第 49 卷 第 6 期：2286-2298 高电压技术 Vol.49, No.6: 2286-2298
2023 年 6 月 30 日 High Voltage Engineering June 30, 2023
DOI: 10.13336/j.1003-6520.hve.20230040

基于自适应强化探索悲观 Q 的多智能体协同 AGC 算法

席 磊[1,2]，全 悦[1]，刘治洪[1]，解立辉[1]
（1. 三峡大学电气与新能源学院，宜昌 443002；
2. 梯级水电站运行与控制湖北省重点实验室（三峡大学），宜昌 443002）

摘 要：双碳目标下规模化可再生能源和柔性负荷的接入，使得电力系统中新能源占比日益增大。而传统的控制方法无法充分调动源–网–荷–储各部分的能动性，给电网带来愈来愈差的控制性能。因此本文从自动发电控制的角度，提出一种自适应强化探索悲观 Q 的多智能体协同算法，以提高源网荷储协同系统的控制性能。算法中所采用的悲观 Q 学习通过选择多个动作值估计器中最小动作值，不仅能够解决传统 Q 学习在动作探索过程中动作值的估计偏差，而且能够控制动作值估计偏差从正到负的变化，有助于提高算法的控制精度。同时自适应强化探索策略的引入，代替了传统 Q 学习中 ε-贪婪策略，能够避免重复和不平衡的探索。通过对改进的 IEEE 标准两区域负荷频率控制模型和源网荷储协同系统模型进行仿真，验证了所提算法的有效性，且与传统强化学习相比，具有更高的 CPS 性能、更小的频率偏差、更小的区域控制误差和更快的收敛速度。

关键词：悲观 Q 学习；自动发电控制；源网荷储；强化学习；新能源；多智能体

Multi-agent Collaborative AGC Algorithm Based on Self-adaptive Reinforcement-exploration Maxmin Q

XI Lei[1,2], QUAN Yue[1], LIU Zhihong[1], XIE Lihui[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: Under the carbon peak and neutrality targets, the proportion of new energy in the power system is increasing. The traditional control method is unable to properly mobilize the initiative of each part of source-gird-load-storage, so that the control performance of power gird is getting worse. Therefore, from the perspective of automatic generation control, this paper proposes a multi-agent collaborative algorithm based on Self-Adaptive Reinforcement-Exploration Maxmin Q in order to improve the control performance of source-gird-load-storage collaborative system. The Maxmin Q-learning adopted in the proposed algorithm mitigates the estimation bias of action-value by using the minimum action-value from multiple action-value estimates. Furthermore, the bias of action-value estimation can be controlled from positive to negative to enhance the control accuracy. Meanwhile, the introduction of Self-Adaptive Reinforcement-Exploration strategy replaces the ε-greedy strategy to avoid repetitive and unbalanced exploration. The simulations verify the effectiveness of the algorithm, where the improved IEEE standard two-area load-frequency control model and the model of "Source-Grid-Load-Storage" collaborative system are built. Compared with the traditional reinforcement learnings, the proposed algorithm achieves higher CPS performance, lower frequency deviation, lower area control error, and faster convergence.

Key words: maxmin Q-learning; automatic generation control; source grid load storage; reinforcement learning; new energy; multi-agent

基金资助项目：国家自然科学基金(52277108)。
Project supported by National Natural Science Foundation of China (52277108).

0　引言

“双碳”目标加速新型电力系统快速发展[1]。分布式可再生能源及柔性负荷的规模化接入使传统集中式电网结构模式逐步走向分散式结构模式。传统控制方法无法充分调动源–网–荷–储各部分的能动性来促进分布式能源[2-4]的消纳、维护电力系统安全稳定。而人工智能技术的进步，推动了电力系统的智能化，通过源–网–荷–储协同实现能源更绿色、更高效利用[5-7]。本文从基于人工智能的自动发电控制(automatic generation control，AGC)角度进行探索，以实现源网荷储系统最优协调控制。

目前可查文献中基于人工智能的 AGC [8-9]尤以强化学习应用最为广泛。文献[8]将控制性能指标(control performance standards，CPS)标准下的 AGC 系统看作一个典型的不确定随机系统，转化为基于马尔可夫决策过程理论的强化学习问题，进而应用 Q 学习来解决此问题，实现了 AGC 系统的动态优化决策。文献[9]提出改进的 R(λ)学习算法，使用在线“模仿学习”代替 Q 学习的预学习过程，克服了强化学习控制器需要另外搭建仿真模型来进行预学习收敛的缺陷。

而在面向 AGC 的诸多强化学习算法中以 Q 学习最为经典[10]，经过多年发展，Q 学习及其衍生算法[11-15]以自成体系。针对非马尔可夫环境下火电占优的互联电网，文献[11]引入随机最优控制中 Q(λ)学习的“后向估计”原理，有效解决火电机组大时滞环节带来的延时回报问题。针对大电网环境，文献[12]提出基于人工情感 Q 学习的自动发电控制，使智能体拥有机械逻辑与人工情感，其人工情感可生成连续动作代替传统强化学习中的离散动作，有效解决维度灾难问题。而学者们研究发现 Q 学习及其衍生算法存在动作值探索过程中的动作值估值偏差问题[13]，其动作值估计偏差会直接影响 AGC 系统的控制性能。

文[14]将双估计量法引入到 Q 学习中，提出了双 Q 学习(double Q-learning，DQL)来解决动作值估值偏差问题。DQL 的提出使得 AGC 控制器性能得到进一步提升。针对接入储能系统的多区域互联电网，文献[15]在 DQL 基础上引入采样参数 σ 来统一离策略和在策略的优缺点，实现了更优的协同控制。文献[16]针对大型风电并网，提出乐观双 Q 算法，在减缓动作值高估偏差的同时避免陷入局部最优解，有效提高了控制性能。

然而上述方法是有针对性地解决Q学习的动作值估计偏差问题，实际在真实复杂的环境里，动作值高估或者低估都有可能更利于智能体学习[17]。文献[17]研究了高估和低估偏差对学习性能的影响，提出悲观 Q 学习(maxmin Q-learning)。悲观一词的翻译取自预测决策理论中的悲观准则(又名 max-min 准则)。跟以往的单纯地去解决 Q 学习算法体系的动作值估计偏差问题不同，悲观 Q 学习提供了一个超参数，通过整定超参数来灵活地控制偏差，进而达到灵活地去选择是否需要智能体高估或低估动作值。

而悲观 Q 的收敛性并不优秀，尤其当超参数值变大时，收敛性会进一步恶化，难以满足 AGC 系统实时控制。要提高其收敛性，方法之一是改善动作探索策略，即智能体如何达到探索新知识与利用旧知识之间的平衡。传统的 ε-贪婪策略[18]很容易实现，而且理论上实验的次数是无限的，那么所有动作都会被无限次尝试，就会保证估计值收敛到真实值。而实际上在探索过程中动作选择是随机的，那么就有可能永远选择到最糟糕的动作。文献[19]提出了改进的 ε-贪婪策略，该方法在探索过程中，会考虑动作选择概率的估计值代替平均概率来选择动作，但该方法的问题在于很难确定估计值。文献[20]则舍弃 ε-贪婪策略，将禁忌搜索算法与 Q 学习相结合，提出基于禁忌搜索想法的 Q 学习算法来平衡探索与利用之间的关系，但该方法在不同的环境需要重新设置禁忌表长度和特赦准则，限制了其适用性。文献[21]通过自适应强化探索策略(self-adaptive reinforcement-exploration strategy，SARE)使智能体减少探索期间无意义的探索，并在利用期间更好地利用信息，平衡了探索与利用之间的矛盾问题。

因此，本文首次在 AGC 系统中应用悲观 Q，并融入 SARE 策略，提出了一种新颖的基于自适应强化探索悲观 Q 学习的多智能体协同算法，即 SARE–悲观 Q。所提算法不仅能够解决 Q 学习算法体系的动作值估计偏差问题，而且很好地平衡了探索与利用之间的矛盾问题。通过对改进的 IEEE 标准两区域负荷频率控制模型和源网荷储协同系统模型进行仿真，验证所提算法的有效性。

1　SARE–悲观 Q 学习

1.1　悲观 Q

悲观Q学习的核心思想为保持N个动作值估计

器，将这些估计器定义为 Q^i，i={1, …, N}。定义 $\boldsymbol{Q}^{\min}$ 为悲观动作值估计器，其计算式如下：

$$\boldsymbol{Q}^{\min}(s,a)=\min_{i\in\{1,\ldots,N\}}\boldsymbol{Q}^{i}(s,a) \tag{1}$$

为了在线进行训练，SARE–悲观 Q 拥有一个经验池 M，M 储存每一次循环经历的状态 s_{M}、动作 a_{M}、奖励 r_{M} 与下一个状态 s'_{M}。在一次循环中，随机选择 k 个(k≤N)动作值估计器，同时在 M 中选择多组数据，k 个动作值估计器使用这些数据进行更新，其动作值更新公式如下：

$$\begin{aligned}\boldsymbol{Q}^{i}(s_{\mathrm{M}},a_{\mathrm{M}})=\boldsymbol{Q}^{i}(s_{\mathrm{M}},a_{\mathrm{M}})+\alpha\big(r_{\mathrm{M}}+\\ \gamma\max_{a'}\boldsymbol{Q}^{\min}(s'_{\mathrm{M}},a')-\boldsymbol{Q}^{i}(s_{\mathrm{M}},a_{\mathrm{M}})\big)\end{aligned} \tag{2}$$

式中：α 为学习率；γ 为折扣因子。其动作值更新公式中所使用的目标值为 $\boldsymbol{Q}^{\min}$ 中的最大动作估计值。当 N=1 时，更新过程相当于原始的 Q 学习，此时动作值有高估偏差。随着 N 值的增加，高估偏差会逐渐减小；当 N>1 时，悲观 Q 从对动作值的高估逐步转为低估。所以，悲观 Q 创造了超参数 N 来调节估计偏差。

对于 N 和动作值估计偏差之间的数学关系参考附录 A。

1.2 SARE 策略

为了更好地解决探索和利用的矛盾问题，SARE 策略构造了行为效用跟踪 E，以记录每个状态对未来的影响，其定义如下：

$$\boldsymbol{E}_{\mathrm{t}}(a_i)=\begin{cases}\beta\boldsymbol{E}_{t-1}(a_i)+ev(s_t)1(r_t>0), & a_i=a_{t-1}=a_{t-2}\\ ev(s_t)1(r_t>0), & a_i=a_{t-1}\neq a_{t-2}\\ 0, & \text{其他}\end{cases} \tag{3}$$

式中：t 为当前时刻；β 为行为效用跟踪因子；动作 a_i 为智能体的动作集合的第 i 个动作；a_{t-1} 和 a_{t-2} 分别为上一个时刻和上上个时刻的动作；r_t 为当前状态的即时奖励；e 为探索激励值，其值越大，智能体越会倾向于探索；$1(r_t>0)$是一个用来判断真值的表达式，当且仅当 r_t>0 时为 1，在其他情况下为 0；$v(s_t)$为状态访问函数，描述一个状态是否被访问，其定义如下：

$$v(s_t)=\begin{cases}0, & s_t\in\boldsymbol{V}\\ 1, & \text{其他}\end{cases} \tag{4}$$

式中：V 为状态访问集合，$v(s_t)$当且仅当当前状态第一次被访问时的值为 1。

由于篇幅限制，SARE 策略的实现原理参考附录 B。

1.3 SARE–悲观 Q

SARE–悲观 Q 融合上述两种强化学习方法并加以改进，以达到理想的适配性。

1)在将 AGC 过程描述成马尔可夫决策过程后，拥有较多的状态与动作，若行为效用跟踪单一地用动作来描述，并不能准确地反映在某一状态下动作对未来的影响，本文将行为效用跟踪改为以状态-动作对来描述。同时，本文设计的奖励函数不会得到正数，若奖励函数值在[−5,0)之内，则表明此刻电力系统处于理想状态，因此判断即时奖励的条件设置为大于等于−5，其计算式如下：

$$\boldsymbol{E}_t(s_{t-1},a_i)=\begin{cases}\beta\boldsymbol{E}_{t-1}(s_{t-1},a_i)+e\times v(s_t)\times1(r_t\geq-5), & a_i=a_{t-1}=a_{t-2}\\ e\times v(s_t)\times1(r_t\geq-5), & a_i=a_{t-1}\neq a_{t-2}\\ 0, & \text{其他}\end{cases} \tag{5}$$

2）每次更新过程中，本文设计智能体随机选择每个动作的概率为

$$\boldsymbol{P}(s_{t-1},a_i)=\frac{1+\boldsymbol{E}(s_{t-1},a_i)}{n+\sum_1^n\boldsymbol{E}(s_{t-1},a_j)} \tag{6}$$

式中：$\boldsymbol{E}(s_{t-1},a_i)$为在 s_{t-1} 状态下动作 a_i 的行为效用跟踪值；n 为所有动作的个数。

3）值得一提的是，SARE 策略是设计在分幕式任务上，而描述为马尔科夫决策过程的 AGC 问题是一个持续性任务，因此本文将最大幕数设计为最大迭代次数 I_m，其计算式如下：

$$\varepsilon_t=1-\sin\left(\left(\frac{t}{I_{\mathrm{m}}}\right)^2\frac{\pi}{2}\right),\quad t\leq\theta I_{\mathrm{m}} \tag{7}$$

式中：ε_t 为当前时刻的探索系数；I_{m} 为训练的最大迭代次数；θ 为自适应探索因子，其值决定了第一阶段何时结束，过小探索不充分，过大无法收敛，经过实验，取值范围在[0,0.5]能达到理想效果。

4）同样，本文将达到最小值 $\varepsilon_{\min}$ 的幕数设计为达到最小值 $\varepsilon_{\min}$ 的迭代次数 I_1，其计算式如下：

$$\varepsilon_t=\begin{cases}\varepsilon_t+\dfrac{t-\theta I_{\mathrm{m}}}{I_1-\theta I_{\mathrm{m}}}(\varepsilon_{\min}-\varepsilon_1)-iR, & \\ \qquad\varepsilon_t+\dfrac{t-\theta I_{\mathrm{m}}}{I_1-\theta I_{\mathrm{m}}}(\varepsilon_{\min}-\varepsilon_1)-iR\geq\varepsilon_{\min} & \\ \varepsilon_{\min}, \quad \text{其他} & \end{cases} \tag{8}$$

式中：ε_1 为第一阶段的最终值；I_1 为 ε 达到最小值的目标迭代次数；R 为智能体每 20 次成功到达目标状态的概率；i 为 R 的利用因子；$\varepsilon_{\min}$ 是设置的探索

因子最小值。通过这样的设计，R 值越高，探索因子 ε 衰减地越快，从而更好地利用信息，加速收敛的同时保证算法效率。

SARE–悲观 Q 学习的整体控制框架见图 1。

3　基于 SARE–悲观 Q 学习的 AGC 系统设计

3.1　基于 SARE–悲观 Q 的 AGC 原理

为衡量基于 SARE–悲观 Q 的 AGC 控制器的控制性能，引入北美电力可靠性委员会提出的 CPS 标准[22]且频率偏差(Δf)的合格率范围为±0.2 Hz。

SARE–悲观 Q 包括智能体、状态、动作、奖励和环境。对于基于 SARE–悲观 Q 的 AGC 系统来说，智能体为 AGC 控制器，环境为电力系统，智能体通过动作与环境进行交互。为获得理想的控制效果，所设置的离散动作集不可过于细化或过于粗略，前者会导致状态−动作对维数过高，后者不利于智能体进行准确控制，两者都会使得 AGC 系统控制性能下降。经过大量实验，本文设置动作集为 A={−50，−40，−30，−20，−10，0，10，20，30，40，50} MW，共 11 个动作。在动作与环境的作用下，智能体会产生新的状态，本文以区域控制误差(area control error，ACE)作为状态输入，划分为 13 个状态，如表 1 所示。最后环境会给出一个即时奖励(奖励函数如 3.2 节所示)。

本文设计的源网荷储协同系统(4.2 节所示)划分为多个区域，采用分布式结构[23]，每个区域应用一个智能体，各智能体间为平级关系。因存在多个智能体与环境进行交互，构成多智能体系统(multi-agent system，MAS)。在 MAS 中，每个智能体需要交互分享信息来更新动作策略。本文设计各智能体的奖励函数相同，即各智能体目标一致，隶属于完全合作任务。因 SARE–悲观 Q 的单个智能体要维护 N 个动作值估计器，若采用群体强化学习方法，会导致状态与动作空间维数爆炸[24]。因此本文采用自主 Q 学习方法(independent Q-learning，IQL)[25]，每个智能体的目标为使所有智能体奖励最大化。

同时，源网荷储协同系统包括了可控柔性负荷。柔性负荷的特点之一就在于其输出与消纳功率的不确定性。若按传统等比例容量功率分配，过低会无法充分利用柔性负荷的优势，过高会导致功率指令与柔性负荷实际输出功率严重不匹配，危害电力系统安全。因此，在发出总调节功率指令后，智能体会进行一次协同功率分配，本文设计计算式如下：

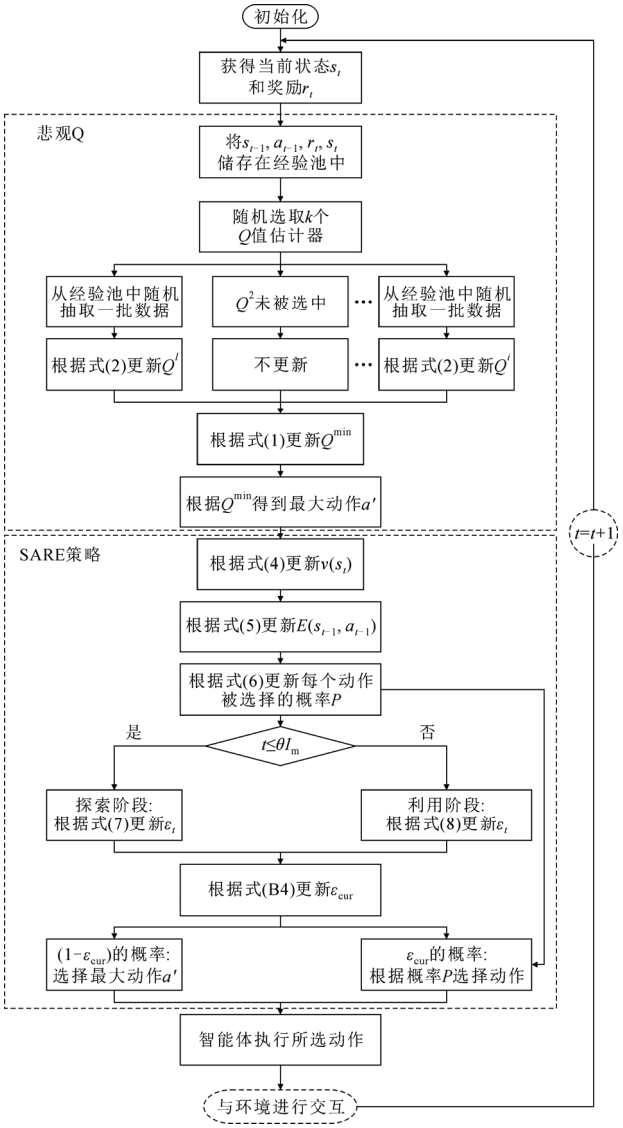


图 1　SARE–悲观 Q 整体控制框架

Fig.1　The overall control framework of SARE-Maxmin Q

表 1　SARE–悲观 Q 状态划分区间

Table 1　Partition of SARE-Maxmin Q state space

状态	区间	状态	区间
1	|ACE|<1MW	8	−10MW≤ACE<−1MW
2	1MW<ACE≤10MW	9	−20MW≤ACE<−10MW
3	10MW<ACE≤20MW	10	−30MW≤ACE<−20MW
4	20MW<ACE≤30MW	11	−40MW≤ACE<−30MW
5	30MW<ACE≤40MW	12	−50MW≤ACE<−40MW
6	40MW<ACE≤50MW	13	ACE<−50MW
7	ACE>50MW		

$$\Delta P_{\mathrm{F}}=\begin{cases}\dfrac{\left(E_{\mathrm{c}}-E_{\mathrm{cmin}}\right)}{E_{\mathrm{cmax}}}\Delta P, & \Delta P>0\\ \dfrac{E_{\mathrm{cmax}}-\left(E_{\mathrm{c}}-E_{\mathrm{cmin}}\right)}{E_{\mathrm{cmax}}}\Delta P, & \Delta P<0\\ 0, & \Delta P=0\end{cases}\tag{9}$$

$$\Delta P_{\mathrm{T}}=\begin{cases}0, & \Delta P=0\\ \Delta P-\Delta P_{\mathrm{F}}, & \text{其他}\end{cases} \tag{10}$$

式中：ΔP_{F} 为柔性负荷输出功率指令；ΔP_{T} 为传统机组输出指令；ΔP 为总调节功率指令；E_{c} 为柔性负荷的现有总储能；E_{cmin} 和 E_{cmax} 分别为柔性可控负荷总储能的下限和上限。

通过这样的设计，源网荷储协同系统形成了区域与区域之间、区域内发电侧与负荷侧之间的协同控制，源网荷储协同系统控制框架如图 2 所示。

3.2 奖励函数

对于新型电力系统的评价，除了可靠性要求，环保性要求也为极其重要的一环。因此本文设置智能体的目标为最小化 ACE 和碳排放(carbon emission，CE)[26]。区域电网的 CE 值等于机组输出功率与机组 CE 强度系数的乘积。将 ACE 和 CE 的量纲进行归一化处理，并将其线性加权，则智能体的综合目标奖励函数为：

$$R_t=-\eta\left(\delta(t)\right)^2-\frac{(1-\eta)\left(\sum_{k=1}^{m}B_k\left(\Delta P_k(t)\right)\right)}{1000}, \tag{11}$$

$$\Delta P_k^{\min}(t)\le\Delta P_k(t)\le\Delta P_k^{\max}(t)$$

式中：$\delta(t)$为 t 时刻 ACE 的瞬时值；$\Delta P_k(t)$为时刻第 k 机组的实际输出功率；η 和 $1-\eta$ 分别为 ACE 和 CE 的权重，将 ACE 和 CE 看做具有相同的权重，故 η 值取为 0.5；m 为区域机组总数；B_k 为第 k 机组的 CE 强度系数，kg/(kW·h)；$\Delta P_k^{\max}(t)$和$\Delta P_k^{\min}(t)$分别为第 k 机组容量的上下限，各类型发电机组 CE 强度系数为：

$$B_k=\begin{cases}0.87, & \Delta P_k>600\\ 0.89, & 300<\Delta P_k<600\\ 0.99, & \Delta P_k<300\end{cases} \tag{12}$$

3.3 参数设置

1）学习因子(0<α<1)，决定了赋予算法更新部分的信任度。α(接近 1)越大，算法的收敛速度越快，而较小的 α 可以增强算法的稳定性。

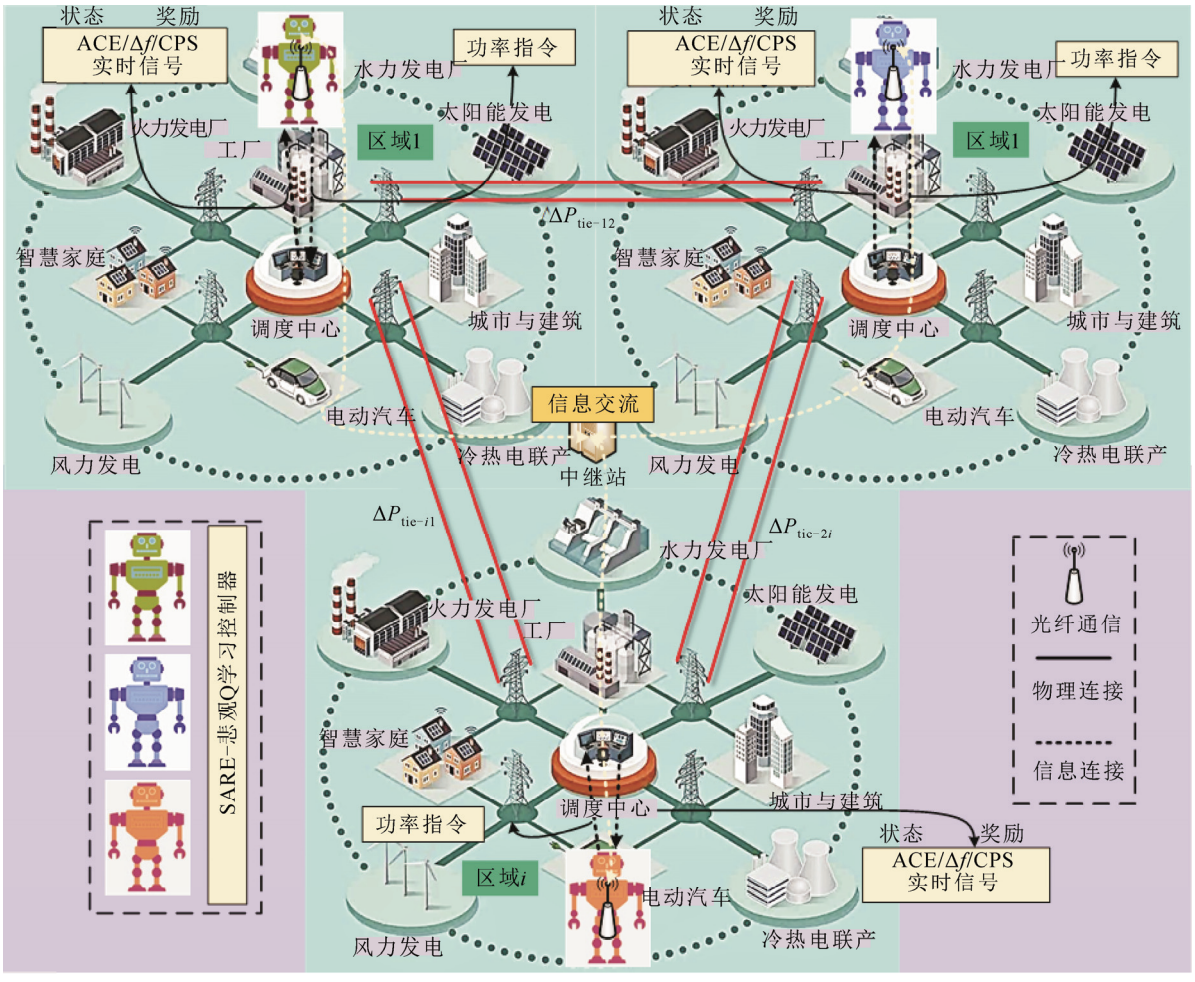


图 2 “源−网−荷−储”协同系统控制框架

Fig.2 Control framework of “Source-Grid-Load-Storage” collaborative system

2）折扣因子 $\gamma(0<\gamma<1)$，权衡立即奖励和后期奖励之间的重要性，考虑 γ 取值应接近 1。

3）自适应探索因子 $\theta(0<\theta<1)$，决定了智能体探索环境的时间。

4）动作值估计器个数 N，随着其值变大，动作值估计偏差呈现由正到负的变化。在预学习阶段选用 N=2、N=4、N=6 和 N=8 进行对比。在正式运行中，选用在预学习阶段获得最优控制效果的 N 值。

5）状态访问次数利用率 μ 和 φ，决定智能体探索新状态的喜好，其值越大，智能体越倾向于探索。

6）行为效用跟踪因子 $\beta(0<\beta<1)$，决定了行为效用跟踪值的衰减速度。

所有参数设置如表 2 所示。

4　仿真研究

4.1　改进的 IEEE 标准两区域 LFC 模型

本文在 IEEE 标准两区域负荷频率控制(load frequency control，LFC)模型基础上融入了冷热电联供系统(combined cooling heating and power, CCHP)[27]与飞轮储能模型[28]。CCHP 集制冷、供热及发电于一体，在高效利用能源资源与低碳运行两方面表现出巨大的优势。飞轮储能具有瞬时响应与精确跟踪的特点[29]，在参与电网调频具有显著的技术优势。改进后两区域 LFC 模型及机组参数如图 3 所示，虚线表示虚拟信息的传输，实线表示物理信息的传输。

4.1.1　预学习

强化学习简单来说就是一种试错学习，SARE–悲观 Q 控制器需要经过随机探索和试错学习过程，即预学习，才可以将其投入在线运行。经过大量实验证明，正弦负荷扰动能够很好地帮助智能体进行预学习，对两区域模型引入周期为 1 200 s，幅值为 1 000 MW 的连续正弦信号，引导智能体做出最优决策。基于 SARE–悲观 Q 控制器的特点，本文实验了不同 N 值下 SARE–悲观 Q 的性能。图 4 给出了 SARE–悲观 Q(N=2、N=4、N=6 和 N=8)在负荷扰动下两区域预学习性能指标以及它们的差异化。

为简述实验结果，SARE–悲观 Q(N=2)简写为 N=2，其他名称同理。由图 4(a)可见 N=2 在 3 000 s 左右能够跟踪负荷，N=4 在 5 000 s 左右能够跟踪负荷，N=6 在 6 000 s 左右能够跟踪负荷，而 N=8 在 8 000 s 左右能够跟踪负荷。

表 2　参数设置表

Table 2　Parameters for algorithm

参数	符号	数值
学习率	α	0.1
自适应探索因子	θ	0.1
折扣因子	γ	0.95
动作值估计器个数	N	2、4、6、8
探索激励因子	e	1
成功率加速利用因子	i	0.1
达到 $\varepsilon_{\min}$ 的目标迭代次数	I_1	1000
最大迭代次数	I_m	1250
状态访问次数利用率	μ 和 φ	0.02 和 0.0002
行为效用跟踪因子	β	0.75

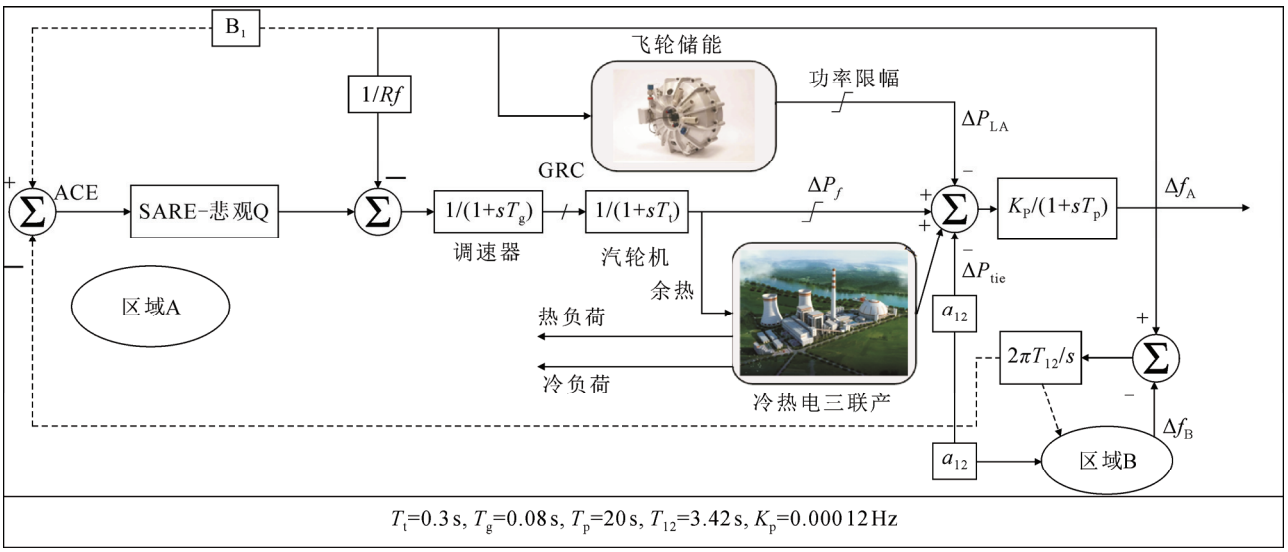


图 3　改进 IEEE 标准两区域 LFC 模型

Fig.3　Improved IEEE standard two-area LFC model

由图 4(b)可知，在预学习频率稳定期间，N=2 时Δf在±0.005 Hz 以内，N=4 时Δf在±0.004 Hz 以内，N=6 时Δf在±0.005 1 Hz 以内，N=8 时的Δf在±0.010 6 Hz 以内，区域内都具有合格的Δf。由图 4(c)可知，N=2 的 10-min ACE(10 min 内 ACE 的平均值)曲线趋于稳定在 4 MW 左右，N=4 稳定在 1 MW 左右，N=6 稳定在 2 MW 左右，而 N=8 稳定在 6 MW 左右。

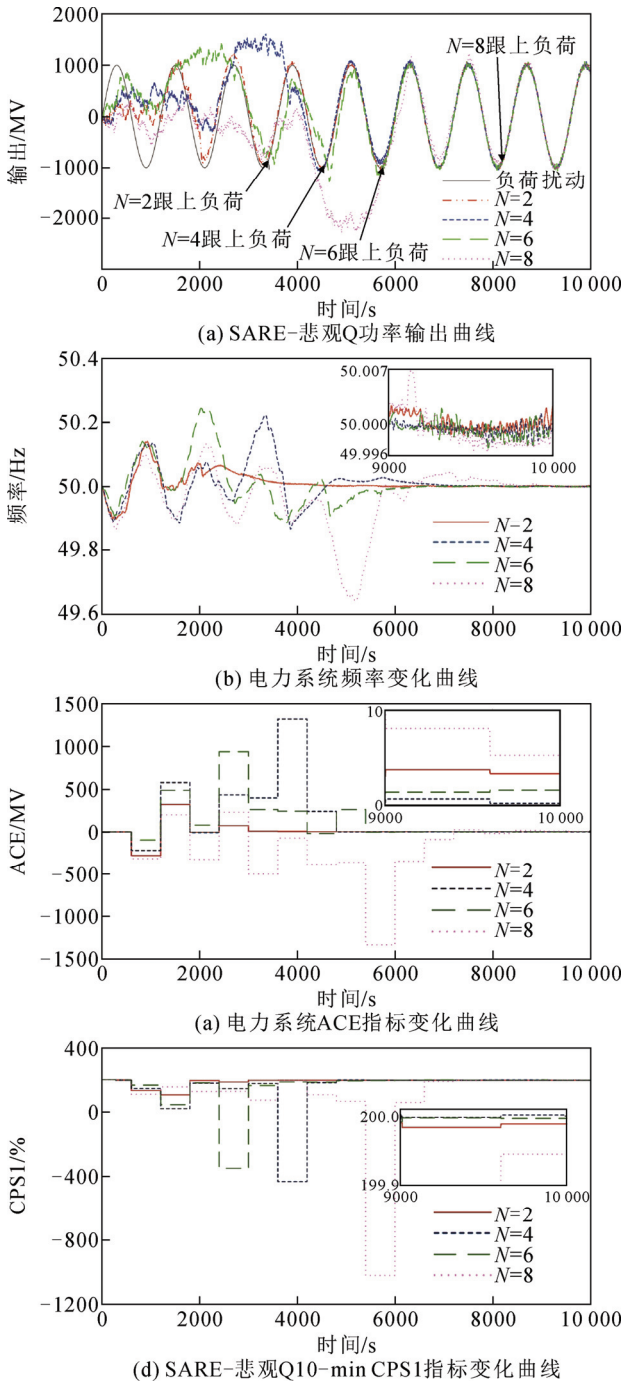


图 4 SARE–悲观 Q 预学习性能指标

Fig.4 Performance indexes of SARE-Maxmin Q in the pre-learning

CPS1 为在一个评估期间内，要求控制区域满足相应要求[22]的指标。由图 4(d)可知，N=2 的 10-min CPS1(10 min 内 CPS1 的平均值)曲线最终稳定在 199.98 %左右，N=4 的 CPS1 曲线最终稳定在 200.002 %左右，N=6 的 CPS1 曲线最终稳定在 199.99 %左右，而 N=8 稳定在 199.94 %左右，区域内都具有合格的 CPS1。

由图 5 可知，N=2 在 3 000 s 以内收敛，N=4 时算法在 4 500 s 以内收敛，N=6 时算法在 6 500 s 以内收敛，N=8 时在 8 000 s 以内收敛。对比分析了 Q[3]、DQL[8]、悲观 Q(N=4)[11]3 种算法的收敛性，可以看出 Q 收敛时间约为 20 000 s，DQL 收敛时间约为 28 000 s，悲观 Q 收敛时间约为 25 000 s。以 N=4 为例，SARE–悲观 Q 的收敛速度快于其他算法 77.5 %~83.93 %。

由分析可知，SARE–悲观 Q 在经过充分的学习探索后，其控制策略已达到最优，并且可以得知，在 N=4 时，SARE–悲观 Q 的控制效果最优。此外，随着 N 值的增加，SARE–悲观 Q 的收敛性会逐步变差。为了拥有更好的控制效果将 SARE–悲观 Q(N=4)控制器投入在线运行。

4.1.2 方波扰动实验

引入方波负荷扰动，模拟负荷持续突增与平稳的极端情况，测试算法的可靠性。对基于 Q、DQL、悲观 Q、SARE–悲观 Q 4 种算法(相同参数数值保持一致，以保证对比的有效性)的控制器进行仿真对比分析。图 6 为 4 种控制器面对方波扰动时的功率输出效果，图 7 为 4 种 AGC 控制器的控制性能指标(以 B 区域为例)。

由图 6 可知，SARE–悲观 Q 控制器在面对方波扰动时的稳定性最好，悲观 Q 其次。DQL 与 Q 都由于极端的负荷突变，导致小段时间难以跟上负荷。

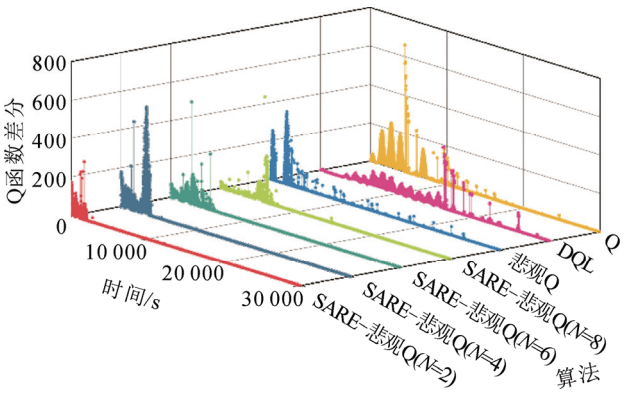


图 5 各算法预学习收敛效果

Fig.5 Convergence of each algorithms in the pre-learning

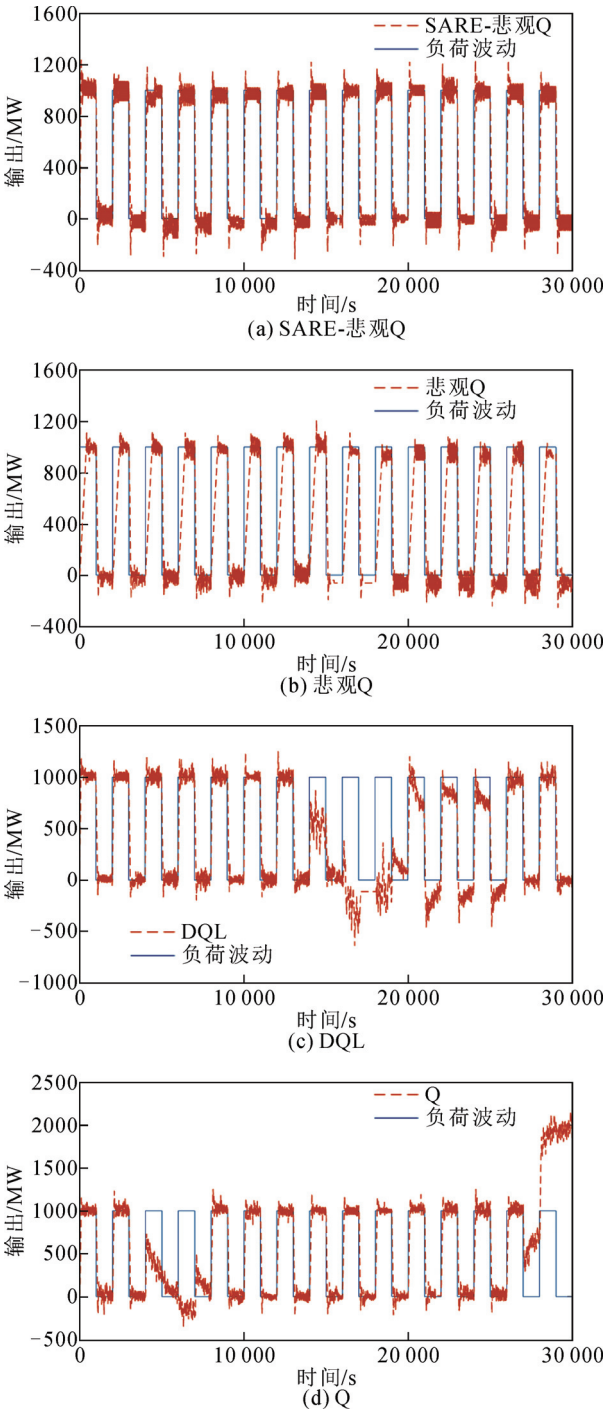


图 6　方波扰动下基于不同算法的控制器输出曲线

Fig.6　Output curve of controller based on different algorithms under square wave

图 7 为上述 4 种智能算法的各项性能考核指标，$|\Delta f|$为整个仿真周期中Δf绝对值的平均值，|ACE|为 ACE 绝对值的平均值。相较于其他算法，SARE–悲观 Q 能降低|ACE|69.37%~76.12%，减少$|\Delta f|$ 56.14%~67.46%，提高 CPS1 指标 9.97%~21.19%。仿真结果表明，SARE–悲观 Q 在能源出力不确定，负荷极端波动的情况下仍能保持稳定的控制效果。

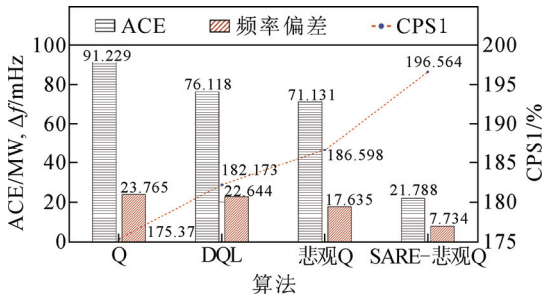


图 7　方波扰动下各算法性能指标对比图

Fig.7　Comparison of performance indexes of each algorithm under square wave

4.2　源网荷储协同系统模型

本文搭建的源网荷储协同系统包括传统火水电机组、分布式能源(风、光伏发电)、CCHP、柔性可控负荷(电动汽车)[30]和储能模型(电动汽车与飞轮储能)。其中，传统机组及风光发电使用实际西南电网数据[31]，风光发电仅作负荷处理。

电动汽车模型考虑了用户使用电动汽车的随机性，采用蒙特卡洛方法模拟了可控电动汽车数量的动态变化；同时为防止电池的过度充放电，模型中增加了功率和能量约束，只有满足此约束的可控电动汽车才能响应系统调频信号。可控电动汽车为满足所设定的充电状态(state of charge，SOC)的电动汽车，只有可控电动汽车能向电网提供调频服务。源网荷储协同系统模型及其机组参数参考附录图 C1。

根据中国公安部 2022 年统计数据[32]，得到中国电动汽车保有量在汽车总保有量中的占比，用该占比乘以各省份民用汽车保有量[33]，最后得到西南各省份预估电动汽车数量。在源网荷储协同系统中，四川区域有 347 913 辆，重庆有 134 356 辆，西藏有 16 205 辆电动汽车。

考虑到实际分布式能源接入的强随机性，通过引入无规则随机负荷信号模拟真实电网环境，进行 24 h 实时仿真，验证 SARE–悲观 Q 的实际工程应用效果。分别对 Q、DQL、悲观 Q、DWoLF-PHC(λ)[34]，SARE–悲观 Q 等 5 种算法进行仿真。图 8 为四川区域风、光伏发电的输出功率曲线，图 9 为四川区域电动汽车总储能变化曲线，图 10 为 SARE–悲观 Q 控制器的输出曲线(取四川区域作为展示)。可以看出 SARE–悲观 Q 的机组输出能够对随机扰动进行快速且精确的跟踪。

5 种算法最终实验结果由表 C1 给出，为更直观地展现出 SARE–悲观 Q 的性能优越性，将实验结

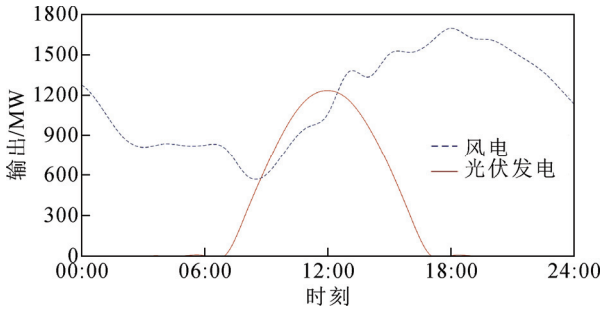


图 8 风、光伏发电功率输出曲线

Fig.8 Wind power and photovoltaic power generation output curve

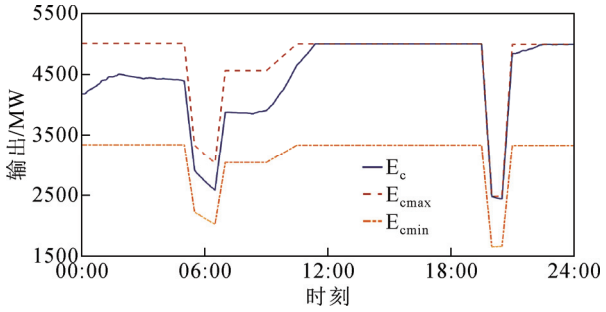


图 9 电动汽车总储能变化曲线

Fig.9 Dynamic change curve of the total storage energy of electric vehicles

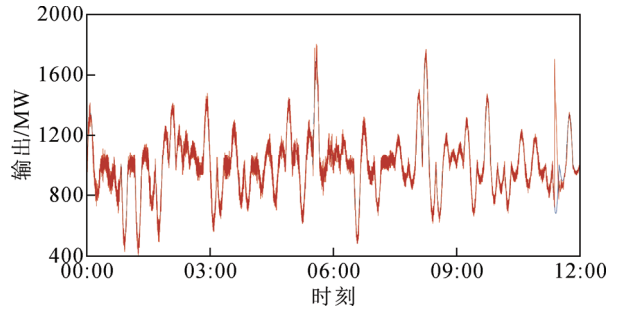


图 10 SARE–悲观 Q 功率输出曲线

Fig.10 Power output curve of SARE-Maxmin Q

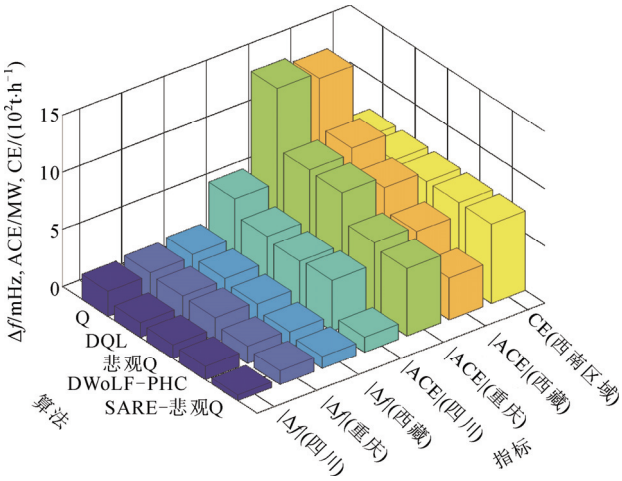


图 11 各算法性能指标对比图

Fig.11 Comparison of performance indexes of each algorithm

果绘制成了图 11 与图 12。图 11 中的 CE 为整个西南区域碳排放总和，分别为 6 966.716、6 958.732、6 946.718、6 953.072 和 6 945.649 t/h，其中 SARE–悲观的每小时碳排放量最低。

在表 C1 中，以四川地区为例，各算法下的$|\Delta f|$分别为 0.002 373、0.002 185、0.002 154、0.001 511 和 0.001 201 Hz，相较其他算法，SARE–悲观 Q 的$|\Delta f|$降低了 44.24 %~49.39 %。各种算法的 CPS1 分别是 199.56 %、199.618 %、199.651 %、199.632 %、199.693 %。各算法|ACE|分别为 14.252、8.952、8.82、6.29、5.879 MW，SARE–悲观 Q 的|ACE|降低了 33.34 %~58.75 %。

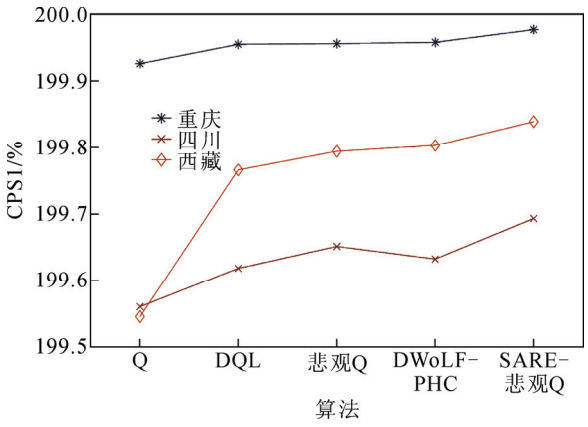


图 12 各算法 CPS1 指标对比曲线图

Fig.12 Comparison curve of CPS1 index of each algorithm

5 结论

1）SARE–悲观 Q 解决了传统强化学习算法中固有的“动作值高估”问题。该算法可通过整定超参数 N，获得最优协同控制策略，最大程度地保证 AGC 机组实行最优最快调节。

2）该算法所采用的 SARE 策略，能够避免重复和不平衡的探索，并在利用期间更好地利用信息，显著提高了算法的收敛能力。

3）实验结果表明 SARE–悲观 Q 能表现出优于其他控制方法的强适应性，且能够有效实现源网荷储系统的协同控制。

在本文的研究中发现，对于确定 SARE–悲观 Q 中的超参数 N 需要做大量的实验，才能找到适应当前环境的最优性能。通过一些人工智能技术(例如元

学习)去自动学习超参数 N 或者实现更强的泛化能力是下一步的研究重点。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 韩肖清，李廷钧，张东霞，等. 双碳目标下的新型电力系统规划新问题及关键技术[J]. 高电压技术，2021，47(9)：3036-3046.
HAN Xiaoqing, LI Tingjun, ZHANG Dongxia, et al. New issues and key technologies of new power system planning under double carbon goals[J]. High Voltage Engineering, 2021, 47(9): 3036-3046.

[2] 舒印彪，张智刚，郭剑波，等. 新能源消纳关键因素分析及解决措施研究[J]. 中国电机工程学报，2017，37(1)：1-8.
SHU Yinbiao, ZHANG Zhigang, GUO Jianbo, et al. Study on key factors and solution of renewable energy accommodation[J]. Proceedings of the CSEE, 2017, 37(1): 1-8.

[3] 刘永梅，王金丽，杨红磊，等. 计及柔性负荷调节能力的有源配电网动态优化方法[J]. 高电压技术，2021，47(1)：73-80.
LIU Yongmei, WANG Jinli, YANG Honglei, et al. Dynamic optimal method of distribution network in consideration of flexible load adjustment capability[J]. High Voltage Engineering, 2021, 47(1): 73-80.

[4] 杨子民，彭小圣，郎建勋，等. 基于集群动态划分与 BLSTM 深度学习的风电集群短期功率预测[J]. 高电压技术，2021，47(4)：1195-1203.
YANG Zimin, PENG Xiaosheng, LANG Jianxun, et al. Short-term wind power prediction based on dynamic cluster division and BLSTM deep learning method[J]. High Voltage Engineering, 2021, 47(4): 1195-1203.

[5] 李建林，郭兆东，马速良，等. 新型电力系统下“源网荷储”架构与评估体系综述[J]. 高电压技术，2022，48(11)：4330-4341.
LI Jianlin, GUO Zhaodong, MA Suliang, et al. Overview of the “Source-grid-load-storage” architecture and evaluation system under the new power system[J]. High Voltage Engineering, 2022, 48(11): 4330-4341.

[6] 罗凤章，张　旭，杨　欣，等. 基于深度学习的综合能源配电系统负荷分析预测[J]. 高电压技术，2021，47(1)：23-32.
LUO Fengzhang, ZHANG Xu, YANG Xin, et al. Load analysis and prediction of integrated energy distribution system based on deep learning[J]. High Voltage Engineering, 2021, 47(1): 23-32.

[7] 董旭柱，华祝虎，尚　磊，等. 新型配电系统形态特征与技术展望[J]. 高电压技术，2021，47(9)：3021-3035.
DONG Xuzhu, HUA Zhuhu, SHANG Lei, et al. Morphological characteristics and technology prospect of new distribution system[J]. High Voltage Engineering, 2021, 47(9): 3021-3035.

[8] 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 K W. Q-learning based dynamic optimal CPS control methodology for interconnected power systems[J]. Proceedings of the CSEE, 2009, 29(19): 13-19.

[9] 余　涛，梁海华，周　斌. 基于 R(λ)学习的孤岛微电网智能发电控制[J]. 电力系统保护与控制，2012，40(13)：7-13.
YU Tao, LIANG Haihua, ZHOU Bin. Smart power generation control for microgrids islanded operation based on R(λ) learning[J]. Power System Protection and Control, 2012, 40(13): 7-13.

[10] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054-1054.

[11] 余　涛，周　斌，陈家荣. 基于多步回溯 Q(λ)学习的互联电网随机最优 CPS 控制[J]. 电工技术学报，2011，26(6)：179-186.
YU Tao, ZHOU Bin, CHAN K W. Stochastic optimal CPS control for interconnected power grids using multi-step backtrack Q(λ) learning[J]. Transactions of China Electrotechnical Society, 2011, 26(6): 179-186.

[12] 殷林飞，郑宝敏，余　涛. 人工情感 Q 学习的互联电网自动发电控制算法[J]. 控制理论与应用，2016，33(12)：1650-1657.
YIN Linfei, ZHENG Baomin, YU Tao. Artificial emotionnal Q-learning for automatic generation control of interconnected power grids[J]. Control Theory & Applications, 2016, 33(12): 1650-1657.

[13] THRUN S, SCHWARTZ A. Issues in using function approximation for reinforcement learning[C]// Proceedings of the Fourth Connectionist Models Summer School. Hillsdale, USA: Erlbaum, 1993.

[14] VAN HASSELT H. Double Q-learning[C]// Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver British, Canada: Curran Associates Inc, 2010.

[15] XI L, ZHOU L P, XU Y C, et al. A multi-step unified reinforcement learning method for automatic generation control in multi-area interconnected power grid[J]. IEEE Transactions on Sustainable Energy, 2021, 12(2): 1406-1415.

[16] XI L, LI H K, ZHU J Z, et al. A novel automatic generation control method based on the large-scale electric vehicles and wind power integration into the grid[J/OL]. IEEE Transactions on Neural Networks and Learning System, 2022: 1-11. https://doi. org/10.1109/TNNLS. 2022.3194247.

[17] LAN Q F, PAN Y C, FYSHE A, et al. Maxmin Q-learning: controlling the estimation bias of Q-learning[C]// Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview. net, 2020: 1-20.

[18] WATKINS C J C H. Learning from delayed rewards[D]. Cambridge, UK: University of Cambridge, 1989.

[19] TIJSMA A D, DRUGAN M M, WIERING M A. Comparing exploration strategies for Q-learning in random stochastic mazes[C]// Proceedings of 2016 IEEE Symposium Series on Computational Intelligence. Athens, Greece: IEEE, 2016: 1-8.

[20] ZHANG X G, LIU Z J. An optimized Q-learning algorithm based on the thinking of Tabu search[C]// Proceedings of 2008 International Symposium on Computational Intelligence and Design. Wuhan, China: IEEE, 2008: 533-536.

[21] ZHANG L P, TANG L, ZHANG S L, et al. A self-adaptive reinforcement-exploration Q-learning algorithm[J]. Symmetry, 2021, 13(6): 1057.

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

[23] 宋成铭. 基于多智能体系统的 AGC、AVC 协调控制[D]. 北京：华北电力大学，2019：24-25.
SONG Chengming. Coordination control of AGC and AVC based on multi-agent system[D]. Beijing, China: North China Electric Power University, 2019: 24-25.

[24] CHENG X Y, YUAN X H, PAN L H, et al. Reinforcement learning in simulation RoboCup soccer[C]//Proceedings of 2004 International Conference on Machine Learning and Cybernetics. Shanghai, China: IEEE, 2004: 26-29.

[25] TAMPUU A, MATIISEN T, KODELJA D, et al. Multiagent cooperation and competition with deep reinforcement learning[J]. PLoS One, 2017, 12(4): e0172395.

[26] 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)：4150-4161.
XI Lei, YU Lu, FU Yimu, et al. Automatic generation control based on

deep reinforcement learning with exploration awareness[J]. Proceedings of the CSEE, 2019, 39(14): 4150-4161.

[27] 席　磊，张　乐，黄悦华，等. 基于贪婪策略的多层自动发电控制[J]. 中国电机工程学报，2020，40(16)：5204-5216.
XI Lei, ZHANG Le, HUANG Yuehua, et al. Multiple level automatic generation control based on the greedy strategy[J]. Proceedings of the CSEE, 2020, 40(16): 5204-5216.

[28] 席　磊，刘宗静，孙梦梦，等. 面向多区域能源互联的多智能体协同 AGC 策略[J]. 中国科学：技术科学，2021，51(6)：673-683.
XI Lei, LIU Zongjing, SUN Mengmeng, et al. The multiagent cooperative AGC strategy for multiregional energy interconnection[J]. SCIENTIA SINICA Technologica, 2021, 51(6): 673-683.

[29] 罗耀东，田立军，王　垚，等. 飞轮储能参与电网一次调频协调控制策略与容量优化配置[J]. 电力系统自动化，2022，46(9)：71-82.
LUO Yaodong, TIAN Lijun, WANG Yao, et al. Coordinated control strategy and optimal capacity configuration for flywheel energy storage participating in primary frequency regulation of power grid[J]. Automation of Electric Power Systems, 2022, 46(9): 71-82.

[30] ZHANG Q, LI Y, LI C, et al. Real-time adjustment of load frequency control based on controllable energy of electric vehicles[J]. Transactions of the Institute of Measurement and Control, 2020, 42(1): 42-54.

[31] 中国电力企业联合会. 中国电力统计年鉴-2021[M]. 北京：中国统计出版社，2021：20-29.
China Electricity Council. China electricity statistical yearbook-2021[M]. Beijing, China: China Statistics Press, 2021: 20-29.

[32] 孔大为. 全国汽车保有量达 3.15 亿量[N]. 人民公安报，2022-10-09(02).
KONG Dawei. The car parc is reached 315 million in China[N]. The People's Public Security Newspaper, 2022-10-09(02).

[33] 国家统计局. 中国统计年鉴-2021[M]. 北京：中国统计出版社，2021: 354-355.
National Bureau of Statistics. China statistical yearbook-2021[M]. Beijing, China: China Statistics Press, 2021: 354-355.

[34] XI L, YU T, YANG B, et al. A novel multi-agent decentralized win or learn fast policy hill-climbing with eligibility trace algorithm for smart generation control of interconnected complex power grids[J]. Energy Conversion and Management, 2015, 103: 82-93.



席　磊
1982—，男，博士，教授，博导
主要从事自动发电控制、人工智能、电力系统网络攻击与防御方面的研究工作
E-mail: xilei2014@163.com

XI Lei
Ph.D., Professor



全　悦
1999—，男，硕士生
主要从事自动发电控制、人工智能方面的研究工作
E-mail: Quanyue_1999@163.com

QUAN Yue



刘治洪
1998—，男，硕士生
主要从事自动发电控制、人工智能方面的研究工作
E-mail: liuzhihong3396@163.com

LIU Zhihong



解立辉(通信作者)
1983—，男，博士生
主要从事自动发电控制、人工智能方面的研究工作
E-mail: lihui20042004@163.com

XIE Lihui
Ph.D. candidate
Corresponding author

收稿日期　2023-01-09　修回日期　2023-03-13　　编辑　曹昭君

附录 A

$$Q^i(s,a)=Q^*(s,a)+e^i(s,a) \tag{A1}$$

式中，$Q^i(s,a)$为第 i 个 Q 值估计器所估计的动作价值，$Q^*(s,a)$为理论上实际的动作价值，$e^i(s,a)$为第 i 个 Q 值估计器拥有的随机近似误差。假设 M 为 s' 状态下的所有动作数，定义一组数据状态 s，动作 a，奖励 r，下一时刻的状态 s'的预估偏差 Z_{MN}为：

$$Z_{MN}=\gamma\left(\max_{a'}Q^{\min}(s',a')-\max_{a'}Q^*(s',a')\right) \tag{A2}$$

式中：

$$Q^{\min}(s,a)=Q^*(s,a)+\min_{i\in\{1,\ldots,N\}}e^i(s,a) \tag{A3}$$

式(3)定义了悲观动作值估计器 $Q^{\min}(s,a)$；式(2)中 $\max_{a'}Q^{\min}(s',a')$ 表示在状态 s'下所有动作中，选择动作 a'会获得最大 $Q^{\min}$ 值，$\max_{a'}Q^*(s',a')$ 同理。

式(4)和式(5)展示了期望估计偏差 $E[Z_{MN}]$是如何与悲观 Q 中动作值估计器的数量 N 相联系的：

$$\begin{aligned}E[Z_{MN}]&=\gamma\tau(1-2t_{MN})\\ t_{MN}&=\frac{M(M-1)\ldots1}{\left(M+\frac{1}{N}\right)\left(M-1+\frac{1}{N}\right)\ldots\left(1+\frac{1}{N}\right)}\end{aligned} \tag{A4}$$

$E[Z_{MN}]$会随着 N 值的增加而减少：

$$\begin{aligned}E[Z_{M,N=1}]&=\gamma\tau\frac{M-1}{M+1}\\ E[Z_{M,N=\infty}]&=-\gamma\tau\end{aligned} \tag{A5}$$

因此，可以通过改变悲观 Q 中估计器的数量来控制偏差。具体来说，随着 N 的增加，估计偏差可以从正减少为负。需要注意的是，当 t_{MN}=1/2 时，$E[Z_{MN}]$=0。这表明，通过选择 N 使 t_{MN}≈1/2，可以将偏差减少接近到 0。

附录 B

1、每次更新过程中，智能体随机选择每个动作的概率为：

$$P(a_i)=\frac{1+\boldsymbol{E}(a_i)}{n+\sum_1^n\boldsymbol{E}(a_j)} \tag{B1}$$

2、在不同的阶段实时调整勘探系数 ε，以达到不同的要求，ε 的衰减分为两个阶段，第一个阶段为智能体初始训练阶段，未知环境下，初始训练阶段主要任务是探索环境，ε 衰减可以相对缓慢，定义如下：

$$\varepsilon_t=1-\sin\left(\left(\frac{t}{T}\right)^2\times\frac{\pi}{2}\right),\quad t\le\theta T \tag{B2}$$

在训练周期 $t>\theta T$ 时，进入第二阶段，重点从探索转为利用，具体公式如下：

$$\varepsilon_t=\begin{cases}\varepsilon_t+\dfrac{t-\theta T}{T_1-\theta T}(\varepsilon_{\min}-\varepsilon_1)-i\times R,\\ \qquad\varepsilon_t+\dfrac{t-\theta T}{T_1-\theta T}(\varepsilon_{\min}-\varepsilon_1)-i\times R\ge\varepsilon_{\min}\\ \varepsilon_{\min},\quad\text{其他}\end{cases} \tag{B3}$$

3、为了提高目标位置附近的探索效率，减少初始位置附近无效的探索，根据该状态的访问次数，对探索因子提出了一种动态的调整方法，如式：

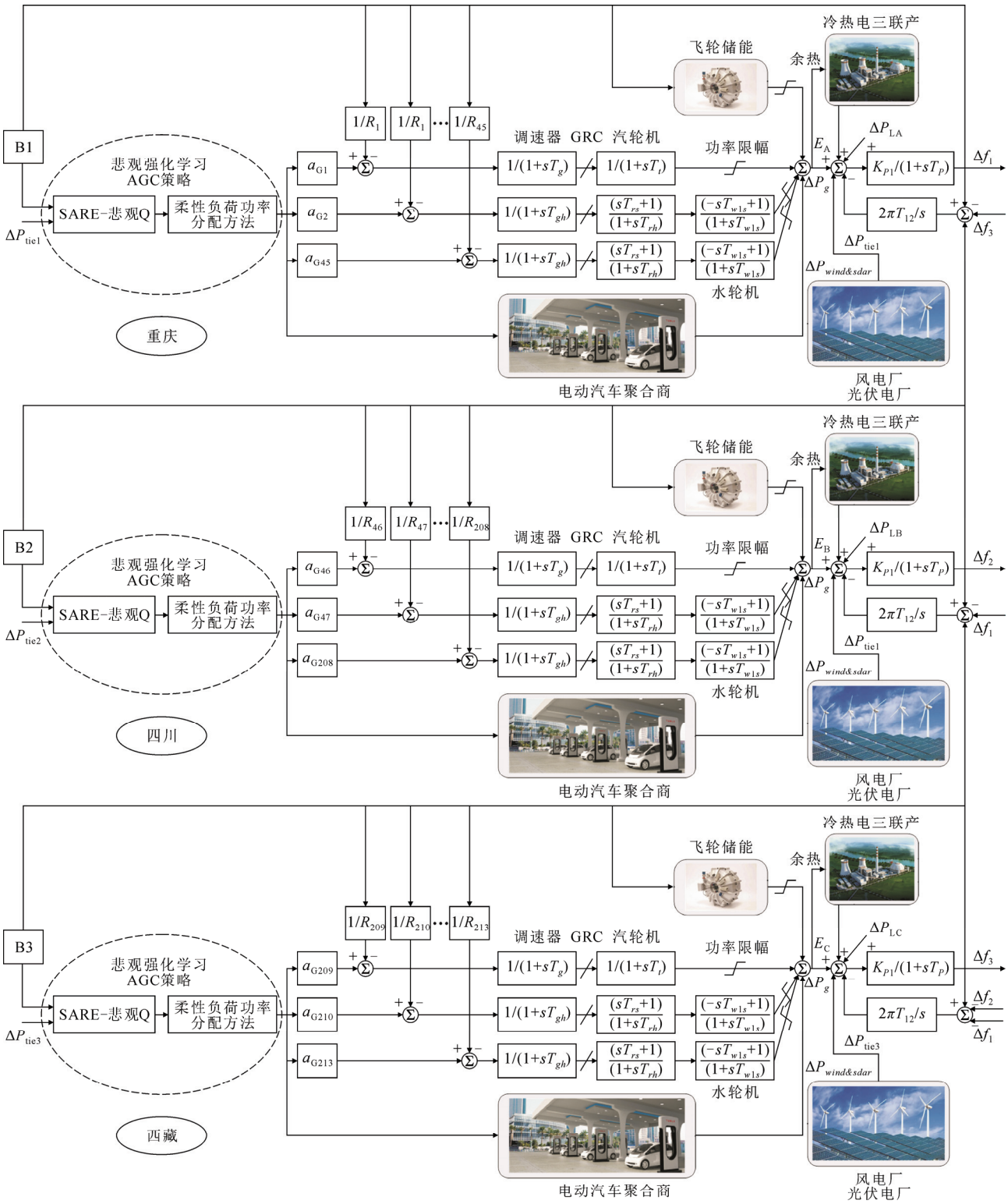
$$\varepsilon_{cur}=\begin{cases}\varepsilon_t+\mu, & x=0\\ \varepsilon_t-\mu\sin\left(\varphi x\times\dfrac{\pi}{2}\right) & \text{其他}\end{cases} \tag{B4}$$

附录 C

表 C1　“源−网−荷−储”协同系统仿真性能指标对比表

Table C1　Comparison of simulation performance indexes in “Source-Grid-Load-Storage” collaborative system

区域	算法	Δf/Hz	|ACE|/MW	CPS1/%	CE/(t·h^{-1})
重庆	Q	0.002165	5.981	199.926	2445.413
	DQL	0.001276	4.552	199.955	2450.424
	悲观 Q	0.001198	4.239	199.956	2447.969
	DWoLF-PHC(λ)	0.001151	4.417	199.958	2451.051
	SARE−悲观 Q	0.000482	1.302	199.977	2445.939
四川	Q	0.002373	14.252	199.56	1949.781
	DQL	0.002185	8.952	199.618	1944.994
	悲观 Q	0.002154	8.82	199.651	1941.444
	DWoLF-PHC(λ)	0.001511	6.29	199.632	1937.402
	SARE−悲观 Q	0.001201	5.879	199.693	1939.346
西藏	Q	0.002556	13.698	199.545	2571.522
	DQL	0.002297	9.429	199.766	2563.314
	悲观 Q	0.001924	7.837	199.794	2557.305
	DWoLF-PHC(λ)	0.00133	5.848	199.803	2564.619
	SARE−悲观 Q	0.001021	3.638	199.839	2560.364



T_p=20 s, $T_{12,\,12,\,31}$=15.9, 7.96, 15.9 s, $K_{p1,\,2,\,3}$=0.0027, 0.0029, 0.0027 Hz

图 C1　源网荷储协同系统模型

Fig.C1　The model of “Source-Grid-Load-Storage” collaborative system