文章编号：0258-8013 (2009) 19-0013-07　　中图分类号：TM 732　　文献标志码：A　　学科分类号：470·40

# 基于 Q 学习的互联电网动态最优 CPS 控制

余涛[1]，周斌[1]，陈家荣[2]

(1. 华南理工大学电力学院，广东省 广州市 510640；2. 香港理工大学电机工程系，中国香港特别行政区)

# Q-learning Based Dynamic Optimal CPS Control Methodology for Interconnected Power Systems

YU Tao[1], ZHOU Bin[1], CHAN Ka-Wing[2]

(1. College of Electric Power, South China University of Technology, Guangzhou 510640, Guangdong Province, China;
2. Department of Electrical Engineering, Hong Kong Polytechnic University, Hong Kong SAR, China)

**ABSTRACT:** The NERC's control performance standard (CPS) based automatic generation control (AGC) problem is a stochastic multistage decision problem, which can be suitably modeled as a reinforcement learning (RL) problem based on Markov decision process (MDP) theory. The paper chose the Q-learning method as the RL algorithm regarding 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. An applicable semi-supervisory pre-learning method was introduced to enhance the stability and convergence ability of Q-learning controllers. Two cases show that the proposed controllers can obviously enhance the robustness and adaptability of AGC systems while the CPS compliances are ensured.

**KEY WORDS:** automatic generation control; Q-learning; Markov decision process; control performance standard; optimal control

**摘要：**控制性能标准(control performance standard，CPS)下互联电网自动发电控制(automatic generation control，AGC)系统是一个典型的不确定随机系统，应用基于马尔可夫决策过程(Markov decision process，MDP)理论的 Q 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。将 CPS 值作为包含 AGC 的电力系统“环境”所给的“奖励”，依靠 Q 值函数与 CPS 控制动作形成的闭环反馈结构进行交互式学习，学习目标为使 CPS 动作从环境中获得的长期积累奖励值最大。提出一种实用的半监督群体预学习方法，解决了 Q 学习控制器在预学习试错阶段的系统镇定和快速收敛问题。仿真研究表明，引入基于 Q 学习的 CPS 控制可显著增强整个 AGC 系统的鲁棒性和适应性，有效提高了 CPS 的考核合格率。

**关键词：**自动发电控制；Q 学习；马尔可夫决策过程；控制性能标准；最优控制

## 0　引言

CPS标准为北美电力可靠性委员会于 1997 年正式推出的联络线功率与系统频率偏差模式下互联电网自动发电控制的最新控制性能标准。2001 年后我国各电网均开始试行CPS考核标准。CPS标准注重中长期AGC性能指标，完整的CPS指标由CPS1和CPS2 指标组成，国内一般采用 10 min考核周期，其详细数学描述见文献[1]。CPS标准更注重AGC系统的长期收益，从根本上改变了传统AGC的控制思想，如何设计适应CPS标准下AGC系统的快速动态优化控制策略成为一个全新的理论研究课题。现有关于CPS控制策略的设计多数为经典PI控制结构[2-5]，其中我国南瑞集团高宗和等学者在CPS控制工程实用化方面做出了重要贡献。文献[6]引入模糊控制原理对CPS控制策略进行了研究，在满足CPS合格率的前提下，减少了机组调节损耗。传统PI控制和模糊控制可保证对受控对象存在的模型不确定性具有较高的鲁棒性，但在最优化设计方面还存在一定欠缺。

实际上，CPS标准下的互联电网AGC系统应被看作一个“不确定的随机系统”，数学模型为高斯–马尔可夫随机过程模型[7]更恰当。要深入揭示其基础规律，运用随机系统的最优控制理论是可行途径，基于最优随机控制理论的预测控制和基于马尔

**基金项目：**国家自然科学基金项目(50807016)；中国香港特别行政区研究资助局项目(RGC No. PolyU G-U494)；广东省自然科学基金项目资助(06300091)。

Project Supported by National Natural Science Foundation of China (50807016).

可夫决策理论的Q学习控制都是值得尝试的思路。文献[8]所提出的“Wedge-Shaped”控制规律[9]与模型预测控制方法相结合，实现了一种新型CPS优化控制策略，有效减少了机组反调次数。文献[10]引入自适应控制理论实现了对文献[5]提出的CPS控制器增益的自动调整，弥补了PI控制在适应性和鲁棒性上的不足，并可实现对AGC机组在“放松控制”和“收紧控制”2 个方向的自适应调整。文献[11]提出了一种基于传统PI控制与Q学习控制混合的CPS 自校正控制方法，提高了文献[10]控制策略的在线学习和动态优化能力。

在文献[11]基础上，本文进一步从理论上完善基于 Q 学习的最优 CPS 控制策略，并提出一种实用的半监督群体预学习方法，解决Q学习控制器在预学习贪婪试错阶段的系统镇定和快速收敛问题。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Q学习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 AGC 控制系统鲁棒性和适应性的同时，提高互联电网 CPS 考核合格率。

## 1 控制原理

### 1.1 Q 学习算法

Q学习以离散时间马尔可夫决策过程(discrete time Markov decision process，DTMDP)为数学基础，与监督学习、统计模式识别和人工神经网络不同，不需要精确的历史训练样本及系统先验知识，是一种基于值函数迭代的在线学习和动态最优技术[12]。Q学习算法通过直接优化一个可迭代计算的状态–动作对值函数$Q(s,a)$，在线寻求最优策略使得期望折扣报酬总和最大。Q学习的值函数满足下式：

$$Q(s,a)=R(s,s',a)+\gamma\sum_{s'\in S}P(s'|s,a)\max_{a\in A}Q(s',a) \quad (1)$$

式中：$s$、$s'$分别为当前状态和下一时刻的状态；$P(s'|s,a)$为状态 $s$ 在控制动作 $a$ 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，即奖励函数值；$0<\gamma<1$，为折扣因子。文献[13-14]给出了 Q 学习算法中各参数选取的一般原则，根据 CPS 标准下的 AGC 控制目标，为体现下一个控制动作的奖励函数值对于折扣报酬总和值函数 $Q(s,a)$的重要性，所选$\gamma$值应接近 1。通过仿真研究可知，$\gamma$值在 0.8~0.98 都有较优的控制效果，本文取值为 0.9。

Q学习算法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，控制器或智能体通过此次试探学习获得的经验即$[s_k, a, r, s_{k+1}]$样本，更新Q值迭代公式如下：

$$\begin{cases}Q^{k+1}(s_k,a_k)=Q^k(s_k,a_k)+\alpha[R(s_k,s_{k+1},a_k)+\\ \qquad\qquad \gamma\max\limits_{a'\in A}Q^k(s_{k+1},a')-Q^k(s_k,a_k)]\\ Q^{k+1}(\tilde{s},\tilde{a})=Q^k(\tilde{s},\tilde{a}),\ \forall(\tilde{s},\tilde{a})\neq(s_k,a_k)\end{cases} \quad (2)$$

式中 $0<\alpha<1$，称为学习因子。$\alpha$指明了要给改善的更新部分多少信任度，较大的$\alpha$值会加快学习算法的收敛速度，而较小的$\alpha$值能保证控制器的搜索空间，从而提高Q学习收敛的稳定性。考虑到负荷扰动的随机性，所选$\alpha$值应接近 0。仿真研究显示，$\alpha$值在 0.001~0.1 范围内都具有良好的收敛特性，文中算例取值为 0.01。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。强化学习面临着探索和利用的权衡问题，定义控制器在当前状态下总是选择具有最高$Q$值的动作，称为贪婪策略$\pi^*$，如下式：

$$\pi^*(s)=\arg\max_{a\in A}Q^k(s,a) \quad (3)$$

但是总是选择最高 $Q$ 值的动作会导致智能体总是沿着相同的路径，并未充分搜索空间中的其他动作，往往收敛于局部最优。

本文采用一种基于概率分布选择动作的追踪算法[14]来构造动作选择策略。该策略在学习初始阶段，控制器从随机开始选择动作，即初始化使得各状态下任意可行动作被选择的概率相等；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式(4)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。

$$\begin{cases}P_s^{k+1}(a_g)=P_s^k(a_g)+\beta[1-P_s^k(a_g)]\\ P_s^{k+1}(a)=P_s^k(a)(1-\beta),\forall a\in A,a\neq a_g\\ P_{\tilde{s}}^{k+1}(a)=P_{\tilde{s}}^k(a),\forall a\in A,\forall\tilde{s}\in S,\tilde{s}\neq s\end{cases} \quad (4)$$

式中 $0<\beta<1$，$\beta$值的大小决定了动作搜索的速度，$\beta$值越接近 1 说明控制动作策略越趋于贪婪策略。仿真比较研究显示，$\beta$值在 0.3~0.6 范围内都能很好地平衡 Q 学习控制器的动作搜索与经验强化问题。本

文算例$\beta$取值为 0.5。$P_s^k(a)$ 为第 $k$ 次迭代时状态 $s$ 下选择动作$a$的概率；$a_g$为由式(3)得到的贪婪动作策略。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。

### 1.2　基于 Q 学习的最优 CPS 控制原理

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，CPS 指标不仅是一个对互联电网 AGC 长期性能的奖惩考核指标，也可以看作衡量电力系统控制品质好坏的一个重要“环境指标”。基于 Q 学习的控制系统通过试错与环境进行交互式学习，从长期的观点构造控制策略，以期从环境获得的长期积累奖励值最大，与 CPS 控制的长期收益最大的特性十分吻合，因此，融入 CPS 指标作为 Q 学习的奖励函数很恰当合理。

本文提出基于 Q 学习算法的互联电网动态最优 CPS 控制系统示意图如图 1 所示。

由图 1 可知，任$i$区域电网的“Q学习控制器”的输入为来自“ACE/Δ$F$/CPS实时监测数据库及长期历史记录数据库”当前系统环境的状态量(State)及计算出的奖励值(Reward)，Q学习控制器则实现在线学习和给出最优控制信号，控制动作为该区域电网调度端所下达的AGC总调节指令$\Delta P_{\text{ord-Q-i}}$。为了说明CPS控制器设计的过程，本文简单归纳了CPS标准的考核方法，其CPS状态相空间图如图 2 所示：

1）$C_{\text{CPS1}}\geq 200\%$，$C_{\text{CPS2}}$为任意值，则CPS指标合格。

2）$100\%\leq C_{\text{CPS1}}<200\%$，$C_{\text{CPS2}}$合格，则CPS指标合格。

3）$C_{\text{CPS1}}<100\%$，CPS指标不合格。

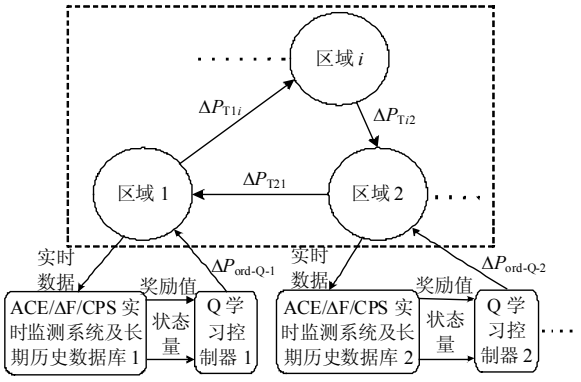



**图 1　基于 Q 学习的互联电网最优 CPS 控制示意图**

**Fig. 1　Q-learning based optimized CPS control structure**

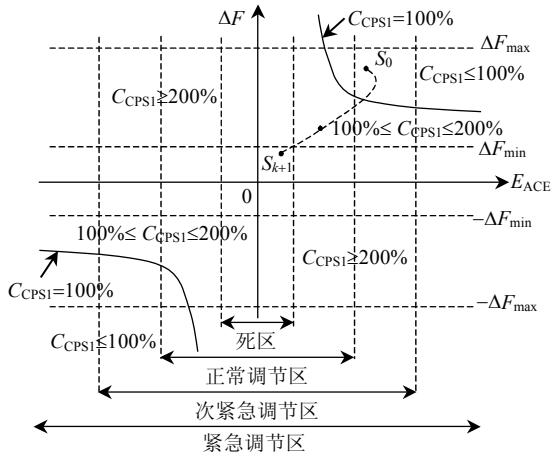



**图 2　CPS 控制区域的示意图**

**Fig. 2　Schematic diagram of CPS control area**

因此，可根据CPS考核标准来定义某$i$区域电网的奖励函数$R_i$如下：

$$\begin{cases} R_i(k)=\sigma_i, \quad \sigma_i\geq 0, & C_{\text{CPS1}i}(k)\geq 200 \\ R_i(k)=-\{\lambda_{1i}[E_{\text{ACE}i}(k)-E^*_{\text{ACE}i}]^2+\mu_{1i}[a_{\text{ord}-i}(k)- \\ \qquad a^*_{\text{ord}-i}]^2\}, & C_{\text{CPS1}i}(k)\in[100,200) \\ R_i(k)=-\{\lambda_{2i}[C_{\text{CPS1}i}(k)-C^*_{\text{CPS1}i}]^2+\mu_{2i}[a_{\text{ord}-i}(k)- \\ \qquad a^*_{\text{ord}-i}]^2\}, & C_{\text{CPS1}i}(k)<100 \end{cases} \tag{5}$$

式中：$\sigma_i$为任意非负数，本文取 0；$C_{\text{CPS1}i}(k)$与$E_{\text{ACE}i}(k)$分别为CPS1 和ACE在第$k$步迭代时刻的瞬时值；$a_{\text{ord-}i}(k)$为$k$时刻的控制动作集$A$的指针；$a^*_{\text{ord}-i}$即功率控制动作为 0 时的指针，引入动作变化项，是为了限制控制器输出功率指令频繁大幅度升降引起的系统振荡和经济代价；$\lambda_1$、$\lambda_2$和$\mu_1$、$\mu_2$分别为状态输入和控制动作的优化权值，其意义相当于线性二次型调节器(linear quadratic regulator，LQR)控制性能指标中的Q和R权值参数[15]；$C^*_{\text{CPS1}i}$为CPS1指标控制期望值，若追求 CPS 高合格率则可取200%，若对电网实施松弛控制则可选 CPS1 的日或月平均值；$E^*_{\text{ACE}i}$为 ACE 控制期望值，从提高 CPS2 指标、减少无意交换电量和避免 ACE 频繁过零的角度，可取 ACE 调节死区值。

基于DTMDP模型的Q学习算法设计CPS控制器需要分析系统特性，以确定状态空间离散集$S$和控制动作集$A$。如果离散化程度太细，会使$\boldsymbol{Q}$矩阵维数过高且导致AGC频繁反复发送调节指令，这与CPS标准减轻AGC机组压力的初衷相悖，反之，减少状态动作维数和离散化类别又不利于改善CPS指标，因此合理安排状态与动作空间离散化十分重要。集合$S$即马尔可夫链状态空间，控制器以区域电网$C_{\text{CPS1}}$和$E_{\text{ACE}}$值作为输入变量构造状态空间集

$S$，本文根据CPS考核特点[16]对图 2 所示的CPS状态相空间进行分区，先将$C_{\mathrm{CPS1}}$划分为$(-\infty, 0)$,[0,100],[100,105],[105,110],[110,115],…[185, 190],[190, 195],[195, 200],[200, $+\infty$]，再以$E_{\mathrm{ACE}}$为横坐标判断第 1、3 象限，则二维输入空间被量化为 46 个不同状态。控制动作集合即为一组离散的AGC功率调节指令$\Delta P_{\mathrm{ord}}$，如何量化动作信号$\Delta P_{\mathrm{ord}}$值需视系统各类型机组容量而定，具体如何选取可参考下文的仿真算例。

在确定了奖励函数、输入状态空间和控制动作集后，即可进行 Q 学习控制器在线自学习与动态优化，其步骤如下：

1）初始化各参数，令 $k$=0。

2）观察当前状态 $S$(0)。

3）由动作概率分布在控制集中选择动作 $a(k)$。

4）观察下一时刻的状态 $S(k+1)$。

5）由式(5)得到一个奖励信号 $R(k)$。

6）根据式(2)更新 $\boldsymbol{Q}$ 矩阵。

7）按照式(3)计算贪婪动作$a_g(k)$。

8）根据式(4)更新动作概率分布。

9）$k=k+1$，返回步骤 3）。

### 1.3 Q 学习控制器的半监督群体预学习方法

事实上，根据 1.2 节设计的Q学习控制器是无法直接投入到真实环境中运行的，原因是Q学习初期阶段会进行大量盲目的试错学习，会导致控制系统不稳定，这不仅危害到实际系统的安全稳定性，还会导致Q学习算法由于无法寻找到系统稳定和动态优化的搜索路径而长期无法收敛，因此对Q学习控制器进行预学习必不可少[17]。

对于含有多个 Q 学习控制器的互联电网最优 CPS 控制问题，存在群体预学习问题，因此，本文提出了一种半监督群体预学习方法：

1）搭建一个受控对象的数字仿真系统，用仿真系统代替真实环境，所有区域电网的 CPS 控制器均采用 PI 控制，并获得一个稳定的仿真环境。

2）选择需要预学习的某一个 Q 学习控制器，与本区域原 PI 控制构成一种附加控制结构，如图 3 所示，并按 1.2 节所介绍方法进行试错学习，直至控制系统收敛稳定。

3）采用 PI 增益参数线性递减原则，逐步让 Q 学习控制器在线学习，减少对 PI 镇定的依赖，最终获得一个纯 Q 学习控制器(即 PI 增益参数全为零)。

4）重复 2）和 3）步，逐个获得整个互联电网所有区域电网纯 Q 学习控制器结构。

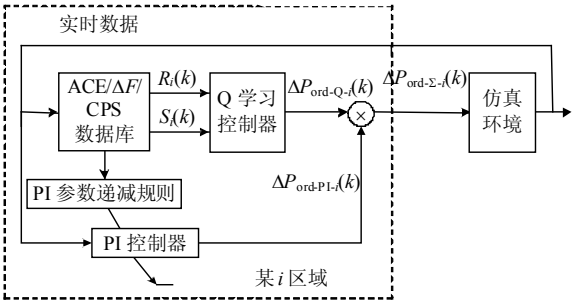



**图 3 半监督群体预学习过程示意图**

**Fig. 3 Semi-supervisory group pre-learning procedure**

5）进行受控对象标称参数模型的群体预学习，获得满意结果后结束预学习过程。

6）预学习结束后，保留当前的 $\boldsymbol{Q}$ 值矩阵和 $\boldsymbol{P}$ 概率矩阵数值，即可将所有 Q 学习控制器投入真实环境运行。

由于在步骤 2）至 4）中需要 PI 控制器来进行辅助镇定和矫正控制系统，所以称此预学习方法为半监督群体预学习方法。

## 2 仿真算例研究

### 2.1 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模型[17]作为研究对象，结构框图如图 4，系统模型相关参数见表 1，系统基准容量取 5 000 MW。本文使用Simulink进行建模仿真研究，其中Q学习算法和控制器由S-function模块编写。算例中Q学习控制器以$C_{\mathrm{CPS1}}$和$E_{\mathrm{ACE}}$实时值作为状态输入，控制器输出动作离散集为：$A$={−500, −300, −100, −50, −20, −10, −5, 0, 5, 10, 20, 50, 100, 300, 500}MW。学习步长一般为AGC控制周期，标准算例中取 3 s。

应用本文所提出的半监督群体预学习方法分别对 A 区域和 B 区域的 Q 学习控制器进行预学习。

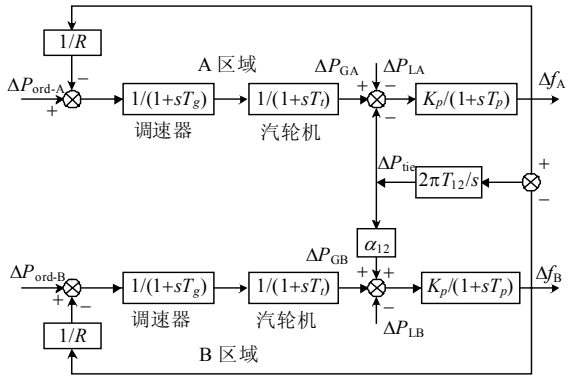



**图 4 两区域互联系统负荷频率控制模型**

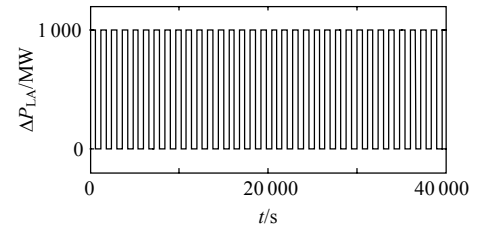
**Fig. 4 Two-area power system LFC model**

**表 1　两区域互联系统模型参数**

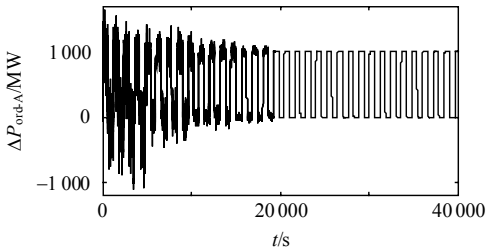
**Tab. 1　System parameters for the two-area LFC model**

| $T_g$/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 |

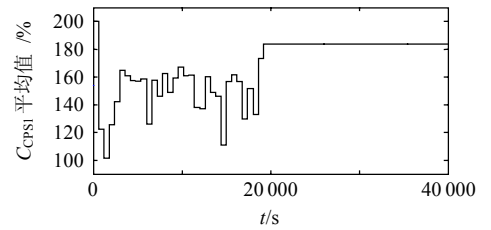
图 5 给出了区域AQ学习控制器的典型学习收敛过程，所选周期为 10 min的连续阶跃负荷扰动发生在区域A，扰动波形如图 5(a)所示。由图 5(b)可见，在经历约 20 000 s后，Q学习控制器的输出已经接近负荷扰动。图 5(c)和 5(d)为$C_{CPS1}$和$E_{ACE}$的 10 min平均值在学习过程的变化曲线，图中显示CPS1 和CPS2考核指标也将趋向一个稳定值，这说明Q学习控制器已逼近一个确定性最优CPS控制策略。在两区域控制器均完成了足够迭代次数的预学习后(即$\boldsymbol{Q}$矩阵已接近于最优值$\boldsymbol{Q}^*$)[12]，则可将Q学习控制器投入真实环境运行。



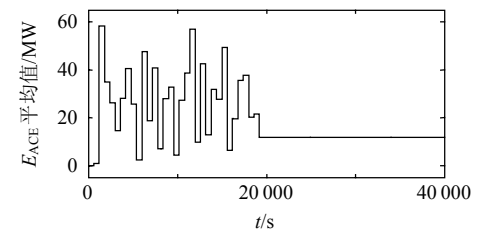

(a) 负荷扰动




(b) Q 学习控制器输出曲线




(c) 区域 A 10 min CPS1 考核指标曲线




(d) 区域 A10 min ACE 平均值指标曲线

**图 5　Q 学习控制器的预学习过程**

**Fig. 5　Learning procedure of Q learning controller**

为了进一步说明不同奖励函数对 Q 学习控制器性能的影响，本文给出了奖励函数式(5)中的几组典型权值参数取值：

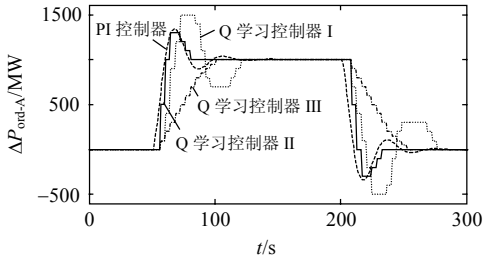
1）Q学习控制器Ⅰ：$\lambda_1$=1、$\lambda_2$=50、$\mu_1$=1、$\mu_2$=1。

2）Q学习控制器Ⅱ：$\lambda_1$=1、$\lambda_2$=50、$\mu_1$=10、$\mu_2$=10。

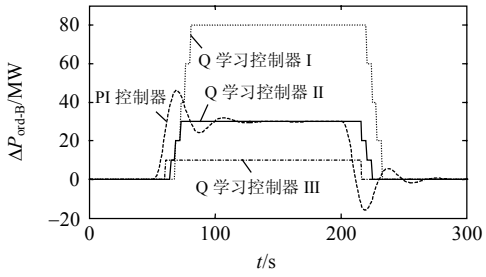
3）Q学习控制器Ⅲ：$\lambda_1$=1、$\lambda_2$=50、$\mu_1$=50、$\mu_2$=50。

其中，区域A和区域B两控制器的奖励函数采用相同权值系数；$\lambda_2/\lambda_1$权值比为定值，由于$C_{CPS1}$与$E_{ACE}$量纲上存在差别，为使得式(5)各分段上奖励函数保持一致性，从而保证学习过程中奖惩的公平性，本算例取值 50。引入方波扰动模拟系统故障停机和甩负荷，结果如图 6 所示。由仿真可知，不同奖励函数对Q学习控制器的优化结果影响十分显著，选择恰当的奖励函数(Q学习控制器Ⅱ)可以获得与良好整定的PI控制器性能相当的控制效果。

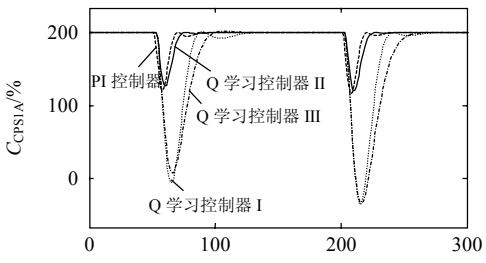
另一方面，$\lambda_1$、$\lambda_2$和$\mu_1$、$\mu_2$与LQR性能指标函数中的$Q$和$R$权值的意义十分类似，由LQR理论[18]与仿真试验分析可得出结论：权值$\lambda$和$\mu$的相对关系反映了CPS控制过程中，对$C_{CPS1}$和$C_{CPS2}$指标偏差和AGC机组调节压力之间的侧重度；当比值$\lambda/\mu$降低时，控制动作输出被降低，AGC系统



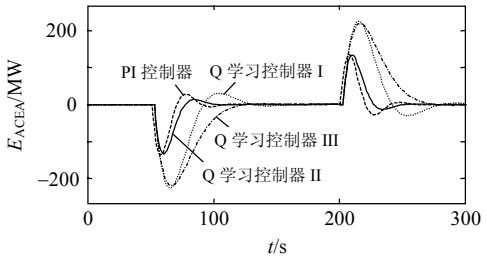

(a) 区域 AQ 学习控制器输出曲线




(b) 区域 BQ 学习控制器输出曲线




(c) 区域 A CPS1 变化曲线




(d) 区域 A ACE 变化曲线

**图 6 基于 Q 学习的最优 CPS 控制仿真试验**

**Fig. 6 Simulation test of Q learning control**

的调节压力及成本也随之下降，AGC 系统向“放松”方向靠近；反之，则向“收紧”方向靠近，控制过程中偏重于电网公司 CPS 考核指标和效益的提高。因此，在 CPS 控制策略设计中，权值比$\lambda/\mu$的选取应根据上述规律并结合实际运行工况对电网 CPS 考核指标和 AGC 系统调节压力及经济代价之间进行协调和平衡。由于 Q 学习控制器是一种具有在线学习和动态优化能力的智能控制器，奖励函数可在线修正，它完全具备了让电网调度人员在线调整比值$\lambda/\mu$，实现在线“收紧”和“放松”控制 AGC 系统的功能。

### 2.2 南方电网实例仿真研究

为结合实际电网进一步研究Q学习控制的适应性与鲁棒性，本文选择更为复杂的南方电网作为研究和对象。所用仿真模型为广东省电力调度中心实际工程项目搭建的详细全过程动态仿真模型[10]。对比PI控制器参数来源于南瑞公司提供的基于PI控制原理的实际CPS控制策略[5]。Q学习控制器的学习步长即调度端AGC控制周期为 4 s，根据广东电网负荷扰动特点及AGC系统联络线偏差控制的“Wedge-Shaped”控制规律[9]，输出动作离散集为$A$={−1000, −600, −300, −100, −50, −20, −10, −5, 0, 5, 10, 20, 50, 100, 300, 600, 1000}MW，其奖励函数采用 2.1 节中Q学习控制器Ⅱ形式。

考虑到复杂电网 AGC 系统是一个典型的随机系统，为了验证所推荐控制器在各种复杂扰动下的适应性和对系统模型参数摄动的鲁棒性，本文采取采样周期较长的有限带宽白噪声负荷扰动进行统计性鲁棒试验，白噪声扰动是功率谱密度在整个频域内均匀分布的噪声扰动，理论上涵盖了种种功率毛刺的扰动情况。本算例进行了以下仿真设计：

1）在广东电网和其他各省网加以采样时间为 15 min、幅值不超过 15 00 MW(对应广东电网最大单一故障——直流单极闭锁)的有限带宽白噪声负荷扰动。

2）对南方电网各省负荷频率响应系数分别加入 10%和 20%幅值的白噪声参数扰动。

由于Q学习控制器是追求长期收益最大，因此，必须通过长期数据统计手段才能获得Q学习控制器的客观评价。选择一天 24 h、以 10 min为CPS考核时段的指标，广东电网统计性试验指标汇总表如表 2 所示。其中，$|\Delta F|$、$|A_{CE}|$、$C_{\mathrm{CPS1}}$为考核值的 24 h 平均值，$C_{\mathrm{CPS2}}$、$C_{\mathrm{CPS}}$为 24 h内考核合格率百分数，CPS2 考核标准阀限值$L_{10}$取南方电网总调推荐值为 288 MW。

由表 2 可知，在标称参数下，良好整定的 PI 控制器与 Q 学习控制器的性能十分接近。但是，在系统模型参数出现扰动后，Q 学习控制器在各个指标上均明显超过 PI 控制器。特别是随着负荷频率系数的参数扰动变化范围从 0%(对应表 2 中的“标称参数”)、10%增至 20%时，Q 学习控制器与整定良好的 PI 控制器的 CPS 总合格率差距从 0.82%、4.57%提升至 5.48%，体现出 Q 学习控制器的高适应性和高鲁棒性能，充分说明了该控制器的在线学习和动态优化能力。

**表 2 广东电网仿真试验 CPS 指标对照表**

**Tab. 2 CPS Compliance of Guangdong Power Grid**

| 指标 | 标称参数 | | 10%白噪声扰动 | | 20%白噪声扰动 | |
|---|---|---|---|---|---|---|
| | PI 控制 | Q 学习 | PI 控制 | Q 学习 | PI 控制 | Q 学习 |
| $\|\Delta F\|$/Hz | 0.049 7 | 0.048 5 | 0.060 1 | 0.051 3 | 0.089 2 | 0.071 6 |
| $\|A_{CE}\|$/MW | 183.23 | 178.93 | 224.13 | 191.74 | 347.14 | 267.86 |
| $C_{\mathrm{CPS1}}$/% | 146.90 | 151.28 | 120.02 | 137.93 | 90.68 | 107.85 |
| $C_{\mathrm{CPS2}}$/% | 94.58 | 94.71 | 89.28 | 92.33 | 75.21 | 81.67 |
| $C_{\mathrm{CPS}}$/% | 89.67 | 90.49 | 83.27 | 87.84 | 72.59 | 78.07 |

## 3 结论

CPS 标准下的 AGC 控制策略是电网“节能调度”一个核心内容，设计适应性高、鲁棒性强的最优 CPS 控制策略需要解决以下 2 个核心问题：1）必须满足电网在复杂运行方式下的 CPS 考核合格率；2）最大程度地减轻 AGC 机组的调节压力，降低频繁控制动作所带来的机组损耗及经济代价，即实现松弛控制。引入 Q 学习控制算法设计 CPS 控制具有以下优点：

1）CPS 标准下的 AGC 最优控制是一个典型的动态优化控制问题，引入基于严格随机最优控制——马尔可夫决策理论的 Q 学习方法具有很高可行性，应用科学的统计性试验对南方电网实例研究显示，Q 学习控制器具有很高的适应性和鲁棒性。

2）CPS 考核指标实时值与 Q 学习算法中的奖

励函数密切相关，如图 6 所示，通过调节奖励函数中各权值比 $\lambda/\mu$，可以直观、有效地实现调度员在线调整 AGC 系统的松弛程度，实现 CPS 松弛控制从而达到节能调度的目的。

在研究中也发现：1）Q 学习算法中的奖励信号来自于 CPS 指标，若能够引入更为广泛的节能和经济调度指标形成综合奖励信号，应可获得更佳的 AGC 控制效果；2）Q 学习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。

## 参考文献


[1] Jaleeli N，Vanslyck L S．NERC's new 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] 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 Yuezhong，Zhang Wangjun．Research on control performance standard based control strategy for AGC[J]．Power System Technology，2004，28(21)：75-79(in Chinese)．

[4] 李滨，韦化，农蔚涛．基于现代内点理论的互联电网控制性能评价标准下的 AGC 控制策略[J]．中国电机工程学报，2008，28(25)：56-61．
Li Bin，Wei Hua，Nong Weitao．AGC control strategy under control performance standard for interconnected power grid based on optimization theory[J]．Proceedings of the CSEE，2008，28(25)：56-61(in Chinese)．

[5] 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 Zonghe，Teng Xianliang，Tu Liqun．Hierarchical AGC mode and CPS control strategy for interconnected power systems [J]．Automation of Electric Power Systems，2004，28(1)：78-81(in Chinese)．

[6] Feliachi A，Rerkpreedapong D．NERC compliant load frequency control design using fuzzy rules[J]．Electric Power Systems Research，2005，73(1)：101-106．

[7] 胡奇英，刘建庸．马尔可夫决策过程引论[M]．西安：西安电子科技大学出版社，2000：160-239．

[8] Nedzad A，Ali F，Dulpichet R．CPS1 and CPS2 compliant wedge-shaped model predictive load frequency control [C]．Proceedings of IEEE Power Engineering Society General Meeting，Piscataway，USA，2004．

[9] Jalceli N，VanSlyck L S．Tie-lie bias prioritized energy control [J]．IEEE Trans. on Power Systems，1995，10(1)：51-59．

[10] 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]．中国电机工程学报，2008，28(13)：45-49．
Yu Tao，Chen Liang，Cai Guanglin．CPS statistic information self-learning methodology based adaptive automatic generation control[J]．Proceedings of the CSEE，2008，28(13)：45-49(in Chinese)．

[11] Yu T，Zhou B．A novel self-tuning CPS controller based on Q-learning method[C]．In Proceedings of IEEE Power and Energy Society General Meeting，Pennsylvania，USA，2008．

[12] Watkins J C H，Peter D．Q-learning[J]．Machine Leaning，1992(8)：279-292．

[13] 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，2001：126-155．

[14] Sutton R S，Barto A G．Reinforcement learning：an introduction [M]．Cambridge：MIT Press，1998：87-160．

[15] 张采，周孝信，蒋林．学习方法整定电力系统非线性控制器参数[J]．中国电机工程学报，2000，20(4)：1-5．
Zhang Cai，Zhou Xiaoxin，Jiang Lin．The adjustment of the parameters of power system non-linnar controller by learning algorithm[J]．Proceedings of the CSEE，2000，20(4)：1-5 (in Chinese)．

[16] North American Electric Reliability Council (NERC)．BAL-001-0 control performance standards：USA [S/OL]．2005-2-8 [2005-4-1]，http://standard.nerc.net/．

[17] Ahamed T P I，Rao P S N，Sastry P S．A reinforcement learning approach to automatic generation control[J]．Electric Power Systems Research，2002，63(1)：9-26．

[18] 张洪钺，王青．最优控制理论与应用[M]．北京：高等教育出版社，2006：68-90．


余涛

**收稿日期**：2009-04-27。

**作者简介**：

余涛(1974—)，男，博士，副教授，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，taoyu1@scut.edu.cn；

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，healbe@163.com；

陈家荣(1965—)，男，博士，助理教授，主要研究方向为电力系统优化算法、在线安全评估和实时仿真系统，eekwchan@polyu.edu.hk。

(编辑　吕鲜艳)