第 39 卷 第 11 期 自 动 化 学 报 Vol. 39, No. 11
2013 年 11 月 ACTA AUTOMATICA SINICA November, 2013

基于数据的自学习优化控制: 研究进展与展望

刘德荣 [1] 李宏亮 [1] 王 鼎 [1]

摘 要 自适应动态规划 (Adaptive dynamic programming, ADP) 方法可以解决传统动态规划中的 "维数灾" 问题, 已经成为控制理论和计算智能领域最新的研究热点. ADP 方法采用函数近似结构来估计系统性能指标函数, 然后依据最优性原理来获得近优的控制策略. ADP 是一种具有学习和优化能力的智能控制方法, 在求解复杂非线性系统的最优控制问题中具有极大的潜力. 本文对 ADP 的理论研究、算法实现、相关应用等方面进行了全面的梳理, 涵盖了最新的研究进展, 并对 ADP 的未来发展趋势进行了分析和展望.

关键词 自适应动态规划, 近似动态规划, 强化学习, 神经网络, 智能控制

引用格式 刘德荣, 李宏亮, 王鼎. 基于数据的自学习优化控制: 研究进展与展望. 自动化学报, 2013, **39**(11): 1858−1870

DOI 10.3724/SP.J.1004.2013.01858

Data-based Self-learning Optimal Control: Research Progress and Prospects

LIU De-Rong[1] LI Hong-Liang[1] WANG Ding[1]

Abstract Adaptive dynamic programming (ADP) method can solve the problem of "curse of dimensionality" in the traditional dynamic programming, and has recently become a hot topic in the field of control theory and computational intelligence. For ADP method, a function approximation structure is used to estimate the performance index function, and then the approximate optimal control policy can be obtained based on the principle of optimality. As a kind of intelligent control methods with learning and optimization capabilities, ADP has great potential in solving the optimal control problem of complex nonlinear systems. This paper presents a comprehensive survey on the theoretical research, algorithm development, and related applications of ADP, which covers the latest research progress. It also analyzes and predicts the future development trend of ADP.

Key words Adaptive dynamic programming (ADP), approximate dynamic programming, reinforcement learning, neural networks, intelligent control

Citation Liu De-Rong, Li Hong-Liang, Wang Ding. Data-based self-learning optimal control: research progress and prospects. *Acta Automatica Sinica*, 2013, **39**(11): 1858−1870

在当今的社会生活和工业领域存在着大量的复杂系统, 如电力系统、交通系统、制造系统、化工过程系统、通信网络系统等. 这些复杂系统通常具有高度的非线性、未知的动态特性、模型的不确定性等, 难于建立精确的数学模型. 随着科技的发展, 人们对控制系统的性能也提出了越来越高的要求. 而传统的经典控制理论和现代控制理论都要求建立精确的数学模型, 这给控制理论的发展带来了巨大的挑战. 与此同时, 随着传感、通信等信息技术的广泛应用, 这些复杂系统每天都会产生和存储着海量的在线和离线数据, 这些数据蕴含着丰富的系统运行信息. 因而, 在难于建立精确的机理模型情况下, 如何发展基于数据的控制理论来解决复杂非线性系统的优化与控制, 已经成为控制理论界亟待解决的热点问题.

20 世纪 50 年代, 美国数学家 Bellman 为解决非线性动态系统的最优控制问题而提出了动态规划 (Dynamic programming, DP) 理论[1]. DP 是一种非线性的最优化规划方法, 并且适用于随机系统, 因此是解决复杂非线性系统最优控制问题的有效方法. 其核心思想是最优性原理, 可以将多级决策控制问题转化为一系列单级决策控制问题. 最优控制理论已经成为现代控制系统设计的主要方法之一, 可以使控制器在给定的系统性能指标下达到最优, 已经在工业控制、空间技术、经济管理和决策等领域获得了广泛的应用, 并取得了显著的成效. 对于被控对象为线性系统、性能指标为二次型函数的线性二次型问题, 主要是求解 Riccati 方程, 已经成为最优控

收稿日期 2013-06-28 录用日期 2013-09-02
Manuscript received June 28, 2013; accepted September 2, 2013
国家自然科学基金 (61034002, 61233001, 61273140) 资助
Supported by National Natural Science Foundation of China (61034002, 61233001, 61273140)
庆祝《自动化学报》创刊 50 周年专刊约稿
Invited Articles for the Special Issue for the 50th Anniversary of Acta Automatica Sinica
1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190
1. The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

制理论中十分成熟的部分. 然而对于实际中大量存在的非线性系统, 则需要求解 Hamilton-Jacobi-Bellman (HJB) 方程, 该方程是非线性偏微分或差分方程, 求解十分困难. 同时, 随着系统维数的增加, 存储量和计算量也随之以指数形式增加, 即 "维数灾" 问题[1]. 传统动态规划法是按照时间阶段逆向计算, 而动态系统的状态是根据系统方程按照时间正向顺序计算, 即只能通过离线方式进行求解, 并且一般要求已知被控系统的数学模型. 因此, 传统 DP 方法很难适用于模型未知的复杂非线性系统.

强化学习 (Reinforcement learning, RL)[2] 是机器学习的一种重要方法, 与监督学习和无监督学习并列为三大机器学习方法之一. 它模拟动物脑的学习机制, 强调在与环境的交互过程中在线地进行学习, 通过环境对不同决策的评价性反馈信号来修改自身的策略, 从而实现决策的优化, 其优势在于可以求解先验信息较少或不需要先验信息的复杂优化决策问题. 因此, RL 在解决模型未知的复杂非线性系统最优控制问题中可以起到十分重要的作用. RL 与 DP 有着密切的关系, 文献 [3] 说明 RL 是通过离线求解的基于模型的 DP 方法的在线无模型版本.

传统的 RL 研究主要针对离散状态和控制空间的系统, 而大部分实际控制系统中的状态和控制都是连续的, 因此通常需要采用函数近似结构. 目前大部分研究都采用了神经网络作为函数近似结构, 也有部分研究采用了模糊技术. 人工神经网络是对生物神经元细胞网络的抽象模拟, 具有自适应、自学习、容错效应、并行处理等能力, 可以对大量数据进行学习, 从而实现函数逼近、系统辨识等功能, 在智能控制领域中获得了大量的应用. 传统的神经网络在闭环反馈控制中的应用主要集中于通过神经网络进行系统辨识和基于辨识模型的控制, 是自适应控制理论的一种扩展. 它通过对在线测量数据学习从而补偿未知系统动力性特性、干扰、建模误差, 不要求已知系统动力学模型, 但通常不能获得最优控制器.

自适应动态规划 (Adaptive dynamic programming, ADP) 理论[4] 融合了 DP、RL 以及函数近似等方法, 其本质就是利用在线或离线数据, 采用函数近似结构来估计系统性能指标函数, 然后依据最优性原理来获得近优的控制策略. 传统控制方法是基于系统模型来设计控制器, 而 ADP 方法基于数据来学习控制器. ADP 通过采用生物启发的学习机制来在线实时求解优化控制问题, 不需要已知系统模型信息, 并且适用于非线性系统和不确定性问题. ADP 是一种基于数据的具有学习和优化能力的智能控制方法, 在解决模型未知的复杂非线性系统的最优控制问题中具有极大的潜力. 从学科的角度来看, ADP 依托于现代控制理论和计算智能理论, 是一种新的智能控制理论与方法.

ADP 的基本思想 (见图 1) 来源于 RL 中的执行-评判结构[2], 其中评判网络用来近似最优性能指标函数, 执行网络用来近似最优控制策略. 评判网络和执行网络的结合相当于一个智能体. 该方法可以有效解决传统 DP 的局限, 可以实现在线学习并且不要求已知系统模型. 与 ADP 类似的名称还有: 自适应评判设计 (Adaptive critic designs, ACDs)[5]、近似动态规划 (Approximate dynamic programming)[6]、神经元动态规划 (Neuro-dynamic programming)[7]. 近十年来, ADP 方法已经成为智能控制与计算智能领域的研究热点. 美国国家科学基金会分别于 2002 年和 2006 年召开了 ADP 方法的研讨会. IEEE 计算智能协会于 2008 年专门成立了 ADP 与 RL 技术委员会, 并分别于 2007、2009、2011 和 2013 年召开了 ADP 与 RL 国际研讨会. 此外, 许多国际期刊还出版了关于 ADP 研究的专刊[8−10]. 目前已经出版了数本 ADP 方法的专著[11−17] 及多篇综述性论文[18−23].

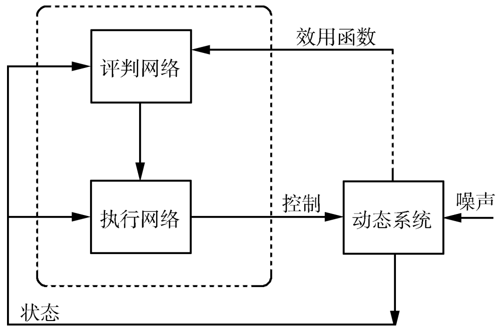


图 1 ADP 基本原理图

Fig. 1 Basic idea of ADP

为促进 ADP 研究的进一步发展, 本文对 ADP 的理论研究、算法实现、相关应用进行了全面的梳理, 其涵盖了最新的研究进展, 并对 ADP 的未来发展趋势进行了分析和预测. 本文后续内容安排如下: 1) 从理论研究对 ADP 进行总结; 2) 从算法实现来对 ADP 进行总结; 3) 介绍 ADP 在微分博弈、网络控制等现代控制理论中的应用; 4) 介绍 ADP 在工业中的应用; 5) 对全文总结并探讨 ADP 未来可能的研究方向.

1 ADP 的理论研究

ADP 的理论研究主要涉及收敛性和稳定性的证明. 该部分按被控系统是离散系统或连续系统分别展开.

1.1 离散系统

对于离散系统, 迭代 ADP 算法是近五年的研究热点, 目前已经建立了相对完善的理论. Al-Tamimi 等提出了一种贪婪迭代 ADP 算法[24] 以解决离散系统的最优控制问题, 给出了严谨的收敛性证明, 并且该方法不要求初始稳定的控制策略. Dierks 等证明了在考虑神经网络误差的情况下该算法的收敛性[25]. Zhang 等通过该算法解决了最优跟踪问题[26], 提出了新的性能指标函数以获得更好的跟踪效果. 文献 [27] 提出一种基于协状态的迭代 ADP 算法解决了带有控制器饱和约束的近似最优控制问题, 通过引入非二次型的性能指标函数来解决控制约束问题. 文献 [28] 还将迭代 ADP 算法用于求解带有时滞的离散非线性系统的最优跟踪问题. Wang 等发展了带有误差边界的贪婪迭代 ADP 算法以解决终端时间不固定的有限时间最优控制问题[29]. Heydari 等将迭代 ADP 方法扩展到终端时间固定的有限时间最优控制问题中, 其中也考虑了控制器的饱和约束[30]. Wang 等将迭代 ADP 方法推广到一般的非仿射形式的非线性系统[31]. 此外, Chen 等基于广义 HJB 方程离线求解仿射非线性离散系统的最优控制[32], 该方法将值函数进行了泰勒展开.

为讨论方便, 我们考虑如下仿射非线性离散动态系统:

$$\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k) + g(\boldsymbol{x}_k)\boldsymbol{u}_k,\ \ k = 0, 1, 2, \cdots \tag{1}$$

其中, $\boldsymbol{x}_k \in \mathbf{R}^n$ 是系统状态, $\boldsymbol{u}_k \in \mathbf{R}^m$ 是控制输入. 很多实际系统都可以表示成如上形式, 而以下结果也可以推广到一般非线性形式 $\boldsymbol{x}_{k+1} = F(\boldsymbol{x}_k, \boldsymbol{u}_k)$. 定义系统的性能指标函数为

$$J(\boldsymbol{x}_k) = \sum_{l=k}^{\infty} \gamma^{l-k} U(\boldsymbol{x}_l, \boldsymbol{u}_l) = \sum_{l=k}^{\infty} \gamma^{l-k} (\boldsymbol{x}_l^{\mathrm{T}} Q \boldsymbol{x}_l + \boldsymbol{u}_l^{\mathrm{T}} R \boldsymbol{u}_l) \tag{2}$$

其中, 折扣因子 γ 满足 $0 < \gamma \le 1$, Q 和 R 为正定矩阵. 根据 Bellman 最优性原理, 最优性能指标函数 $J^*(\boldsymbol{x}_k)$ 可以通过求解如下离散 HJB 方程获得:

$$J^*(\boldsymbol{x}_k) = \min_{\boldsymbol{u}_k} \left\{ U(\boldsymbol{x}_k, \boldsymbol{u}_k) + \gamma J^*(\boldsymbol{x}_{k+1}) \right\} \tag{3}$$

对应的最优控制策略 $\boldsymbol{u}^*(\boldsymbol{x}_k)$ 为

$$\boldsymbol{u}^*(\boldsymbol{x}_k) = \arg\min_{\boldsymbol{u}_k} \left\{ U(\boldsymbol{x}_k, \boldsymbol{u}_k) + \gamma J^*(\boldsymbol{x}_{k+1}) \right\} \tag{4}$$

当 $\gamma < 1$ 时, 一般采用 Banach 不动点定理来证明算法的收敛性. 本文主要考虑 $\gamma = 1$ 时的情况, 其收敛性的证明主要利用值函数的单调性[24−31]. 目前主要有策略迭代 (Policy iteration) 和值迭代 (Value iteration) 两种方法.

策略迭代算法是从一个初始稳定的控制策略 $\mu_0(\cdot)$ 开始, 通过求解一系列 Lyapunov 等式, 逐步改进控制策略从而最终达到最优. 策略迭代算法的每一步迭代包括以下两个过程:

策略评价

$$V_{i+1}(\boldsymbol{x}_k) = U\big(\boldsymbol{x}_k, \boldsymbol{\mu}_i(\boldsymbol{x}_k)\big) + V_{i+1}(\boldsymbol{x}_{k+1}) \tag{5}$$

策略提高

$$\boldsymbol{\mu}_{i+1}(\boldsymbol{x}_k) = \arg\min_{\mu} \left\{ U\big(\boldsymbol{x}_k, \boldsymbol{\mu}(\boldsymbol{x}_k)\big) + V_{i+1}(\boldsymbol{x}_{k+1}) \right\} = -\frac{1}{2} R^{-1} g^{\mathrm{T}}(\boldsymbol{x}_k) \frac{\partial V_{i+1}(\boldsymbol{x}_{k+1})}{\partial \boldsymbol{x}_{k+1}} \tag{6}$$

其中, i 表示迭代步数, 不断重复过程 (5) 和 (6) 直到算法收敛到最优性能指标函数和最优控制. 可以证明, 在迭代过程中值函数单调递减, 并且每个控制策略均能使系统稳定.

事实上, 策略评价过程 (5) 也可以通过如下迭代的方式得到:

$$V_{i+1}^{j+1}(\boldsymbol{x}_k) = U\big(\boldsymbol{x}_k, \boldsymbol{\mu}_i(\boldsymbol{x}_k)\big) + V_{i+1}^{j}(\boldsymbol{x}_{k+1}) \tag{7}$$

其中, j 为新的迭代步数. 在某些条件下, 容易证明式 (7) 也是收敛的, 且第 $i+1$ 步的值函数 $V_{i+1}(\boldsymbol{x}_k) = V_{i+1}^{\infty}(\boldsymbol{x}_k)$. 当式 (7) 仅迭代 K 步时, 也可以进行策略更新, 值函数和控制策略同样可以收敛到最优性能指标函数和最优控制, 该算法被称为广义策略迭代 (Generalized policy iteration)[20].

值迭代算法是从任意一个正定的初始值函数 $V_0(\cdot)$ 开始, 不需要初始稳定的控制策略, 但在迭代的过程中不能保证控制器的稳定性. 值迭代算法的值函数更新公式如下:

$$V_{i+1}(\boldsymbol{x}_k) = U\big(\boldsymbol{x}_k, \boldsymbol{\mu}_i(\boldsymbol{x}_k)\big) + V_i(\boldsymbol{x}_{k+1}) \tag{8}$$

策略更新与式 (6) 相同. 容易发现值迭代是广义策略迭代的一种特例, 即式 (8) 相当于式 (7) 仅迭代一步. 文献 [24−31] 所采用的均是这种值迭代方法.

通常来说, 策略迭代算法经过更少的迭代步数即可收敛, 但在每一步策略评价时要花费更多的时间. 与策略迭代相反, 值迭代算法通常需要更多迭代步数才能收敛, 但在每一步值函数更新时计算量更小. 而广义策略迭代算法的性能是策略迭代和值迭代的折中. 以上三种算法在进行策略评价时采用的均是值函数, 还有一类算法是采用 Q 函数[33].

下面仅以基于 Q 函数的策略迭代算法为例进行

说明. Q 函数更新过程为

$$Q_{i+1}\big(\boldsymbol{x}_k,\boldsymbol{\mu}_i(\boldsymbol{x}_k)\big)=U(\boldsymbol{x}_k,\boldsymbol{\mu}_i)+Q_{i+1}\big(\boldsymbol{x}_{k+1},\boldsymbol{\mu}_i(\boldsymbol{x}_{k+1})\big) \tag{9}$$

策略提高过程为

$$\boldsymbol{\mu}_{i+1}(\boldsymbol{x}_k)=\arg\min_{\boldsymbol{\mu}}\big\{Q_{i+1}(\boldsymbol{x}_k,\boldsymbol{\mu})\big\} \tag{10}$$

基于 Q 函数的 ADP 算法与基于值函数的 ADP 算法具有相同的收敛性. 由于 Q 函数是状态 $\boldsymbol{x}$ 和控制 $\boldsymbol{\mu}$ 的函数, 因此最优控制可直接通过 Q 函数获得, 而不需要系统的数学模型.

以上所有 ADP 算法的收敛性分析都是建立在策略评价和策略提高都可以精确求解的前提下. 而对于非线性离散系统, 策略评价和策略提高都难以精确求解. Liu 等分析了值迭代算法在每一步值函数更新过程存在有限近似误差时的收敛特性[34]. 在满足某些条件时, 值函数会收敛到最优值函数的某个邻域内, 并且可以获得近优值函数的上界.

在线 ADP 算法的稳定性分析方面, 主要是通过 Lyapunov 函数法证明闭环系统的稳定性[35]. 在线 ADP 算法已经扩展到输出反馈[36]、控制约束[37]、跟踪[38]、有界噪声[39]、非仿射系统[40] 等问题. Dierks 等提出了基于时间的 ADP 算法[41], 即策略评价和策略提高在每个采样点同时进行更新.

1.2　连续系统

与离散系统相比, ADP 在连续时间动态系统中的研究还很有限. Kleinman 指出连续 Riccati 方程的求解可以通过解一系列的 Lyapunov 方程的方法来获得[42], 该方法比直接解 Riccati 方程要容易. Saridis 等将该思想推广到非线性连续系统[43], 通过迭代的方法解广义 HJB 方程以获得最优控制. Wang 等提出了逐次逼近的方法来求解非线性随机系统最优控制的近似解[44−45], 并给出了详细的理论分析[46]. Beard 等通过在每一步迭代过程中使用 Galerkin 近似的方法找到了广义 HJB 方程的近似解, 并使用一系列的多项式函数作为基函数来近似地解广义 HJB 方程[47]. Murray 等提出了一种迭代 ADP 算法来解决非线性连续系统的最优控制问题, 证明了迭代过程中控制律的稳定性和性能指标函数的收敛性[4]. Abu-Khalaf 等使用基于神经网络的迭代 ADP 算法来获得 HJB 方程的近似解, 解决了带有执行器饱和的非线性系统近似优化控制问题[48]. Cheng 等研究了终端时间固定的非线性系统最优控制问题, 并给出了基于神经网络的广义 HJB 方程的解法[49−50]. Tassa 等采用最小二乘法和神经网络值函数近似方法求解连续 HJB 方程[51].

以上算法都是通过离线方式运行的, 当前的研究焦点是连续系统最优控制的在线自适应求解. Doya 给出了学习与控制的神经机制, 并将 RL 的理论推广到了连续时间、连续状态空间中[3]. Hanselmann 等提出了连续时间 ACDs 方法[52], 以解决用微分方程描述的连续系统模型的最优控制问题. Ferrari 等提出一种约束 ADP 方法来设计具有闭环稳定性的自适应神经网络控制器[53]. Seiffertt 等将时间尺度上的微积分理论应用到 ADP 方法中[54].

考虑如下仿射非线性连续动态系统:

$$\dot{\boldsymbol{x}}=f(\boldsymbol{x})+g(\boldsymbol{x})\boldsymbol{u} \tag{11}$$

其中, $\boldsymbol{x}(t)\in\mathbf{R}^n$ 是系统状态, $\boldsymbol{u}(t)\in\mathbf{R}^m$ 是控制输入. 假设系统在包含原点的紧集 $\Omega\subseteq\mathbf{R}^n$ 上是可镇定的. 定义如下性能指标函数:

$$\begin{aligned}V\big(\boldsymbol{x}(t)\big)=&\int_t^{\infty}r\big(\boldsymbol{x}(\tau),\boldsymbol{u}(\tau)\big)\mathrm{d}\tau= \\ &\int_t^{\infty}\big(\boldsymbol{x}(\tau)^{\mathrm{T}}Q\boldsymbol{x}(\tau)+\boldsymbol{u}(\tau)^{\mathrm{T}}R\boldsymbol{u}(\tau)\big)\mathrm{d}\tau\end{aligned} \tag{12}$$

对于容许控制策略[47]$\boldsymbol{\mu}(\boldsymbol{x})$, 定义如下 Hamilton 函数:

$$H\big(\boldsymbol{x},\boldsymbol{\mu}(\boldsymbol{x}),\nabla V\big)=r\big(\boldsymbol{x},\boldsymbol{\mu}(\boldsymbol{x})\big)+\nabla V^{\mathrm{T}}\big(f(\boldsymbol{x})+g(\boldsymbol{x})\boldsymbol{\mu}(\boldsymbol{x})\big) \tag{13}$$

其中, $\nabla V=\frac{\partial V}{\partial\boldsymbol{x}}$. 最优性能指标函数 $V^*(\boldsymbol{x})$ 满足如下连续时间 HJB 方程:

$$0=\min_{\boldsymbol{\mu}}H\big(\boldsymbol{x},\boldsymbol{\mu}(\boldsymbol{x}),\nabla V^*\big) \tag{14}$$

最优控制策略 $\boldsymbol{\mu}^*(\boldsymbol{x})$ 为

$$\begin{aligned}\boldsymbol{\mu}^*(\boldsymbol{x})=&\arg\min_{\boldsymbol{\mu}}H\big(\boldsymbol{x},\boldsymbol{\mu}(\boldsymbol{x}),\nabla V^*\big)= \\ &-\frac{1}{2}R^{-1}g^{\mathrm{T}}(\boldsymbol{x})\nabla V^*\end{aligned} \tag{15}$$

与离散系统类似, 连续系统策略迭代算法每一步迭代包含如下两个过程:

策略评价

$$0=r\big(\boldsymbol{x},\boldsymbol{\mu}_i(\boldsymbol{x})\big)+\nabla V_i^{\mathrm{T}}\big(f(\boldsymbol{x})+g(x)\boldsymbol{\mu}_i(\boldsymbol{x})\big) \tag{16}$$

策略提高

$$\boldsymbol{\mu}_{i+1}(\boldsymbol{x})=-\frac{1}{2}R^{-1}g^{\mathrm{T}}(\boldsymbol{x})\nabla V_i \tag{17}$$

文献 [48] 中已经证明该方法中值函数单调递减, 每一步迭代产生的控制策略均是稳定的. 该方法需要初始稳定的控制策略, 并且需要精确已知系统的数学模型.

2009 年, Vrabie 等提出了在线的积分 RL 算法来解决线性连续系统的最优控制问题[55], 该方法不需要已知系统的内部动态特性, 并将其推广到了非

线性连续系统[56]. 事实上, 式 (12) 可以表示成如下等价形式:

$$V\big(\boldsymbol{x}(t)\big)=\int_t^{t+T} r\big(\boldsymbol{x}(\tau),\boldsymbol{u}(\tau)\big)\mathrm{d}\tau+V\big(\boldsymbol{x}(t+T)\big) \tag{18}$$

其中, $T>0$ 为任意的时间常数. 该积分形式的公式 (18) 即可作为连续系统积分 RL 算法的不动点等式. 文献 [55] 证明了式 (18) 与式 (12) 是等价的, 但式 (18) 不涉及系统模型信息. 和离散系统类似, 由式 (18) 可以发展出策略迭代、广义策略迭代、值迭代等算法. 基于积分的策略迭代算法主要过程如下:

策略评价

$$V_{j+1}\big(\boldsymbol{x}(t)\big)=\int_t^{t+T} r(\boldsymbol{x}(\tau),\boldsymbol{\mu}_j(\tau))\mathrm{d}\tau+V_{j+1}(\boldsymbol{x}(t+T)) \tag{19}$$

$$V_{j+1}(0)=0$$

策略提高

$$\boldsymbol{\mu}_{j+1}(\boldsymbol{x})=\arg\min_{\boldsymbol{\mu}} H(\boldsymbol{x},\boldsymbol{\mu}(\boldsymbol{x}),\nabla V_{j+1})= -\frac{1}{2}R^{-1}g^{\mathrm{T}}(\boldsymbol{x})\nabla V_{j+1} \tag{20}$$

基于积分的值迭代算法中值函数更新过程如下:

$$V_{j+1}\big(\boldsymbol{x}(t)\big)=\int_t^{t+T} r\big(\boldsymbol{x}(\tau),\boldsymbol{\mu}_j(\tau)\big)\mathrm{d}\tau+V_j\big(\boldsymbol{x}(t+T)\big) \tag{21}$$

策略更新与式 (20) 相同. 基于积分的广义策略迭代算法也可以通过类似方式得到, 在此不再赘述.

前面的方法中, 策略评价和策略提高都是交替进行的. 2010 年, Vamvoudakis 等提出了同时调整评判网络和执行网络的在线同步策略迭代算法 (Synchronous policy iteration)[57], 通过 Lyapunov 函数法可以证明闭环系统的一致最终有界稳定性, 该方法要求已知系统的数学模型. 文献 [41] 中离散系统基于时间的 ADP 方法与 [57] 中的同步策略迭代方法具有相似的思想. 文献 [58] 通过递归神经网络来辨识未知系统的模型, 然后采用同步策略迭代方法在线求解连续非线性系统的最优跟踪问题. 文献 [59] 进一步提出了辨识模型、评判网络和执行网络同时更新的算法.

与离散时间下的 Q 学习类似, 目前已经有连续系统下的 Q 学习方法研究. 文献 [60] 建立了 Q 学习算法与 Pontryagin 极小值原理之间的关系, 并提出了最速下降 Q 学习方法来求解连续系统的近似最优控制, 该方法不需要任何系统模型信息. 文献 [61] 提出了一种新的策略迭代算法, 与文献 [55] 中的方法相比, 该方法可以在线求解系统模型完全未知的连续系统的自适应最优控制器. 文献 [62−63] 提出了积分 Q 学习方法 (Integral Q-learning) 方法, 该方法同时进行策略评价和策略提高. 可以发现, 文献 [61−62] 中的算法具有相同的特性, 并且均需要满足持续激励条件 (Persistence of excitation, PE). 其主要思想是在控制输入中添加扰动信号 $\boldsymbol{e}$, 即:

$$\dot{\boldsymbol{x}}=f(\boldsymbol{x})+g(\boldsymbol{x})(\boldsymbol{u}+\boldsymbol{e}) \tag{22}$$

则式 (19) 中的策略评价过程变为

$$V_{j+1}\big(\boldsymbol{x}(t)\big)=\int_t^{t+T}\Big[r\big(\boldsymbol{x}(\tau),\boldsymbol{\mu}_j(\tau)\big)- \nabla V_{j+1}^{\mathrm{T}}g(\boldsymbol{x})\boldsymbol{e}\Big]\mathrm{d}\tau+V_{j+1}\big(\boldsymbol{x}(t+T)\big) \tag{23}$$

根据策略提高过程 (20), 式 (23) 可以表示为

$$V_{j+1}\big(\boldsymbol{x}(t)\big)=\int_t^{t+T}\Big[r\big(\boldsymbol{x}(\tau),\boldsymbol{\mu}_j(\tau)\big)+2\boldsymbol{\mu}_{i+1}^{\mathrm{T}}R\boldsymbol{e}\Big]\mathrm{d}\tau+ V_{j+1}\big(\boldsymbol{x}(t+T)\big) \tag{24}$$

可见, 式 (24) 中不含有任何模型信息, 并且策略评价和策略提高过程可以同时进行.

2 算法实现

ADP 算法的实现主要涉及算法结构、函数近似、学习算法等几个方面.

目前对 ADP 算法结构的研究主要集中在离散系统, 并已取得了非常丰富的研究成果. Werbos 最早提出了 ADP 的两种基本结构[64], 即启发式动态规划 (Heuristic dynamic programming, HDP) 和二次启发式规划 (Dual heuristic programming, DHP). HDP 是应用最为广泛的结构, 其评判网络输入是状态变量 $\boldsymbol{x}_k$, 评判网络输出是系统的性能指标函数 $J(\boldsymbol{x}_k)$. DHP 结构的评判网络输入是状态变量 $\boldsymbol{x}_k$, 而评判网络输出是性能指标函数对状态的梯度 $\frac{\partial J(\boldsymbol{x}_k)}{\partial \boldsymbol{x}_k}$. DHP 结构避免了评判网络的误差反向传播, 因此比 HDP 结构具有更高的计算精度. Werbos 还给出了两种执行依赖的结构[6], 即执行依赖启发式动态规划 (Action-dependent heuristic dynamic programming, ADHDP) 和执行依赖二次启发式规划 (Action-dependent dual heuristic programming, ADDHP). 这两种结构与 HDP 和 DHP 的区别是把状态变量 $\boldsymbol{x}_k$ 和控制变量 $\boldsymbol{u}_k$ 均作为评判网络的输入. ADHDP 的执行网络的学习目标可以直接通过评判网络的误差反向传播得到, 因而不需要已知系统模型. Prokhorov 等进而提出了两种新的结构[5], 即全局二次启发式规划 (Globalized dual heuristic programming, GDHP) 和执行依赖全局二次启发式规划 (Action-dependent globalized dual heuristic programming, ADGDHP). GDHP 的评判网络输入是 $\boldsymbol{x}_k$, 而评判网络输出是 $J(\boldsymbol{x}_k)$ 和 $\frac{\partial J(\boldsymbol{x}_k)}{\partial \boldsymbol{x}_k}$. Liu 等提出一种迭代 GDHP 方法求解非线性离散系统的最优控制问题[65]. GDHP 虽然结合了 HDP 和 DHP 的优点, 但因需要计算神经网络输出对输入向量和权值向量的二阶混合偏导数 $\frac{\partial J^2}{\partial \boldsymbol{x}\partial w}$ 而使学习过程变得十分复杂. 在最新的文献 [66] 中, Fairbank 等给出了精确计算该二阶混合偏导数的算法, 可以使 GDHP 算法更简单有效地执行. 在文献 [67] 中, Si 等对这六种结构进行了较为全面的分析和总结. 从目前来看, 应用较多的是 HDP、DHP、ADHDP 三种结构, 其他三种结构使用较少.

此外, Padhi 等提出了基于 DHP 的单网络自适应评价设计 (Single network adaptive critic, SNAC) 的新结构[68], 即在 DHP 的基础上去掉了执行网络. Ding 等又进而提出了基于 HDP 的 SNAC 结构[69], 即在 HDP 的基础上去掉了执行网络. 这两种结构都具有结构简单、计算量小等优点, 但只适用于求解仿射非线性系统的最优控制. 文献 [70] 在 ADHDP 的基础上增加了一个参考网络, 其产生的内部目标信号作为评判网络的强化信号, 用于指导评判网络的学习. 传统的强化信号大多是人们根据已有的知识和经验来设计的, 并且多为二进制的值 (如 0 或 -1). 而由参考网络输出的强化信号是自适应生成的, 不需要人工手动地设计, 并且是连续有界的. 该结构虽然在球杆系统中进行了实验验证, 但尚缺少合理的理论解释.

对于状态空间和控制空间连续的情况, ADP 算法的实现需要采用函数近似结构. 目前绝大部分都采用了 BP 神经网络[26, 28, 29, 31], 有的研究也采用了径向基函数神经网络[27]、小波神经网络[71]、支持向量机[72]、递归神经网络[73]、回声状态网络[74] 等. Xu 等将核函数方法引入 ACDs 算法中[75], 提出了基于稀疏内核机的 HDP 和 DHP 方法, 目的是提高算法的泛化性能和学习效率. ADP 的学习算法可以采用最小二乘法[55−56]、梯度下降[28−31]、Levenberg-Marquardt[76] 等各种参数优化方法. 此外, 也有部分研究将 ADP 与计算智能方法进行了结合, 如模糊系统[77]、进化计算[78−79] 等.

理论上, 迭代 ADP 算法需要遍历整个状态空间[26−31], 即每一步值函数和控制策略更新须同时取遍状态空间里的所有状态. 这在实际中是很难做到的. 通常的做法是每步迭代随机在状态空间内取 N 个状态, 当算法迭代到无穷步时, 相当于取遍了整个状态空间. 传统的在线 ADP 算法[67] 实际上是一种特例, 即每一步迭代只取一个状态 ($N = 1$). 而为了在算法停止更新前充分遍历状态空间, 在线 ADP 算法一般需要满足 PE 条件, 该条件相当于 RL 中的探索 (Exploration) 和利用 (Exploitation) 的平衡问题.

3 ADP 与控制理论的结合

Jiang 等采用鲁棒 ADP 方法来计算不确定动态系统的全局稳定和次优控制策略[80], 系统动态特性和系统阶数不需要精确已知. 通过将 ADP 理论和现代非线性控制的技术相结合, 鲁棒 ADP 方法中控制器的设计是从带有干扰抑制的鲁棒控制角度给出的. 通过 Backstepping 技术, 鲁棒控制器可以通过设计产生任意小的 L_2 增益. 在没有噪声输入的情况下, 该鲁棒控制策略也能保留优化性能. 当系统存在乘性和加性噪声时, 文献 [81] 采用鲁棒 ADP 方法来求解随机系统的优化控制问题.

博弈论为研究多人决策问题提供了理想的工具, 而其纳什均衡解一般通过离线迭代的方式计算获得, 并且要求已知系统模型. 作为一种在线优化方法, ADP 理论已经被广泛应用到了微分博弈问题中. 由于零和博弈问题与 H_∞ 鲁棒控制有着密切的关系, 即控制器是最小化参与者而干扰是最大化参与者, 故目前的研究主要集中于应用 ADP 方法求解二人零和微分博弈问题. 对于线性系统, 主要是求解博弈代数 Riccati 方程 (Game algebraic Riccati equation, GARE). Vrabie 等采用一种在线 ADP 方法求解 GARE[82], 其中仅有一个参与者学习和优化它的控制策略. 并进一步采用了积分 RL 方法来求解 GARE[83], 优点是仅要求系统的部分模型. Wu 等提出了一种策略同步更新算法来求解该问题[84], 即控制和干扰同时更新参数, 该方法同样仅要求系统的部分模型. 而对于如下带有干扰输入的非线性连续系统:

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}) + k(\boldsymbol{x})\boldsymbol{d}(\boldsymbol{x}) \tag{25}$$

及值函数

$$V\big(\boldsymbol{x}(t)\big) = \int_t^\infty (\boldsymbol{x}^{\mathrm{T}}Q\boldsymbol{x} + \boldsymbol{u}^{\mathrm{T}}R\boldsymbol{u} - \theta^2\|\boldsymbol{d}\|^2)\mathrm{d}\tau \tag{26}$$

二人零和博弈的纳什均衡解满足如下 Hamilton-Jacobi-Isaacs (HJI) 方程:

$$\begin{aligned} 0 =& \nabla V^{\mathrm{T}}(\boldsymbol{x})f(\boldsymbol{x}) + \frac{1}{4\theta^2}\nabla V^{\mathrm{T}}(\boldsymbol{x})k(\boldsymbol{x})k^{\mathrm{T}}(\boldsymbol{x})\nabla V(\boldsymbol{x}) + \\ & \boldsymbol{x}^{\mathrm{T}}Q\boldsymbol{x} - \frac{1}{4}\nabla V^{\mathrm{T}}(\boldsymbol{x})g(\boldsymbol{x})R^{-1}g^{\mathrm{T}}(\boldsymbol{x})\nabla V(\boldsymbol{x}) \end{aligned} \tag{27}$$

Abu-Khalaf 等解决了带有输入约束的非线性连续系统的 H_∞ 最优状态反馈控制问题[85−86], 该方法为离线迭代, 要求已知系统的数学模型, 并且存在两个迭代环. Zhang 等采用四个执行网络和两个评判网络计算微分博弈问题的鞍点解[87], 系统模型要求精确已知, 该方法也适用于鞍点不存在的情况. Vamvoudakis 等提出了基于同步策略迭代的在线自适应学习算法求解 HJI[88], 系统模型要求精确已知, 值函数和策略更新同时进行. 文献 [89] 采用文献 [84] 中的方法来求解 HJI. Johnson 等采用策略迭代方法在线求解该问题[90], 采用鲁棒动态神经网络来辨识未知系统, 并且所有参数同时调整.

Al-Tamimi 等采用 HDP 和 DHP 两种结构在线求解线性离散系统的零和博弈问题[91]. Q 学习也被用来求解该问题及 H_∞ 鲁棒控制[92−93], 其优点是不需要已知系统的数学模型. 对于非线性离散系统, Mehraeen 等对值函数进行泰勒展开以离线方式求解近似 HJI 方程[94], 并给出了近似 HJI 方程收敛的充分条件. Liu 等采用迭代 ADP 方法求解带有控制约束的非线性离散系统的 H_∞ 鲁棒控制[95].

当多个控制器既竞争又合作时, 则需要求解多人非零和博弈问题. 对于非零和纳什博弈, 主要在于求解耦合代数 Hamilton-Jacobi (HJ) 方程. Vrabie 等采用积分 RL 方法在线求解线性连续系统的非零和微分博弈的纳什均衡解[96]. Vamvoudakis 等采用同步策略迭代算法求解非线性连续系统的多人非零和博弈问题[97], 该方法需要满足 PE 条件. 文献 [98] 进一步将该方法推广到定义在通信图上的多人合作博弈问题, 给出了交互纳什平衡点 (Interactive Nash equilibrium) 的定义, 并设计了分布式的自适应学习算法. Zhang 等采用基于单网络的 ADP 方法来求解非线性连续系统的非零和博弈的纳什平衡点[99], 该方法不需要初始稳定的控制策略, 同时可以保证闭环系统一致最终有界稳定性. Stackberg 博弈也是一种非零和博弈, 它包含分层级的两组参与者: 领导拥有其他参与者的策略信息, 而跟随者不拥有这些信息. Vamvoudakis 等提出一种基于策略迭代的在线自适应最优控制方法来求解线性连续系统的 Stackberg 平衡点[100], 每个控制器均有一个评判网络和执行网络, 同时给出了使整个系统稳定的参数更新算法.

包含多个级联子系统的非线性大系统通常可以表示成如下形式:

$$\dot{\boldsymbol{x}}_i(t) = f_i\big(\boldsymbol{x}_i(t)\big) + g_i\big(\boldsymbol{x}_i(t)\big)\Big(\boldsymbol{u}_i\big(\boldsymbol{x}_i(t)\big) + \Psi_i\big(\boldsymbol{y}(t)\big)\Big) \tag{28}$$

$$\boldsymbol{y}_i(t) = C_i\boldsymbol{x}_i(t)$$

采用 ADP 方法可以设计系统 (28) 的全局渐近稳定的分散控制器. Mehraeen 等采用 ADP 方法来在线求解一类级联非线性动态系统的分散最优控制[101]. 对于每个子系统, 都用一个评判网络来估计整个系统的性能和一个执行网络来提供近优的控制信号. Jiang 等采用鲁棒 ADP 方法来在线求解复杂大系统的分散控制问题[102], 并采用小增益定理来分析系统的稳定性, 该方法适用于模型参数未知的情况. 这类研究的一个典型应用是多机电力系统的控制器设计.

典型的带有随机延时和丢包的网络控制系统可以表示为

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= f\big(\boldsymbol{x}(t)\big) + \delta(t)g\big(\boldsymbol{x}(t)\big)\boldsymbol{u}\big(t-\tau(t)\big) \\ \boldsymbol{y}(t) &= C\boldsymbol{x}(t)\end{aligned} \tag{29}$$

其中

$$\delta(t) = \begin{cases} I, & \text{在时刻 } t \text{ 可以接受到控制输入} \\ 0, & \text{在时刻 } t \text{ 控制输入丢失}\end{cases}$$

I 为单位矩阵, 0 为零矩阵. Xu 等采用基于 Q 学习的 ADP 方法求解带有随机延时和丢包的线性网络控制系统的近似最优控制问题[103]. 文献 [104] 提出一种基于时间的 ADP 方法来求解不确定的非线性网络控制系统 (29) 的随机优化控制问题, 同时也考虑了随机延时和丢包等因素. 之所以称为基于时间的 ADP 方法, 是因为在每个采样时刻均更新值函数和控制输入, 该方法与策略迭代和值迭代算法不同, 可以更好的实现实时控制.

此外, ADP 还与自适应控制[55−57]、模型预测控制[105]、多智能体协调控制[98] 等有着密切的关系.

4 ADP 的应用

ADP 方法由于适合解决复杂非线性系统的最优控制问题, 已经在运动控制、电力系统、交通物流、多媒体处理、无线传感器网络、化工过程等领域取得了大量的成功应用.

Cox 等采用一种线性二次型 ACDs 算法解决 NASA 登陆车的自动降落问题[106]. Enns 等基于直接神经动态规划方法设计了 Apache 直升飞机的跟踪控制系统[107]. Nodland 等利用 ADP 思想设计了自适应优化的输出反馈控制器[108], 实现无人直升飞机的轨迹跟踪控制. Lin 采用基于模糊神经网络的 ACDs 来设计倾斜转弯导弹的自动驾驶系统[109]. Han 等采用 ACDs 来设计敏捷导弹的控制器[110]. Lin 等通过 ACDs 方法使轮式移动机器人具有自动生成控制信号的能力[111]. Liu 等利用 ACDs 思想实现了汽车发动机的自学习控制, 设计了发动机扭矩和空气-燃料比控制器[112]. Shih 等和 Peter 等设计

了基于 ADP 的神经网络跟踪控制器[113−114], 将其应用于火花点火发动机的排放控制, 并且不需要 PE 条件和线性参数化假设.

ADP 理论在电力系统中的应用最为广泛. Venayagamoorthy 所在的研究团队采用 HDP、DHP、ADHDP 等 ADP 思想设计了同步发电机的神经控制器[115]、单机电力系统涡轮发电机的优化控制器[116]、多机电力系统静态补偿器的神经模糊控制器[117] 和优化神经控制器[118]、电力系统广域控制器[119]、大规模风电场和静止同步补偿器的无功协调控制[120] 等, 均比传统的控制器获得了更好的控制效果. 为了使得具有间歇性的新能源达到比较高的渗透率, 电网的管理与控制需要考虑其高可变性和不确定性. 该团队还利用基于 DHP 的 ADP 方法设计了基于广域测量的动态随机优化电流控制算法[121]. Lu 等将直接 HDP 方法应用到大规模电力系统的稳定控制问题[122], 并以我国南方电网为例进行了验证. Huang 等利用 ACDs 思想解决住宅能源管理与控制问题[123], 该智能优化方法可以从用户需求和环境中学习.

ADP 理论在交通和物流系统中也取得了成功的应用[124]. Lin 等基于 ACDs 方法设计了地铁自动列车的自适应最优控制器[125−128], 该算法可以直接从交通数据中进行学习, 不需要已知交通模型. Zhao 等采用 DHP 方法实现高速公路系统闸道的优化协调管理[129], 该方法通过近似的交通模型以离线学习的方式获得控制器. Cai 等采用 ADP 方法设计了自适应交通信号控制器[130], 与时间固定的控制方法相比, 该方法可以大大减少车辆拥堵. Shervais 等利用 DHP 解决物流系统的库存控制和交通策略的同步优化问题[131], 该方法通过定期调整控制策略以适应随机、非静态的环境.

ADP 方法在多媒体处理中也取得了一些应用. Sun 等采用 ACDs 方法来优化便携式视频通讯设备中实时视频编码所消耗的电能[132], 进而延长设备的使用时间. 他们把视频编码器看做一个非线性控制系统, 在率失真约束下, 选择最好的复杂性控制参数配置来最小化电能消耗. Iftekharuddin 采用 HDP 和 DHP 两种方法来获得与图像变形无关的在线自动目标识别系统[133], 该系统不需要特征提取等预处理过程. Venayagamoorthy 等采用 ACDs 方法来设计语音编码中的非均匀最优量化器, 以最大化信噪比[134].

Lee 将基于 ADP 的控制策略用于连续搅拌釜式反应器等化工过程的控制[135−138], 并仅利用过程中产生的输入输出数据, 该方法与传统的模型预测控制相比具有更好的控制效果. Govindhasamy 等将 ADHDP 方法应用于连续搅拌釜式反应器[139]. Iyer 等将一类 ACDs 方法应用于发酵罐的在线优化与控制中[140].

此外, ADP 方法还已经用于通讯服务网络中的呼叫准入控制方法[141−142]、分布式传感器网络的资源管理问题[143].

5 总结与展望

ADP 方法无论在理论研究还是实际应用中都取得了重要的进展, 并且展示出了巨大的发展潜力. 但对于实际中的复杂系统, ADP 目前在理论和算法上仍然有很多问题尚未解决. 通过以上对 ADP 的总结与分析, 本文最后给出如下几个重要的研究方向:

1) ADP 绝大部分理论都是基于值函数和控制策略精确迭代的, 而实际算法中所采用的函数近似结构会导致策略评价和策略提高过程存在近似误差. 因此, 考虑策略评价和策略提高存在近似误差时 ADP 算法的收敛性对于 ADP 理论和实际应用都是十分必要的.

2) ADP 方法目前解决的大多是状态和控制维数较低的系统, 而对于实际复杂系统中高维、连续的状态和控制空间, 目前尚没有有效的解决办法. 针对高维系统, 提出具有收敛速度快和计算复杂度低的 ADP 方法是非常重要的.

3) 目前大部分在线 ADP 算法都需要满足 PE 条件, 而在实际应用中 PE 条件很难验证, 如何提出更容易验证的 PE 条件或直接去掉 PE 条件将对在线 ADP 方法的应用起到重要的作用.

4) ADP 方法与自适应控制、鲁棒控制、模型预测控制、网络控制等现代控制理论都有着密切的联系, 借鉴现代控制理论的重要方法是十分必要的, 同时还可以拓宽 ADP 的适用范围, 解决系统带有噪声、时滞等挑战性问题.

5) 目前, ADP 的研究主要集中于单个控制器的设计, 而对于多个控制器的设计却鲜有涉及, 采用 ADP 方法实现大系统的分散控制和多智能体的分布式协调控制将是一个十分有益的探索.

6) 针对电网、交通等复杂的系统, 在无法获得精确数学模型及不确定环境下, 采用 ADP 方法求解大规模实际优化问题, 取得重大的实际应用.

References

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

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

3 Doya K. Reinforcement learning in continuous time and space. *Neural Computation*, 2000, **12**(1): 219−245

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

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

6 Werbos P J. Approximate dynamic programming for real-time control and neural modeling. *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. New York: Van Nostrand, 1992

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

8 Lewis F L, Huang J, Parisini T, Prokhorov D V, Wunsch D C. Special issue on neural networks for feedback control systems. *IEEE Transactions on Neural Networks*, 2007, **18**(4): 969−972

9 Lewis F L, Lendaris G, Liu D. Special issue on adaptive dynamic programming and reinforcement learning in feeedback control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 896−897

10 Ferrari S, Sarangapani J, Lewis F L. Special issue on approximate dynamic programming and reinforcement learning. *Journal of Control Theory and Applications*, 2011, **9**(3): 309

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

12 Si J, Barto A G, Powel W B, Wunsch D C. *Handbook of Learning and Approximate Dynamic Programming*. Piscataway, NJ: IEEE, 2004

13 Powel W B. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken, NJ: Wiley, 2007

14 Lucian B, Robert B, Bart S, Damien E. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, FL: CRC Press, 2010

15 Lewis F L, Liu D R. *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ: Wiley, 2013

16 Shan Q H, Liu D R, Luo Y H, Wang D. *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, UK: Springer, 2013

17 Xu Xin. *Reinforcement Learning and Approximate Dynamic Programming*. Beijing: Science Press, 2010
(徐昕. 增强学习与近似动态规划. 北京: 科学出版社, 2010)

18 Lewis F L, Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 2009, **9**(3): 32−50

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

20 Lewis F L, Vrabie D, Vamvoudakis K. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 2012, **32**(6): 76−105

21 Xu Xin, Shen Dong, Gao Yan-Qing, Wang Kai. Learning control of dynamical systems based on Markov decision processes: research frontiers and outlooks. *Acta Automatica Sinica*, 2012, **38**(5): 673−687
(徐昕, 沈栋, 高岩青, 王凯. 基于马氏决策过程模型的动态系统学习控制: 研究前沿与展望. 自动化学报, 2012, **38**(5): 673−687)

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

23 Geist M, Pietquin O. Algorithmic survey of parametric value function approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(6): 845−867

24 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

25 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

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

27 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

28 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 Transactions on Neural Networks*, 2011, **22**(12): 1851−1862

29 Wang F Y, Jin N, Liu D E, 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

30 Heydari A, Balakrishnan S N. Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(1): 145−157

31 Wang D, Liu D R, Wei Q L, Zhao D B, Jin N. Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica*, 2012, **48**(8): 1825−1832

32 Chen Z, Jagannathan S. Generalized Hamilton-Jacobi-Bellman formulation-based neural network control of affine nonlinear discrete time systems. *IEEE Transactions on Neural Networks*, 2008, **19**(1): 90−106

33 Watkins C J C H, Dayan P. Q-Learning. *Machine Learning*, 1992, **8**: 279−292

34 Liu D R, Wei Q L. Finite-approximation-error based optimal control approach for discrete-time nonlinear systems. *IEEE Transactions on Cybernetics*, 2013, **43**(2): 779−789

35 Seong C Y, Widrow B. Neural dynamic optimization for control systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2001, **31**(4): 482−513

36 Lewis F L, Vamvoudakis K G. Reinforcement learning for partially observable dynamic processes: adaptive dynamic programming using measured output data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2011, **41**(1): 14−25

37 He P A, Jagannathan S. Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2007, **37**(2): 425−436

38 Yang L, Si J, Tsakalis K S, Rodriguez A A. Direct heuristic dynamic programming for nonlinear tracking control with filtered tracking error. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2009, **39**(6): 1617−1622

39 Yang Q M, Jagannathan S. Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2012, **42**(2): 377−390

40 Yang Q M, Vance J B, Jagannathan S. Control of nonaffine nonlinear discrete-time systems using reinforcement-learning-based linearly parameterized neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 994−1001

41 Dierks T, Jagannathan S. Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(7): 1118−1129

42 Kleinman D. On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, 1968, **13**(1): 114−115

43 Saridis G N, Lee C S. An approximation theory of optimal control for trainable manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, 1979, **9**(3): 152−159

44 Wang F Y, Saridis G N. Suboptimal control for nonlinear stochastic systems. In: Proceedings of the 31st IEEE Conference on Decision and Control. Tucson, Arizona, USA: IEEE, 1992. 1856−1861

45 Saridis G N, Wang F Y. Suboptimal control for nonlinear stochastic systems. *Control Theory and Advanced Technology*, 1994, **10**(4): 847−871

46 Wang F Y, Saridis G N. On successive approximation of optimal control of stochastic dynamic systems. *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*. Boston, MA: Kluwer, 2002. 333−386

47 Beard R W, Saridis G N, Wen J T. Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation. *Automatica*, 1997, **33**(12): 2159−2177

48 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

49 Cheng T, Lewis F L, Abu-Khalaf M. Fixed-final-time-constrained optimal control of nonlinear systems using neural network HJB approach. *IEEE Transactions on Neural Networks*, 2007, **18**(6): 1725−1736

50 Cheng T, Lewis F L, Abu-Khalaf M. A neural network solution for fixed-final time optimal control of nonlinear systems. *Automatica*, 2007, **43**(3): 482−490

51 Tassa Y, Erez T. Least squares solutions of the HJB equation with neural network value-function approximators. *IEEE Transactions on Neural Networks*, 2007, **18**(4): 1031−1041

52 Hanselmann T, Noakes L, Zaknich A. Continuous time adaptive critics. *IEEE Transactions on Neural Networks*, 2007, **18**(3): 631−647

53 Ferrari S, Steck J E, Chandramohan R. Adaptive feedback control by constrained approximate dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 982−987

54 Seiffertt J, Sanyal S, Wunsch D C. Hamilton-Jacobi-Bellman equations and approximate dynamic programming on time scales. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 918−923

55 Vrabie D, Pastravanu O, Abu-Khalaf M, Lewis F L. Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, 2009, **45**(2): 477−484

56 Vrabie D, Lewis F L. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 2009, **22**(3): 23−246

57 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.

58 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. *Transactions on Neural Networks*, 2011, **22**(12): 2226−2236

59 Bhasin S, Kamalapurkar R, Johnson M, Vamvoudakis K G, Lewis F L, Dixon, W E. A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, 2013, **49**(1): 82−92

60 Mehta P, Meyn S. Q-learning and Pontryagin′s minimum principle. In: Proceedings of the 48th IEEE Conference on Decision and Control. Shanghai, China: IEEE, 2009. 3598−3605

61 Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 2012, **48**(10): 2699−2704

62 Lee J Y, Park J B, Choi Y H. Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems. *Automatica*, 2012, **48**(11): 2850−2859

63 Lee J Y, Park J B, Choi Y H. Integral reinforcement learning with explorations for continuous-time nonlinear systems. In: Proceedings of the 2012 IEEE World Congress on Computational Intelligence. Brisbane, Australia: IEEE, 2012. 1042−1047

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

65 Liu D R, Wang D, Zhao D B, Wei Q L, Jin N. Neural-network-based optimal control for a class of unknown discrete-time nonlinear systems using globalized dual heuristic programming. *IEEE Transactions on Automation Science and Engineering*, 2012, **9**(3): 628−634

66 Fairbank M, Alonso E, Prokhorov D. Simple and fast calculation of the second-order gradients for globalized dual heuristic dynamic programming in neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(10): 1671−1676

67 Si J, Wang Y T. Online learning control by association and reinforcement. *IEEE Transactions on Neural Networks*, 2001, **12**(2): 264−276

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

69 Ding J, Balakrishnan S N. Approximate dynamic programming solutions with a single network adaptive critic for a class of nonlinear systems. *Journal of Control Theory and Applications*, 2011, **9**(3): 370−380

70 Ni Z, He H B, Wen J Y. Adaptive learning in tracking control based on the dual critic network design. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(6): 913−928

71 Jin N, Liu D R, Huang T, Pang Z Y. Discrete-time adaptive dynamic programming using wavelet basis function neural networks. In: Proceedings of the 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning. Honolulu, HI: IEEE, 2007. 135−142

72 Deb A K, Jayadeva, Gopal M. SVM-based tree-type neural networks as a critic in adaptive critic designs for control. *IEEE Transactions on Neural Networks*, 2007, **18**(4): 1016−1030

73 Eaton P H, Prokhorov D V, Wunsch D C II. Neurocontroller alternatives for fuzzy ball-and-beam systems with nonuniform nonlinear friction. *IEEE Transactions on Neural Networks*, 2000, **11**(2): 432−435

74 Koprinkova-Hristova P, Oubbati M, Palm G. Adaptive critic design with echo state network. In: Proceedings of the 2010 IEEE International Conference on Systems Man and Cybernetics. Sofia, Bulgaria: IEEE, 2010. 1010−1015

75 Xu X, Hou Z S, Lian C Q, He H B. Online learning control using adaptive critic designs with sparse kernel machines. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(5): 762−775

76 Fu J, He H, Zhou X. Adaptive learning and control for MIMO system based on adaptive dynamic programming. *IEEE Transactions on Neural Networks*, 2011, **22**(7): 1133−1148

77 Mohagheghi S, Venayagamoorthy G K, Harley R G. Fully evolvable optimal neurofuzzy controller using adaptive critic designs. *IEEE Transactions on Fuzzy Systems*, 2008, **16**(6): 1450−1461

78 Kulkarni R V, Venayagamoorthy G K. Adaptive critics for dynamic optimization. *Neural Networks*, 2010, **23**(5): 587−591

79 Kang Qi, Wang Lei, An Jing, Wu Qi-Di. Approximate dynamic programming based parameter optimization of particle swarm systems. *Acta Automatica Sinica*, 2010, **36**(8): 1171−1181
(康琦, 汪镭, 安静, 吴启迪. 基于近似动态规划的微粒群系统参数优化研究. 自动化学报, 2010, **36**(8): 1171−1181)

80 Jiang Y, Jiang Z P. Robust adaptive dynamic programming with an application to power systems. *IEEE Transactions on Neural Networks*, 2013, **24**(7): 1150−1156

81 Jiang Y, Jiang Z P. Approximate dynamic programming for optimal stationary control with control-dependent noise. *IEEE Transactions on Neural Networks*, 2011, **22**(12): 2392−2398

82 Varbie D, Lewis F L. Adaptive dynamic programming algorithm for finding online the equilibrium solution of the two-player zero-sum differential game. In: Proceedings of the 2010 International Joint Conference on Neural Networks. Barcelona, Spain: IEEE, 2010. 1−8

83 Varbie D, Lewis F L. Adaptive dynamic programming for online solution of a zero-sum differential game. *Journal of Control Theory and Applications*, 2011, **9**(3): 353−360

84 Wu H, Luo B. Simultaneous policy update algorithms for learning the solution of linear continuous-time H_∞ state feedback control. *Information Sciences*, 2013, **222**(10): 472−485

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

86 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

87 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

88 Vamvoudakis K G, Lewis F L. Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *International Journal of Robust and Nonlinear Control*, 2012, **22**(13): 1460−1483

89 Wu H N, Luo B. Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear H_∞ control. *IEEE Transactions on Neural Networks*, 2012, **23**(12): 1884−1895

90 Johnson M, Bhasin S, Dixon W E. Nonlinear two-player zero-sum game approximate solution using a policy iteration algorithm. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference. Orlando, USA: IEEE, 2011. 142−147

91 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

92 Al-Tamimi A, Lewis F L, Abu-Khalaf M. Modelfree Q-learning designs for linear discretet-time zero-sum games with application to H_∞ control. *Automatica*, 2007, **43**(3): 473−481

93 Kim J H, Lewis F L. Model-free H_∞ control design for unknown linear discrete-time systems via Q-learning with LMI. *Automatica*, 2010, **46**(8): 1320−1326

94 Mehraeen S, Dierks T, Jagannathan S, Crow M L. Zero-sum two-player game theoretic formulation of affine nonlinear discrete-time systems using neural networks. In: Proceedings of the 2010 International Joint Conference on Neural Networks. Barcelona, Spain: IEEE, 2010. 1−8

95 Liu D R, Li H L, Wang D. H_∞ control of unknown discrete-time nonlinear systems with control constraints using adaptive dynamic programming. In: Proceedings of the 2012 International Joint Conference on Neural Networks. Brisbane, Australia: IEEE, 2012. 3056−3061

96 Vrabie D, Lewis F L. Integral reinforcement learning for online computation of feedback Nash strategies of nonzero-sum differential games. In: Proceedings of the 49th IEEE Conference on Decision and Control. Atlanta, GA: IEEE, 2010. 3066−3071

97 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

98 Vamvoudakis K G, Lewis F L, Hudas G R. Multi-agent differential graphical games: online adaptive learning solution for synchronization with optimality. *Automatica*, 2012, **48**(8): 1598−1611

99 Zhang H G, Cui L L, Luo Y H. Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP. *IEEE Transactions on Cybernetics*, 2013, **43**(1): 206−216

100 Vamvoudakis K G, Lewis F L, Johnson M, Dixon W E. Online learning algorithm for Stackelberg games in problems with hierarchy. In: Proceedings of the 51st IEEE Conference on Decision and Control. Maui, Hawaii, USA: IEEE, 2012. 1883−1889

101 Mehraeen S, Jagannathan S. Decentralized optimal control of a class of interconnected nonlinear discrete-Time systems by using online Hamilton-Jacobi-Bellman formulation. *IEEE Transactions on Neural Networks*, 2011, **22**(11): 1757−1769

102 Jiang Y, Jiang Z P. Robust adaptive dynamic programming for large-scale systems with an application to multimachine power systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2012, **59**(10): 693−697

103 Xu H, Jagannathan S, Lewis F L. Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses. *Automatica*, 2012, **48**(6): 1017−1030

104 Xu H, Jagannathan S. Stochastic optimal controller design for uncertain nonlinear networked control system via neuro dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(5): 471−484

105 Bertsekas D P. Dynamic programming and suboptimal control: a survey from ADP to MPC. *European Journal of Control*, 2005, **11**(4−5): 310−334

106 Cox C J, Stepniewski S W, Jorgensen C C, Saeks R. On the design of a neural network autolander. *International Journal of Robust and Nonlinear Control*, 1999, **9**: 1071−1096

107 Enns R, Si J. Helicopter trimming and tracking control using direct neural dynamic programming. *IEEE Transactions on Neural Networks*, 2003, **14**(4): 929−939

108 Nodland D, Zargarzadeh H, Jagannathan S. Neural-network-based optimal adaptive output feedback control of a helicopter UAV. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(7): 1061−1073

109 Lin C. Adaptive critic autopilot design of bank-to-turn missiles using fuzzy basis function networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **35**(2): 929−939

110 Han D, Balakrishnan S N. State-constrained agile missile control with adaptive-critic-based neural networks. *IEEE Transactions on Control Systems Technology*, 2002, **10**(4): 481−489

111 Lin W S, Chang L H, Yang P C. Adaptive critic anti-slip control of wheeled autonomous robot. *Automatica*, 2008, **44**(11): 2716−2723

112 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

113 Shih P, Kaul B C, Jagannathan S, Drallmeier J A. Reinforcement-learning-based output-feedback control of nonstrict nonlinear discrete-time systems with application to engine emission control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2009, **39**(5): 1162−1179

114 Peter S, Brian C K, Jagannathan S, Drallmeier J A. Reinforcement-learning-based dual-control methodology for complex nonlinear discrete-time systems with application to spark engine EGR operation. *IEEE Transactions on Neural Networks*, 2008, **19**(8): 1369−1388

115 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

116 Liu W X, Venayagamoorthy G K, Wunsch D C. 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): 1377−385

117 Mohagheghi S, Venayagamoorthy G K, Harley R G. Adaptive critic design based neuro-fuzzy controller for a static compensator in a multimachine power system. *IEEE Transactions on Power Systems*, 2006, **21**(4): 1744−1754

118 Mohagheghi S, Valle Y, Venayagamoorthy G K, Harley R G. A proportional-integrator type adaptive critic design-based neurocontroller for a static compensator in a multimachine power system. *IEEE Transactions on Industry Applications Electronics*, 2007, **54**(1): 86−96

119 Ray S, Venayagamoorthy G K, Chaudhuri B, Majumder R. Comparison of adaptive critics and classical approaches based wide area controllers for a power system. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 1002−1007

120 Qiao W, Harley R G, Venayagamoorthy G K. Coordinated reactive power control of a large wind farm and a STATCOM using heuristic dynamic programming. *IEEE Transactions on Energy Conversion*, 2009, **24**(2): 493−503

121 Liang J Q, Venayagamoorthy G K, Harley R G. Wide-area measurement based dynamic stochastic optimal power flow control for smart grids with high variability and uncertainty. *IEEE Transactions on Smart Grid*, 2012, **3**(1): 59−69

122 Lu C, Si J N, 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

123 Huang T, Liu D E. A self-learning scheme for residential energy system control and management. *Neural Computing and Applications*, 2013, **2**(2): 259−269

124 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): 676−681
(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): 676−681)

125 Lin W S, Sheu J W. Metro traffic regulation by adaptive optimal control. *EEE Transactions on Intelligent Transportation System*, 2011, **12**(4): 1064−1073

126 Lin W S, Sheu J W. Optimization of train regulation and energy usage of metro lines using an adaptive-optimal-control algorithm. *IEEE Transactions on Automation Science and Engineering*, 2011, **8**(4): 855−864

127 Sheu J W, Lin W S. Adaptive optimal control for designing automatic train regulation for metro line. *IEEE Transactions on Control Systems Technology*, 2012, **20**(5): 1319−1327

128 Sheu J W, Lin W S. Energy-saving automatic train regulation using dual heuristic programming. *IEEE Transactions on Vehicular Technology*, 2012, **61**(4): 1503−1514

129 Zhao D B, Bai X R, Wang F Y, Xu J, Yu W. DHP method for ramp metering of freeway traffic. *EEE Transactions on Intelligent Transportation System*, 2011, **12**(4): 990−999

130 Cai C, Wong C K, Heydecker B G. Adaptive traffic signal control using approximate dynamic programming. *Transportation Research Part C*, 2009, **17**(5): 456−474

131 Shervais S, Shannon T T, Lendaris G G. Intelligent supply chain management using adaptive critic learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 2003, **33**(2): 235−244

132 Sun Z, Chen X, He Z Z. Adaptive critic design for energy minimization of portable video communication devices. *IEEE Transactions on Circuits and Systems for Video Technology*, 2010, **20**(1): 37−37

133 Iftekharuddin K M. Transformation invariant on-line target recognition. *IEEE Transactions on Neural Networks*, 2011, **22**(6): 906−918

134 Venayagamoorthy G K, Zha W. Comparison of nonuniform optimal quantizer designs for speech coding with adaptive critics and particle swarm. *IEEE Transactions on Industry Applications*, 2007, **43**(1): 238−244

135 Lee J M, Lee J H. Approximate dynamic programming-based approaches for input-output data-driven control of nonlinear processes. *Automatica*, 2005, **41**(7): 1281−1288

136 Lee J M, Lee J H. An approximate dynamic programming based approach to dual adaptive control. *Journal of Process Control*, 2009, **19**(1): 85−864

137 Lee J M, Kaisare N S, Lee J H. Choice of approximator and design of penalty function for an approximate dynamic programming based control approach. *Journal of Process Control*, 2006, **16**(2): 135−156

138 Lee J M, Lee J H. Value function-based approach to the scheduling of multiple controllers. *Journal of Process Control*, 2008, **18**(6): 533−542

139 Govindhasamy J J, McLoone S F, Irwin G W. Second-order training of adaptive critics for online process control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, **35**(2): 381−385

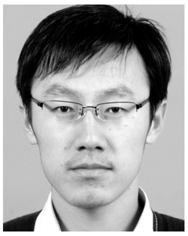
140 Iyer M S, Wunsch D C. Dynamic re-optimization of a fed-batch fermentor using adaptive critic designs. *IEEE Transactions on Neural Networks*, 2001, **12**(6): 1433−1444

141 Marbach P, Mihatsch O, Tsitsiklis J N. Call admission control and routing in integrated services networks using neuro-dynamic programming. *IEEE Journal on Selected Areas in Communications*, 2000, **18**(2): 197−208

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

143 Williams J L, Fisher J W, Willsky A S. Approximate dynamic programming for communication-constrained sensor network management. *IEEE Transactions on Signal Processing*, 2007, **55**(8): 3995−4003

刘德荣 中国科学院自动化研究所研究员. 主要研究方向为智能控制理论及应用, 自适应动态规划, 人工神经网络, 计算神经科学, 电力系统运行与控制. 本文通信作者. E-mail: derong.liu@ia.ac.cn
(**LIU De-Rong** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers intelligent control, adaptive dynamic programming, neural networks, computational neuroscience, and power systems. Corresponding author of this paper.)



李宏亮 中国科学院自动化研究所博士研究生. 主要研究方向为机器学习, 自适应动态规划, 神经网络, 微分博弈, 电力系统运行与控制.
E-mail: hongliang.li@ia.ac.cn
(**LI Hong-Liang** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers machine learning, adaptive dynamic programming, neural networks, differential game, and power systems.)

王 鼎 中国科学院自动化研究所助理研究员. 主要研究方向为自适应动态规划, 神经网络与学习系统, 智能控制.
E-mail: ding.wang@ia.ac.cn
(**WANG Ding** Assistant professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers adaptive dynamic programming, neural networks and learning systems, and intelligent control.)