自适应动态规划综述

张化光 [1, 2] 张欣 [3] 罗艳红 [1] 杨 珺 [1]

摘 要 自适应动态规划 (Adaptive dynamic programming, ADP) 是最优控制领域新兴起的一种近似最优方法, 是当前国际最优化领域的研究热点. ADP 方法利用函数近似结构来近似哈密顿 – 雅可比 – 贝尔曼 (Hamilton-Jacobi-Bellman, HJB) 方程的解, 采用离线迭代或者在线更新的方法, 来获得系统的近似最优控制策略, 从而能够有效地解决非线性系统的优化控制问题. 本文按照 ADP 的结构变化、算法的发展和应用三个方面介绍 ADP 方法. 对目前 ADP 方法的研究成果加以总结, 并对这一研究领域仍需解决的问题和未来的发展方向作了进一步的展望.

关键词 自适应动态规划, 神经网络, 非线性系统, 稳定性

引用格式 张化光, 张欣, 罗艳红, 杨珺. 自适应动态规划综述. 自动化学报, 2013, **39**(4): 303−311

DOI 10.3724/SP.J.1004.2013.00303

An Overview of Research on Adaptive Dynamic Programming

ZHANG Hua-Guang[1, 2] ZHANG Xin[3] LUO Yan-Hong[1] YANG Jun[1]

Abstract Adaptive dynamic programming (ADP) is a novel approximate optimal control scheme, which has recently become a hot topic in the field of optimal control. As a standard approach in the field of ADP, a function approximation structure is used to approximate the solution of Hamilton-Jacobi-Bellman (HJB) equation. The approximate optimal control policy is obtained by using the offline iteration algorithm or the online update algorithm. This paper gives a review of ADP in the order of the variation on the structure of ADP scheme, the development of ADP algorithms and applications of ADP scheme, aiming to bring the reader into this novel field of optimization technology. Furthermore, the future studies are pointed out.

Key words Adaptive dynamic programming (ADP), neural networks (NNs), nonlinear systems, stability

Citation Hua-Guang Zhang, Xin Zhang, Yan-Hong Luo, Jun Yang. An overview of research on adaptive dynamic programming. *Acta Automatica Sinica*, 2013, **39**(4): 303−311

动态系统在自然界中是普遍存在的, 对于动态系统的稳定性分析长期以来一直是研究热点, 且已经提出了一系列方法. 然而控制科技工作者往往在保证控制系统稳定性的基础上还要求其最优性. 本世纪 50 ∼ 60 年代, 在空间技术发展和数字计算机实用化的推动下, 动态系统的优化理论得到了迅速的发展, 形成了一个重要的学科分支: 最优控制. 它在空间技术、系统工程、经济管理与决策、人口控制、多级工艺设备的优化等许多领域都有越来越广泛的应用. 1957 年 Bellman 提出了一种求解最优控制问题的有效工具: 动态规划 (Dynamic programing, DP) 方法[1]. 该方法的核心是贝尔曼最优性原理, 即: 多级决策过程的最优策略具有这种性质, 不论初始状态和初始决策如何, 其余的决策对于由初始决策所形成的状态来说, 必定也是一个最优策略. 这个原理可以归结为一个基本的递推公式, 求解多级决策问题时, 要从末端开始, 到始端为止, 逆向递推. 该原理适用的范围十分广泛, 例如离散系统、连续系统、线性系统、非线性系统、确定系统以及随机系统等.

下面分别就离散和连续两种情况对 DP 方法的基本原理进行说明. 首先考虑离散非线性系统. 假设一个系统的动态方程为

$$\boldsymbol{x}(k+1) = F(\boldsymbol{x}(k), \boldsymbol{u}(k), k), \quad k = 0, 1, \cdots \quad (1)$$

其中, $\boldsymbol{x} \in \mathbf{R}^n$ 为系统的状态向量, $\boldsymbol{u} \in \mathbf{R}^m$ 为控制输入向量. 系统相应的代价函数 (或性能指标函数)

收稿日期 2012-07-19 录用日期 2012-10-29
Manuscript received July 19, 2012; accepted October 29, 2012
国家重点基础研究发展计划 (973 计划) (2009CB320601), 国家自然科学基金 (61034005, 61104099, 61104010), 辽宁省教育厅科技研究项目 (LT2010040) 资助
Supported by National Basic Research Program of China (973 Program) (2009CB320601), National Natural Science Foundation of China (61034005, 61104099, 61104010), and Science and Technology Research Program of the Education Department of Liaoning Province (LT2010040)
本文为黄琳院士约稿
Recommended by Academician HUANG Lin
1. 东北大学信息科学与工程学院 沈阳 110819 2. 东北大学流程工业综合自动化国家重点实验室 沈阳 110819 3. 中国石油大学 (华东) 信息与控制工程学院 青岛 266580
1. School of Information Science and Engineering, Northeastern University, Shenyang 110819 2. State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819 3. College of Information and Control Engineering, China University of Petroleum, Qingdao 266580
该文的英文版同时发表在 *Acta Automatica Sinica*, vol. 39, no. 4, pp. 303−311, 2013.

形式为

$$J(\boldsymbol{x}(i),i)=\sum_{k=i}^{\infty}\gamma^{k-i}l(\boldsymbol{x}(k),\boldsymbol{u}(k),k) \tag{2}$$

其中, 初始状态 $\boldsymbol{x}(k)=\boldsymbol{x}_k$ 给定, $l(\boldsymbol{x}(k),\boldsymbol{u}(k),k)$ 是效用函数, γ 为折扣因子且满足 $0<\gamma\leq 1$. 控制目标就是求解容许决策 (或控制) 序列 $\boldsymbol{u}(k)$, $k=i$, $i+1,\cdots$, 使得代价函数 (2) 最小.

根据贝尔曼最优性原理, 始自第 k 时刻任意状态的最小代价包括两部分, 其中一部分是第 k 时刻内所需最小代价, 另一部分是从第 $k+1$ 时刻开始到无穷的最小代价累加和, 即

$$\begin{aligned}J^*(\boldsymbol{x}(k))=\ &\min_{\boldsymbol{u}(k)}\{l(\boldsymbol{x}(k),\boldsymbol{u}(k))+\\ &\gamma J^*(\boldsymbol{x}(k+1))\}\end{aligned} \tag{3}$$

相应的 k 时刻的控制策略 $\boldsymbol{u}(k)$ 也达到最优, 表示为

$$\begin{aligned}\boldsymbol{u}^*(k)=\ &\arg\min_{\boldsymbol{u}(k)}\{l(\boldsymbol{x}(k),\boldsymbol{u}(k))+\\ &\gamma J^*(\boldsymbol{x}(k+1))\}\end{aligned} \tag{4}$$

接下来, 考虑连续非线性 (时变) 动态 (确定) 系统的最优控制问题. 考察如下的连续时间系统:

$$\dot{\boldsymbol{x}}(t)=F(\boldsymbol{x}(t),\boldsymbol{u}(t),t),\quad t\geq t_0 \tag{5}$$

其中, $F(\boldsymbol{x},\boldsymbol{u},t)$ 为任意连续函数. 求一容许控制策略 $\boldsymbol{u}(t)$ 使得代价函数 (或性能指标函数)

$$J(\boldsymbol{x}(t),t)=\int_t^{\infty}l(\boldsymbol{x}(\tau),\boldsymbol{u}(\tau))\mathrm{d}\tau \tag{6}$$

最小. 我们可以通过离散化的方法将连续问题转换为离散问题, 然后通过离散动态规划方法求出最优控制, 当离散化时间间隔趋于零时, 两者必趋于一致. 通过应用贝尔曼最优性原理, 可以得到 DP 的连续形式为

$$\begin{aligned}-\frac{\partial J^*}{\partial t}=\ &\min_{\boldsymbol{u}\in U}\left\{l(\boldsymbol{x}(t),\boldsymbol{u}(t),t)+\right.\\ &\left.\left(\frac{\partial J^*}{\partial \boldsymbol{x}(t)}\right)^{\mathrm{T}}F(\boldsymbol{x}(t),\boldsymbol{u}(t),t)\right\}=\\ &l(\boldsymbol{x}(t),\boldsymbol{u}^*(t),t)+\\ &\left(\frac{\partial J^*}{\partial \boldsymbol{x}(t)}\right)^{\mathrm{T}}F(\boldsymbol{x}(t),\boldsymbol{u}^*(t),t)\end{aligned} \tag{7}$$

可以看出, 上式是 $J^*(\boldsymbol{x}(t),t)$ 以 $\boldsymbol{x}(t)$、t 为自变量的一阶非线性偏微分方程, 在数学上称其为哈密顿–雅可比–贝尔曼 (Hamilton-Jacobi-Bellman, HJB) 方程.

如果系统是线性的且代价函数是状态和控制输入的二次型形式, 那么其最优控制策略是状态反馈的形式, 可以通过求解标准的黎卡提方程得到. 如果系统是非线性系统或者代价函数不是状态和控制输入的二次型形式, 那么就需要通过求解 HJB 方程进而获得最优控制策略. 然而, HJB 方程这种偏微分方程的求解是一件非常困难的事情. 此外, DP 方法还有一个明显的弱点: 随着 $\boldsymbol{x}$ 和 $\boldsymbol{u}$ 维数的增加, 计算量和存储量有着惊人的增长, 也就是我们平常所说的 "维数灾" 问题[1−2]. 为了克服这些弱点, Werbos 首先提出了自适应动态规划 (Adaptive dynamic programming, ADP) 方法的框架[3], 其主要思想是利用一个函数近似结构 (例如神经网络、模糊模型、多项式等) 来估计代价函数, 用于按时间正向求解 DP 问题.

近些年来, ADP 方法获得了广泛的关注, 也产生了一系列的同义词, 例如: 自适应评价设计[4−7]、启发式动态规划[8−9]、神经元动态规划[10−11]、自适应动态规划[12] 和增强学习[13] 等. 2006 年美国科学基金会组织的 "2006 NSF Workshop and Outreach Tutorials on Approximate Dynamic Programming" 研讨会上, 建议将该方法统称为 "Adaptive/Approximate dynamic programming". Bertsekas 等在文献 [10−11] 中对神经元动态规划进行了总结, 详细地介绍了动态规划、神经网络的结构和训练算法, 提出了许多应用神经元动态规划的有效方法. Si 等总结了 ADP 方法在交叉学科的发展, 讨论了 DP 和 ADP 方法与人工智能、近似理论、控制理论、运筹学和统计学的联系[14]. 在文献 [15] 中, Powell 展示了如何利用 ADP 方法求解确定或者随机最优化问题, 并指出了 ADP 方法的发展方向. Balakrishnan 等在文献 [16] 中从有模型和无模型两种情况出发, 对之前利用 ADP 方法设计动态系统反馈控制器的方法进行了总结. 文献 [17] 从要求初始稳定和不要求初始稳定的角度对 ADP 方法做了介绍. 本文将基于我们的研究成果, 在之前研究的基础上, 概述 ADP 方法的最新进展.

1 ADP 的结构发展

为了执行 ADP 方法, Werbos 提出了两种基本结构: 启发式动态规划 (Heuristic dynamic programming, HDP) 和二次启发式规划 (Dual heuristic programming, DHP), 其结构如图 1 和图 2 所示[4].

HDP 是 ADP 方法最基础并且应用最广泛的结构, 其目的是估计系统的代价函数, 一般采用三个网络: 评价网、控制网和模型网. 评价网的输出用来估计代价函数 $J(\boldsymbol{x}(k))$; 控制网用来映射状态变量和控制输入之间的关系; 模型网用来估计下一时刻的系统状态. 而 DHP 方法则是估计系统代价函数的

梯度. DHP 的控制网和模型网的定义与 HDP 相同, 而其评价网的输出是代价函数的梯度 $\frac{\partial J(\boldsymbol{x}(k))}{\partial \boldsymbol{x}(k)}$.

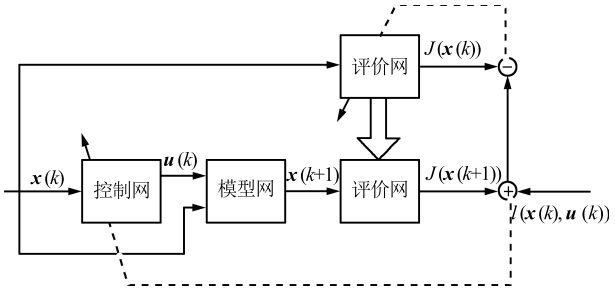


图 1　HDP 结构图

Fig. 1　The HDP structure

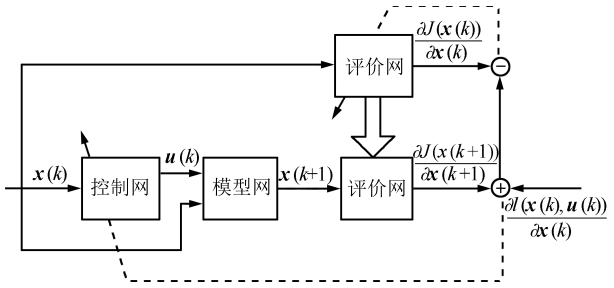


图 2　DHP 结构图

Fig. 2　The DHP structure

Werbos 进一步提出了两种改进结构: 控制依赖启发式动态规划 (Action dependent HDP, ADHDP) 和控制依赖二次启发式规划 (Action dependent DHP, ADDHP). 这两种方法与 HDP 和 DHP 的主要区别是, 评价网的输入不再只有系统状态, 而且包含控制输入. 在此基础上, Prokhorov 等提出了两种新的结构: 全局二次启发式规划 (Globalized DHP, GDHP) 和控制依赖全局二次启发式规划 (Action dependent globalized DHP, ADGDHP)[18]. 其特点是评价网不但估计系统的代价函数本身同时也估计代价函数的梯度. 上述的 ADP 结构都能用来求解最优控制策略, 但是计算速度和计算精度不同. HDP 相对简单计算速度较快, 但是计算精度比较低. GDHP 计算精度高但是其计算过程需要更长的时间. 具体的比较情况在文献 [18] 进行了详细的讨论.

ADP 方法进一步发展, 舍弃了评价网和控制网并存的双网络结构, Padhi 等提出了单网络自适应评价 (Single network adaptive critic, SNAC) 的方法, 其结构如图 3 所示[19].

SNAC 方法省略了控制网, 只留下评价网. 评价网的输出为代价函数的梯度, 定义为协状态向量 $\lambda(k) = \frac{\partial J(\boldsymbol{x}(k))}{\partial \boldsymbol{x}(k)}$. 这种单网结构减少了计算量而且消除了控制网的近似误差. 不过这种方法的实现前提是最优控制策略可以通过状态向量和协状态向量显式表达. 因此, 此方法只能用来求解一般二次型代价函数线性系统或者仿射非线性系统的最优控制问题. 由于连续系统的 ADP 方法是在离散系统 ADP 方法的基础上发展起来的, 所以其结构图与离散系统的结构图大致相同, 只不过其中变量的更迭是在连续空间中进行的.

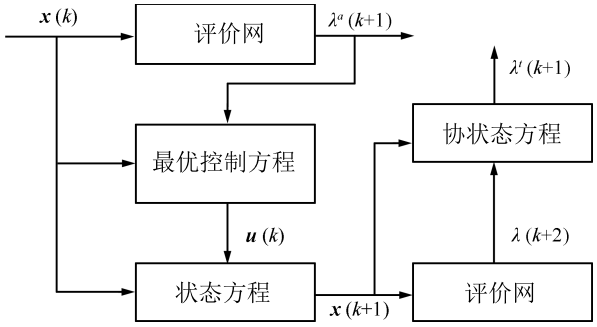


图 3　SNAC 结构图

Fig. 3　The SNAC structure

2　ADP 的算法发展

ADP 的算法经历了一个由离线迭代到在线实现的发展过程. 其理论研究主要涉及稳定性分析和收敛性证明.

2.1　离线迭代算法

2002 年, Murray 等首次提出了针对连续系统的迭代 ADP 算法[12]. 考虑一个连续微分方程

$$\dot{x} = f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}, \quad x(t_0) = x_0 \tag{8}$$

相应代价函数如式 (6), 其中 $l(\boldsymbol{x}, \boldsymbol{u}) = Q(\boldsymbol{x}) + \boldsymbol{u}^{\mathrm{T}} R(\boldsymbol{x})\boldsymbol{u}$. 那么最优控制策略可以表示为

$$\boldsymbol{u}^*(\boldsymbol{x}) = -\frac{1}{2}R^{-1}(\boldsymbol{x})g^{\mathrm{T}}(\boldsymbol{x})\left(\frac{\partial J^*(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)^{\mathrm{T}} \tag{9}$$

由于 $J^*(\boldsymbol{x})$ 需要通过求解 HJB 方程 (7) 得到, 而偏微分方程很难求出解析解, 所以提出了下边的迭代方法. 首先给定一个初始稳定的控制策略, 之后在下面两个式子之间进行迭代.

$$J_i(x_0) = \int_{t_0}^{+\infty} l\left(x_{i-1}, u_{i-1}\right) \mathrm{d}t \tag{10}$$

$$u_i(\boldsymbol{x}) = -\frac{1}{2}R^{-1}(\boldsymbol{x})g^{\mathrm{T}}(\boldsymbol{x})\left(\frac{\partial J_i(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)^{\mathrm{T}} \tag{11}$$

Murray 等在文献 [12] 中给出了系统的稳定性和迭代的收敛性的证明. 这是第一次从数学上证明了从初始稳定的控制策略开始进行迭代的迭代算法能够保证系统的稳定性和迭代性能指标的收敛性, 是

ADP 理论的巨大突破. 随后, Abu-Khalaf 等研究了具有饱和约束的连续非线性系统的最优控制问题[20], 提出了一个基于广义 HJB 方程的迭代 ADP 算法, 得到了近似最优饱和控制器, 并严格证明了该算法的收敛性. 与文献 [12] 中的 ADP 迭代算法不同的是, 文献 [20] 采用的是策略迭代的算法, 每次迭代更新的都是策略方程. 而文献 [12] 采用的是值迭代的算法, 每次迭代更新的是值函数.

对于离散时间系统, Lewis 等提出了一种不要求初始稳定控制策略的迭代 ADP 算法[21−23]. 考虑如下的离散系统:

$$\boldsymbol{x}(k+1)=f(\boldsymbol{x}(k))+g(\boldsymbol{x}(k))\boldsymbol{u}(k) \tag{12}$$

相应的代价函数如式 (2), 其中 $\gamma=1$, $l(\boldsymbol{x},\boldsymbol{u})=\boldsymbol{x}^{\mathrm{T}}(k)Q\boldsymbol{x}(k)+\boldsymbol{u}^{\mathrm{T}}(k)R\boldsymbol{u}(k)$, Q 和 R 是正定矩阵. 控制目标就是寻找最优的控制策略使得代价函数最小. 该迭代算法从初始值函数 $V_0(\cdot)=0$ 开始, 在控制策略和值函数之间进行迭代.

$$u_i(\boldsymbol{x}(k))=-\frac{1}{2}R^{-1}g^{\mathrm{T}}(\boldsymbol{x}(k))\left(\frac{\partial V_i(\boldsymbol{x}(k))}{\partial \boldsymbol{x}(k)}\right)^{\mathrm{T}} \tag{13}$$

$$\begin{aligned}V_{i+1}(\boldsymbol{x}(k))=&\boldsymbol{x}(k)^{\mathrm{T}}Q\boldsymbol{x}(k)+\\&u_i^{\mathrm{T}}(\boldsymbol{x}(k))Ru_i(\boldsymbol{x}(k))+V_i(\boldsymbol{x}(k+1))\end{aligned} \tag{14}$$

其中, $\boldsymbol{x}(k+1)=f(\boldsymbol{x}(k))+g(\boldsymbol{x}(k))u_i(\boldsymbol{x}(k))$. Zhang 等首次在理论上证明了迭代的控制策略收敛到最优控制策略, 值函数序列收敛到最优的代价函数, 即由所有容许控制策略得到的代价函数里的最小值, 同时证明了这个最优的代价函数满足 HJB 方程. 即当 $i\to\infty$ 时, $V_\infty(\boldsymbol{x}(k))=J^*(\boldsymbol{x}(k))$ 和 $u_\infty(\boldsymbol{x}(k))=u^*(\boldsymbol{x}(k))$.

随后, 文献 [22] 应用迭代 HDP 算法解决了一类离散系统的最优跟踪控制问题. 由于跟踪问题可能导致现存的代价函数趋于无穷大, 所以重新定义了一个新型的代价函数. 通过系统变换, 将最优跟踪问题转化为了最优调节问题, 然后利用迭代 HDP 算法来获得近似最优跟踪控制器.

在前面研究成果的基础之上, 针对非线性时滞系统, 文献 [24−26] 提出了基于 ADP 方法的近似最优控制方案. 考虑离散时间时滞系统:

$$\begin{aligned}&\boldsymbol{x}(k+1)=f(\boldsymbol{x}(k-\sigma_0),\cdots,\boldsymbol{x}(k-\sigma_m))+\\&\qquad g(\boldsymbol{x}(k-\sigma_0),\cdots,\boldsymbol{x}(k-\sigma_m))\boldsymbol{u}(k)\\&\boldsymbol{x}(k)=\lambda(k),\ -\sigma_m\le k\le 0\end{aligned} \tag{15}$$

其中, $\lambda(k)$ 是初始状态, σ_i, $i=0,1,\cdots,m$ 是时间延迟, 且满足 $0=\sigma_0<\sigma_1<\cdots<\sigma_m$ 为非负整数. 相应的代价函数如式 (2) 所示. 控制目标就是寻找最优的控制策略使得系统的代价函数最小. 首先, 给定初始代价函数 $V_0(\cdot)=0$, 任意给定初始状态 $\lambda(k)$ 和初始控制 $\beta(k)$, 从 $i=0$ 开始寻找最优控制, 基于 HDP 方法在控制策略、值函数和系统状态三者之间进行迭代.

$$\begin{aligned}u_i(k)=&\arg\inf_{v(k)}\{\boldsymbol{x}^{\mathrm{T}}(k)Q\boldsymbol{x}(k)+\boldsymbol{u}^{\mathrm{T}}(k)R\boldsymbol{u}(k)+\\&V_i(\boldsymbol{x}(k+1))\}\end{aligned} \tag{16}$$

$$\begin{aligned}V_{i+1}(x_i(k))=&x_i(k)^{\mathrm{T}}Qx_i(k)+u_i(k)^{\mathrm{T}}Ru_i(k)+\\&V_i(x_{i-1}(k+1))\end{aligned} \tag{17}$$

$$\begin{aligned}&x_i(t+1)=\\&\begin{cases}f(x_{i+1}(t-\sigma_0),\cdots,x_{i+1}(t-\sigma_m))+\\\quad g(x_{i+1}(t-\sigma_0),\cdots,x_{i+1}(t-\sigma_m))\times\\\quad u_{i+1}(t),\quad t\ge k\\f(x_i(t-\sigma_0),\cdots,x_i(t-\sigma_m))+\\\quad g(x_i(t-\sigma_0),\cdots,x_i(t-\sigma_m))\times\\\quad u_i(t),\quad 0\le t<k\end{cases}\\&x_i(t)=\lambda(t),\quad -\sigma_m\le t\le 0\end{aligned} \tag{18}$$

并且证明了这种迭代算法的收敛性. 当 $i\to\infty$ 时, $\boldsymbol{u}_\infty(k)=\boldsymbol{u}^*(k)$, $V_\infty(\boldsymbol{x}(k))=J^*(\boldsymbol{x}(k))$ 和 $\boldsymbol{x}_\infty(k)=\boldsymbol{x}^*(k)$.

ADP 方法也被用来按时间正向求解微分对策问题[27−32]. 在系统的最优化设计中, 经常会出现一方面要求控制变量使得性能指标取极小, 另一方面在干扰影响较大时, 考虑干扰信号使性能指标取极大, 或者是一方追逐, 一方逃逸这样的情况. 这就提出了动态系统的双边最优化问题, 即微分对策问题. 下面就二人零和微分对策问题, 简单描述如何利用迭代 ADP 方法求解微分对策问题. 考虑如下系统:

$$\dot{\boldsymbol{x}}=f(\boldsymbol{x})+g(\boldsymbol{x})\boldsymbol{u}+k(\boldsymbol{x})\boldsymbol{w} \tag{19}$$

相应的代价函数为

$$J(\boldsymbol{x},\boldsymbol{u},w,t)=\int_t^\infty l(\boldsymbol{x}(\tau),\boldsymbol{u}(\tau),\boldsymbol{w}(\tau))\mathrm{d}\tau \tag{20}$$

控制器 $\boldsymbol{u}$ 的目标是使得系统的代价函数 (20) 最小, 而控制器 $\boldsymbol{w}$ 则希望代价函数达到最大. 定义上值函数和下值函数分别为

$$\overline{V}(\boldsymbol{x})=\inf_{\boldsymbol{u}\in U[t,\infty)}\sup_{\boldsymbol{w}\in W[t,\infty)}J(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w}) \tag{21}$$

$$\underline{V}(\boldsymbol{x})=\sup_{\boldsymbol{w}\in W[t,\infty)}\inf_{\boldsymbol{u}\in U[t,\infty)}J(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w}) \tag{22}$$

相应的控制对分别定义为 $(\overline{\boldsymbol{u}},\overline{\boldsymbol{w}})$ 和 $(\underline{\boldsymbol{u}},\underline{\boldsymbol{w}})$. 那么 $\overline{V}(\boldsymbol{x})=J(\boldsymbol{x},\overline{\boldsymbol{u}},\overline{\boldsymbol{w}})$, $\underline{V}(\boldsymbol{x})=J(\boldsymbol{x},\underline{\boldsymbol{u}},\underline{\boldsymbol{w}})$. 如果 $\overline{V}(\boldsymbol{x})$

和 $\underline{V}(\boldsymbol{x})$ 存在, 并且满足 $\overline{V}(\boldsymbol{x}) = \underline{V}(\boldsymbol{x}) = V^*(\boldsymbol{x})$, 我们就认为该二人零和微分对策问题的最优控制对 $(\boldsymbol{u}^*, \boldsymbol{w}^*)$ 存在, 即鞍点存在. 在假设鞍点存在的条件下, Lewis 等结合 H_∞ 控制, 采用迭代 ADP 的方法分别研究了离散线性系统和连续仿射非线性系统的二人零和微分对策问题[27−29]. 该迭代方法分为内环迭代和外环迭代, 首先给定一个稳定的控制 u_j, 控制内环迭代, 更新 w^i, 当 w^i 收敛之后再外环更新 u_{j+1}, 之后再进行内环迭代, 直到值函数收敛到最优值, u_j 收敛到 $\boldsymbol{u}^*$, w^i 收敛到 $\boldsymbol{w}^*$. 文献 [30] 对有限时域的非仿射非线性二人零和微分对策问题进行了研究. 将非仿射非线性对策问题分解成一系列线性对策问题进行处理. 值得注意的是, 上述研究都是基于鞍点存在进行的, 但是在实际中一些非线性二人零和问题的鞍点是不存在的, 即 $\overline{V}(\boldsymbol{x}) \neq \underline{V}(\boldsymbol{x})$. 导致我们只能获得其混合最优解, 文献 [31] 首次讨论了在鞍点不存在情况下, 如何利用迭代 ADP 的方法求解混合最优解 $V^0(\boldsymbol{x})$, $\underline{V}(\boldsymbol{x}) \leq V^0(\boldsymbol{x}) \leq \overline{V}(\boldsymbol{x})$, 并且此方法也适用于鞍点存在的情况.

在实际应用中, 通常需要系统在一个有限时间内达到某种性能指标, 如实现系统镇定或者跟踪某个目标轨迹. 而现有的基于 ADP 方法的研究成果绝大多数研究的是无限时域近似最优控制问题. 为了处理有限时域问题, 文献 [33−34] 提出了一种基于 ADP 方法的有限时域最优控制方案. 该方案通过迭代方法求得最优控制策略, 使得系统的代价函数在一个 ε 界内无限接近其最优值, 并且能够求解出最优的控制步数. 有限时域最优控制问题为 ADP 方法的研究开辟了一个新领域, 有待进一步深入研究, 如在连续系统、时滞系统中实现有限时间镇定或者跟踪控制.

2.2 在线自适应算法

近年来, 一些学者提出了一些新的 ADP 算法, 这些算法不再是采取离线迭代, 而是采取在线自适应的方式获得最优控制的解[35−37]. 克服了迭代算法需要离线计算, 一旦系统发生变化, 需要重新离线计算的缺点.

Vamvoudakis 等在文献 [35−36] 中基于策略迭代提出了一种在线自适应方法, 用于求解连续非线性系统的最优控制问题, 并在理论上证明了这种在线自适应算法的稳定性. 这种在线自适应的方法也在离散系统中得到了研究. 文献 [37−38] 采用在线自适应方法, 分别研究了一类离散仿射非线性系统的最优镇定问题和最优跟踪问题.

下面我们针对连续系统的情况对这种在线自适应方法的基本原理进行概括. 离散系统的基本思想和连续系统类似, 考虑到篇幅问题, 这里不再给予详细说明.

考虑一个连续仿射非线性系统, 其系统动态描述如式 (8), 相应代价函数如式 (6). 相应的哈密顿函数为

$$H(\boldsymbol{x}, \boldsymbol{u}, J_{\boldsymbol{x}}) = l(\boldsymbol{x}, \boldsymbol{u}) + J_{\boldsymbol{x}}^{\mathrm{T}}(f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}) \quad (23)$$

其中, $J_{\boldsymbol{x}}$ 为代价函数 $J(\boldsymbol{x})$ 对 $\boldsymbol{x}$ 的偏导. 当控制策略和代价函数均取最优值时, 满足 HJB 方程, 从而 $H(\boldsymbol{x}, \boldsymbol{u}^*, J_{\boldsymbol{x}}^*) = 0$.

通常采用神经网络 (Neural networks, NNs) 构造评价网和控制网, 分别表示为

$$J(\boldsymbol{x}) = W_c^{\mathrm{T}}\phi_c(\boldsymbol{x}) + \varepsilon_c \quad (24)$$

$$\boldsymbol{u}(\boldsymbol{x}) = W_a^{\mathrm{T}}\phi_a(\boldsymbol{x}) + \varepsilon_a \quad (25)$$

其中, W_c 和 W_a 是神经网络的目标权值, $\phi_c(\cdot)$ 和 $\phi_a(\cdot)$ 是激活函数, ε_c 和 ε_a 是神经网络的有界近似误差.

评价网的实际输出表示为 $\hat{V}(\boldsymbol{x}) = \hat{W}_c^{\mathrm{T}}\phi_c(\boldsymbol{x})$, 将其带入哈密顿函数 (23), 可得 $H(\boldsymbol{x}, \boldsymbol{u}, \hat{W}_c) = e_c$. 评价网的目标就是使得 $e_c = 0$, 从而满足 HJB 方程, 使得评价网的输出能够逼近代价函数的最优值. 控制网的实际输出表示为 $\hat{\boldsymbol{u}}(\boldsymbol{x}) = \hat{W}_a^{\mathrm{T}}\phi_a(\boldsymbol{x})$, 控制网的目标就是使得控制网的实际输出逼近由评价网输出决定的近似最优控制策略 $-R^{-1}g^{\mathrm{T}}(\boldsymbol{x})\nabla\phi_c^{\mathrm{T}}(\boldsymbol{x})W_c/2$, 其中 $\nabla\phi_c(\boldsymbol{x})$ 是 $\phi_c(\boldsymbol{x})$ 对 $\boldsymbol{x}$ 的偏导. 定义两者之差等于 e_a, 控制网的目标就是使 $e_a = 0$. 基于上述思想设计评价网和控制网的权值更新规则, 使得评价网和控制网的权值能够同步更新. 采用在线自适应的方式调节神经网络权值, 随着时间的推移, 神经网络的权值最终收敛, 使得评价网的输出逐渐逼近最优代价, 控制网的输出逼近最优控制策略. 通过 Lyapunov 定理, 可以证明这种在线自适应方法权值的收敛性和系统的稳定性.

为了放松对系统模型完全或者部分已知的要求, Dierks 等在文献 [39] 中针对离散仿射非线性系统提出了一种在线系统辨识方案, 利用神经网络的一致逼近性, 采用神经网络辨识结构去重构系统动态, 其示意图如图 4 所示, 进而采用 ADP 方法去求解最优控制策略. 针对非仿射非线性系统, Zhang 等在文献 [40] 中提出了一个数据驱动的鲁棒近似最优跟踪策略. 利用可获得的数据建立数据驱动模型, 用来重构系统动态. 在建立的数据驱动模型基础上采用在线自适应的方法求解近似最优跟踪策略. 并且首次设计了一个新型的鲁棒项, 保证了跟踪误差渐近收敛到零.

同样, 利用 ADP 方法求解微分对策问题也开始向在线学习的方向发展, 文献 [41] 针对多人非零和微分对策问题, 基于策略迭代提出了一种在线自适应的控制方案. 这种多人非零和微分对策问题中的各个控制器之间既合作又竞争, 产生了相互耦

合的哈密顿–雅可比 (Hamilton-Jacobi, HJ) 方程. Vamvoudakis 等所提出的方案可以通过自适应的方法在线实时地近似最优策略和 Nash 平衡点, 对于每个控制器都有相应的评价网和控制网, 这些网络的更新是同步的, 并且保证整个闭环系统的稳定性.

值得注意的是, 上述方法均采用评价网和控制网两个神经网络, 而且为了保证在线运行时系统的稳定性, 往往要求给定一个初始稳定的控制. 为了放松这两个条件, 文献 [42] 提出了一个单网络的在线控制方案, 用来处理连续仿射非线性系统的最优控制问题. 该方案只使用一个评价网去近似系统的代价函数, 省略了控制网. 最优控制策略的估计值可以通过方程 $\hat{\boldsymbol{u}} = -R^{-1}g^{\mathrm{T}}(\boldsymbol{x})\nabla\phi_c^{\mathrm{T}}(\boldsymbol{x})W_c/2$ 和评价网的输出直接计算出来. 并且文献 [42] 通过采用一种新型的参数训练方法, 克服了对初始稳定控制的要求, 在初始控制不是容许控制的情况下, 能够保证在线学习过程中系统状态的有界性.

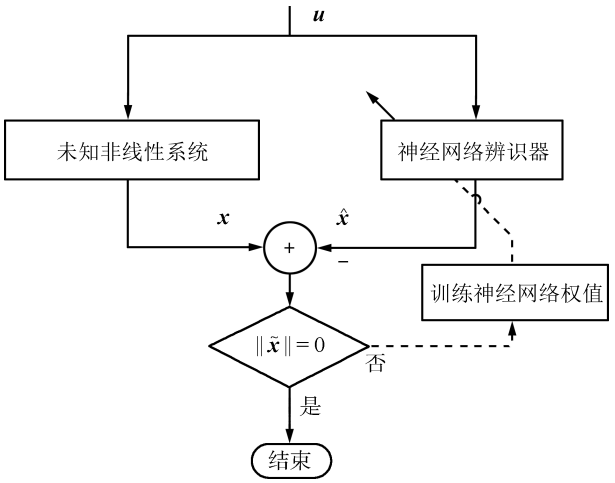


图 4 神经网络辨识结构

Fig. 4 Neural network identification structure

3 ADP 方法的应用

ADP 方法与现存的其他最优控制方法相比, 具有独特的算法和结构. 其克服了经典变分理论不能处理控制变量具有闭集约束条件的最优控制问题的缺点. 同极大值原理一样, ADP 方法不仅适用于处理带有开集性约束条件的最优控制问题, 而且也适用于处理带有闭集约束条件的最优控制问题. 但是极大值原理只给出了最优控制问题的必要条件. 而 DP 方法和 ADP 方法所给出的却是最优控制问题的充分条件. 然而, 由于 DP 方法中 HJB 方程难以求解及 “维数灾” 问题, DP 方法的直接应用变得十分困难. 因此, ADP 方法作为 DP 方法的近似解法, 克服了 DP 方法局限性, 从而更适合应用在具有强耦合、强非线性、高复杂性的系统中.

电力系统属于一类难以控制的高复杂性多变量非线性系统, 这种系统的特点是动态特性随着负载的变化而发生明显改变, 同时又要保证系统在工况时变情况下 (甚至是故障情况下) 的稳定性. 特别是随着智能电网的提出, 传统的线性化方法已经不能完全满足新的需求. 目前迫切需要开发出具有全局优化, 协调控制的智能节点 (包括断路器、重合装置、变电站等). 基于 ADP 方法的优化控制为此提供了理论基础, 近几年来已有一些成功的应用. 文献 [43] 将 HDP 方法应用到单机电力系统汽轮发电机的实时控制中, 克服了传统的基于频域中相位补偿理论的超前滞后补偿器无法保证在电力系统实际运行中的性能的缺点. 文献 [44] 将 ADP 方法应用到同步发电机控制中, 取代了传统的自动电压调节器. 文献 [45] 将 DHP 方法应用到多机电力系统的发电机励磁控制. 文献 [46] 采用 DHDP 方法实现了静止无功补偿器附加阻尼控制. 文献 [47] 提出一种基于 IPIDNN 的 DHDP 结构和算法, 该结构可以利用已有的 PID 参数指导初值选取, 并通过对 4 机 2 区系统中静止无功补偿器附加阻尼控制的仿真来表明该算法能够充分抑制互联电网的区间低频振荡.

关于智能交通系统的研究也是优化控制领域的热点问题. 智能交通系统的控制系统是既包括受交叉口信号灯调节的街区交通系统, 也包括市内快速路并通过出入口匝道与街区路网耦合在一起构成的城市交通路网的复杂的非线性大系统. 近年来, 基于 ADP 方法在单路口交通信号的优化控制和快速路入口匝道的信号控制方面已得到初步的应用[48−53]. 目前, 先进的城市交通信号控制与管理系统是采用分层递阶的分布式控制方案. 通过多 Agent 之间的通讯获得上游或下游的交通状态, 并用来构造自身的性能指标函数, 从而在自身性能指标函数学习优化的过程中实现相互之间的协调和整体性能优化[54].

此外, 基于 ADP 方法的最优化控制还在导航系统[55]、飞行器[56−57]、通讯系统[58] 等领域有着成功的应用, 并且正在以较快的速度蓬勃发展.

4 ADP 方法的未来发展方向

ADP 方法作为一个新兴的近似最优求解方法, 对其研究才刚刚起步. 下面将简单介绍现有 ADP 方法的不足之处和研究热点, 并期望能通过下边的介绍使读者把握它的发展趋势.

1) 新型 ADP 算法的提出. 当前, ADP 方法尚处于发展阶段, 现有的各种算法都有其不足之处, 急需针对这些算法的缺点, 提出新算法.

2) 有限时间 ADP 算法的研究. 由于在实际应用中, 通常需要系统在有限时间内达到某种性能指

标, 探索有限时间最优控制问题的求解仍是一个难点.

3) 基于输出反馈的 ADP 方法研究. 目前大多数 ADP 方法的结果主要集中在状态反馈方面, 而在输出反馈方面的结果比较少尚未成熟.

4) 在线自适应算法的完善. 由于迭代算法本身需要一个较长的离线计算时间, 系统一旦改变, 需要重新离线计算. 通过设计权值更新率, 采用自适应的方法在线运行是 ADP 方法发展的一个必然趋势.

5) ADP 方法应用于大时滞不确定系统甚至变时滞不确定系统的近似最优控制的研究. 含有时滞的系统的控制问题属于无穷维系统的控制问题, ADP 方法对其最优控制的研究有待于进一步研究.

5 结论

非线性系统的最优控制一直是控制领域研究的热点和难点之一. ADP 方法作为一种近似求解最优控制问题的新方法, 结合了神经网络、自适应评价设计、增强学习和经典动态规划等理论, 克服了 DP 方法的 "维数灾" 问题, 能够获得近似最优的闭环反馈控制律, 因而被认为是解决非线性系统最优控制的有效方法, 受到了不少研究者的关注. 因此, 进一步探讨 ADP 理论及其算法, 对更深入地解决非线性系统的最优控制问题有着重要的理论意义和应用价值. ADP 方法的研究方兴未艾, 希望通过本文介绍, 使读者对该方法有一个初步的认识, 并能将此方法应用到科学和工程领域中的各种优化问题中去.

References

1 Bellman R E. *Dynamic Programming*. Princeton: Princeton University Press, 1957

2 Dreyfus S E, Law A M. *The Art and Theory of Dynamic Programming*. New York: Academic Press, 1977

3 White D A, Sofge D A. *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*. New York: Van Nostrand Reinhold, 1992

4 Werbos P J. Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook*, 1977, **22**: 25−38

5 Werbos P J. *A Menu of Designs for Reinforcement Learning over Time*. Cambridge, MA: MIT Press, 1990. 67−95

6 Widrow B, Gupta N, Maitra S. Punish/reward: learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 1973, **3**(5): 455−465

7 Chen Zong-Hai, Wen Feng, Wang Zhi-Ling. Neural network control of nonlinear systems based on adaptive critic. *Control and Decision*, 2007, **22**(7): 765−768, 773
(陈宗海, 文峰, 王智灵. 基于自适应评价的非线性系统神经网络控制. 控制与决策, 2007, **22**(7): 765−768, 773)

8 Lendaris G G, Paintz C. Training strategies for critic and action neural networks in dual heuristic programming method. In: Proceedings of the 1997 IEEE International Conference on Neural Networks. Houston, USA: IEEE, 1997. 712−717

9 Werbos P J. Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks*, 1990, **3**(2): 179−189

10 Bertsekas D P, Tsitsiklis J N. *Neuro-Dynamic Programming*. Belmont: Athena Scientific, 1996

11 Bertsekas D P. Dynamic programming and optimal control. *Approximate Dynamic Programming (Fourth edition) II*. Belmont: Athena Scientific, 2012

12 Murray J J, Cox C J, Lendaris G G, Saeks R. Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and reviews*, 2002, **32**(2): 140−153

13 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press, 1998

14 Si J, Barto A G, Powell W B, Wunsch D. *Handbook of Learning and Approximate Dynamic Programming*. Hoboken: Wiley-IEEE Press, 2004

15 Powell W B. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Princeton: Wiley, 2007

16 Balakrishnan S N, Ding J, Lewis F L. Issues on stability of ADP feedback controllers for dynamical systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 913−917

17 Wang F Y, Zhang H G, Liu D R. Adaptive dynamic programming: an introduction. *IEEE Computational Intelligence Magazine*, 2009, **4**(2): 39−47

18 Prokhorov D V, Wunsch D C II. Adaptive critic designs. *IEEE Transactions on Neural Networks*, 1997, **8**(5): 997−1007

19 Padhi R, Unnikrishnan N, Wang X H, Balakrishnan S N. A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems. *Neural Networks*, 2006, **19**(10): 1648−1660

20 Abu-Khalaf M, Lewis F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 2005, **41**(5): 779−791

21 Al-Tamimi A, Lewis F L, Abu-Khalaf M. Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 943−949

22 Zhang H G, Wei Q L, Luo Y H. A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 937−942

23 Zhang H G, Luo Y H, Liu D R. Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks*, 2009, **20**(9): 1490−1503

24 Wei Q L, Zhang H G, Liu D R, Zhao Y. An optimal control scheme for a class of discrete-time nonlinear systems with time delays using adaptive dynamic programming. *Acta Automatica Sinica*, 2010, **36**(1): 121−129

25 Song R Z, Zhang H G, Luo Y H, Wei Q L. Optimal control laws for time-delay systems with saturating actuators based on heuristic dynamic programming. *Neurocomputing*, 2010, **73**(16−18): 3020−3027

26 Zhang H G, Song R Z, Wei Q L, Zhang T Y. Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming. *IEEE Transaction on Neural Networks*, 2011, **22**(12): 1851−1862

27 Al-Tamimi A, Abu-Khalaf M, Lewis F L. Adaptive critic designs for discrete-time zero-sum games with application to H_∞ control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2007, **37**(1): 240−247

28 Abu-Khalaf M, Lewis F L, Huang J. Policy iterations on the Hamilton-Jacobi-Isaacs equation for H_∞ state feedback control with input saturation. *IEEE Transactions on Automatic Control*, 2006, **51**(12): 1989−1995

29 Abu-Khalaf M, Lewis F L, Huang J. Neurodynamic programming and zero-sum games for constrained control systems. *IEEE Transactions on Neural Networks*, 2008, **19**(7): 1243−1252

30 Zhang X, Zhang H G, Wang X Y, Luo Y H. A new iteration approach to solve a class of finite-horizon continuous-time nonaffine nonlinear zero-sum game. *International Journal of Innovative Computing, Information and Control*, 2011, **7**(2): 597−608

31 Zhang H G, Wei Q L, Liu D R. An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games. *Automatica*, 2011, **47**(1): 207−214

32 Wei Q L, Zhang H G, Cui L L. Data-based optimal control for discrete-time zero-sum games of 2-D systems using adaptive critic designs. *Acta Automatica Sinica*, 2009, **35**(6): 682−692

33 Wang F Y, Jin N, Liu D R, Wei Q L. Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear systems with ε-error bound. *IEEE Transactions on Neural Networks*, 2011, **22**(1): 24−36

34 Lin Xiao-Feng, Zhang Heng, Song Shao-Jian, Song Chun-Ning. Adaptive dynamic programming with ε-error bound for nonlinear discrete-time systems. *Control and Decision*, 2011, **26**(10): 1586−1590, 1595
(林小峰, 张衡, 宋绍剑, 宋春宁. 非线性离散时间系统带 ε 误差限的自适应动态规划. 控制与决策, 2011, **26**(10): 1586−1590, 1595)

35 Vamvoudakis K G, Vrabie D, Lewis F L. Online policy iteration based algorithms to solve the continuous-time infinite horizon optimal control problem. In: Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning. Nashville, USA: IEEE, 2009. 36−41

36 Vamvoudakis K G, Lewis F L. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 2010, **46**(5): 878−888

37 Dierks T, Jagannthan S. Optimal control of affine nonlinear discrete-time systems. In: Proceedings of the 17th Mediterranean Conference on Control and Automation. Thessaloniki, Greece: IEEE, 2009. 1390−1395

38 Dierks T, Jagannathan S. Optimal tracking control of affine nonlinear discrete-time systems with unknown internal dynamics. In: Proceedings of the 48th IEEE Conference on Decision and Control and Conference on Chinese Control. Shanghai, China: IEEE, 2009. 6750−6755

39 Dierks T, Thumati B T, Jagannathan S. Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence. *Neural Networks*, 2009, **22**(5−6): 851−860

40 Zhang H G, Cui L L, Zhang X, Luo Y H. Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks*, 2011, **22**(12): 2226−2236

41 Vamvoudakis K G, Lewis F L. Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*, 2011, **47**(8): 1556−1569

42 Dierks T, Jagannathan S. Optimal control of affine nonlinear continuous-time systems. In: Proceedings of the 2010 American Control Conference (ACC). Baltimore, USA: IEEE, 2010. 1568−1573

43 Liu W X, Venayagamoorthy G K, Wunsch D C II. A heuristic-dynamic-programming-based power system stabilizer for a turbogenerator in a single-machine power system. *IEEE Transactions on Industry Applications*, 2005, **41**(5): 1377−1385

44 Park J W, Harley R G, Venayagamoorthy G K. Adaptive-critic-based optimal neurocontrol for synchronous generators in a power system using MLP/RBF neural networks. *IEEE Transactions on Industry Applications*, 2003, **39**(5): 1529−1540

45 Venayagamoorthy G K, Harley R G, Wunsch D C. Dual heuristic programming excitation neurocontrol for generators in a multimachine power system. *IEEE Transactions on Industry Applications*, 2003, **39**(2): 382−394

46 Lu C, Si J, Xie X R. Direct heuristic dynamic programming for damping oscillations in a large power system. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 1008−1013

47 Sun Jian, Liu Feng, Si J, Guo Wen-Tao, Mei Sheng-Wei. An improved approximate dynamic programming and its application in SVC control. *Electric Machines and Control*, 2011, **15**(5): 95−102
(孙健, 刘锋, Si J, 郭文涛, 梅生伟. 一种改进的近似动态规划方法及其在 SVC 的应用. 电机与控制学报, 2011, **15**(5): 95−102)

48 Bazzan A L C. A distributed approach for coordination of traffic signal agents. *Autonomous Agents and Multi-Agent Systems*, 2005, **10**(1): 131−164

49 Zhao Dong-Bin, Liu De-Rong, Yi Jian-Qiang. An overview on the adaptive dynamic programming based urban city traffic signal optimal control. *Acta Automatica Sinica*, 2009, **35**(6): 677−681
(赵冬斌, 刘德荣, 易建强. 基于自适应动态规划的城市交通信号优化控制方法综述. 自动化学报, 2009, **35**(6): 677−681)

50 Ray S, Venayagamoorthy G K, Chaudhuri B, Majumder R. Comparison of adaptive critic-based and classical wide-area controllers for power systems. *IEEE Transactions Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 1002−1007

51 Li T, Zhao D B, Yi J Q. Heuristic dynamic programming strategy with eligibility traces. In: Proceedings of the 2008 American Control Conference. Seattle, USA: IEEE, 2008. 4535−4540

52 Bai X R, Zhao D B, Yi J Q, Xu J. Coordinated control of multiple ramp metering based on DHP(λ) controller. In: Proceedings of the 11th IEEE International Conference on Intelligent Transportation Systems. Beijing, China: IEEE, 2008. 351−356

53 Cai C. An approximate dynamic programming strategy for responsive traffic signal control. In: Proceedings of the 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning. Honolulu, USA: IEEE, 2007. 303−310

54 Li T, Zhao D B, Yi J Q. Adaptive dynamic programming for multi-intersections traffic signal intelligent control. In: Proceedings of the 11th IEEE International Conference on Intelligent Transportation Systems. Beijing, China: IEEE, 2008. 286−291

55 Bertsekas D P, Homer M L, Logan D A, Patek S D, Sandell N R. Missile defense and interceptor allocation by neuro-dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 2000, **30**(1): 42−51

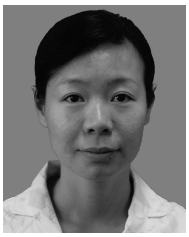
56 Ferrari S, Stengel R F. Online adaptive critic flight control. *Journal of Guidance, Control, and Dynamics*, 2004, **27**(5): 777−786

57 Liu D R, Javaherian H, Kovalenko O, Huang T. Adaptive critic learning techniques for engine torque and air-fuel ratio control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 988−993

58 Liu D R, Zhang Y, Zhang H G. A self-learning call admission control scheme for CDMA cellular networks. *IEEE Transactions on Neural Networks*, 2005, **16**(5): 1219−1228

张化光 东北大学信息科学与工程学院教授. 主要研究方向为自适应动态规划, 模糊控制, 网络控制及混沌控制. 本文通信作者. E-mail: zhanghuaguang@mail.neu.edu.cn
(**ZHANG Hua-Guang** Professor at the School of Information Science and Engineering, Northeastern University. His research interest covers adaptive dynamic programming, fuzzy control, network control and chaos control. Corresponding author of this paper.)

张 欣 中国石油大学 (华东) 信息与控制工程学院讲师. 主要研究方向为近似动态规划, 神经网络, 自适应控制, 微分对策及其工业应用.
E-mail: jackie_zx@yahoo.com.cn
(**ZHANG Xin** Lecturer at the College of Information and Control Engineering, China University of Petroleum. Her research interest covers approximate dynamic programming, neural networks, adaptive control, game theory, and their industrial application.)



罗艳红 东北大学信息科学与工程学院副教授. 主要研究方向为自适应动态规划, 最优控制, 神经网络控制.
E-mail: neuluo@gmail.com
(**LUO Yan-Hong** Associate professor at the School of Information Science and Engineering, Northeastern University. Her research interest covers adaptive dynamic programming, approximate optimal control, and neural network control.)

杨 珺 东北大学信息科学与工程学院讲师. 主要研究方向为非线性系统鲁棒控制, 模糊控制, 网络控制.
E-mail: yangjun@mail.neu.edu.cn
(**YANG Jun** Lecturer at the School of Information Science and Engineering, Northeastern University. His research interest covers robust control, fuzzy control and network control of nonlinear systems.)