自动化学报
ACTA AUTOMATICA SINICA

带有输入时滞的非线性系统基于学习的输出反馈控制

刘思彤 高伟男 姜钟平

Learning-based Output-feedback Control for Nonlinear Systems With Input Time-delay

LIU Si-Tong, GAO Wei-Nan, JIANG Zhong-Ping

在线阅读 View online: https://doi.org/10.16383/j.aas.c250101

带有输入时滞的非线性系统基于学习的输出反馈控制

刘思彤 [1] 高伟男 [1] 姜钟平 [2]

摘 要 针对具有输入时滞的非线性系统直接自适应最优控制问题, 提出一种新的数据驱动输出反馈控制方法. 该方法通过融合 Q 学习与值迭代和策略迭代, 在学习过程中无需依赖系统动力学知识. 在系统满足一致可观性的条件下, 提出一种基于输出数据和带有时滞的输入数据的系统状态重构方法, 基于值迭代和策略迭代来学习自适应最优控制策略. 最后, 将该方法应用于范德波尔振荡器这一经典非线性系统的控制, 并通过仿真结果充分验证了该方法的有效性.

关键词 最优控制, 输出反馈, 时滞, 自适应动态规划

引用格式 刘思彤, 高伟男, 姜钟平. 带有输入时滞的非线性系统基于学习的输出反馈控制. 自动化学报, 2025, **51**(10): 1001−1009

DOI 10.16383/j.aas.c250101 **CSTR** 32138.14.j.aas.c250101

Learning-based Output-feedback Control for Nonlinear Systems With Input Time-delay

LIU Si-Tong[1] GAO Wei-Nan[1] JIANG Zhong-Ping[2]

Abstract This paper proposes a new data-driven output-feedback control method to address the direct adaptive optimal control problem for nonlinear systems with input time-delay. The combination of Q-learning with value iteration (VI) and policy iteration (PI) enables the learning process to be conducted without any knowledge of the system dynamics. Under the condition that the system is uniformly observable, we propose a novel method to reconstruct the state of the system based on output data and input data with time-delay. We then present two iterative methods, VI and PI, to learn the adaptive optimal control policy. Finally, the proposed methods are applied to the classical nonlinear system control——Van der Pol oscillator. The simulation results demonstrate the effectiveness of the proposed methods.

Key words Optimal control, output-feedback, time-delay, adaptive dynamic programming

Citation Liu Si-Tong, Gao Wei-Nan, Jiang Zhong-Ping. Learning-based output-feedback control for nonlinear systems with input time-delay. *Acta Automatica Sinica*, 2025, **51**(10): 1001−1009

最优控制问题的核心目标在于设计能够使代价函数最小的最优控制策略. 然而, 现有最优控制研究往往依赖于系统动力学知识[1]. 随着现代工程系统动态愈加复杂, 获取完整的动力学信息变得尤为困难. 为应对这一挑战, 近年来涌现出多种数据驱动方法[2−5], 这些方法通过利用系统的输入输出数据直接进行控制策略的学习和优化, 无需依赖精确的系统模型. 其中无模型自适应控制[6−8] 方法计算量小, 适合实时控制, 适用于工业过程控制、机器人控制等场景. 针对非线性系统的最优控制问题, 传统方法常将其转化为求解哈密顿−雅可比−贝尔曼方程, 但直接求解该方程复杂度高、计算量大. 自适应动态规划方法作为一种数据驱动技术, 无需依赖系统模型设计最优控制器, 并且能有效降低计算复杂度. 因此, 该方法受到学者们的广泛关注[9−18]. 与直接求解哈密顿−雅可比−贝尔曼方程的方式不同, 自适应动态规划方法通过迭代的方式来逼近最优控制策略, 从而有效减轻计算负担[19−22].

策略迭代 (Policy iteration, PI) 和值迭代 (Value iteration, VI) 是强化学习和动态规划中的两种典型迭代方法, 它们能够在学习过程中同步逼近最优值函数及最优控制策略. 其中, 策略迭代需从可接受的控制策略出发, 且每次迭代生成的策略均保持可接受性[23−24]. 然而, 在模型动态未知的情况下, 获取此类可接受的初始策略往往颇具挑战. 与策略迭代不同, 值迭代无需可接受的初始控制策略, 允

收稿日期 2025-03-13 录用日期 2025-08-04
Manuscript received March 13, 2025; accepted August 4, 2025
国家重点研发计划 (2024YFA1012702), 国家自然科学基金 (62373090, 62521001), 辽宁省兴辽英才计划 (XLYC2403177) 资助
Supported by National Key Research and Development Program of China (2024YFA1012702), National Natural Science Foundation of China (62373090, 62521001), and Liaoning Revitalization Talents Program (XLYC2403177)
本文责任编委 王金枝
Recommended by Associate Editor WANG Jin-Zhi
1. 东北大学流程工业综合自动化全国重点实验室 沈阳 110819 中国 2. 纽约大学坦登工程学院电子与计算机工程系 纽约 NY11201 美国
1. State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China 2. Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, New York NY11201, USA

许从任意控制策略开始学习, 这一特性使得值迭代在缺乏可接受初始控制策略的无模型场景中更具优势[25−26].

Q 学习是一种求解动态规划问题的迭代方法. 具体而言, Q 学习通过求解 Q 函数来寻求最优的值函数和最优的控制策略. 值得注意的是, Q 学习不依赖于系统模型信息, 通过利用所有状态和输入信息来更新迭代 Q 函数, 从而在实际应用中展现出更高的有效性. Q 学习可以与策略迭代和值迭代方法相结合, 在完全无系统模型信息的动态系统中探寻最优反馈控制策略[23, 27].

大多数现有非线性系统自适应动态规划方法的研究假设系统全状态信息可测. 然而在许多复杂系统中, 完整测量系统状态信息往往成本高昂或甚至根本不能实现, 在系统动态未知时, 往往不能设计观测器估计系统状态. 因此, 如何利用输入输出信息重构状态是一个理论挑战, 亟需研究基于输出反馈的自适应动态规划方法. 另外, 通讯时滞在实际应用中非常常见, 且时滞如处理不当会严重影响控制性能, 甚至导致系统不稳定, 因此带有时滞的控制问题已成为学者们关注的主要领域[28−31]. 对于线性系统, 系统输入输出与状态之间有相对明确的线性对应关系[32−33]. 但对于系统模型未知、含有输入时滞的非线性系统, 寻找这种关系更为复杂. 为应对以上挑战, 本文针对含有输入时延的非线性离散时间系统, 基于自适应动态规划和非线性最优控制理论, 设计输出反馈自适应最优控制器.

本文的主要贡献如下:

1) 针对一致可观的非线性离散时间仿射系统, 基于系统输入和输出数据, 解决带有时滞的非线性系统状态重构难题.

2) 将 Q 学习与策略迭代和值迭代相结合, 提出无需系统模型信息的迭代方法, 可以针对带有时滞的非线性系统设计输出反馈自适应最优控制.

3) 将所提出方法应用于范德波尔振荡器, 通过仿真验证了方法的有效性.

本文结构安排如下: 第 1 节阐述离散时间非线性系统的最优控制问题, 并回顾相关基础知识; 第 2 节提出状态重构方法, 以及基于模型的值迭代与策略迭代算法; 第 3 节提出数据驱动的值迭代与策略迭代方法; 第 4 节将所提方法应用于范德波尔振荡器, 通过仿真验证其有效性; 最后, 第 5 节对全文进行总结.

1 问题阐述与预备知识

考虑一类带有输入时滞的非线性离散时间仿射系统

$$\begin{cases} x_{k+1} = f(x_k) + g(x_k)u_{k-d} \\ y_k = h(x_k) \end{cases} \tag{1}$$

其中, $k \in \mathbf{N}^+$, $x_k \in \mathbf{R}^n$ 是状态, $d \in \mathbf{N}^+$ 代表输入时延步数, $u_{k-d} \in \mathbf{R}$ 是控制输入. 函数 $f : \mathbf{R}^n \to \mathbf{R}^n$, $g : \mathbf{R}^n \to \mathbf{R}^n$ 和函数 $h : \mathbf{R}^n \to \mathbf{R}$ 都是局部利普希茨函数并且满足条件 $f(0) = 0$, $h(0) = 0$.

本文的目标是基于输出反馈来构建一个自适应最优控制策略, 该策略可以使以下代价函数最小:

$$J = \sum_{k=d}^{\infty}[C(y_k) + u_{k-d}^2] \tag{2}$$

其中, $C : \mathbf{R} \to \mathbf{R}$ 是一个径向无界的正定函数.

基于非线性最优控制理论, 可以找到最优的价值函数 $V^*(x_k)$ 和对应的最优控制策略 $u_{k-d}^* = u^*(x_k)$, 形式如下:

$$\begin{cases} V^*(x_k) = \min\limits_{u} \left(C(y_k) + u^2 + V^*(x_{k+1})\right) \\ u^*(x_k) = \arg\min\limits_{u} \left(C(y_k) + u^2 + V^*(x_{k+1})\right) \end{cases} \tag{3}$$

最优的控制策略可以具体表述为

$$u_{k-d}^* = u^*(x_k) = -\frac{1}{2}g^{\mathrm{T}}(x_k)\frac{\partial V^*(x_{k+1})}{\partial x_{k+1}} \tag{4}$$

为确保最优控制策略与最优值函数存在, 可作出如下假设.

假设 1. 以下的哈密顿−雅可比−贝尔曼方程存在唯一解, 即正定值函数 V^*:

$$V^*(x_k) = C(y_k) + V^*(x_{k+1}) + \frac{1}{4}\frac{\partial V^{*\mathrm{T}}(x_{k+1})}{\partial x_{k+1}}g(x_k)g^{\mathrm{T}}(x_k)\frac{\partial V^*(x_{k+1})}{\partial x_{k+1}} \tag{5}$$

2 基于模型的输出反馈值迭代和输出反馈策略迭代

在本节中, 首先针对一致可观的非线性系统 (1) 提出一种基于历史输入输出数据的状态重构方法. 接着, 介绍基于模型的值迭代和策略迭代算法, 旨在逼近输出反馈最优控制策略.

2.1 一致可观的非线性系统状态重构

令 $U_{[k-n,\,k-1]}$ 和 $Y_{[k-n,\,k-1]}$ 分别代表由输入输出数据构成的向量, 它们的具体形式为

$$\begin{cases} U_{[k-n,\,k-1]} = [u_{k-n},\ u_{k-n+1},\ \cdots,\ u_{k-1}]^{\mathrm{T}} \\ Y_{[k-n,\,k-1]} = [y_{k-n},\ y_{k-n+1},\ \cdots,\ y_{k-1}]^{\mathrm{T}} \end{cases} \tag{6}$$

为便于表述, 定义

$$F^u(x) = f(x) + g(x)u \tag{7}$$

那么, 系统状态就可以通过如下的输入输出序列进行表示

$$x_k = \Phi(x_{k-n-d},\ U_{[k-n-2d,\ k-d-1]}) \tag{8}$$

$$Y_{[k-n-d,\ k-d-1]} = H(x_{k-n-d},\ U_{[k-n-2d,\ k-d-1]}) \tag{9}$$

其中

$$\Phi(x_{k-n-d},\ U_{[k-n-2d,\ k-d-1]}) = F^{u_{k-d-1}} \circ F^{u_{k-d-2}} \circ \cdots \circ F^{u_{k-n-2d}}(x_{k-n-d})$$

$$H(x_{k-n-d},\ U_{[k-n-2d,\ k-d-1]}) = \begin{bmatrix} h(x_{k-n-d}) \\ \vdots \\ h \circ F^{u_{k-2d-2}} \circ \cdots \circ F^{u_{k-n-2d}}(x_{k-n-d}) \end{bmatrix}$$

上述式中的“$\circ$”代表复合函数算子. 下面给出局部一致可观的概念.

定义 1[34]**.** 系统 (1) 在集合 $\boldsymbol{X} \subset \mathbf{R}^n$, $\boldsymbol{Y} \subset \mathbf{R}$ 和 $\boldsymbol{U} \subset \mathbf{R}$ 上是局部一致可观的充要条件是映射 $H^*: \boldsymbol{X} \times \boldsymbol{U} \to \boldsymbol{Y} \times \boldsymbol{U}$ 是单射. 当 $\boldsymbol{X}$, $\boldsymbol{U}$ 和 $\boldsymbol{Y}$ 分别是具有对应维度的实数集合, 则该系统是一致可观的.

注 1. 可观的线性时不变系统和非线性严格反馈系统都是一致可观的. 与持续激励条件不同, 非线性系统是否满足 (局部) 一致可观条件, 本质上主要由系统动态决定.

为确保状态可以由输入输出数据唯一确定, 作出以下假设.

假设 2[2]**.** 以 u_k 作为输入, y_k 作为输出, 则 x_k 作为状态的系统 (1) 是局部一致可观的.

在假设 2 的条件下, 可以通过以下引理利用输入输出数据来重构系统的状态.

引理 1. 在假设 2 的条件下, 存在一个函数 $\Theta: \mathcal{Z} \to \boldsymbol{X}$, 使得对于任意的输入 $U_{[k-n-d,\ k-1]} \in \boldsymbol{U}$ 和输出 $Y_{[k-n,\ k-1]} \in \boldsymbol{Y}$, 有

$$x_{k+d} = \Theta(z_k) \tag{10}$$

其中, $z_k = [Y^{\mathrm{T}}_{[k-n,\ k-1]} \quad U^{\mathrm{T}}_{[k-n-d,\ k-1]}]^{\mathrm{T}}$, $\mathcal{Z} = \boldsymbol{Y} \times \boldsymbol{U}$.

证明. 在假设 2 和引理 1 条件下, 映射 H 是一种单射, 这意味着对于输入序列 $U_{[k-n-2d,\ k-d-1]} \in \boldsymbol{U}$ 和输出序列 $Y_{[k-n-d,\ k-d-1]}$ 而言, 总存在唯一的 $x_{k-n-d} \in \boldsymbol{X}$ 满足式 (9). 根据式 (8) 可以进一步看出 x_k 是由 x_{k-n-d} 所唯一确定的. 由此可得式 (8) 和式 (9) 的解是唯一的. 换言之, x_k 和 x_{k-n-d} 是由 $U_{[k-n-2d,\ k-d-1]}$ 和 $Y_{[k-n-d,\ k-d-1]}$ 唯一确定的. 因此, 存在一个函数 $\Theta: \mathcal{Z} \to \boldsymbol{X}$ 使得对于施加到系统的任意输入 $U_{[k-n-d,\ k-1]}$ 和输出 $Y_{[k-n,\ k-1]}$, 都有以下关系成立

$$x_{k+d} = \Theta(z_k)$$

□

基于哈密顿–雅可比–贝尔曼方程 (5) 的解及状态重构的结果 (10), 可以设计出最优的输出反馈控制策略:

$$\bar{u}_k^* = \bar{u}^*(z_k) = u^*(\Theta(z_k)) = u^*(x_{k+d}) \tag{11}$$

这与式 (4) 是完全等价的.

时滞系统的输出反馈控制器设计思路如图 1 所示. 首先, 在状态重构器中基于系统的动态输入输出数据重构系统状态; 其次, 在状态预估器中基于重构的状态和系统模型预估系统状态; 最后, 基于预估状态设计反馈控制输入. 为摆脱对系统准确模型信息的需求, 将在第 3 节中基于系统输入输出数据和自适应优化控制方法直接设计动态输出反馈控制器.

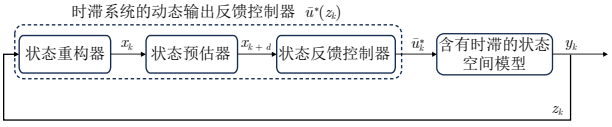


图 1　时滞系统的输出反馈控制器设计思路

Fig. 1　Design approach for output feedback controllers with time-delay system

2.2　基于模型的输出反馈值迭代和输出反馈策略迭代

由于直接从哈密顿–雅可比–贝尔曼方程求解最优的输出反馈控制策略 (11) 的计算代价极高, 本文采用值迭代和策略迭代方法来逼近最优控制策略及其对应的值函数. 为在后续的数据驱动方法中完全不依赖系统的模型信息, 将 Q 学习与策略迭代和值迭代相结合. 定义 Q 函数如下:

$$\mathcal{Q}_i(x_{k+d},\ u_k) = C(y_{k+d}) + u_k^2 + V_i(x_{k+d+1}) \tag{12}$$

$$\bar{\mathcal{Q}}_i(z_k,\ u_k) = C(y_{k+d}) + u_k^2 + V_i(\Theta(z_{k+1})) \tag{13}$$

其中, i 代表迭代次数. 基于引理 1, 可以发现 $\mathcal{Q}_i(x_{k+d},\ u_k)$ 与 $\bar{\mathcal{Q}}_i(z_k,\ u_k)$ 等价. 接下来, 设计值迭代方法和策略迭代方法, 以逼近最优的输出反馈控制策略和值函数.

算法 1. 模型的输出反馈值迭代算法

1. 选择 $\bar{\mathcal{Q}}_0$ 作为初始的 Q 函数, 选取 $\epsilon > 0$, 令 $i \leftarrow 0$;
2. 更新 Q 函数 $\bar{\mathcal{Q}}_{i+1}$

$$\bar{\mathcal{Q}}_{i+1}(z_k,\ u_k) = \bar{\mathcal{Q}}_i(z_{k+1},\ \bar{u}_i(z_{k+1})) + C(y_{k+d}) + u_k^2$$

3. 通过下式更新策略

$$\bar{u}_{i+1}(z_k) = \arg\min_{u_k} \bar{\mathcal{Q}}_{i+1}(z_k,\ u_k)$$

4. $i \leftarrow i+1$;

5. 重复执行步骤 2 ~ 4, 直至 $\bar{\mathcal{Q}}_{i+1}(z_k,\ u_k) - \bar{\mathcal{Q}}_i(z_k,\ u_k) < \epsilon$.

引理 2. 考虑步骤 2 ~ 4 产生的 $\bar{\mathcal{Q}}_i$ 和 $\bar{u}_i$. 对于任意的 $i = 1,\ 2,\ \cdots$, 满足以下条件:

1) $\bar{\mathcal{Q}}_{i+1}(z,\ u) \geq \bar{\mathcal{Q}}_i(z,\ u)$;

2) 对于每个固定的 $(z,\ u)$, $\{\bar{\mathcal{Q}}_{i+1}(z,\ u)\}_{i=0}^{\infty}$ 和 $\{\bar{u}_i(z)\}_{i=0}^{\infty}$, 分别收敛到 $\bar{\mathcal{Q}}^*(z,\ u)$ 和 $\bar{u}^*(z)$.

注 2. 在文献 [2] 中已经证明, 对于无时滞的系统, 值迭代算法下的 Q 函数序列是单调不减的, 并且随着 i 趋于无穷, Q 函数收敛到 $\bar{\mathcal{Q}}^*$, 同时 $\bar{u}_i$ 收敛到 $\bar{u}^*$. 基于引理 1, 本文所提出的输出反馈值迭代算法在每一代中获得的 Q 函数和相应的控制策略与文献 [2] 是完全等价的, 所以对于带时滞的系统的 Q 函数同样是单调不减的, 并随着 i 趋于无穷, Q 函数收敛到 $\bar{\mathcal{Q}}^*$, 同时 $\bar{u}_i$ 收敛到 $\bar{u}^*$.

算法 2. 模型的输出反馈策略迭代算法

1. 选择一个初始可接受的控制策略 $\bar{u}_0$, 选取 $\epsilon > 0$, 令 $i \leftarrow 0$;

2. 更新 Q 函数 $\bar{\mathcal{Q}}_i$

$$\bar{\mathcal{Q}}_i(z_k,\ u_k) = \bar{\mathcal{Q}}_i(z_{k+1},\ \bar{u}_i(z_{k+1})) + C(y_{k+d}) + \bar{u}_i^2(z_k)$$

3. 通过下式更新策略

$$\bar{u}_{i+1}(z_k) = \arg\min_{u_k} \bar{\mathcal{Q}}_i(z_k,\ u_k)$$

4. $i \leftarrow i+1$;

5. 重复执行步骤 2 ~ 4, 直至 $\bar{\mathcal{Q}}_{i+1}(z_k,\ u_k) - \bar{\mathcal{Q}}_i(z_k,\ u_k) < \epsilon$.

定理 1. 算法 2 中步骤 2 ~ 4 产生的 $\bar{\mathcal{Q}}_i$ 和 $\bar{u}_i$, 随着 i 趋于无穷大, Q 函数和控制策略会单调收敛到最优的 Q 函数和最优的控制策略.

证明. 在文献 [35] 中已经证明了由策略迭代产生的值函数和控制策略会单调收敛到最优的值函数和最优的控制策略. 由此可得

$$C(y_{k+d+1}) + u_{i+1}^2(x_{k+d+1}) + V_i(F^{u_{i+1}(x_{k+d+1})}(x_{k+d+1})) \leq V_i(x_{k+d+1}) \tag{14}$$

根据式 (12) 和式 (13), 可得

$$C(y_{k+d}) + \bar{u}_{i+1}^2(z_k) + \bar{\mathcal{Q}}_i(z_{k+1},\ \bar{u}_{i+1}(z_{k+1})) \leq \bar{\mathcal{Q}}_i(z_k,\ u_k) \tag{15}$$

因此, 可得

$$\bar{\mathcal{Q}}_{i+1}(z_k,\ u_k) \leq \bar{\mathcal{Q}}_i(z_k,\ u_k) \tag{16}$$

换言之, $\{\bar{\mathcal{Q}}_i(z,\ u)\}_{i=0}^{\infty}$ 是一个单调递减的序列. 又因为 Q 函数是正定函数且存在下界, 因此输出反馈的策略迭代会收敛. 在假设 1 的条件下, 可以得出结论, 算法会收敛到哈密顿–雅可比–贝尔曼方程的解, 即最优的值函数和最优控制策略. □

3 数据驱动的输出反馈值迭代和输出反馈策略迭代

在本节中, 将数据驱动的思想应用于输出反馈值迭代和输出反馈策略迭代方法设计. 该方法不依赖于系统动力学知识, 并且能够同时学习最优的 Q 函数和最优的控制策略. 为实现数据驱动的输出反馈值迭代和输出反馈策略迭代, 通过以下公式近似 Q 函数

$$\hat{\mathcal{Q}}_i(z,\ u) = \Psi^{\mathrm{T}}(z,\ u)\hat{s}_i \tag{17}$$

$$\Psi = \begin{bmatrix} \psi_1(z,\ u) \\ \psi_2(z,\ u) \\ \vdots \\ \psi_{N_1}(z,\ u) \end{bmatrix} \tag{18}$$

其中, $\{\psi_j(z,\ u)\}_{j=1}^{N_1}$ 是一组线性无关的光滑基函数, N_1 是一个足够大的整数. $\hat{s}_i \in \mathbf{R}^{N_1}$ 是对应于光滑基函数的权重. 可以使用 $\hat{\mathcal{Q}}_i(z,\ u)$ 作为原 Q 函数的近似, 来替代 $\bar{\mathcal{Q}}_i(z,\ u)$.

3.1 数据驱动的输出反馈值迭代

将原 Q 函数用它的近似替代后可得

$$\Psi^{\mathrm{T}}(z_k,\ u_k)\hat{s}_{i+1} = \hat{\mathcal{Q}}_i(z_{k+1},\ \hat{u}_i(z_{k+1})) + C(y_{k+d}) + u_k^2 \tag{19}$$

$$\hat{u}_i(z_k) = \arg\min_{u_k} \hat{\mathcal{Q}}_i(z_k,\ u_k) \tag{20}$$

各基函数的权重可以根据最小二乘法求解, 具体形式为

$$\hat{s}_{i+1} = [\Omega_v^{\mathrm{T}}\Omega_v]^{-1}\Omega_v^{\mathrm{T}}\mathcal{R}_v \tag{21}$$

其中

$$\Omega_v = \begin{bmatrix} \Psi^{\mathrm{T}}(z_k,\ u_k) \\ \Psi^{\mathrm{T}}(z_{k+1},\ u_{k+1}) \\ \vdots \\ \Psi^{\mathrm{T}}(z_{k+l},\ u_{k+l}) \end{bmatrix}$$

$$\mathcal{R}_v = \begin{bmatrix} C(y_{k+d}) + u_k^2 \\ C(y_{k+d+1}) + u_{k+1}^2 \\ \vdots \\ C(y_{k+d+l}) + u_{k+l}^2 \end{bmatrix}$$

其中, $l \in \mathbf{N}$ 并且 $l > N_1$. 为确保式 (21) 有唯一解, 作出如下假设.

假设 3. 存在一个 $l^* > 0$ 和 $\delta > 0$ 使得对于任意的 $l > l^*$ 和任意的 z_k, u_k 恒有下式成立

$$\frac{1}{l+1}\sum_{k=0}^{l}\Psi^{\mathrm{T}}(z_k,\ u_k)\Psi(z_k,\ u_k) \geq \delta I_{N_1} \tag{22}$$

算法 3. 数据驱动的输出反馈值迭代算法

1. 选择一个充分小的终止界限 $\epsilon > 0$, 使用一个控制输入采集在线的输入输出数据, $\bar{\mathcal{Q}}_0 = C(y_{k+d}) + u_k^2$, 并且令 $i \leftarrow 0$;
2. 从式 (21) 求解 $\hat{s}_{i+1}$;
3. 通过式 (20) 更新控制策略;
4. $i \leftarrow i + 1$;
5. 重复执行步骤 2 ~ 4, 直至 norm $(\hat{s}_i - \hat{s}_{i-1}) < \epsilon$.

引理 3. 对于在任意一次迭代过程中得到的 Q 函数 $\hat{\mathcal{Q}}_i$ 和控制策略 $\hat{u}_i$, 总存在一个紧集 $\boldsymbol{D}_z \times \boldsymbol{D}_u \subset \mathcal{Z} \times \mathbf{R}$ 使得 $\lim_{N_1\to\infty}\hat{\mathcal{Q}}_{i+1}(z,\ u) = \tilde{\mathcal{Q}}_{i+1}(z,\ u)$, $\lim_{N_1\to\infty}\hat{u}_{i+1}(z) = \tilde{u}_{i+1}(z)$, $\forall (z,\ u) \in \boldsymbol{D}_z \times \boldsymbol{D}_u$. 其中

$$\begin{cases}\tilde{\mathcal{Q}}_{i+1}(z_k,\ u_k) = \hat{\mathcal{Q}}_i(z_{k+1},\ \hat{u}_i(z_{k+1})) + C(y_{k+d}) + u_k^2\\ \tilde{u}_{i+1}(z_k) = \arg\min\limits_{u_k}\tilde{\mathcal{Q}}_{i+1}(z_k,\ u_k)\end{cases} \tag{23}$$

证明. 令 $\tilde{\mathcal{Q}}_{i+1}$ 取如下形式

$$\tilde{\mathcal{Q}}_{i+1}(z,\ u) = \sum_{j=1}^{\infty}\psi_i(z,\ u)\tilde{s}_{i+1,\,j} \tag{24}$$

即, $\tilde{\mathcal{Q}}_{i+1}$ 由无数个基函数加权求和组成, 于是有 $e_{i,\,k} = \Psi(z_k,\ u_k)\Lambda_i + \zeta_{i,\,k}$, 其中

$$\begin{cases}\Lambda_i = [\tilde{s}_{i+1,\,1},\ \tilde{s}_{i+1,\,2},\ \cdots,\ \tilde{s}_{i+1,\,N_1}]^{\mathrm{T}} - \\ \qquad [\hat{s}_{i+1,\,1},\ \hat{s}_{i+1,\,2},\ \cdots,\ \hat{s}_{i+1,\,N_1}]^{\mathrm{T}}\\ \zeta_{i,\,k} = \sum\limits_{j=N_1+1}^{\infty}\tilde{s}_{i+1,\,j}\psi_j(z_k,\ u_k)\end{cases} \tag{25}$$

因为权重是通过最小二乘得到的, 所以有下式成立

$$\sum_{k=0}^{l}e_{i,\,k}^2 \leq \sum_{k=0}^{l}\zeta_{i,\,k}^2 \tag{26}$$

又因为

$$\sum_{k=0}^{l}\Lambda_i^{\mathrm{T}}\Psi^{\mathrm{T}}(z_k,\ u_k)\Psi(z_k,\ u_k)\Lambda_i = \sum_{k=0}^{l}(e_{i,\,k} - \zeta_{i,\,k})^2$$

所以有下式成立

$$|\Lambda_i|^2 \leq \frac{4}{\delta}\max_{1\leq k\leq l}\zeta_{i,\,k}^2 \tag{27}$$

显然, 当 N_1 趋于无穷大时, 可得 $\lim_{N_1\to\infty}\zeta_{i,\,k} = 0$. 因此, 可进一步看出

$$\lim_{N_1\to\infty}\tilde{\mathcal{Q}}_{i+1}(z,\ u) - \hat{\mathcal{Q}}_{i+1}(z,\ u) = \lim_{N_1\to\infty}\sum_{j=N_1+1}^{\infty}\tilde{s}_{i+1,\,j}\psi_j(z,\ u) = 0 \tag{28}$$

即, $\lim_{N_1\to\infty}\hat{\mathcal{Q}}_{i+1}(z,\ u) = \tilde{\mathcal{Q}}_{i+1}(z,\ u)$. 同样可得 $\lim_{N_1\to\infty}\hat{u}_{i+1}(z) = \tilde{u}_{i+1}(z)$. □

下面给出数据驱动输出反馈值迭代算法的收敛性证明的主要结果.

定理 2. 在假设 3 的条件下, 对于任意的 $\epsilon > 0$ 总存在正整数 i^* 与 N_1^* 和一个紧集 $\boldsymbol{D}_z \times \boldsymbol{D}_u \subset \mathcal{Z} \times \mathbf{R}$, 使得下式恒成立

$$\begin{cases}\left|\sum\limits_{j=1}^{N_1}\hat{s}_{i^*,\,j}\psi_j(z,\ u) - \bar{\mathcal{Q}}^*(z,\ u)\right| \leq \epsilon\\ |\hat{u}_{i^*}(z) - \bar{u}^*(z)| \leq \epsilon\end{cases} \tag{29}$$

式中, $(z,\ u) \in \boldsymbol{D}_z \times \boldsymbol{D}_u$, $N_1 > N_1^*$.

证明. 对于每一个 $i \geq 0$ 和给定的 $\epsilon > 0$, 总存在一个足够大的正整数 N_1^*, 当 $N_1 > N_1^*$ 时有

$$\begin{cases}\left|\hat{\mathcal{Q}}_{i+1}(z,\ u) - \bar{\mathcal{Q}}_{i+1}(z,\ u)\right| \leq \epsilon\\ |\hat{u}_{i+1}(z) - \bar{u}_{i+1}(z)| \leq \epsilon\end{cases} \tag{30}$$

当 $i = 0$ 时, 总有 $\tilde{\mathcal{Q}}_0(z,\ u) = \bar{\mathcal{Q}}_0(z,\ u)$ 与 $\tilde{u}_0(z) = \bar{u}_0(z)$.

假设对于 $i = j \geq 1$ 有

$$\begin{cases}\lim\limits_{N_1\to\infty}\hat{\mathcal{Q}}_j(z,\ u) = \bar{\mathcal{Q}}_j(z,\ u)\\ \lim\limits_{N_1\to\infty}\hat{u}_j(z) = \bar{u}_j(z)\end{cases} \tag{31}$$

定义

$$R_{j+1,\,k} = \hat{\mathcal{Q}}_j(z_{k+1},\ \hat{u}_{j+1}(z_{k+1})) - \tilde{\mathcal{Q}}_j(z_{k+1},\ \tilde{u}_{j+1}(z_{k+1})) \tag{32}$$

通过基于模型的输出反馈值迭代的算法 3 的步骤 2 与式 (23), 可得

$$R_{j+1,\,k} = \tilde{\mathcal{Q}}_{j+1}(z_k,\ u_k) - \bar{\mathcal{Q}}_{j+1}(z_k,\ u_k) \tag{33}$$

基于引理 3, $\lim_{N_1\to\infty}R_{j+1,\,k} = 0$, 于是可得

$$\begin{cases}0 = \tilde{\mathcal{Q}}_{j+1}(z,\ u) - \bar{\mathcal{Q}}_{j+1}(z,\ u)\\ 0 = \tilde{u}_{j+1}(z) - \bar{u}_{j+1}(z)\end{cases}$$

在引理 3 的基础上, 可以进一步得到:

$$\begin{cases}\lim\limits_{N_1\to\infty}\hat{\mathcal{Q}}_{j+1}(z,\ u)=\bar{\mathcal{Q}}_{j+1}(z,\ u)\\ \lim\limits_{N_1\to\infty}\hat{u}_{j+1}(z)=\bar{u}_{j+1}(z)\end{cases}\tag{34}$$

即, 当 N_1 趋于无穷大时, 计算得到的结果与基于模型的算法得到的结果完全一样的. 所以对于任意的 $\epsilon>0$, 存在 $N_1^*>0$, 当 $N_1>N_1^*$ 时, 有

$$\left|\sum_{j=1}^{N_1}\hat{s}_{i,\,j}\psi_j(z,\ u)-\bar{\mathcal{Q}}_i(z,\ u)\right|\le\epsilon\tag{35}$$

基于引理 2 可知, 数据驱动的算法同样可以收敛到最优解, 即

$$\begin{cases}\left|\sum\limits_{j=1}^{N_1}\hat{s}_{i^*,\,j}\psi_j(z,\ u)-\bar{\mathcal{Q}}^*(z,\ u)\right|\le\epsilon\\ |\hat{u}_{i^*}(z)-\bar{u}^*(z)|\le\epsilon\end{cases}\tag{36}$$

□

3.2 数据驱动的输出反馈策略迭代

将 Q 函数的近似值代入策略迭代中, 可得

$$C(y_{k+d})+\hat{u}_i^2(z_k)\ =\ \left[\Psi^{\mathrm{T}}(z_k,\ u_k)-\Psi^{\mathrm{T}}(z_{k+1},\ \hat{u}_i(z_{k+1}))\right]\hat{s}_i\tag{37}$$

$$\hat{u}_{i+1}(z_k)=\arg\min_{u_k}\hat{\mathcal{Q}}_i(z_k,\ u_k)\tag{38}$$

每一代的权重可以通过最小二乘法计算, 具体形式为

$$\hat{s}_i=[\Omega_p^{\mathrm{T}}\Omega_p]^{-1}\Omega_p^{\mathrm{T}}\mathcal{R}_p\tag{39}$$

其中

$$\Omega_p=\begin{bmatrix}\Psi^{\mathrm{T}}(z_k,\ u_k)-\Psi^{\mathrm{T}}(z_{k+1},\ \hat{u}_i(z_{k+1}))\\ \Psi^{\mathrm{T}}(z_{k+1},\ u_{k+1})-\Psi^{\mathrm{T}}(z_{k+2},\ \hat{u}_i(z_{k+2}))\\ \vdots\\ \Psi^{\mathrm{T}}(z_{k+l},\ u_{k+l})-\Psi^{\mathrm{T}}(z_{k+l+1},\ \hat{u}_i(z_{k+l+1}))\end{bmatrix}$$

$$\mathcal{R}_p=\begin{bmatrix}C(y_{k+d})+\hat{u}_i^2(z_k)\\ C(y_{k+d+1})+\hat{u}_i^2(z_{k+1})\\ \vdots\\ C(y_{k+d+l})+\hat{u}_i^2(z_{k+l})\end{bmatrix}$$

其中, $l\in\mathbf{N}$ 并且 $l>N_1$. 假设 4 确保式 (39) 有唯一解.

假设 4. 存在 $l^*>0$, 当 $l>l^*$ 时恒有下式成立

$$\mathrm{rank}(\Omega_p)=N_1\tag{40}$$

算法 4. 数据驱动策略迭代算法

1. 使用一个初始的控制输入来采集在线数据, 选择一个充分小的终止界限 $\epsilon>0$, 选取一个初始可接受的控制策略 $\bar{u}_0$, 并且令 $i\leftarrow 0$;
2. 从式 (39) 求解 $\hat{s}_i$;
3. 通过式 (38) 更新控制策略;
4. $i\leftarrow i+1$;
5. 重复执行步骤 2 ~ 4, 直至 $\mathrm{norm}\,(\hat{s}_i-\hat{s}_{i-1})<\epsilon$.

注 3. 在数据驱动的输出反馈值迭代算法的收敛证明中, 已证明当基函数的数量足够多时, 数据驱动方法所得到的解与基于模型的方法一致. 即, 数据驱动的输出反馈策略迭代同样能够收敛到最优解.

注 4. 本文提出的输出反馈策略迭代和值迭代的算法复杂度为 $\mathrm{O}(N_1 l)$.

注 5. 本文提出的输出反馈策略迭代本质上是一种"认知−决策−行动"一体化的方法. 具体而言, 先通过动态的输入输出数据认知系统的实时状态和存在的时滞, 然后决策当前控制策略的价值, 并基于价值进行策略更新改进, 最后在系统中执行更新的控制策略.

4 仿真结果

在本节中, 以经典的非线性系统, 即范德波尔振荡器为例来验证所提出控制方法的有效性. 系统由如下的方程描述:

$$\begin{cases}\dot{x}_1=x_2\\ \dot{x}_2=-x_1+\tau x_2-x_2^3+u\\ y=x_1\end{cases}$$

其中, 参数 $\tau\in[-0.5,\ -0.3]$. 使用欧拉方法将该系统离散化为离散时间系统, 同时引入一步输入时滞

$$\begin{cases}x_{1,\,k+1}=x_{1,\,k}+Tx_{2,\,k}\\ x_{2,\,k+1}=x_{2,\,k}+T(-x_{1,\,k}+\tau x_{2,\,k}-x_{2,\,k}^3+u_{k-1})\\ y_k=x_{1,\,k}\end{cases}$$

其中, $T=0.1$ 为采样周期.

通过使用在线的输入输出数据, 采用数据驱动的输出反馈值迭代方法来计算最优的控制策略. 设置 $\tau=-0.4$ 并且初始状态为 $x_0=[0,\ 0.1]^{\mathrm{T}}$. 初始的控制策略选取为一系列不同频率的正弦信号. 输入输出数据从第 1 步采样到第 150 步. 代价函数选取为 $J=\sum_{k=1}^{\infty}(2y_k^2+u_{k-1}^2)$. z_k 定义为 $[y_{k-1},\ y_{k-2},\ u_{k-1},\ u_{k-2},\ u_{k-3}]^{\mathrm{T}}$. 为方便起见, 定义 $y_{k-1}=y_1$, $y_{k-2}=y_2$, $u_{k-i}=u_i$, $\forall i\in\{1,\ 2,\ 3\}$ 及 $u_k=u_0$. 总共选取 43 个 Q 函数的光滑基函数, 其具体形式如式 (41) 所示.

$$\begin{aligned}\Psi = [&y_1^2,\ y_1y_2,\ y_2^2,\ u_3^2,\ y_1^3,\ y_1^2y_2,\ y_1y_2^2,\ y_2^3,\\ &y_1^4,\ y_1^3y_2,\ y_1^2y_2^2,\ y_1y_2^3,\ y_2^4,\ y_1^5,\ y_1^4y_2,\\ &y_1^3y_2^2,\ y_1^2y_2^3,\ y_1y_2^4,\ y_2^5,\ y_1^6,\ y_1^5y_2,\ y_1^4y_2^2,\\ &y_1^3y_2^3,\ y_1^2y_2^4,\ y_1y_2^5,\ y_2^6,\ u_0^2,\ u_0y_1,\ u_0y_2,\\ &u_0u_1,\ u_0u_2,\ u_0u_3,\ u_0y_1^2,\ u_0y_2^2,\ u_0u_1^2,\\ &u_0u_2^2,\ u_0u_3^2,\ u_0y_1^3,\ u_0y_2^3,\ u_0y_1^4,\ u_0y_2^4,\\ &u_0y_1^5,\ u_0y_2^5]^{\mathrm{T}} \end{aligned} \tag{41}$$

选取 ϵ 为 0.1. 在迭代 348 次后算法完成收敛. 在第 151 步应用所学到的最优控制策略. 最优的输出反馈控制策略如下:

$$\begin{aligned}u_k = &-9.375\ 2y_{k-1} + 9.412\ 2y_{k-2} - 0.016\ 4u_{k-1} +\\ &0.124\ 8u_{k-2} - 0.111\ 3u_{k-3} + 2.731\ 7y_{k-1}^2 -\\ &0.220\ 3y_{k-2}^2 - 0.051\ 7u_{k-1}^2 - 0.084\ 3u_{k-2}^2 -\\ &0.061\ 1u_{k-3}^2 + 17.419\ 0y_{k-1}^3 - 19.837\ 6y_{k-2}^3 -\\ &2.290\ 4y_{k-1}^4 - 7.467\ 6y_{k-2}^4 + 8.188\ 5y_{k-1}^5 +\\ &4.047\ 0y_{k-2}^2\end{aligned}$$

另外, 将本文提出的输出反馈值迭代算法与文献 [2] 所提出的算法进行仿真对比. 为公平起见, 设置完全相同的学习参数, 仅删除有关 u_3 的基函数项. 图 2 展示了 Q 函数中权重范数的收敛情况. 图 3 展示了使用本文方法所学到的控制策略后闭环系统的输入输出轨迹. 图 4 展示了使用文献 [2] 中方法所学到的控制策略下闭环系统的输入输出轨迹. 由于文献 [2] 并未考虑时滞的存在, 导致闭环系统状态持续震荡. 图 5 ~ 8 描述输出反馈值迭代算法迭代前后 Q 函数的变化.

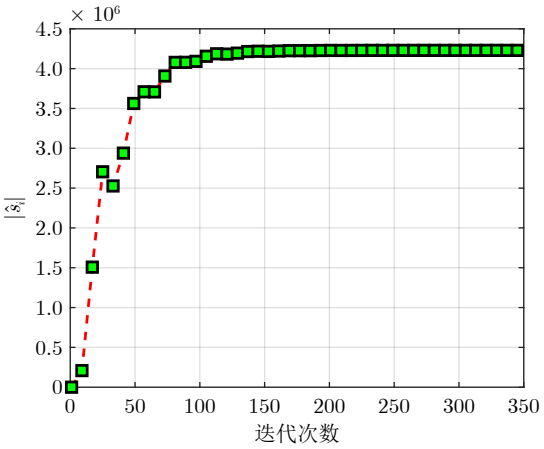


图 2　权重范数随迭代的变化情况

Fig. 2　Variation of weighting paradigm with iteration

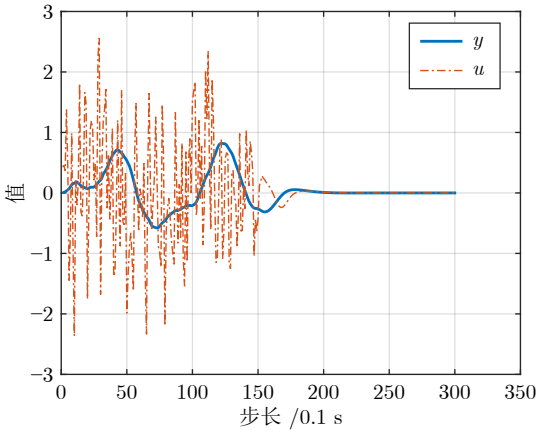


图 3　考虑时滞的输出反馈下闭环系统的输入输出轨迹

Fig. 3　The input-output trajectories of the closed-loop system under output feedback with time-delay

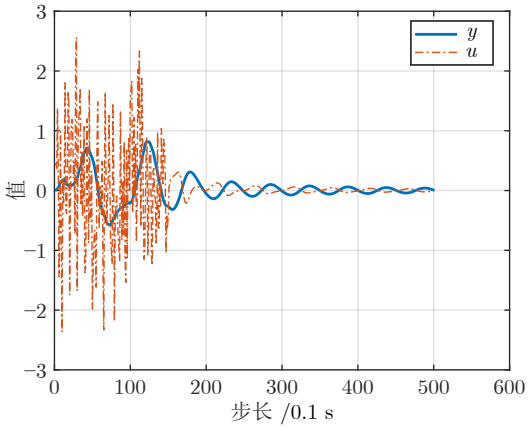


图 4　不考虑时滞的输出反馈下闭环系统的输入输出轨迹

Fig. 4　The input-output trajectories of the closed-loop system under output feedback without time-delay

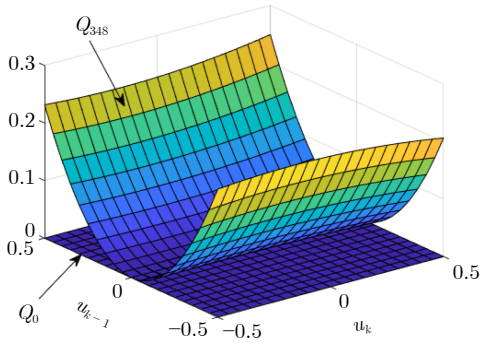


图 5　Q 函数迭代前后对比 (除 u_k, u_{k-1} 外其他参数为 0)

Fig. 5　Comparison of the Q function before and after iteration (with other parameters set to 0 except for u_k and u_{k-1})

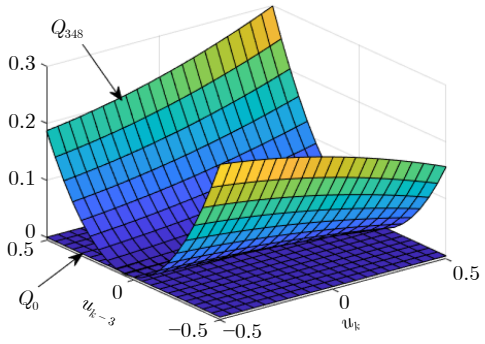


图 6 Q 函数迭代前后对比 (除 u_k, u_{k-3} 外其他参数为 0)

Fig. 6 Comparison of the Q function before and after iteration (with other parameters set to 0 except for u_k and u_{k-3})

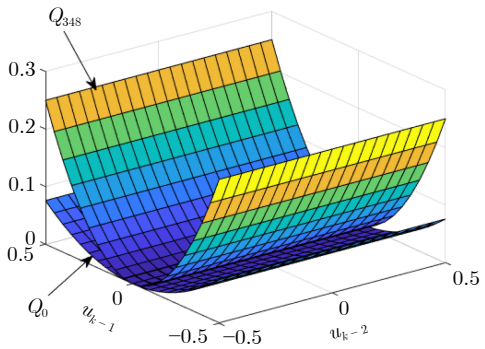


图 7 Q 函数迭代前后对比 (除 u_{k-1}, u_{k-2} 外其他参数为 0)

Fig. 7 Comparison of the Q function before and after iteration (with other parameters set to 0 except for u_{k-1} and u_{k-2})

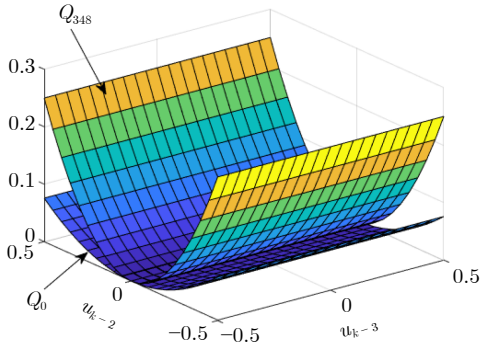


图 8 Q 函数迭代前后对比 (除 u_{k-2}, u_{k-3} 外其他参数为 0)

Fig. 8 Comparison of the Q function before and after iteration (with other parameters set to 0 except for u_{k-2} and u_{k-3})

5 结束语

针对具有输入时滞的非线性系统直接自适应最优控制问题, 本文提出一种新的数据驱动输出反馈控制方法. 通过利用输出和带有时滞的输入数据构建系统状态, 并将其作为反馈, 成功克服了含有输入时滞的输出反馈自适应最优控制策略设计这一挑战. 同时, 将 Q 学习与两种自适应动态规划方法相结合, 使得算法学习过程中无需依赖系统动力学信息. 仿真结果验证了该方法的有效性.

References

1 Lewis F L, Vrabie D L, Syrmos V L. *Optimal Control*. Hoboken: Wiley, 2012.

2 Gao W N, Jiang Z P, Chai T Y. Bridging the gap between reinforcement learning and nonlinear output-feedback control. In: Proceedings of the 43rd Chinese Control Conference (CCC). Kunming, China: IEEE, 2024. 2425−2431

3 Gao W N, Jiang Z P. Data-driven cooperative output regulation of multi-agent systems under distributed denial of service attacks. *Science China Information Sciences*, 2023, **66**(9): Article No. 190201

4 Hou Z S, Wang Z. From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 2013, **235**: 3−35

5 Bu X H, Yu Q X, Hou Z S, Qian W. Model free adaptive iterative learning consensus tracking control for a class of nonlinear multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(4): 677−686

6 Hou Z S, Xiong S S. On model-free adaptive control and its stability analysis. *IEEE Transactions on Automatic Control*, 2019, **64**(11): 4555−4569

7 Liu S D, Lin G, Ji H H, Jin S T, Hou Z S. A novel enhanced data-driven model-free adaptive control scheme for path tracking of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2025, **26**(1): 579−590

8 Xiong S S, Hou Z S. Model-free adaptive control for unknown MIMO nonaffine nonlinear discrete-time systems with experimental validation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(4): 1727−1739

9 Dong Yu-Chen, Gao Wei-Nan, Jiang Zhong-Ping. Cooperative optimal output regulation for multi-agent systems based on distributed adaptive internal model. *Acta Automatica Sinica*, 2025, **51**(3): 678−691
(董昱辰, 高伟男, 姜钟平. 基于分布式自适应内模的多智能体系统协同最优输出调节. 自动化学报, 2025, **51**(3): 678−691)

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

11 Vrabie D, Vamvoudakis K G, Lewis F L. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London: Institution of Engineering and Technology, 2013.

12 Jiang Y, Jiang Z P. Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 2015, **60**(11): 2917−2929

13 Wei Q L, Liu D R, Liu Y, Song R Z. Optimal constrained self-learning battery sequential management in microgrid via adaptive dynamic programming. *IEEE/CAA Journal of Automatica Sinica*, 2017, **4**(2): 168−176

14 Zhang H G, Liu D R, Luo Y H, Wang D. *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London: Springer, 2013.

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

16 Liu D R, Wei Q L, Wang D, Yang X, Li H L. *Adaptive Dynamic Programming With Applications in Optimal Control*. Cham:

Springer, 2017.

17 Yang Y L, Modares H, Vamvoudakis K G, He W, Xu C Z, Wunsch D C. Hamiltonian-driven adaptive dynamic programming with approximation errors. *IEEE Transactions on Cybernetics*, 2022, **52**(12): 13762−13773

18 Gao W N, Jiang Z P. Learning-based adaptive optimal output regulation of linear and nonlinear systems: An overview. *Control Theory and Technology*, 2022, **20**(1): 1−19

19 Liu D R, Wei Q L. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(3): 621−634

20 Li C, Liu D R, Wang D. Data-based optimal control for weakly coupled nonlinear systems using policy iteration. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018, **48**(4): 511−521

21 Zhao B, Wang D, Shi G, Liu D R, Li Y C. Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018, **48**(10): 1725−1735

22 Gao W N, Jiang Z P, Chai T Y. Resilient control under denial-of-service and uncertainty: An adaptive dynamic programming approach. *IEEE Transactions on Automatic Control*, 2025, **70**(6): 4085−4092

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

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

25 Wei Q L, Liu D R, Lin H Q. Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems. *IEEE Transactions on Cybernetics*, 2016, **46**(3): 840−853

26 Gao W N, Mynuddin M, Wunsch D C, Jiang Z P. Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(10): 5229−5240

27 Gao W N, Huang M Z, Jiang Z P, Chai T Y. Sampled-data-based adaptive optimal output-feedback control of a 2-degree-of-freedom helicopter. *IET Control Theory and Applications*, 2016, **10**(12): 1440−1447

28 Liu D R, Wei Q L. Adaptive dynamic programming for a class of discrete-time non-affine nonlinear systems with time-delays. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Barcelona, Spain: IEEE, 2010. 1−6

29 Xiao F, Shi Y, Ren W. Robustness analysis of asynchronous sampled-data multiagent networks with time-varying delays. *IEEE Transactions on Automatic Control*, 2018, **63**(7): 2145−2152

30 Liu Y, Zhang H G, Yu R, Xing Z X. H_∞ tracking control of discrete-time system with delays via data-based adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(11): 4078−4085

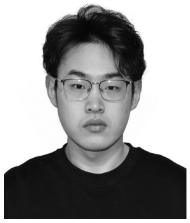
31 Gao W N, Jiang Z P. Adaptive optimal output regulation of time-delay systems via measurement feedback. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(3): 938−945

32 Gao W N, Jiang Y, Jiang Z P, Chai T Y. Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming. *Automatica*, 2016, **72**: 37−45

33 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

34 Moraal P E, Grizzle J W. Observer design for nonlinear systems with discrete-time measurements. *IEEE Transactions on Automatic Control*, 1995, **40**(3): 395−404

35 Heydari A. Analyzing policy iteration in optimal control. In: Proceedings of the American Control Conference (ACC). Boston, USA: IEEE, 2016. 5728−5733



刘思彤 东北大学流程工业综合自动化全国重点实验室硕士研究生. 2023 年获得东北大学机械工程专业学士学位. 主要研究方向为自适应动态规划, 输出反馈, 最优控制和强化学习. E-mail: 2370761@stu.neu.edu.cn

(**LIU Si-Tong** Master student at the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. He received his bachelor degree in mechanical engineering from Northeastern University in 2023. His research interest covers adaptive dynamic programming, output feedback, optimal control, and reinforcement learning.)

高伟男 东北大学流程工业综合自动化全国重点实验室教授. 2017 年获得美国纽约大学博士学位. 主要研究方向为人工智能, 自适应动态规划, 优化控制和输出调节. 本文通信作者. E-mail: gaown@mail.neu.edu.cn

(**GAO Wei-Nan** Professor at the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. He received his Ph.D. degree from New York University, USA in 2017. His research interest covers artificial intelligence, adaptive dynamic programming, optimal control, and output regulation. Corresponding author of this paper.)

姜钟平 欧洲科学院外籍院士, 美国纽约大学杰出教授, IEEE Fellow, IFAC Fellow. 1993 年获得法国巴黎高等矿业大学自动控制与数学专业博士学位. 主要研究方向为稳定性理论, 鲁棒/自适应/分布式非线性控制, 鲁棒自适应动态规划, 强化学习及其在信息、机械和生物系统中的应用. E-mail: zjiang@nyu.edu

(**JIANG Zhong-Ping** Foreign Member of the Academia Europaea (Academy of Europe), Institute Professor at Tandon School of Engineering, New York University, USA, IEEE Fellow, IFAC Fellow. He received his Ph.D. degree in automatic control and mathematics from the Ecole des Mines de Paris, France in 1993. His research interest covers stability theory, robust/adaptive/distributed nonlinear control, robust adaptive dynamic programming, reinforcement learning and their applications in information, mechanical, and biological systems.)