Reactive Power Optimization Calculation Based on Multi-step Q(λ) Learning Algorithm

HU Xi-bing, YU Tao

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

Email:xikeren@163.com

Abstract: In order to pursue greater economic benefits, the operation of power systems increasingly close to the critical stability, increasing the possibility of instability of the system. Thus security has become the focus of modern power system. Take the security of the power system operation for study and establish a reactive power optimization model aimed at constraint variable stability margin. A multi-step predictable Q(λ) learning algorithm based on Q learning algorithm of reinforcement learning is proposed, which with its good backtracking ability, continuously try and backtrack, getting the long-term maximum value of reward to find the optimal action. It is with advantages of online learning capability and convergence speed. This algorithm is compared with other algorithms in IEEE14 standard example and achieves good results, which proves that multi-step Q(λ) learning algorithm is feasible and efficiency for reactive power optimization.

Key words: Power System; Reactive Power Optimization; Reinforcement Learning; Multi-step Q(λ) Algorithm.

基于多步回溯 Q(λ)学习算法的无功优化研究

胡细兵，余　涛

华南理工大学电力学院，广东省 广州市 510640

Email:xikeren@163.com

摘　要：为了追求更大的经济利益，电力系统的运行日益接近其稳定临界，增加了系统失稳的可能性，安全性已成为现代电力系统研究的重点。本文以电力系统运行的安全性为研究对象，从而建立了以约束变量稳定裕度为目标的无功优化模型。在强化学习 Q-学习算法的基础上提出了一种具备多步预见能力的 Q(λ)学习算法，利用其良好的回溯能力，经过不断的试错、回溯，从而获得长期最大奖励值来寻找最优的动作，具有在线学习能力强和收敛速度快的优点，通过其在标准 EEE14 节点的算例中与其他算法进行比较，取得了良好的效果，验证了多步 Q(λ)学习算法在处理无功优化问题的可行性和有效性。

关键词： 电力系统；无功优化；强化学习；多步 Q(λ)学习

0 引言

电力系统无功优化是保证系统安全、经济运行的一项有效手段，是提高电压质量、改善系统无功分布、降低网络损耗的一项重要措施[1]。最初关于无功优化的研究仅局限于运行的经济性[2]，世界上几次大的电力系统事故后，人们逐渐认识到，一次大面积停电事故造成的经济损失，可能超过几十年经济调度收益[3]。

本质上，无功优化问题是一个离散、有约束、非线性组合优化问题。目前，研究所采用的方法包括单纯形法、二次规划法、动态规划法等经典算法这些方法各自都有一定的优越性和适应性，但是一般需假设各控制变量是连续的，目标函数是可微的，且易陷入局部最优解[4]。随着计算机技术的发展，以遗传算法、Tabu搜索、模拟退火算法、粒子群算法、免疫算法等为代表的现代启发式算法[5]，已应用到电力系统无功优化领域，这类随机优化方法通过对优化变量的随机组合来获取全局最优解,适宜求解各种离散优化问题,但普遍寻优速度慢,一定条件下易陷入局部最优值[6]。

近年来，人工智能中的强化学习[7](RL)取得了快

基金项目：国家自然科学基金面上项目（50807016）；广东省自然科学基金项目（9151064101000049）；中央高校基本科研业务费专项资金资助

978-1-935068-17-4 © 2010 SciRes.

速的发展，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[8]。RL 中的经典单步 Q-学习算法在文[9]中已被成功应用于电力系统的无功优化控制中，取得了较好的效果。笔者研究中发现，该算法的在线学习能力弱，学习速度较慢，难以满足实际电网实时控制的要求。本文提出了一种多步回溯 Q(λ)学习算法，显式地利用资格迹对将来多步决策的在线强化信息进行高效地回溯操作，提高了算法收敛速度，取得良好的效果。

1 多步 Q(λ)学习的原理

RL 是学习通过探索未知环境获得的经验来寻找最优动作的一个过程。RL 假设这个“环境”可以描述为一个状态集合 S 以及一个动作集合 A，运行空间被分为了离散的学习步长，在每个学习步长中，Agent 观察当前“环境”的状态 s(s∈S)，然后选择一个趋于增大长期期望值函数的动作 a∈A。[7]

RL 的本质目的在于寻找一种策略 π，其目的在于使得每个状态 s 的值函数达到最大。这里的值函数对应状态动作对的函数，可以定义为：

$$Q^{\pi}(s,a)=E\{r_1+\gamma r_2+...+\gamma^{i-1}r_i+...|s_0=s,a_0=a\} \quad (1)$$

其中 r 为立即奖励，γ 是一个参数，$\gamma\in[0,1]$称为折扣率，决定将来奖励信号对现在的作用。强化学习的本质不要求当前的立即奖励达到最大，而是希望最终的奖励折扣总和达到最大。

多步 Q(λ)学习(Multi-step Q(λ) learning)[10]是基于离散马尔可夫决策过程的经典 Q-学习[8]结合了 TD(λ)算法[7]多步回报的思想。多步 Q(λ)值函数的回溯更新规则利用资格迹来获取算法行为的频度和渐新度两种启发信息，从而考虑了未来控制决策的影响。资格迹[12]用于解决延时强化学习的时间信度分配问题，第 k 步迭代时刻的矩阵形式即 $e_k(s,a)$，是对过去所访问状态与动作信息的一种临时记录。对任何状态－动作对而言，资格迹都将以时效性按指数 λ^k 衰减。一旦执行非贪婪探索动作时，迹则可以复位设置为零。资格迹更新公式定义如下：

$$I_{xy}=\begin{cases}1 & x=y\\ 0 & \text{其 他}\end{cases} \quad (2)$$

$$e_k(s,a)=I_{ss_k}\cdot I_{aa_k}+\begin{cases}\gamma\lambda e_{k-1}(s,a) & Q_{k-1}(s_k,a_k)=\max\limits_a Q_{k-1}(s_k,a)\\ 0 & \text{其他}\end{cases} \quad (3)$$

式中，I_{xy} 是迹特征函数；$0<\gamma<1$，为折扣因子；λ 为迹衰退系数。

资格迹 λ-回报算法的“后向估计”机理提供了一个逼近最优值函数 Q^*的渐进机制，而这类对所有状态－动作对 Q 值的高效持续更新是以提高算法复杂度和增加计算量为代价的。设 Q_k 代表 Q^*估计值的第 k 次迭代值，Q(λ)学习迭代更新公式如下：

$$\delta_k=R(s_k,s_{k+1},a_k)+\gamma\max_{a'}Q_k(s_{k+1},a')-Q_k(s_k,a_k) \quad (4)$$

$$Q_{k+1}(s,a)=Q_k(s,a)+\alpha\delta_k e_k(s,a) \quad (5)$$

式中，0<α<1，称为学习因子；$R(s_k,s_{k+1},a_k)$是第 k 步迭代时刻环境由状态 s_k 经动作 a_k 转移到 s_{k+1} 后的奖励函数值；Q(s,a)代表 s 状态下执行动作 a 的 Q 值函数，其实现方式均采用 lookup 查表法。

多步 Q(λ)学习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 π^*，如下式：

$$\pi^*(s)=\arg\max_{a\in A}Q^k(s,a) \quad (6)$$

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而往往收敛于局部最优，通常采用由概率矢量法派生的追踪算法或 boltzmann 分布法[11]。

2 无功优化模型

目标函数参照文献[9]，本文无功优化的研究对象是系统运行的安全行，以约束变量的稳定裕度为目标函数，目标函数表示如下：

$$\min f=\min\left(\frac{1}{n}\sum_{j=1}^{n}\left|\frac{2z_j-z_{j\max}-z_{j\min}}{z_{j\max}-z_{j\min}}\right|\right) \quad (7)$$

式中，n 表示约束变量的个数，z_j 第 j 个约束变量的值，(z_{jmin},z_{jmax})表示它的上下限。

电力系统约束条件分成两部分，一个是等式约束，另一是不等式约束，不再详述。

3 多步 Q(λ)学习算法在无功优化中的应用

3.1 多步 Q(λ)学习算法的计算流程

RL 的一个特点就是通过判断状态的变化来反映所需要处理的事件的特征。基于上述分析，基于多步 Q(λ)学习的无功优化计算中，通过当前电力系统运行特征，在线寻找最优策略。

无功优化计算中的动作是约束变量中的控制变量，动作 a 的个数与初始条件有关，例如发电机的机端电压的调整范围，变压器分接头的档位，以及可投切电容器的组数等均有关，一般是这些动作数的乘积。在当前电力系统“厂网分开”的背景下，电网侧无功优化的动作一般选取调压变压器和可投切电容器两种模式，动作 a 的空间为：

$$A=\prod_{i=1}^{m}a_i \quad (10)$$

式中，a_i 为第 i 个动作变量的动作空间，一个有 m 个动作变量。

根据第 2 节中的式(8)和式(9)可知，无功优化计算中的约束条件分为等式约束和不等式约束，而等式约束的本质是潮流计算，只需要每一次潮流计算迭代结果收敛，即满足等式约束。因而在使用多步 Q(λ)学习的无功优化计算中，根据 Q(λ)学习的策略选择一个动作后，通过观察每一次潮流计算的结果，判断不满足不等式约束的个数，从而来修正状态，确定 s'的值，并给出立即奖励的值 r。

应用多步 Q(λ)学习算法的无功优化流程如下：

Step1：初始化；

Step2：预判断下一个状态，执行一个动作，从潮流计算的结果观察当前的立即奖励，并修正下一个状态；

Step3：根据学习策略向前观察下一探索动作；

Step4：更新资格迹和 Q 值；

Step5：判断下一个探索动作是否是最优动作，对资格迹进行重新赋值，并更新概率 P 矩阵；

Step6：判断 s'是否是终结点，如果不是，则将下一个状态和选择的动作，赋给当前的状态和动作，返回 Step2。

3.2. 具体动作策略选择及参数设置

本文采用一种基于概率分布选择动作的追踪算法[12]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。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式(11)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。

$$\begin{cases} P_s^{k+1}(a_g)=P_s^k(a_g)+\beta(1-P_s^k(a_g)) \\ P_s^{k+1}(a)=P_s^k(a)(1-\beta) \quad \forall a\in A,\quad a\neq a_g \\ P_s^{k+1}(a)=P_s^k(a) \quad \forall a\in A,\quad \forall \tilde{s}\in S,\quad \tilde{s}\neq s \end{cases} \tag{11}$$

式中 0<β<1，β 值的大小决定了动作搜索的速度，$P_s^k(a)$ 代表第 k 次迭代时状态 s 下选择动作 a 的概率，a_g 为由(6)式得到的贪婪动作策略。

多步 Q(λ)学习算法应用在无功优化中所涉及的参数有四个：γ，γ 为折扣因子；α，α 称为学习因子；β，β 称为动作收敛速度；λ，λ 为迹衰退系数。

无功优化问题根据自身的特点，由于多步 Q(λ)学习算法中两个连续动作所对应的潮流计算结果之间没有关联，因而后续动作所产生的奖励对前步没有太大的影响。故折扣因子 γ 取值接近 0。

学习因子 α，α 指明了要给改善的更新部分多少信任度，一般来讲，较大的 α 值会加快学习算法的收敛速度，而较小的 α 值能保证控制器的搜索空间从而提高学习收敛的稳定性，

β 为动作搜索的速度因子，β 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 β 值在 0.8～1.0 范围内都能很好地平衡 Q(λ)学习控制器的动作搜索与经验强化问题。

迹衰退系数 λ，λ 越大，迹衰减越慢，表明控制器能回溯到过去越远的信息，对于本文，控制器不需要回溯到非常远的信息，仿真比较显示 λ 在 0.2～0.5 范围都有较优的回溯特性，。

考虑到无功优化问题的特点，通过仿真经验，本文 γ 取 0.0005，α 取为 0.99，β 取值为 0.9，λ 取 0.4。

4 算例分析

为了验证多步 Q(λ)学习的正确性和可行性，本文在 MATLAB6.5 仿真平台上，在 2.0GHz、1GRAM 的计算机上对 IEEE14 算例进行了仿真。

IEEE14 节点系统的线路参数和负荷参数参照文献[12]，此时的负荷在初始值保持恒定，与文献[9]类似，多步 Q(λ)学习的动作变量包括了 3 个变压器分接头，每个有 16 个档位，以及在节点 9 处的 9 组无功补偿，动作向量的总个数为 16×16×16×9=36864 个。

对应的状态，选择一个动作后，通过判断潮流计算中的各发电机的无功出力，平衡机的有功出力，各 PQ 节点的电压是否在其所对应的约束范围之内，最终得出不满足不等式约束的个数，在 IEEE14 节点的多步 Q(λ)学习中，发电机 1(平衡节点)的有功出力，判断发电机 1、2、3、6 的无功出力，以及负荷节点(节点 4、5，7—14)的电压，对应状态个数 1+4+10=15。

奖励值也与潮流计算结果中是否满足不等式约束的个数有关，由于在多步 Q(λ)学习中，奖励函数的值越大越好，而现实中，目标函数 f 要求越小越佳，因而立即奖励函数 r：

$$r=-(f+Kn) \tag{13}$$

其中，f 为目标函数，n 为不满足不等式约束的个数，K 为系数，为了配合数量级，仿真中取 K=10。最终形成的 Q 矩阵为(15，36864)维。以初始 $Q(s,a)$=0 的多步 Q(λ)学习算法的仿真结果如下图 1 所示：

其经过 36954 步收敛，耗时约 50s，计算结果与文[9]和文[12]中的结果比较，统计如下表 1 所示。

978-1-935068-17-4 © 2010 SciRes.

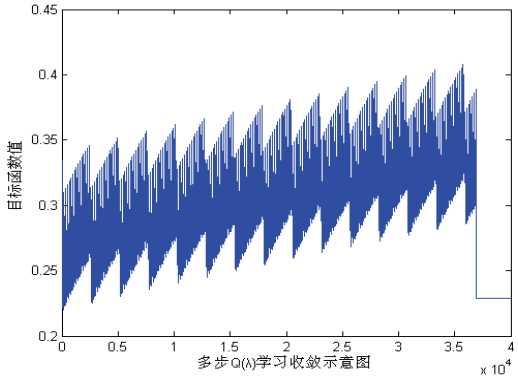

Figure.1 convergence results based on multi-step Q(λ) learning
图 1 基于多步 Q(λ)学习的无功优化计算收敛示意图

Tab.1 reactive power optimization results comparison of different algorithms
表 1 不同算法的无功优化结果比较

算法	Q(λ)学习	Q 学习[9]	概率潮流法[12]
目标函数	0.2291	0.2268	0.6113
学习步数	36954	39000	-
动作变量			
t56	0.97	1.03	0.94
t49	0.90	0.97	0.97
t47	0.91	0.90	0.98
b9	0.27	0.12	0.12
约束变量			
Qg2	0.05	**-0.02**	0.08
Qg3	0.30	0.28	0.31
Qg6	0.01	0.31	0.12
T23	0.20	0.20	0.20
T56	0.30	0.34	0.39
V4	0.97	0.98	0.97
V5	0.98	0.99	0.97
V7	1.05	1.04	0.98
V8	1.05	1.04	0.98
V9	1.04	1.01	0.97
V10	1.03	1.00	0.97
V11	1.01	1.00	0.98
V12	0.99	0.99	0.98
V13	0.99	0.98	0.98
V14	1.00	0.98	0.95

可知，基于 RL 的标准 Q-学习算法和多步 Q(λ)学习的在 IEEE14 中的无功优化目标函数结果比常规的概率潮流法均有明显的提高；多步 Q(λ)学习和标准 Q-学习的目标函数值基本一致，但是前者的收敛速度比后者有明显的提高，增幅约 5.24%，其计算结果中的约束变量没有越限，而后者中发电机 2 无功出力越限，因此，多步 Q(λ)学习算法比 Q-学习算法更为有效。

5 结论

本文综合考虑到电力系统运行的特点，以运行安全性为出发点，以系统中相关约束变量的稳定裕度为目标函数。基于强化学习理论的多步回溯 Q(λ)学习由于其自身的可回溯性，因而具有良好的在线学习能力。本文中通过其在 IEEE14 节点中与标准 Q-学习算法以及常规的概率潮流法进行比较，计算结果显示，其能取得更优的效果，从而验证了多步 Q(λ)学习在无功优化计算中具有计算收敛速度快，收敛精度高的优点，为解决无功优化问题提供来一种全新的有效方法。

References (参考文献)

[1] CHEN HENG. Steady-state analysis of power system[M]. Beijing: China Electric Power Press, 1995.
陈珩.电力系统稳态分析[M].北京:中国电力出版社，1995.

[2] XIANG Tie-yuan, ZHOU Qing-shan, LI Fu-peng. Research On Niche Genetic Algorithm For Reactive Power Optimization[J]. Proceedings of the CSEE, 2005, 25(17):48-51.
向铁元，周青山，李富鹏，等. 小生境遗传算法在无功优化中的应用研究[J] . 中国电机工程学报，2005，25(17)：48-51.

[3] LI Wen-chen. Safe operation of power systems-Model and Algorithm[M]. Congqing: Chongqing University Press. 1988.
李文沅.电力系统安全经济运行—模型与算法[M].重庆:重庆大学出版社. 1988.

[4] YANG Li-xi, WANG Kai, CHENG Jie. Application of Modified Plant Growth Simulation Algorithm in Solution of Reactive Power Optimization Problem [J]. High Voltage Engineering, 2009,25(17):48-51.
杨丽徙，王锴，程杰.应用改进模拟植物生长算法求解无功优化问题[J]. 高电压技术，2009，25(17)：48-51.

[5] XIONG Hu-gang, CHENG Hao-zhong, LI Hong-zhong. Multi-objective Reactive Power Optimization Based on Immune Algorithm [J]. Proceedings of the CSEE, 2006,11(26):102-108.
熊虎岗,程浩忠,李宏仲.基于免疫算法的多目标无功优化[J]. 中国电机工程学报，2006,11(26)，102-108.

[6] LIU Shukui, LI Qi, CHEN Weirong. Multi-objective Reactive Power Optimization Based On Modified Particle Swarm Optimization Algorithm[J]. Electric Power Automation Equipment, 2009, 29(11):31-36.
刘述奎，李奇，陈维荣等.改进粒子群优化算法在电力系统多目标无功优化中应用[J]. 电力自动化设备,2009,29(11):31-36.

[7] ZHANG Ru-bo. Theory and Application of reinforcement learning[M]. Harbin: Harbin Engineering University Press, 2001.
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.

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

[9] John G. Vlachogiannis, Nikos D. Hatziagyriou. Reinforcement Learning for Reactive Power Control[J]. IEEE Transactions on Power Systems, 2004, 19(3): 1317-1325.

[10] Jing Peng，R J Williams. Incremental Multi-Step Q-Learning [J]. Machine Leaning, 1996, 22: 283-290.

[11] Richard S. Sutton, Andrew G. Barto. Reinforcement Learning：An Introduction [M]. Cambridge: MIT Press, 1998: 87-160.

[12] R. N. Allan and M. R. G. Al-Shakarchi, Probabilistic techniques in a.c. load-flow analysis, in IEE Proc. Gener., Transm., Distrib., vol. 124, Feb. 1977, pp. 154-16.