基于强化学习的互联电网CPS自校正控制

余 涛，周 斌

（华南理工大学电力学院，广东 广州 510640）

摘要： AGC是一个动态多级决策问题——马尔可夫决策过程(MDP)，应用强化学习算法可有效地实现控制策略的在线学习和动态优化决策。引入Q学习算法作为强化学习核心算法，将CPS值看作包含AGC的电力系统“环境”所给的“奖励”，依靠奖励值Q函数与CPS控制动作形成的闭环控制结构实现在线学习。学习目标是使CPS控制动作从环境获得的长期积累奖励值最大，从而快速自动地在线优化CPS控制系统的输出。仿真研究显示，引入强化学习自校正控制后显著增强了整个AGC系统的鲁棒性和适应性，有效提高了CPS考核合格率。

关键词： 强化学习；Q学习算法；自动发电控制；CPS标准；自校正控制

Reinforcement learning based CPS self-tuning control methodology for interconnected power systems

YU Tao， ZHOU Bin

(College of Electric Power, South China of Technology, Guangzhou 510640, China)

Abstract: The automatic generation control (AGC) problem is a stochastic multistage decision problem, which can be modeled as a Markovian Decision Process (MDP). The paper introduces the Q-learning method as the core algorithm of reinforcement learning (RL), and regards the CPS values as the rewards from the interconnected power systems. By regulating a closed-loop CPS control rule to maximize the total reward in the procedure of on-line learning, the optimal CPS control strategy can be gradually obtained. The case study shows that after adding the RL control, the robustness and adaptability of AGC system is enhanced obviously and the CPS compliance is ensured.

This work is supported by National Natural Science Foundation of China(No.50807016) and Natural Science Funds of Guangdong Province (No. 06300091).

Key words: reinforcement learning; Q-learning algorithm; automatic generation control; CPS; self-tuning control

中图分类号： TM71; TP181 文献标识码：A 文章编号： 1674-3415(2009)10-0033-06

0 引言

北美电力可靠性委员会(NERC)于1997年推出联络线功率与系统频率偏差(TBC)模式下互联电网自动发电控制(AGC)系统的最新控制性能标准(CPS)/扰动控制标准(DCS)[1]，取代原先的A标准和B标准后，如何设计与CPS/DCS标准相符的AGC控制策略成为国内外专家关注的焦点[2~7]。我国高宗和等学者在CPS标准下AGC分层控制和控制策略从工程实用化方面做出了重要贡献[8,9]。

现有CPS标准下的AGC控制策略大多数为基于工程经验、增益固定的PI控制结构，往往难以满足AGC控制系统对高适应性和高鲁棒性的要求。

基金项目： 国家自然科学基金项目（50807016）；广东省自然科学基金博士启动基金项目（06300091）

本文提出一种基于强化学习RL(Reinforcement Learning)的CPS自校正控制策略。该控制策略无需对被控对象的运行条件和动态特性做任何近似假设，在线学习算法对电网结构、参数和运行方式具有更佳的适应性。通过对标准两区域互联系统及以南方电网为实例的仿真研究显示，该强化学习自校正控制器能够快速自动地在线优化CPS控制系统的输出，显著增强了AGC控制系统鲁棒性和适应性的同时，提高了互联电网CPS考核合格率。

1 CPS标准简介

基于CPS标准的AGC系统目标是通过自动调节系统有功出力，维持电网频率和本控制区域净交换功率控制在允许范围内，即把由负荷变化或机组出力波动产生的区域控制偏差(ACE)限定在一定范围内，使CPS考核指标达到合格的要求。NERC的

CPS 考核标准由两部分组成：CPS1 是统计 ACE 变化量与频率偏差关系的标准，用于提高频率控制质量；CPS2 作用是限制大的不可接受且不可预见的系统潮流。

CPS1 要求对于某 i 区域电网在某一段考核时段(如 10 min)内有

$$\alpha_{\mathrm{CF1}}=\frac{\sum(E_{\mathrm{AVE-min}}\cdot\Delta F_{\mathrm{AVE}})}{10B_i\cdot n}\le\varepsilon_1^2 \tag{1}$$

式中：$E_{\mathrm{AVE\text{-}min}}$ 为 1 min 内 ACE 的平均值；ΔF_{AVE} 为 1 min 频率偏差的平均值；B_i 为控制区域的频率偏差系数；ε_1 为互联电网对全年 1 min 频率平均偏差均方根的控制目标值；n 为该时段内的分钟数。则这一段时段 CPS1 的指标的统计公式为

$$C_{\mathrm{CPS1}}=(2-\beta_{\mathrm{CF1}})\times100\% \tag{2}$$

式中：$\beta_{\mathrm{CF1}}=\alpha_{\mathrm{CF1}}/\varepsilon_1^2$。CPS2 要求考核时段(10 min) ACE 的平均值的绝对值控制在规定的范围 L_{10} 以内，即

$$\left|\sum E_{\mathrm{AVE-min}}\right|/10\le L_{10} \tag{3}$$

式中：$L_{10}=1.65\cdot\varepsilon_{10}\cdot\sqrt{(10B_i)\cdot(10B_s)}$；$B_i$ 和 B_s 分别为该区域电网和整个互联电网的频率偏差系数；ε_{10} 为互联电网对全年 10 min 频率平均偏差的均方根值的控制目标值。CPS 的考核合格标准可参考 NERC 标准文件[1]。

2　控制原理

2.1 强化学习自校正控制算法

强化学习[10]又称再励学习、评价学习，是基于马尔可夫决策过程(MDP)模型[11]的一种学习控制(Learning Control)技术，通常是由优化一个状态的值函数(value function)的途径来学习最优控制策略：$S\rightarrow A$，使 agent 选择的动作从环境获得的长期积累奖励值最大。强化学习模型主要由 world 和 agent 两大模块构成，world 由一组环境所有可能的状态集合即状态空间 S 来描述，动作空间 A 为 agent 可能产生的动作集合。根据强化学习算法的不同，值函数有不同的形式，目前较有影响的强化学习算法有 TD 算法、Q 学习算法、Sarsa 算法、Dyan 算法[12]等。本文将基于 Q 学习算法(Q-learning algorithm)提出一种互联电网 CPS 在线自校正控制方法。

Q 学习算法是一种不依赖于对象模型的强化学习算法[13]，其目的是通过试错(trial-and-error)与环境交互获得策略的改进，从长期的观点构造控制策略。它通过直接优化一个可迭代计算的状态－动作对值函数 $Q(s,a)$来在线寻求最优策略使得期望折扣报酬总和最大。Tsitsiklis 等人则证明了 Q 学习算法的收敛特性[14]。Q 学习的值函数满足式(4)：

$$Q(s,a)=R(s,s',a)+\gamma\sum_{s'\in S}P(s'|s,a)\max_{a\in A}Q(s',a) \tag{4}$$

式中：s, s'分别代表当前状态和下一时刻的状态，γ 为折扣因子，一般取值在 0.5～0.98 范围内，$P(s'|s,a)$ 为状态 s 在动作 a 发生后转移到状态 s'的概率，$R(s,s',a)$为环境由状态 s 经过动作 a 转移到状态 s'后给出的立即强化信号(immediate reinforcement)。

Q 学习算法是利用迭代计算的方法求取最优 Q 值函数的估计值，设 Q^k 代表最优值函数 Q^*的第 k 次迭代值，agent 通过此次试探学习获得的经验(experience)即$<s_k, a, r, s_{k+1}>$样本，更新 Q 值迭代公式如下：

$$\begin{aligned}&Q^{k+1}(s_k,a_k)=Q^k(s_k,a_k)+\\&\alpha[R(s_k,s_{k+1},a_k)+\gamma\max_{a'\in A}Q^k(s_{k+1},a')-Q^k(s_k,a_k)]\\&Q^{k+1}(\tilde{s},\tilde{a})=Q^k(\tilde{s},\tilde{a})\qquad\forall(\tilde{s},\tilde{a})\ne(s_k,a_k)\end{aligned} \tag{5}$$

式中：$0<\alpha<1$，称为学习因子，α 指明了要给改善的更新部分多少信任度。Q 函数的实现主要采用 lookup 表格的方法来表示，$Q(s,a)(s\in S, a\in A)$代表 s 状态下执行动作 a 的 Q 值，表的大小等于 $S\times A$ 的笛卡尔乘积中元素的个数，表中 Q 值的初始化可任意给定，一般初值都设为 0，且在训练中 Q 值不会下降且保持在 0 和最优值 Q^*区间内。

Q 学习算法中动作选择策略是控制算法的关键，本文中动作策略即 AGC 功率调节指令。强化学习面临着搜索(exploration)和利用(exploitation)的权衡问题，本文采用一种常用的 Boltzmann 分布探索法来构造动作选择策略。该策略即在学习初始阶段 agent 从随机开始选择动作，初始化使得各状态下任意可行动作被选择的概率相等。然后在学习过程中随着 Q 值函数表格的变化，各状态下动作概率分布按照式(6)进行更新，有较高 Q 值的动作被赋予较高的概率，而且所有动作的概率都非零。

$$P_s^k(a_i)=\frac{e^{Q^k(s,a_i)/T}}{\sum_{a\in A}e^{Q^k(s,a)/T}} \tag{6}$$

式中：T 为温度因子，其值随着学习进行而逐渐衰减，$P_s^k(a)$代表第 k 次迭代时状态 s 下选择动作 a 的概率。在经过足够迭代次数的探索和利用之后，Q^k 将会以概率 1 收敛于最优值函数 Q^*，最终得到一个 Q^*矩阵表示的最优控制策略。

2.2 基于强化学习的 CPS 自校正控制的结构

互联电网 AGC 系统中的 CPS 指标控制过程是一个动态多级决策问题，强化学习算法正是以马尔可夫决策过程进行建模研究。CPS 指标不仅是一个对互联电网的 AGC 性能的奖惩考核指标，也可以看作衡量一个电力系统控制品质好坏的一个重要“环境指标”。因此，将 CPS 指标作为强化学习的奖惩依据是很恰当合理的。

现有 CPS 控制系统一般采用以 ACE 和ΔF 作为输入的 PI 控制结构，本文提出一种在当前 CPS 控制器基础上实现基于强化学习的 CPS 自校正控制结构，如图 1 所示。

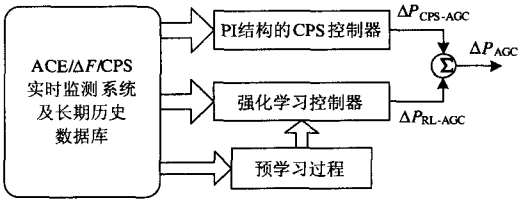

图 1　基于强化学习的 CPS 自校正控制结构

Fig. 1　PI/RL self-tuning CPS control structure

（1）CPS 控制器。对电网 CPS1 和 CPS2 指标的实时控制。一般采用协调良好的 PI 负反馈控制结构及其改进措施，如南方电网目前使用 NARI 提出的 CPS 调节功率分量与比例积分控制分量相结合构成区域总调节分量[8, 9]，在最新 CPS 考核标准下对 AGC 控制策略进行了全面改进。

（2）强化学习控制器。作为附加校正环节由不断变化的输入信号来对 CPS 控制器的输出进行动态补偿。该控制器在运行过程一直处于在线学习状态，在电网结构、参数或运行方式等发生变化时，RL 控制器仍可进行在线自学习以调整最优控制策略，从而根据强化学习机理在线优化功率调节指令以增强 AGC 系统的自适应和自学习能力。

（3）预学习过程。强化学习经过一定时间的反复试错学习后，即可称为 PI/RL 控制器。RL 控制器在应用时需要经过一段随机动作探索的预学习过程，这种预学习是通过 RL 控制器与 CPS 控制器组合体相互作用来实现的。预学习过程完毕后，RL 控制器便进入平稳在线学习阶段，控制指令不会出现初期学习过程中的大幅随机输出指令现象。

（4）ACE/ΔF/CPS 实时监测系统及长期历史数据库。主要用来实时采集监视电网当前的 ACE、ΔF 和 CPS1 等瞬时值和平均值，其数据作为 PI/RL 各子控制器提供系统的状态反馈量。同时记录和统计每日、每月和每年的 CPS 完成率，并存入长期历史数据库中。

2.3 RL 控制器的设计

基于 Q 学习算法设计自校正控制器，首先需要分析系统特性以确定状态集合 S 和动作集合 A。集合 S 即马尔可夫链状态空间，本控制策略根据 CPS1 和 ACE 值来构造状态集合 S。先将 CPS1 划分为三个不同状态，即$(-\infty, 100)$、$[100, 200)$、$[200, +\infty)$，ACE 的符号即 ACE_{sign} 的值为$\{-1, 0, 1\}$，分别代表 ACE 值为负、0 和正，其中 ACE_{sign} 的作用是为了区分引起 CPS1 指标波动的原因，此二维输入将状态空间分为 9 个不同状态。A 集合为一系列离散的功率调节指令ΔP，如何量化动作信号ΔP 值需视系统机组容量而定。评价奖惩函数 $R(s,s',a)$由第 k 步时刻的立即强化信号 r 由被控变量 CPS1 或 ACE 的方差以及相应动作变化量平方的比例项来确定，即

$$\begin{cases} R=0 & CPS1\geq 200 \\ R=-[\lambda_1(ACE)^2+\mu_1(a_k-a_{k-1})^2] & 100\leq CPS1<200 \\ R=-[\lambda_2(CPS1-CPS1^*)^2+\mu_2(a_k-a_{k-1})^2] & CPS1<100 \end{cases} \tag{7}$$

该 RL 自校正控制器的设计建立在 PI 控制基础上，主要针对 CPS1 和 CPS2 指标进行评价。CPS1 指标在 200(%)以上总是合格的，故评价函数 R 赋最高值 0；由于 CPS 考核中，当 100%<CPS1<200%时主要对 CPS2 指标进行考核，这时评价函数 R 用 ACE 值来确定；而 CPS1 值在 100(%)以下时考核是不合格的，评价函数此时则针对 CPS1 值进行惩罚。在这里 $CPS1^*$为设定的理想值，本文算例中取 200(%)；引入动作变化项是为了限制功率控制信号的波动，以免控制器输出功率指令频繁大幅度升降引起的系统振荡。方程中 a_k 值是动作集 A 的指针，而不是实际的输出值，λ_1、λ_2 和 μ_1、μ_2 分别为式(7)中平衡前后各平方项的权重值。

综上，强化学习控制器的学习过程是由一系列训练样本经验来实现的，其基于 Q 学习算法的 CPS 自校正学习具体步骤如下，其中第 3 步到第 9 步重复执行既定的次数称为学习周期(episode)。

①初始化各参数，令 k=0；

②观察当前状态 s_0，即 $CPS1_0$ 与 ACE_0；

③根据动作概率分布选择并执行一个动作 a_k；

④观察下一时刻的状态 s_{k+1}；

⑤由式(7)得到一个立即强化信号 r_k；

⑥根据式(5)更新 Q 矩阵；

⑦根据式(6)更新动作概率分布；

⑧k=k+1，返回步骤 3。

3　仿真研究

3.1 标准两区域互联系统的仿真研究

以典型的 IEEE 两区域互联系统的负荷频率控制(LFC)模型作为研究对象，如图 2 所示。其中系统 A 与系统 B 模型结构及参数完全一样，所用 CPS 控制策略采用传统 PI 控制结构，并分别附加强化学习自校正控制器进行仿真研究。系统模型相关参数取自文献[15]，见表 1，系统基准容量为 5 000 MW。

表 1　两区域互联系统模型参数

Tab. 1　System parameters for the two-area LFC model

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

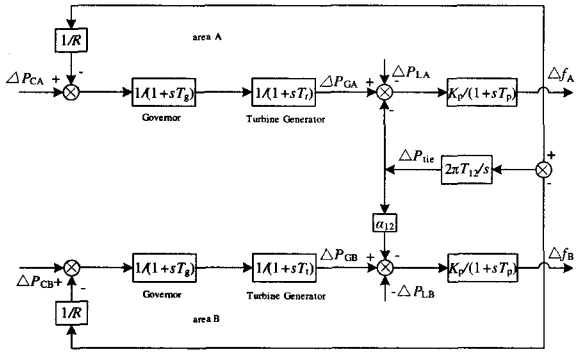

图 2　两区域互联系统负荷频率控制模型

Fig. 2　The two-area power system LFC model

算例中 RL 控制器以 CPS1 和 ACE 作为输入，作为校正控制环节且考虑到控制系统学习收敛速度，该 RL 控制器所允许的输出动作离散集为：A={-500, -300, -100, 0, 100, 300, 500}，单位是 MW，被选择的动作指令与 PI 控制器的输出信号相加。学习步长一般取 AGC 系统控制周期，该标准算例中取 2 s。文中使用 Simulink 仿真工具进行建模研究，RL 算法和控制器由 S-function 模块编写。

本文采用有规律的方波负荷扰动进行试错学习，以此进行 RL 控制器的预学习，典型学习收敛过程如图 3 所示。所选固定负荷扰动发生在 A 区域，强化学习控制器在初始阶段的 15 000 s 左右都属于预学习过程。图 3(c)和图 3(d)给出了 PI/RL 控制器与仅由 PI 控制器下以 10 min 为考核时段的 CPS1 和 CPS2 平均值曲线，经过大约 7 000 次迭代学习，就可以明显看到 RL 控制器提高 CPS1 和 CPS2 考核指标的效果。随着进一步学习训练，预学习过程结束时 RL 控制器会逼近一个确定性 CPS 控制策略。上述学习过程的负荷扰动选择在本区域，若选择外网即 B 区域的负荷扰动，方法与 A 区域的扰动类似。

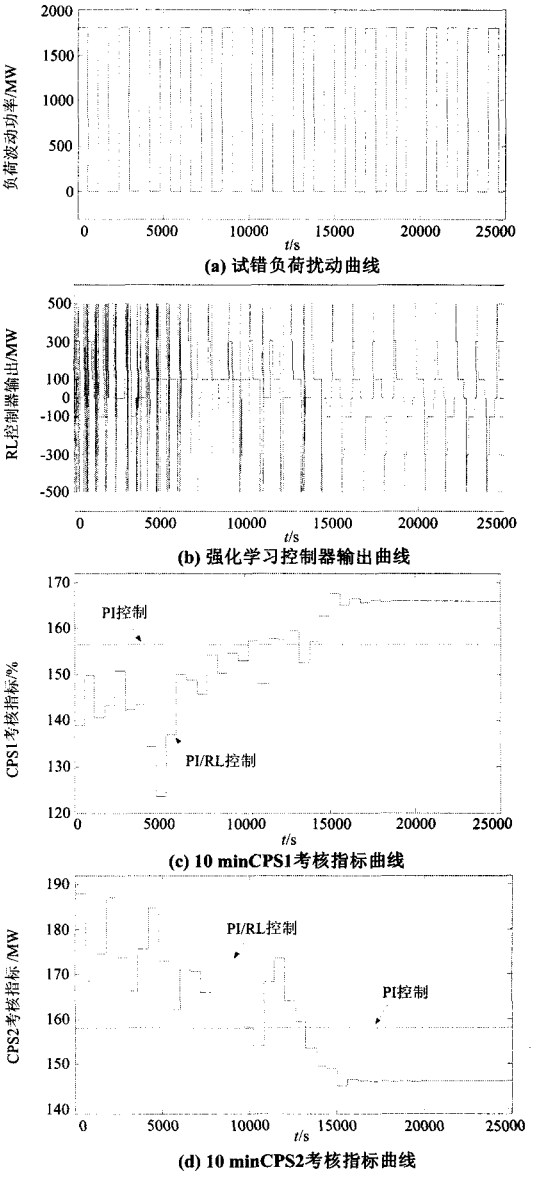

图 3　RL 控制器的学习过程

Fig. 3　Learning procedure of RL controller

在一系列预学习过程完毕后，则可将 RL 控制器与原有的 PI 控制器构成 PI/RL 控制器投入正常运行。算例中通过引入几组较大扰动样本进行仿真验证，仿真研究将 PI/RL 控制器与单一的 PI 控制器进行动态性能比较，如图 4 所示。图 4(b)给出了 PI/RL 控制器下各个子控制器的输出，可以看出 RL 控制器已经“学会了”在大部分时间保持不变，即输出为 0，只有在设定值(CPS1 和 CPS2)发生大幅变化时，RL 控制器输出就会有较大变动来校正 PI 控制器的输出。图 4(c)中该控制方法较单一的 PI 控制对 CPS1 指标有明显的提高，以 CPS1 一分钟平均值为例，PI 控制下出现了 5 次 CPS1 值在 100 以下，而 PI/RL 控制下

只有 1 次。图 4(d)和图 4(e)显示了 PI/RL 控制方法在 CPS2 指标和系统频率方面具有更好的动态特性，同时有利于提高 CPS2 以及整个 CPS 的考核合格率。

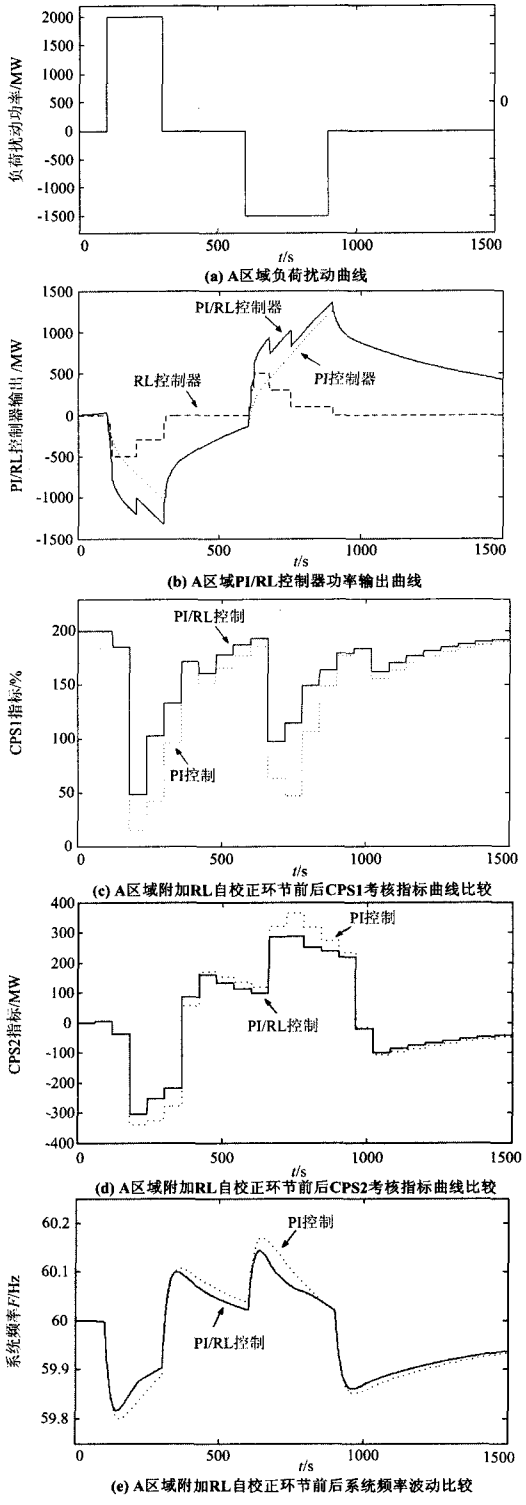

图 4　基于 RL 的 CPS 自校正控制仿真试验

Fig. 4　Simulation test of PI/RL self-tuning control

3.2 南方电网实例仿真研究

为结合实际电网进一步研究强化学习自校正控制机理，仿真对象选择南方电网。在联络线功率与系统频率偏差(TBC)模式下，结合 NARI 所提出的 AGC 分层控制和 CPS 控制策略作为研究目标[8,9]。应用广东省电力调度中心已用的 AGC 系统模型，所推荐的 RL 自校正控制器设在广东电网，其它省电网仍用原有 CPS 控制策略。RL 控制器的设计同两区域互联系统模型，学习步长即 AGC 控制周期为 4 s，输出动作离散集为 A={−1000, −600, −300, −100, −50, 0, 50, 100, 300, 600, 1000}。

在经过足够迭代次数的预学习之后，我们分别在广东电网和外网加以周期性负荷扰动进行对照性统计仿真试验。广东电网内负荷扰动为周期 25 min 的正负波动的方波，正方波幅值为 1 500 MW、脉宽为 7.5 min，负方波幅值为−1000 MW、脉宽为 5 min。外网负荷扰动为周期 5 min、幅值 500 MW、脉宽 2 min 的方波负荷。在一天 24 h 内，以 10 min 为 CPS 考核时段的指标汇总表见表 2。其中，$|\Delta F|$、|ACE|、CPS1 为平均值，CPS2、CPS 为考核合格率百分数，CPS2 考核标准阀限值 L_{10} 为 288 MW。由基于强化学习的 CPS 自校正控制在南方电网的仿真统计试验可知，该方法较现有 CPS 控制策略对于 CPS 各指标都有较为明显的改善。

表 2　广东电网仿真试验 CPS 指标对照表

Tab. 2　The case study of PI/RL self-tuning control in Guangdong Power Grid

指标	PI 控制	PI/RL 控制
$\|\Delta F\|$/Hz	0.059 6	0.043 1
|ACE|/MW	281.505 6	207.497 5
CPS1	112.601 3	149.114 3
CPS2/(%)	86.61%	94.17%
CPS/(%)	66.67%	83.34%

4　结论

综上，基于强化学习的互联电网 CPS 自校正控制方法具有以下特点：

（1）RL 控制器的设计不依赖于电网模型，其在线自学习的特性非常适合于多变量、非线性、运行工况随负荷时刻变化的现代 AGC 系统，可有效提高 AGC 控制系统的鲁棒性和适应性，有利于提高互联电网 CPS 标准下的考核合格率。

（2）在现有 PI 控制结构的 CPS 控制系统的基础上添加一个强化学习自校正控制环节，无需对原有 AGC 控制策略进行更改，控制策略和工程实现十分简便。

（3）RL 算法是多智能体系统（MAS）的一个重要基础，应用 RL 算法有利于推动基于 MAS 架构的 AGC 系统的构建。

在研究中笔者也发现：（1）Q 学习算法中的奖励信号来自于 CPS 指标，若引入更为广泛的节能和经济调度指标形成综合奖励信号，可获得更佳的 AGC 控制效果。（2）RL 控制器的控制动作离散集区间较大，较易形成过调，后续研究中应考虑采用模糊控制方法对输入输出信号模糊化，并通过模糊推理来确定强化信号。

参考文献

[1] Jaleeli N，Vanslyck L S．NERC's Mew Control Performance Standards[J]．IEEE Trans on Power Systems，1999，14(3)：1091-1099.

[2] Yao M，Shoults R R，Kelm R．AGC Logic Based on NERC's New Control Performance Standard and Disturbance Control Standard[J]．IEEE Trans on Power Systems，2000，15(2)：855-857.

[3] Feliachi A，Rerkpreedapong D．NERC Compliant Load Frequency Control Design Using Fuzzy Rules[J]．Electric Power Systems Research，2005，73(1)：101-106.

[4] Makarov Y，Hawkins D．New AGC Algorithms[A]．In:Erican EPRI Infrastructure Integration & Markets Product Line Council Meeting[C].California(USA)：2002.

[5] 李正，敬东，赵强，等．CPS/DCS 标准在大区互联电网 AGC 控制策略中的应用[J]．电力系统及其自动化学报，2003，15(12)：27-32，48.
LI Zheng，JING Dong，ZHAO Qiang，et al．Application of AGC Control Strategy Based on CPS/DCS Standard in Inter-connectioned Power Grid [J]．Proceedings of the EPSA，2003，15(12)：27-32，48.

[6] 唐悦中，张王俊．基于 CPS 的 AGC 控制策略研究[J]．电网技术，2004，28(21)：75-79.
TANG Yue-zhong，ZHANG Wang-jun．Research on Control Performance Standard Based Control Strategy for AGC[J]．Power System Technology，2004，28(21)：75-79.

[7] 余涛，陈亮，蔡广林．基于 CPS 标准统计信息自学习机理的 AGC 自适应控制[J]．电机工程学报，2008，28(13)：45-49.
YU Tao，CHEN Liang，CAI Guang-lin．CPS Statistic Information Self-learning Methodology Based Adaptive Automatic Generation Control[J]．Proceedings of the CSEE，2008，28(13)：45-49.

[8] 高宗和，滕贤亮，涂力群．互联电网 AGC 分层控制与 CPS 控制策略[J]．电力系统自动化，2004，28(1)：78-81.
GAO Zong-he，TENG Xian-liang，TU Li-qun．Hierarchical AGC Mode and CPS Control Strategy for Interconnected Power Systems[J]．Automation of Electric Power Systems，2004，28(1)：78-81.

[9] 高宗和，滕贤亮，张小白．互联电网 CPS 标准下的自动发电控制策略[J]．电力系统自动化，2005，29(19)：40-44.
GAO Zong-he，TENG Xian-liang，ZHANG Xiao-bai．CPS Control Strategy for Interconnected Power Systems[J]．Automation of Electric Power Systems，2005，29(19)：40-44.

[10] Sutton R S，Barto A G．Reinforcement Learning：an Introduction[M]．Cambridge：MIT Press，1998.

[11] Mine H，Osaki S．Markovian Decision Processes[M]．New York：Eisevier，1970.

[12] 张汝波．强化学习理论及应用[M]．哈尔滨：哈尔滨工程大学出版社，2001.

[13] Watkins J C H，Dayan Peter．Q-learning[J]．Machine Leaning，1992，8：279-292.

[14] Tsitsiklis，John N．Asynchronous Stochastic Approximation and Q-learning [J]．Machine Leaning，1994，16(3)：185-202.

[15] Ray G，Prasad A N，Prasad G D．A New Approach to the Design of Robust Load Frequency Controller for Large Scale Power Systems[J]．Electric Power System Research，1999，51：13-22.

收稿日期：2008-06-27

作者简介：

余　涛(1974-)，男，副教授，博士，主要研究方向为复杂电力系统的非线性控制理论和仿真研究；

周　斌(1984-)，男，硕士研究生，主要研究方向为电力系统优化控制方法。E-mail:zhou.bin@mail.scut.edu.cn