综合频率偏差因子分布曲线



（b）CPS1与BAAL指标协调配合下的综合频率偏差因子分布曲线

图2　综合频率偏差因子在时间上的分布曲线

Figure 2　The distribution curve of the integrated frequency deviation factor over time

由图2(a)可知，在仅考虑CPS1指标时，只要l_1曲线面积S_{AEMB}小于面积S_{ABCD}，系统仍然满足频率控制性能指标要求，但此时系统频率长期越限将导致系统电能质量降低，严重影响系统中各类设备的安全运行。在仅考虑BAAL指标时，只要l_2曲线面积S_{AEKB}小于面积S_{ABCD}，系统仍然满足频率控制性能指标要求，但此时系统频率为满足BAAL要求会频繁出现频率“垂直坠落”与“尖端振荡”现象，也即图2(a)中l_2曲线在a、b、c、d点出现频率陡降，在o、p、q点出现锯齿波峰。这一过程中频率可能瞬时下降到频率安全临界值以下，出现频率偏差严重越限情况。此外，系统频率偏差方向会出现频繁的瞬时反向情况，此时同步机组在短时间前后频繁接收相反的频率偏差信号会极具加大机组磨损，影响机组使用寿命。如图2(b)所示，l_3曲线在G点过后频率会出现频率持续越限趋势，但是短期频率持续越限后BAAL指标将占主导，频率变化将转入反向过程。值得注意的是，这一反向过程虽然由BAAL性能指标主导，但是由于CPS指标对频率长时间尺度统计具有滚动均方根的积分特性，会削弱“垂直坠落”与“尖端振荡”现象，从而让频率偏差变化变得平滑。

可见，计及CPS1和BAAL多维控制标准协同评价的本质即是：让两类指标在不同时间维度上各司其职，但是在评价效果上互相牵制。若两类控制评价标准紧密配合对系统频率进行约束，则既能保证系统长期的频率控制质量又能保障系统短期的频率控制安全。

2　计及多维控制标准协同评价的智能AGC控制策略

如图3所示，本文构建了基于多目标协同奖励函数强化学习频率控制策略的AGC控制模型。该模型主要包括系统调速器及发电机等效模块、系统频率偏差动态模型、智能频率控制器。其中R为区域m的等效机组调差系数；T_g为区域m的调速器时间常数；T_t为区域m的等效发电机时间常数；M为区域m的电力系统等值惯性系数；D为区域m的电力系统等值阻尼系数；ΔP_{tie}、ΔX_g、ΔP_g、ΔP_d、ΔP_Σ分别为区域m的联络线交换功率偏差、调节阀位置变化量、发电机输出功率变化量、负荷扰动变化量、机组总的调节指令。

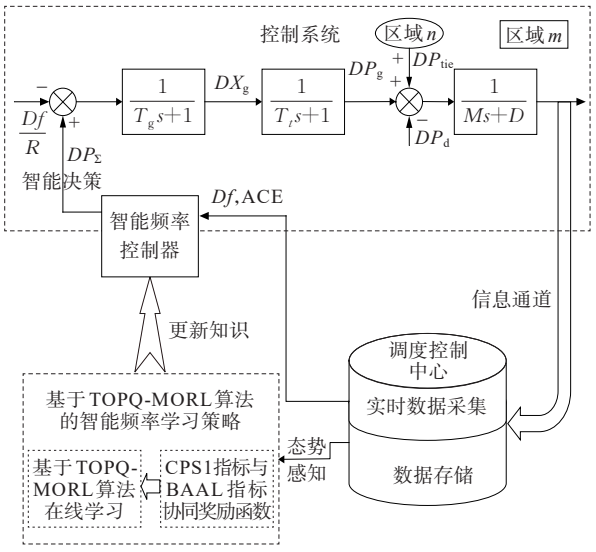


图3　多尺度标准协同评价的智能AGC控制策略

Figure 3　Intelligent AGC control strategy for collaborative evaluation of multi-scale standards

频率控制器智能学习阶段:强化学习是基于系统与环境反复交互作用的一种自学习框架,本文采用多目标协同奖励函数强化学习策略对智能频率控制器进行学习训练,该策略主要包括CPS1指标与BAAL指标协同奖励函数和基于TOPQ-MORL的智能频率控制学习算法两部分。

1) CPS1指标与BAAL指标协同奖励函数。MORL的基本思想是智能大脑与环境的反复交互作用,仅从所在环境中由自身经历产生反馈的奖励信息来学会执行任务并不断地进行自我改进[28]。因此,奖励函数是智能体与外部未知动态环境进行交互的唯一信号,表征智能体对于系统到达该状态的满意程度,一般都以需要被优化的目标作为奖励函数的设置。本文在多目标强化学习过程中引入CPS1指标和BAAL指标作为立即奖励函数来引导TOPQ-MORL学习,并用动态协调因子表征不同指标对环境状态变化的感知程度。

2) TOPQ-MORL[29]智能频率控制学习。对基于多维频率控制性能标准协同奖励函数的MORL学习,在状态发送变化后会给出这一状态变化过程中环境的实时反馈奖励(奖励也有正负之分),基于这一实时奖励来更新CPS1指标与BAAL指标两个目标各自的状态-动作集。然后采用不同策略(本文采用TOPQ)对更新的状态-动作集进行动作搜索,选择出新的动作,从而实现环境状态的进一步变化,周而复始迭代学习,最终动作即为同时满足CPS和BAAL指标环境反馈特征信息的动作。

频率控制器在线部署阶段:学习成熟的频率控制器在每个AGC控制周期中接收能量管理系统(energy management system,EMS)中的SCADA数据库实时采集频率偏差Δf、ACE、CPS、BAAL等数据,做出实时的频率控制动作。

2.1 CPS1指标与BAAL指标协同奖励函数

与传统Q学习不同,多目标Q学习针对多个目标构造出强化学习任务,然后同时对这些任务进行迭代优化学习。同时考虑BAAL指标和CPS1指标的即时奖励函数分别为

$$\begin{aligned} R_1(s,s',a)&=(A_{\mathrm{CE}}(t)-B_{\mathrm{AAL}}(t))^2\\ R_2(s,s',a)&=(C^*_{\mathrm{PS1}}-C_{\mathrm{PS1}}(t))^2\\ R_i(s,s',a)&\leftarrow\lambda_i R_i(s,s',a),i\in\{1,2\}\end{aligned}\tag{6}$$

其中,$R_i(s,s',a)$为第i个目标由状态s经过动作a转移到状态s'所获得的即时奖励值;$A_{\mathrm{CE}}(t)=\Delta P_{\mathrm{tie}}(t)+(-10B_m)\Delta F(t)$为当前时刻的区域控制偏差实时值,$\Delta P_{\mathrm{tie}}(t)$为$t$时刻的联络线公里偏差,$\Delta F(t)$为$t$时刻系统频率偏差;$s$表示$t$时刻系统状态,$s'$表示$t+1$时刻状态,$a$为系统从$s$到状态$s'$时的系统动作。$B_{\mathrm{AAL}}(t)$为$t$时刻的BAAL瞬时值,$C_{\mathrm{PS1}}(t)$为$t$时刻的CPS1瞬时值,$C^*_{\mathrm{PS1}}$为目标值,一般取200%,状态集划分见表1~3[16]。

表1 功率生成Q控制器的状态划分

Table 1 State set of power generation Q controller

状态	A_{CE}状态划分区间	状态	A_{CE}状态划分区间
1	$\in(-\infty,-100)$	7	$\in(20,40]$
2	$\in[-100,-80)$	8	$\in(40,60]$
3	$\in[-80,-60)$	9	$\in(60,80]$
4	$\in[-60,-40)$	10	$\in(80,100]$
5	$\in[-40,-20)$	11	$\in(100,\infty)$
6	$\in[-20,20]$		

表2 CPS1和BAAL的相关重要性程度

Table 2 Relative importance table of CPS1 and BAAL

重要性程度	$K_{1,2}$	$K_{2,1}$
同等重要	4	4
略微重要	4+1	4-1
重要	4+2	4-2
更重要	4+3	4-3
非常重要	4+4	4-4

表3 两区域互联系统模型参数

Table 3 System parameters for the two-area LFC model

T_g/s	T_t/s	T_p/s	R/(Hz/p.u.)	T_{12}/s	K_p/(Hz/p.u.)
0.08	0.3	20	2.4	0.545	120

λ_i为协同奖励函数的协调因子,协调因子又分为静态协调因子和动态协调因子。本文采用动态协调因子,也即λ_i随着每次状态转移过程而动态变化。这一动态变化过程反映了CPS1指标和BAAL

指标在不同时间维度统计特性的牵制作用，进一步体现了CPS1和BAAL指标对学习过程的协同引导。为此，本文采用考虑决策者偏好以及指标数据之间内在统计规律的组合赋权法[30]来确定动态协调因子的取值，根据组合权重值分别与主观赋权法和客观赋权法所求得的权重值之间的偏差应尽可能小的思想，建立组合权值的优化模型[31]，并用乘法加权求取最终值[32]。

首先，系统由当前ACE实时值对多维控制性能标准的达标情况进行判断（具体如图4所示），并确定当前状态下系统对2种指标的重要性偏好程度（具体过程见表2）。

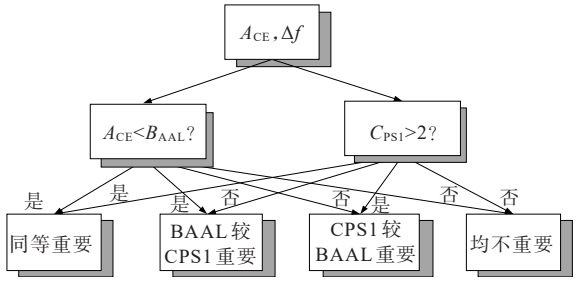


图4　CPS1和BAAL的达标情况评估

Figure 4　The assessment diagram of CPS1 and BAAL

此时，可得每个动作周期下各目标的权重因子：

$$w_i=\frac{K_{j,i}}{K_{j,i}+K_{i,j}},i\neq j \tag{7}$$

为消除主观因素，采用熵值法计算两种指标之间的差异系数β_i：

$$\beta_i=\frac{1+\ln^{-1}(N)\sum_{y=1}^{K}P_{y,i}\ln(P_{y,i})}{\sum_{i=1}^{N}\left(1+\ln^{-1}(N)\sum_{y=1}^{K}P_{y,i}\ln(P_{y,i})\right)} \tag{8}$$

$$P_{y,i}=x_{y,i}/\sum_{y=1}^{K}x_{y,i} \tag{9}$$

式(8)、(9)中，$x_{y,i}$为第i个频率控制性能评价指标在第y个时刻下标准化后的指标值；K为从0到当前时刻t下的第i个频率控制性能评价指标个数；N为目标个数；$P_{y,i}$为计算第i个频率控制性能评价指标在第y个时刻下的指标值占从0到t时刻该指标总数的比重。

最后采用乘法加权确定最终的协调因子，因此，综合式(7)、(8)可得协调因子：

$$\lambda_i=\frac{\sqrt{w_i\beta_i}}{\sum_{i=1}^{N}\sqrt{w_i\beta_i}} \tag{10}$$

2.2　TOPQ-MORL智能频率控制学习算法

多目标强化学习算法（MORL）与传统Q学习算法不同，MORL是针对多个目标状态—动作价值Q矩阵同时进行的迭代优化，各目标均存在与Q学习相对应的以Q表形式存在的状态—动作价值函数$Q_i(s,a)$，其更新公式与传统Q学习的状态—动作价值函数更新相同：

$$Q_i(s,a)=Q_i(s,a)+\alpha\left[R_i(s,s',a)+\gamma\max_{a\in A}Q_i(s',a)-Q_i(s,a)\right] \tag{11}$$

式中，$\alpha(0<\alpha<1)$为学习率，较大的学习率会加快收敛速度，但是失去了较好的搜索空间，为提高Q学习收敛的稳定性，本文取0.01；γ为折扣系数，本文取0.9；$Q_i(s,a)$表示第i个目标状态s下选择动作a的Q值，Q表的大小为$S*A$，初值Q表一般设为0矩阵。

为方便选取满足各目标下的最优动作，本文用$\boldsymbol{Q}_{\mathrm{M}}(s,a)$向量表示$N$个目标分别在状态$s$下选择动作$a$的状态—动作价值函数：

$$\boldsymbol{Q}_{\mathrm{M}}(s,a)=[Q_1(s,a),Q_2(s,a),\cdots,Q_N(s,a)] \tag{12}$$

传统MORL算法动作选择机制为当前状态下总是选择向量$\boldsymbol{Q}_{\mathrm{M}}(s,a)$中最大目标值对应的动作，将其定义为$\pi^*_{Q_{\mathrm{M}}}$，该动作即为当前状态下各个目标的最优动作策略：

$$\pi^*_{Q_{\mathrm{M}}}=\arg\max_a\{\max_i\boldsymbol{Q}_{\mathrm{M}}(s,a)\} \tag{13}$$

但是，上述最优动作选择策略不能保证智能体充分探索整个状态—动作空间，容易陷入局部最优解，导致智能体寻优精度欠佳，因此，本文采用TOPQ策略对$\boldsymbol{Q}_{\mathrm{M}}(s,a)$向量空间进行全局寻优。

TOPQ策略是一种全局最优策略，首先，它通过对各目标当前状态下的Q值进行探索，分别获得各目标当前状态下的最大Q值，用$\boldsymbol{W}_i(s)$表示：

$$\boldsymbol{W}_i(s)=\max_a Q_i(s,a),1\leqslant i\leqslant N \tag{14}$$

然后，智能体在$\boldsymbol{W}_i(s)$的集合中筛选出当前状态下的最大目标Q值，如下所示，并将其定义为

$W_{\max}(s)$：

$$W_{\max}(s)=\max_i \boldsymbol{W}_i(s)=$$

$$\max_i\left\{\max_i Q_i(s,a)\right\},1\leqslant i\leqslant N \tag{15}$$

最后，智能体即可利用筛选得到的最大目标Q值 $W_{\max}(s)$找到最优动作，并对动作空间做出智能决策，确保所选动作为最大Q值对应目标下的全局最优解，此时最优动作为

$$a^*=\arg\max_a W_{\max}(s),1\leqslant i\leqslant N \tag{16}$$

本文基于多目标强化学习提出了计及多维控制标准协同评价的多目标智能频率控制TOPQ-MORL算法，具体算法框架如图5所示。

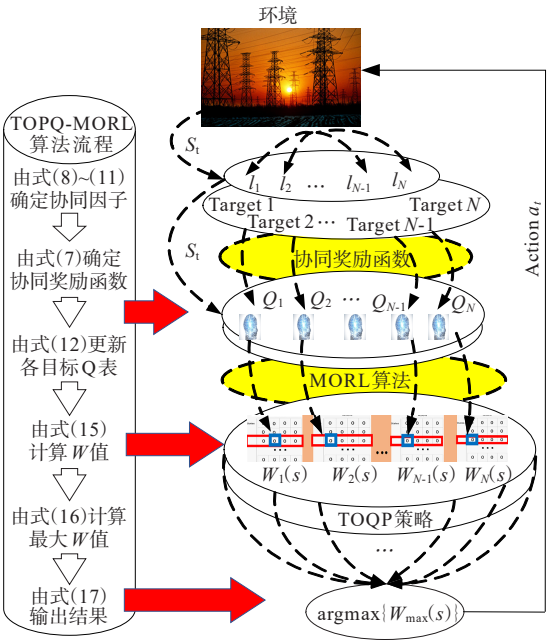


图5 TOPQ-MORL算法框架

Figure 5 The framework of TOPQ-MORL algorithm

3 仿真算例研究

本文算法按以下方式产生样本：在Matlab/Simulink下搭建了典型两区域互联电网AGC负荷频率控制模型[33]（使用Matlab 2018b/Simulink环境运行），该模型系统2个区域的参数设置相同，具体值见表3，仿真模型如图5所示。系统基准容量为1 000 MW，TOPQ-MORL频率控制算法采用Python语言编写，通过S-function模块实现Simulink下的仿真。在预学习阶段，在该模型中对A区域施加周期为1 200 s，幅值为100 MW，时长20 000 s的风电波动来产生训练样本。在实际应用时，可对历史频率跌落事件进行学习，并可在实时运行中补充样本数据学习。在互联电网运行中，需要控制性能评价标准来评价和规范各控制区域的行为[34]。针对单一CPS1目标下所提算法的预学习过程如图6所示。在预学习阶段，对A区域施加周期为1 200 s，幅值为100 MW，时长20 000 s的风电波动，通过使用一个2范数的Q函数矩阵 $\left\|Q_t(s,a)-Q_{t-1}(s,a)\right\|^2\leqslant\zeta$（$\zeta$为一常量）作为预学习达到最优策略的标准[35]。

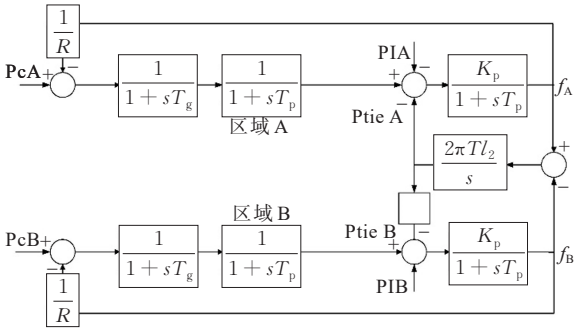
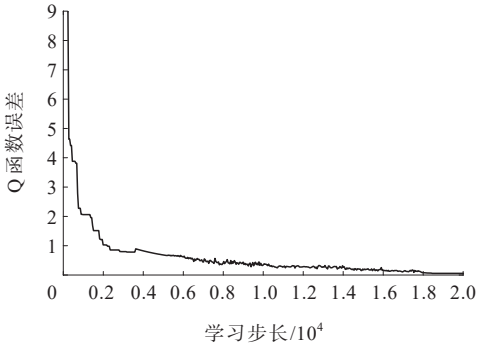
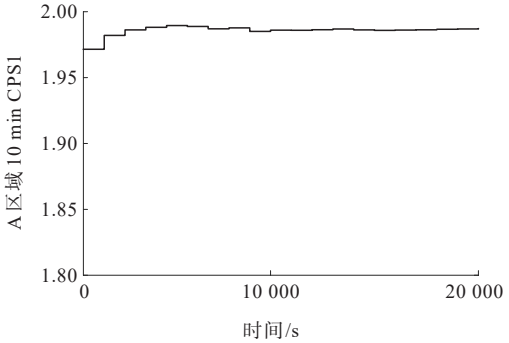


图5 典型两区域仿真模型



（a）CPS1目标下Q函数差分收敛结果



（b）$C_{\text{avg-10-min}}$曲线

图6 预学习过程

Figure 6 Pre-learning process of the TOPQ-MORL algorithm

本文的训练样本规模为 20 000 个样本。人工智能本身是无模型控制，在已经训练好的 AGC 模型中，只需要输入相应的联络线功率以及频率偏差数据，就可以得到相应的输出指标，并不依赖于实际目标电网，所以不需要目标电网系统以及其他区域系统的参数。

本文所搭建的典型两区域互联电网 AGC 负荷频率控制模型，已在两个区域中分别将所有同步发电机等值为一台同步发电机。目前输出是作用在等值发电机上，在实际应用中，可根据机组优先级、分配控制参数等因素，分配到各常规电厂和新能源电站。

为了验证本文所提控制策略的控制性能，本文设置了以下 4 类控制算法。

算法 1：传统的基于 CPS1 频率控制性能评价指标的单目标强化学习智能频率控制算法（CPS1-MORL）。

算法 2：基于多维频率控制性能评价指标协同多目标 Q 函数的传统贪婪策略多目标强化学习智能频率控制算法（CoordinateQ-MORL）。

算法 3：基于多维频率控制性能评价指标协同奖励函数的传统贪婪策略多目标强化学习智能频率控制算法（Greedy-MORL）。

算法 4：本文所提的基于多维频率控制性能评价指标协同奖励函数的改进 TOPQ 策略多目标强化学习智能频率控制算法（TOPQ-MORL）。

由图 6(a)可知，经过大约 20 000 次迭代后，Q 函数趋于稳定，这意味着已学习到了最佳 CPS1 策略。图 6(b)给出了 A 区域每 10 min CPS1 的平均值（$C_{\text{avg-10-min}}$）在预学习过程的变化曲线。可知，$C_{\text{avg-10-min}}$ 在最初学习过程中有一个向上的微小波动，随后几乎保持在一个稳定可接受的值，其值为 199.017%，这说明 TOPQ-MORL 算法已逼近最优 CPS1 控制策略。与此同时，目标 BAAL 对应 Q 矩阵也已收敛。

此外，从算法的学习时间角度分别对 4 种算法做了多次仿真并统计 4 种算法的平均计算时间，具体见表 4。可知，一方面，单目标 CPS1-RL 比多目标 TOPQ-MORL 计算时间会更短。因为 CPS1-MORL 只优化单一 CPS1 目标，而 TOPQ-MORL 需要同时优化 CPS1 以及 BAAL 两个冲突目标，并实时计算协调因子。另一方面，多目标强化学习采用不同动作搜索策略对学习时间也有影响。CoordinateQ-MORL 与 Greedy-MORL 相比，CoordinateQ-MORL 计算时间更短，因为 Weighted Sum Approach 不能充分探索动作集合，大大减少了智能体探索动作集合时间[23]；Greedy-MORL 与 TOPQ-MORL 相比，Greedy-MORL 计算时间同样会更短，因为 TOPQ 相比传统贪婪策略所经历的搜索步骤更多。

表 4　算法平均计算时间比较

Table 4　Simulation results under two different algorithms

算法	计算时间/s
CPS1-MORL	12 031
CoordinateQ-MORL	18 546
Greedy-MORL	20 015
TOPQ-MORL	21 457

上述算法的输出动作离散集 $A=\{-500,-300,-100,-50,-10,0,10,50,100,300,500\}$，共设置 11 个离散动作。学习步长一般为 5 s（AGC 控制周期）。考虑到大规模间歇性新能源接入电网后会显著增强系统随机性，因此引入随机扰动以模拟未知新能源大规模并网环境下，电力系统每一时刻均在随机变化的复杂工况，验证 TOPQ-MORL 能否适应不断变化的电网环境。因此，在 A 区域施加周期为 1 200 s、幅值为 100 MW 的风电随机扰动。表 5～7 中的 $|\Delta f|$，CPS1 为 20 min 的平均值，BAAL 为 20 min 内 ACE 小于 BAAL 指标的个数占比。

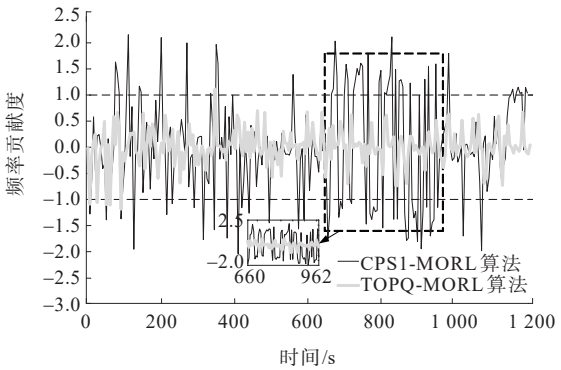
3.1　控制策略性能分析

图 7 给出了算法 1 和算法 4 的频率偏差自身贡献度（$\Delta f/\varepsilon$）与 CPS1 指标变化曲线。本文采用 3ε 阈值进行运算，其中 ε 取 0.01。频率贡献度具有反映采用不同算法频率质量的能力，若频率贡献度越过 ±1 说明此时的频率越过了规定的限值 3ε。

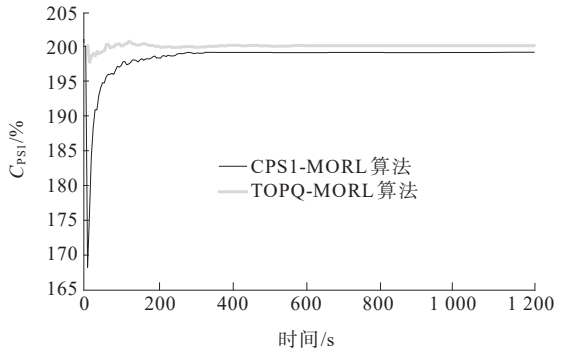
如图 7(a)所示，算法 4 在整个仿真周期下的频率贡献度曲线相较算法 1 而言能更好地被约束在规定的控制范围内。同时，在 660～962 s 时间段内算法 4 的频率贡献度均在规定的控制范围内，而算法 1 的频率贡献度在此区间段连续越限且出现陡降的现象，超过了本文规定的短期指标 BAAL 频率连续越限时长 5 min，会对系统运行安全造成较大影响。主要有 2 个方面原因。一是算法 4 通过实时松弛 2 种指标的权重大小对频率进行控制，在仿真周期内

若出现频率上下波动幅度较大或是出现“频率坠落”现象,将给短期频率控制性能指标BAAL赋予更大权重,致使系统频率在整个时间尺度下均在控制范围以内。若仿真周期内出现频率连续越限的情况,将给长期频率控制性能指标CPS1赋予较大权重。二是由于算法4同时考虑2种指标协同参与评价AGC控制,而算法1仅考虑了长期控制性能指标CPS的影响,忽略了短期频率严重越限问题,以至于算法4的整体频率控制效果要好于算法1。

如图7(b)所示,算法4的CPS1曲线在整个仿真周期内波动幅度较小,且稳定在200%水平。算法1的CPS1曲线在仿真初期波动幅度较大,其最小值低至168%且最终稳定在197%。可知,综合考虑多维度频率性能指标协同配合评价能有效改善系统CPS1控制性能指标。表5给出了算法1和算法4的频率控制性能指标均值。从表5可以看出,采用算法4相较算法1的频率偏差降低了55%,CPS1提高了3%,BAAL达标率也提高了12%。进一步证明了算法4相较算法1有更好的频率控制效果。



(a) 自身贡献度变化曲线



(b) CPS1变化曲线

图7 不同控制算法下频率偏差自身贡献度控制与CPS指标变化曲线

Figure 7 $\Delta f/\varepsilon$ and CPS1 figure of different control algorithms

表5 2种不同控制算法的频率控制指标均值

Table 5 The mean values of frequency control indexes of two different control algorithms

算法	$\lvert\Delta f\rvert$/Hz	CPS1指标/%	BAAL指标/%
CPS1-MORL	0.014 3	196	86.4
TOPQ-MORL	0.006 4	200	98.5

综上所述,引入短期频率控制性能该指标BAAL与CPS1指标协同对系统频率进行约束,能够有效提高系统全时间尺度下的频率质量。

3.2 协同奖励函数对频率控制性能的影响

为验证本文所提协同奖励函数的有效性,表6给出了算法2与算法3的控制性能指标。

表6 不同控制算法的控制效果

Table 6 Simulation results under four different algorithms

算法	$\lvert\Delta f\rvert$/Hz	CPS1指标/%	BAAL指标/%
CoordinateQ-MORL	0.013 2	197	96.2
Greedy-MORL	0.012 9	199	97.2

可以看出,算法3的各项控制性能指标相较算法2而言效果更优。这是因为在多目标状态-动作价值函数之间引入协调因子可能会导致智能体不能充分探索动作集合,一些动作在整个探索周期内都不会被选中,可能导致遗漏关键动作。而采用协同奖励函数可以有效解决上述问题,智能体能充分探索动作空间从而提高频率控制各项性能指标。

综上所述,引入协同奖励函数能够有效提高系统频率质量以及各项频率性能指标。

3.3 不同学习策略对控制性能影响

为验证本文提出的学习策略TOPQ的有效性,图8给出了算法3以及算法4的CPS1变化曲线。

由图8可知,在负荷扰动出现后算法4能使系统的CPS1迅速地回到接近200的水平,相较算法3有更快的收敛速度。同时,算法4下的CPS1值上下波动幅度相较算法3而言较小且最终稳定值要高于算法3。这是因为TOPQ策略从全局考虑对动作进行选择,有效改善了传统贪婪策略容易陷入局部最优解问题。表7给出了算法3和算法4的频率控制性能指标值。可以看出,采用TOPQ策略下的频率偏差绝对值的平均值相较Greedy策略而言降低了50%,同时多维度频率控制性能指标也有所提升。

有效证明了所提策略能够提高控制器的性能，从而使得到的结果更加趋近于全局最优。

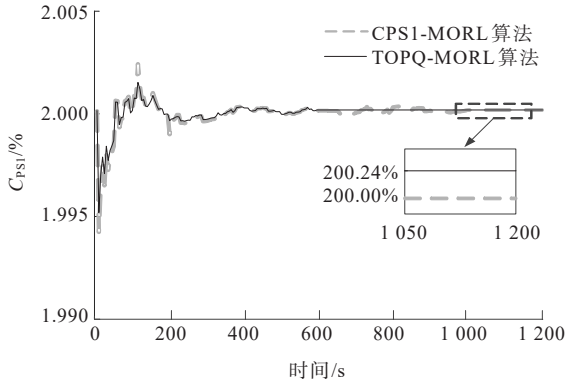


图 8　CPS1 变化曲线

Figure 8　The curve of CPS1

表 7　两种不同控制算法的控制效果

Table 7　Simulation results under two different algorithms

算法	$\lvert\Delta f\rvert$/Hz	CPS1 指标/%	BAAL 指标/%
Greedy-MORL	0.012 9	199	97.2
TOPQ-MORL	0.006 4	200	98.5

综上所述，全局搜索策略 TOPQ 比局部搜索策略 Greedy、CoordinateQ 耗时更多，但是搜索质量更高。因此 TOPQ 是一种以时间换空间的搜索策略，Greedy、CoordinateQ 是一种以空间换时间的搜索策略。

4　结语

本文提出一种多维度评价标准协同评价的基于协同奖励函数多目标强化学习的智能频率控制策略。

仿真结果表明：① TOPQ-MORL 算法相较 CPS1-MORL 算法能够有效提高系统频率在全时间尺度上的质量；②本文所提协同奖励函数能够改善传统多目标 Q 函数线性加权的不足，能有效提高控制器控制性能；③TOPQ 学习策略相较 Greedy 学习策略而言能更好地探索全局动作，从而提高频率各项控制性能指标。总体来看，本文提出的基于 CPS1 和 BAAL 学习准则协同的智能 AGC 控制策略能够有效应对风电等新能源并网时带来的短时功率扰动问题，能有效解决多维度频率控制性能指标在时间尺度上的矛盾，提高了系统稳定性。

本文尽力解决多维度的频率控制性能在时间尺度上的矛盾，因此时间尺度拖延较长，可能存在经济性问题，但在文中暂时没有考虑。在进一步的工作中，可以在回报函数中加入经济性指标进行多目标优化，以求对经济性问题加以改善。

参考文献：

[1] 胡泽春，罗浩成．大规模可再生能源接入背景下自动发电控制研究现状与展望[J]．电力系统自动化，2018，42(8)：2-15.
HU Zechun，LUO Haocheng．Research status and prospect of automatic generation control with integration of large-scale renewable energy[J]．Automation of Electric Power Systems，2018，42(8)：2-15.

[2] 谢小荣，贺静波，毛航银，等．“双高”电力系统稳定性的新问题及分类探讨[J]．中国电机工程学报，2021，41(2)：461-475.
XIE Xiaorong，HE Jingbo，MAO Hangyin，et al．New issues and classification of power system stability with high shares of renewables and power electronics[J]．Proceedings of the CSEE，2021，41(2)：461-475.

[3] 王念，张靖，李博文，等．自动发电控制研究综述[J]．电测与仪表，2020，57(21)：1-8.
WANG Nian，ZHANG Jing，LI Bowen，et al．Research review of automatic generation control[J]．Electrical Measurement & Instrumentation，2020，57(21).

[4] 徐艳春，蒋伟俊，孙思涵，等．含高渗透率风电的配电网暂态电压量化评估方法[J]．中国电力，2022，55(7)：152-162.
XU Yanchun，JIANG Weijun，SUN Sihan，et al．Quantitative assessment method for transient voltage of distribution network with high-penetration wind power[J]．Electric Power，2022，55(7)：152-162.

[5] 杨建宾，谢丽蓉，宋新甫，等．基于可再生能源的碳捕集—电转气协同运行方法[J]．智慧电力，2022，50(12)：70-78.
YANG Jianbin，XIE Lirong，SONG Xinfu，et al．Collaborative operation method of carbon capture-P2G based on renewable energy[J]．Smart Power，2022，50(12)：70-78.

[6] 阮正鑫，张逸，张嫣，等．高比例光伏与配电网超高次谐波交互影响研究[J]．电力工程技术，2021，40(2)：18-25.

RUAN Zhengxin, ZHANG Yi, ZHANG Yan, et al. Interaction of high proportion photovoltaic and supraharmonic in distribution network[J]. Electric Power Engineering Technology, 2021, 40(2): 18-25

[7] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning,1992,8(3-4):279-292.

[8] 谢庆,张煊宇,王春鑫,等.新一代人工智能技术在输变电设备状态评估中的应用现状及展望[J].高压电器,2022,58(11):1-16.
XIE Qing, ZHANG Xuanyu, WANG Chunxin, et al. Application status and prospect of the new generation artificial intelligence technology in the state evaluation of power transmission and transformation equipment[J]. High Voltage Apparatus, 2022,58(11):1-16.

[9] 程乐峰,余涛,张孝顺,等.机器学习在能源与电力系统领域的应用和展望[J].电力系统自动化,2019,43(1):15-31.
CHENG Lefeng, YU Tao, ZHANG Xiaoshun, et al. Machine learning for energy and electric power systems: state of the art and prospects[J]. Automation of Electric Power Systems,2019,43(1):15-31.

[10] 张廷锋,陶熠昆,何凛,等.基于遗传算法的电力巡检机器人作业调度优化方法[J].电网与清洁能源,2022,38(3):68-73.
ZHANG Tingfeng, TAO Yikun, HE Lin, et al. A genetic algorithm-based optimization method for job scheduling of electric power inspection robots[J]. Power System and Clean Energy, 2022, 38(3):68-73.

[11] 梁煜东,陈峦,张国洲,等.基于深度强化学习的多能互补发电系统负荷频率控制策略[J].电工技术学报,2022,37(7):1768-1779.
LIANG Yudong, CHEN Luan, ZHANG Guozhou, et al. Load frequency control strategy of hybrid power generation system: a deep reinforcement learning-based approach[J]. Transactions of China Electrotechnical Society, 2022, 37(7):1768-1779.

[12] 杨丽,孙元章,徐箭,等.基于在线强化学习的风电系统自适应负荷频率控制[J].电力系统自动化,2020,44(12):74-83.
YANG Li, SUN Yuanzhang, XU Jian, et al. Adaptive load frequency control of wind power system based on online reinforcement learning[J]. Automation of Electric Power Systems,2020,44(12):74-83.

[13] 余涛,周斌,陈家荣.基于Q学习的互联电网动态最优CPS控制[J].中国电机工程学报,2009,29(19):13-19.
YU Tao, ZHOU Bin, CHAN Kawing. Q learning based optimal dynamic optimal CPS control methodology for interconnected power systems[J]. Proceedings of the CSEE, 2009,29(19):13-19.

[14] YU T, ZHOU B, CHAN K W, et al. Stochastic optimal relaxed automatic generation control in non-markov environment based on multi-step $Q(\lambda)$ learning[J]. IEEE Transactions on Power Systems,2011,26(3):1272-1282.

[15] YIN L F, YU T, ZHOU L, et al. Artificial emotional reinforcement learning for automatic generation control of large-scale interconnected power grids[J]. IET Generation, Transmission & Distribution,2017,11(9):2305-2313.

[16] 殷林飞,余涛.基于深度Q学习的强鲁棒性智能发电控制器设计[J].电力自动化设备,2018,38(5):12-19.
YIN Linfei, YU Tao. Design of strong robust smart generation controller based on deep Q learning[J]. Electric Power Automation Equipment,2018,38(5):12-19.

[17] 席磊,余璐,付一木,等.基于探索感知思维深度强化学习的自动发电控制[J].中国电机工程学报,2019,39(14):4150-4162.
XI Lei, YU Lu, FU Yimu, et al. Automatic power generation control based on deep reinforcement learning with exploration awareness[J]. Proceedings of the CSEE, 2019, 39(14):4150-4162.

[18] 黄超,卜思齐,陈麒宇,等.元电力:新一代智能电网[J].发电技术,2022,43(2):287-304.
HUANG Chao, BU Siqi, CHEN Qiyu, et al. Meta-power: next-generation smart grid[J]. Power Generation Technology, 2022, 43(2):287-304.

[19] WANG C X, MCCALLEY J D. Impact of wind power on control performance standards[J]. International Journal of Electrical Power & Energy Systems,2013,47:225-234.

[20] NERC. BAL-001-2-real power balancing control performance standard background document[EB/OL]. North America:NERC,2013[2015-02-01].http://www.nerc.com/.

[21] 谈超,戴则梅,滕贤亮,等.北美频率控制性能标准发展分析及其对中国的启示[J].电力系统自动化,2015,39(18):1-7.
TAN Chao, DAI Zemei, TENG Xianliang, et al. Development of frequency control performance standard in North America and its enlightenment to China[J]. Automation of Electric Power Systems,2015,39(18):1-7.

[22] 常烨骁,刘娆,巴宇,等.平衡监管区区域控制偏差限制标准剖析[J].电网技术,2016,40(1):256-262.
CHANG Yekui, LIU Rao, BA Yu, et al. Analysis of balancing authority ACE limit standard of North America [J].Power System Technology,2016,40(1):256-262.

[23] WANG H Z, LEI Z X, ZHANG X, et al. Multiobjective reinforcement learning-based intelligent approach for optimization of activation rules in automatic generation control[J].IEEE Access,2019,7:17480-17492.

[24] VAMPLEW P,YEARWOOD J,DAZELEY R,et al.On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts[C]//Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence. New York:ACM,2008.

[25] LI J W,YU T,ZHANG X S.Coordinated load frequency control of multi-area integrated energy system using multi-agent deep reinforcement learning[J]. Applied Energy,2022,306:117900.

[26] 李瑞群,王若冰,田涛,等.多智能体同时到达多目标点的协同强化学习算法[J].计算机应用与软件,2021,38(9):199-204.
LI Ruiqun,WANG Ruobing,TIAN Tao,et al.Collaborative reinforcement learning algorithm of multi-agent achieving simultaneous multi-objectives[J]. Computer Applications and Software,2021,38(9):199-204.

[27] 部俊锋,李昌卫,刘浩.二次再热机组一次调频能力探讨[J].山东电力技术,2021,48(12):68-71.
BU Junfeng, LI Changwei, LIU Hao. Discussion on the primary frequency control performance of double reheat unit[J].Shandong Electric Power,2021,48(12):68-71.

[28] 赵知劲,朱家晟,叶学义,等.基于多智能体模糊深度强化学习的跳频组网智能抗干扰决策算法[J].电子与信息学报,2021,43:2-9.
ZHAO Zhijin, ZHU Jiasheng, YE Xueyi, et al. Intelligent anti-jamming decision algorithm for frequency hopping network based on multi-agent fuzzy deep reinforcemnet learning[J]. Journal of Electronics & Information Technology,2022,43:2-9.

[29] LIU C M, XU X, HU D W. Multiobjective reinforcement learning: a comprehensive overview[J]. IEEE Transactions on Systems, Man and Cybernetics: Systems, 2015, 45(3): 385-398.

[30] 姜媛媛,张振振,薛生,等.改进组合赋权法的配电网隐患评估[J].科学技术与工程,2020,20(22):9030-9035.
JIANG Yuanyuan, ZHANG Zhenzhen, XUE Sheng, et al. Evaluation of distribution network hidden dangers by improved combination weighting method[J]. Science Technology and Engineering,2020,20(22):9030-9035.

[31] 贺春光,檀晓林,周兴华,等.基于博弈论组合赋权的智能配电网项目投资效益评价[J].电力科学与技术学报,2022,37(1):161-167.
HE Chunguang, TAN Xiaolin, ZHOU Xinghua, et al. Investment benefit evaluation of intelligent distribution network project based on game theory combination weighting[J]. Journal of Electric Power Science and Technology,2022,37(1):161-167.

[32] 赵洪山,李静璇,米增强,等.基于 CRITIC 和改进 Grey-TOPSIS 的电能质量分级评估方法[J].电力系统保护与控制,2022,50(3):1-8.
ZHAO Hongshan,LI Jingxuan,MI Zengqiang,et al.Grading evaluation of power quality based on CRITIC and improved Grey-TOPSIS[J]. Power System Protection and Control,2022,50(3):1-8.

[33] IMTHIAS AHAMED T P,NAGENDRA RAO P,SASTRY P S. A reinforcement learning approach to automatic generation control[J]. Electric Power Systems Research, 2002,63(1):9-26.

[34] 李卫东,刘娆,巴宇.新一代互联电网运行控制性能评价标准设计的理论基础与工作展望[J].电力科学与技术学报,2011,26(1):13-19+26.
LI Weidong, LIU Rao, BA Yu. Theory and prospect of performance evaluation criterion design for new interconnected power grid operation and control[J].Journal of Electric Power Science and Technology, 2011, 26(1): 13-19+26.

[35] SUTTON R S, BARTO A G. Reinforcement learning : an introduction[M].Cambridge:MIT Press,1998:60.