基于折扣广义值迭代的智能最优跟踪及应用验证

王 鼎 [1, 2, 3, 4] 赵明明 [1, 2, 3, 4] 哈明鸣 [5] 乔俊飞 [1, 2, 3, 4]

摘 要 设计了一种基于折扣广义值迭代的智能算法, 用于解决一类复杂非线性系统的最优跟踪控制问题. 通过选取合适的初始值, 值迭代过程中的代价函数将以单调递减的形式收敛到最优代价函数. 基于单调递减的值迭代算法, 在不同折扣因子的作用下, 讨论了迭代跟踪控制律的可容许性和误差系统的渐近稳定性. 为了促进算法的实现, 建立一个数据驱动的模型网络用于学习系统动态信息, 同时构造评判网络和执行网络用于近似迭代代价函数和计算迭代跟踪控制律. 值得注意的是, 我们提出了新颖的停止准则来保证迭代跟踪控制律的有效性. 这种停止准则包含两个条件, 一个条件用来保证迭代跟踪控制律的可用性, 这有利于评估误差系统的渐近稳定性; 而另一个条件用来确保跟踪控制律的近似最优性. 最后, 通过包括污水处理在内的两个应用实例验证了本文提出的近似最优跟踪控制方法的可行性和有效性.

关键词 自适应评判控制, 可容许性, 广义值迭代, 智能最优跟踪, 神经网络

引用格式 王鼎, 赵明明, 哈明鸣, 乔俊飞. 基于折扣广义值迭代的智能最优跟踪及应用验证. 自动化学报, 2022, **48**(1): 182−193

DOI 10.16383/j.aas.c210658

Intelligent Optimal Tracking With Application Verifications via Discounted Generalized Value Iteration

WANG Ding[1, 2, 3, 4] ZHAO Ming-Ming[1, 2, 3, 4] HA Ming-Ming[5] QIAO Jun-Fei[1, 2, 3, 4]

Abstract In this paper, based on the discounted generalized value iteration, an intelligent algorithm is designed to address optimal tracking control problems for a class of complex nonlinear systems. By choosing an appropriate initial value, the iterative cost function converges to the optimum in a monotonically decreasing form. In the light of the monotonically decreasing value iteration algorithm, we discuss the admissibility properties of the iterative tracking control law and the asymptotic stability of the error system with different discounted factors. For facilitating the implementation of the algorithm, a data-driven model network is established to learn the unknown system. The critic and action networks are constructed to approximate the cost function and compute the iterative tracking control law. It is worth noting that a new termination criterion is developed to guarantee the effectiveness of the iterative tracking control law. The termination criterion contains two conditions. The first condition is used to ensure the validity of the tracking control law, which is helpful to evaluate the stability of the error system. The second condition is adopted to guarantee the near-optimal properties of the tracking control law. Finally, two experimental examples are conducted, where a wastewater treatment application is involved, in order to demonstrate the control performance of the proposed near-optimal tracking control method.

Key words Adaptive critic control, admissibility properties, generalized value iteration, intelligent optimal tracking, neural networks

Citation Wang Ding, Zhao Ming-Ming, Ha Ming-Ming, Qiao Jun-Fei. Intelligent optimal tracking with application verifications via discounted generalized value iteration. *Acta Automatica Sinica*, 2022, **48**(1): 182−193

复杂非线性系统的控制与优化广泛存在于工业和生活领域[1−2]. 针对一般的非线性系统, 通常采用 Hamilton-Jacobi-Bellman (HJB) 方程的框架来解决其最优控制问题[3]. 由于这类偏微分方程的解析

收稿日期 2021-07-15 录用日期 2021-11-02
Manuscript received July 15, 2021; accepted November 2, 2021
北京市自然科学基金 (JQ19013), 国家自然科学基金 (61773373, 61890930-5, 62021003), 科技创新 2030—— "新一代人工智能"重大项目 (2021ZD0112302, 2021ZD0112301), 国家重点研发计划 (2018YFC1900800-5) 资助
Supported by Beijing Natural Science Foundation (JQ19013), National Natural Science Foundation of China (61773373, 61890930-5, 62021003), and National Key Research and Development Program of China (2021ZD0112302, 2021ZD0112301, 2018YFC1900800-5)
本文责任编委 刘艳军
Recommended by Associate Editor LIU Yan-Jun
1. 北京工业大学信息学部 北京 100124 2. 计算智能与智能系统北京市重点实验室 北京 100124 3. 北京人工智能研究院 北京 100124 4. 智慧环保北京实验室 北京 100124 5. 北京科技大学自动化学院 北京 100083
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124 3. Beijing Institute of Artificial Intelligence, Beijing 100124 4. Beijing Laboratory of Smart Environmental Protection, Beijing 100124 5. School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083

解难以获取, 于是人们提出许多方法求得 HJB方程的近似解. 其中, 自适应动态规划 (Adaptive dynamic programming, ADP) 整合了动态规划理论、函数近似工具和强化学习机制, 能够获得令人满意的近似最优控制策略[4−5]. 至今, ADP 在解决复杂非线性系统的最优控制问题上已有大量的成果, 例如跟踪控制[6−8], 鲁棒控制[9−11] 和事件触发控制[12−14]等. 根据基本的迭代形式, ADP 算法通常分为值迭代[15] 和策略迭代[16]. 针对一般离散非线性系统, 文献 [15] 详尽地阐明了具有零初始代价函数的值迭代算法收敛性, 而文献 [16] 讨论了策略迭代算法的收敛性. 值得一提的是, 策略迭代算法需要一个初始可容许控制律并且迭代过程中的控制律都能使得系统稳定, 而值迭代过程中的迭代控制律可能是无效的, 即不能保证系统的稳定性. 然而, 复杂非线性系统的初始可容许控制律通常难以获取且策略迭代过程中的计算量较大. 因此, 我们更关注如何改进值迭代过程中迭代控制律的实用性. 传统值迭代算法要求零初始条件并且迭代指标增大到无穷才能保证控制律是可容许的. 但是在实际应用中, 算法必须在有限迭代步骤内找到一个有效的控制律[17]. 因此, 提出合适的停止准则对于算法的实现是至关重要的. 为了保证迭代控制律的可用性以及克服传统值迭代算法的不足, 广义值迭代算法应运而生[18−20]. 广义值迭代算法允许任意一个半正定函数作为初始代价函数, 这使得迭代代价函数的单调性不唯一. 针对非线性系统的最优控制, 文献 [17] 讨论了无折扣广义值迭代框架下迭代控制律的可容许性并提出了一个新的迭代停止准则. 无折扣情况下单调递减的代价函数序列能够保证所有的控制律都是可容许的. 然而有折扣情况下单调递减的代价函数序列无法保证迭代控制律的稳定性. 基于广义值迭代算法, 文献 [20] 进一步指明了折扣因子与系统稳定性的关系. 然而, 在带有折扣因子的广义值迭代算法中, 迭代控制律的可容许性以及折扣因子和初始代价函数的关系还没有研究. 在本文中, 我们旨在进一步研究折扣广义值迭代中迭代控制律的可容许性, 并将广义值迭代算法推广到解决非线性系统的最优跟踪问题中.

非线性系统的跟踪问题一直是工程领域的热点之一. 传统控制方法存在参数固定和自适应能力差的局限, 使其难以应对复杂的外界干扰. ADP 方法具有显著的自适应能力, 已广泛应用于求解复杂未知非线性系统的跟踪问题. 为了实现有效的跟踪, 最优跟踪控制问题通常被转换为关于误差系统的最优调节问题. 文献 [6] 使用贪婪迭代启发式动态规划 (Heuristic dynamic programming, HDP) 算法解决了无限时域的最优跟踪控制问题. 文献 [7] 则提出了一种有限时域的神经最优跟踪控制策略. 基于执行-评判结构, 文献 [8] 提出了一种部分模型未知的自适应最优控制方法, 有效地解决了离散系统的跟踪问题. 文献 [21] 通过对误差系统建模从而解决了带有控制约束的非线性系统跟踪问题. 然而, 这些研究更倾向于仿射系统或者对误差系统进行建模. 仿射系统的稳定控制可以根据其表达式求解, 这有利于实现跟踪控制. 然而, 由于存在复杂的数学模型或者模型信息未知, 非仿射形式的稳定控制往往难以求解. 为了解决非仿射系统的跟踪控制问题, 文献 [22] 使用了一种新的数值方法来求解稳定控制并避免了对误差系统建模. 利用数据驱动思想, 文献 [23] 使用 HDP 技术实现了对污水处理过程中溶解氧和硝态氮浓度的跟踪控制. 文献 [24−25] 运用二次启发式动态规划算法克服了对称和不对称约束情况下的复杂系统跟踪控制问题. 总之, 基于 ADP 的非线性系统最优跟踪控制研究已经取得了很大的进展. 然而, 上述工作都是基于传统的值迭代算法, 并没有讨论迭代过程中误差系统的稳定性和跟踪控制律的可容许性.

基于此, 本文提出一种基于折扣广义值迭代算法的离散时间未知非线性系统近似最优跟踪控制方法. 值得注意的是, 该算法的初始代价函数不为零并且需要满足一定条件使得代价函数序列单调递减. 在不同折扣因子的作用下, 我们讨论了迭代跟踪控制律的可容许性和误差系统的稳定性. 通过收集系统的输入输出样本数据来构造模型网络以评估下一时刻状态和求解稳定控制. 评判网络和执行网络分别用于近似代价函数和跟踪控制律. 此外, 我们建立了一个新的停止准则作为迭代过程停止的依据. 最后, 通过两个仿真实例验证了本文提出算法的控制性能.

在本文中, $\mathbf{R}$ 表示所有实数集. $\mathbf{R}^n$ 表示由全部 n 维实向量组成的欧氏空间. 令 Ω 为$\mathbf{R}^n$ 上的一个紧集. $\mathbf{R}^{n\times m}$ 表示 $n\times m$ 实矩阵组成的空间. I_n 为 $n\times n$ 维单位矩阵. $\mathbf{N}=\{0,1,2,\cdots\}$ 为所有非负整数的集合. $\mathbf{N}^+=\{1,2,\cdots\}$ 为所有正整数的集合.

1 问题描述

考虑一类具有非仿射形式的动态系统

$$x(k+1)=\mathcal{F}(x(k),u(k)),\quad k\in\mathbf{N} \tag{1}$$

其中, $x(k)\in\mathbf{R}^n$ 是状态向量, $u(k)\in\mathbf{R}^m$ 是控制向量. 系统函数 $\mathcal{F}(\cdot)$ 相对于其参数在紧集 Ω 上是可微的. 假设系统 (1) 是可控的, 且其状态和控制量可观测. 考虑跟踪问题, 我们的目标是设计一个反馈控

制策略 $u(x(k))$ 使得原始系统 (1) 跟踪上参考轨迹. 这里, 定义有界参考轨迹为

$$r(k+1)=\mathcal{R}(r(k)),\quad k\in\mathbf{N} \tag{2}$$

其中, $r(k)\in\mathbf{R}^n$是 k时刻的参考轨迹, $\mathcal{R}(\cdot):\mathbf{R}^n\to\mathbf{R}^n$ 是一个可微的函数. 不失一般性, 我们假设存在一个相对于参考轨迹的稳定控制 $u(r(k))$ 满足方程 $r(k+1)=\mathcal{F}(r(k),u(r(k)))$ 并且可以求解. 对于仿射系统, 其稳定控制可以通过状态矩阵和控制矩阵的构造形式来求解. 然而, 对于非仿射系统, 上述稳定控制的求解方法已不适用. 因此, 本文将在后续部分给出非仿射系统稳定控制的求解方法. 为了构造误差系统, 分别定义跟踪误差和跟踪控制律为

$$e(k)=x(k)-r(k) \tag{3}$$

和

$$u(e(k))=u(x(k))-u(r(k)) \tag{4}$$

基于式 (1)~(4), 可以得到如下所示的误差系统动态

$$\begin{aligned}e(k+1)=&\mathcal{F}(e(k)+r(k),\\&u(e(k))+u(r(k)))-\mathcal{R}(r(k))\end{aligned} \tag{5}$$

最优跟踪控制的思想是通过调节跟踪误差系统 (5) 使得误差衰减到零向量, 即 $e(k)\to\mathbf{0}$. 假设误差系统是可控的, 那意味着存在至少一个连续的跟踪控制律 $u(e(k))$ 使得误差系统渐近稳定. 受文献 [6−7, 22] 启发, 针对含有折扣因子的误差系统最优调节问题, 我们定义如下所示的代价函数

$$\mathcal{J}(e(k))=\sum_{l=k}^{\infty}\gamma^{l-k}U(e(l),u(e(l))) \tag{6}$$

其中, $\gamma\in(0,1]$ 是折扣因子, $U(e(l),u(e(l)))\geq 0$ 是效用函数, $U(0,0)=0$. 在本文中, 效用函数选为二次型形式, 即 $U(e(l),u(e(l)))=e^{\mathrm{T}}(l)Qe(l)+u^{\mathrm{T}}(e(l))Ru(e(l))$, 其中 Q 和 R 是正定矩阵. 简洁起见, 效用函数中的二次型重写为 $Q(e(l))+R(u(e(l)))$. 待设计的跟踪控制律不仅需要在 Ω 上使得误差系统稳定, 并且需要使得式 (6) 中的代价函数有界, 即 $u(e(k))$ 是可容许的跟踪控制律[15, 26]. 对于误差系统 (5), 假设存在至少一个可容许的跟踪控制律. 接下来, 式 (6) 中的代价函数可以进一步写为

$$\mathcal{J}(e(k))=U(e(k),u(e(k)))+\gamma\mathcal{J}(e(k+1)) \tag{7}$$

最优跟踪控制问题的核心是找到一个最优跟踪控制策略使得代价函数 (7) 最小, 这种最小的代价函数也称为最优代价函数. 根据 Bellman 最优性原理, 最优代价函数满足如下所示的 HJB 方程

$$\begin{aligned}\mathcal{J}^*(e(k))=&\min_{u(e(k))}\{U(e(k),u(e(k)))+\\&\gamma\mathcal{J}^*(e(k+1))\}\end{aligned} \tag{8}$$

因此, 相应的最优跟踪控制策略为

$$\begin{aligned}u^*(e(k))=&\arg\min_{u(e(k))}\{U(e(k),u(e(k)))+\\&\gamma\mathcal{J}^*(e(k+1))\}\end{aligned} \tag{9}$$

对于本文中的一般非线性系统, 由于最优代价函数和最优跟踪控制策略不能够精确地求解, 我们使用广义值迭代算法来获取其近似解.

2 面向智能最优跟踪的折扣广义值迭代

在本节中, 我们给出带有折扣因子的广义值迭代算法并讨论折扣广义值迭代算法的性质.

2.1 折扣广义值迭代算法推导

基于值迭代思想, 我们构建两个迭代序列, 即代价函数序列 $\{V_i(e(k))\}$和跟踪控制律序列 $\{\nu_i(e(k))\}$, 其中 $i\in\mathbf{N}$ 为迭代指标. 不同于传统的值迭代算法, 广义值迭代算法允许采用任意一个半正定函数进行初始化. 在此, 令初始代价函数为 $V_0(e(k))=e^{\mathrm{T}}(k)\Lambda e(k)$, 其中, Λ 是一个半正定的矩阵. 对于 $i=0,1,\cdots$, 算法的学习过程包括以迭代方式计算跟踪控制律

$$\begin{aligned}\nu_i(e(k))=&\arg\min_{u(e(k))}\big\{U(e(k),u(e(k)))+\\&\gamma V_i(e(k+1))\big\}\end{aligned} \tag{10}$$

和代价函数

$$\begin{aligned}V_{i+1}(e(k))=&\min_{u(e(k))}\big\{U(e(k),u(e(k)))+\\&\gamma V_i(e(k+1))\big\}\end{aligned} \tag{11}$$

为了最小化迭代过程中的代价函数, 迭代跟踪控制律的形式为

$$\nu_i(e(k))=-\frac{\gamma}{2}R^{-1}\left(\frac{\partial e(k+1)}{\partial u(e(k))}\right)^{\mathrm{T}}\frac{\partial V_i(e(k+1))}{\partial e(k+1)} \tag{12}$$

值得一提的是, 本文没有对误差动态系统 (5) 进行建模. 对误差系统进行建模会增大计算量并且引入新的逼近误差. 因此, 为了克服求解 $\dfrac{\partial e(k+1)}{\partial u(e(k))}$ 的困难, 我们基于文献 [22] 引入如下的一个转换公式

$$\frac{\partial e(k+1)}{\partial u(e(k))}=\frac{\partial x(k+1)}{\partial u(x(k))} \tag{13}$$

进而, 式 (12) 中 $e(k+1)$ 相对于 $u(e(k))$ 的偏导数转换为 $\dfrac{\partial x(k+1)}{\partial u(x(k))}$, 后者的获取通过对原系统建立的模型网络来实现, 这样既减少了计算量, 又能避免误差系统建模过程中逼近误差对控制器设计产生的不利影响.

2.2　折扣广义值迭代算法性质

接下来, 我们重点关注折扣广义值迭代算法的性质, 包括单调性、有界性、收敛性和最优性.

引理 1 (单调性). 定义跟踪控制律序列 $\{\nu_i\}$ 和代价函数序列 $\{V_i\}$ 如式 (10) 和式 (11) 所示, $V_0(e(k))=e^{\mathrm{T}}(k)\Lambda e(k)$. 对于所有的 $e(k)\in\Omega$, 如果 $V_0(e(k))\le V_1(e(k))$, 则 $V_i(e(k))\le V_{i+1}(e(k))$, $\forall i\ge 0$; 另一方面, 如果 $V_0(e(k))\ge V_1(e(k))$, 则 $V_i(e(k))\ge V_{i+1}(e(k))$, $\forall i\ge 0$.

引理 2 (有界性). 令 $\pi(e(k))$ 是一个任意的控制策略且 $\pi(0)=0$. 我们定义一个新的迭代代价函数为

$$\mathcal{Z}_{i+1}(e(k))=U(e(k),\pi(e(k)))+\gamma\mathcal{Z}_i(e(k+1)) \quad (14)$$

如果 $\pi(e(k))$ 是可容许控制律, 则 $\lim_{i\to\infty}\mathcal{Z}_i(e(k))$ 有界.

引理 1 和引理 2 的证明可通过与文献 [17] 类似的方法给出, 只需注意折扣因子的存在. 引理 1 中的单调性是至关重要的, 这也是广义值迭代算法和传统值迭代算法的最大区别. 传统值迭代算法中的 $\{V_i\}$ 是一个单调非减序列, 而广义值迭代算法中代价函数序列的单调性不唯一. 事实上, 单调递减的代价函数序列有利于判断系统的稳定性和控制律的可容许性. 无折扣广义值迭代算法的收敛性已在文献 [17−18] 中给出. 接下来, 我们将阐明具有折扣因子的广义值迭代算法的收敛性.

定理 1 (收敛性). 假设条件 $0\le\gamma\mathcal{J}^*(e(k+1))\le\delta U(e(k),u(e(k)))$ $(0<\delta<\infty)$ 一致成立且初始代价函数满足 $0\le\underline{\delta}\mathcal{J}^*(e(k))\le V_0(e(k))\le\overline{\delta}\mathcal{J}^*(e(k))$, 其中 $0\le\underline{\delta}\le 1\le\overline{\delta}<\infty$. 如果跟踪控制律序列 $\{\nu_i\}$ 和代价函数序列 $\{V_i\}$ 按照式 (10) 和式 (11) 进行迭代更新, 且 $V_0(e(k))=e^{\mathrm{T}}(k)\Lambda e(k)$, 则代价函数序列通过以下的不等式一致收敛到最优代价函数

$$\left(1+\frac{\underline{\delta}-1}{(1+\delta^{-1})^i}\right)\mathcal{J}^*(e(k))\le V_i(e(k))\le \left(1+\frac{\overline{\delta}-1}{(1+\delta^{-1})^i}\right)\mathcal{J}^*(e(k)) \quad (15)$$

证明. 首先, 用公式推导来证明不等式的左边部分. 当 $i=0$ 时, $\underline{\delta}\mathcal{J}^*(e(k))\le V_0(e(k))$ 成立. 当 $i=1$ 时, 可以得到

$$\begin{aligned}V_1(e(k))=&\min_{u(e(k))}\{U(e(k),u(e(k)))+\gamma V_0(e(k+1))\}\ge\\&\min_{u(e(k))}\{U(e(k),u(e(k)))+\gamma\underline{\delta}\mathcal{J}^*(e(k+1))\}\ge\\&\min_{u(e(k))}\left\{\left(1-\delta\frac{1-\underline{\delta}}{1+\delta}\right)U(e(k),u(e(k)))+\left(\underline{\delta}+\frac{1-\underline{\delta}}{1+\delta}\right)\gamma\mathcal{J}^*(e(k+1))\right\}=\\&\left(1+\frac{\underline{\delta}-1}{1+\delta^{-1}}\right)\mathcal{J}^*(e(k))\end{aligned} \quad (16)$$

假设不等式 (15) 的左边部分对于 $i-1$ 成立. 对于 i, 可以进一步得到

$$\begin{aligned}V_i(e(k))=&\min_{u(e(k))}\{U(e(k),u(e(k)))+\gamma V_{i-1}(e(k+1))\}\ge\\&\min_{u(e(k))}\left\{U(e(k),u(e(k)))+\left(1+\frac{\underline{\delta}-1}{(1+\delta^{-1})^{i-1}}\right)\gamma\mathcal{J}^*(e(k+1))+\right.\\&\left.\frac{\delta^{i-1}(\underline{\delta}-1)}{(1+\delta)^i}(\delta U(e(k),u(e(k)))-\gamma\mathcal{J}^*(e(k+1)))\right\}=\\&\left(1+\frac{\underline{\delta}-1}{(1+\delta^{-1})^i}\right)\min_{u(e(k))}\left\{U(e(k),u(e(k)))+\gamma\mathcal{J}^*(e(k+1))\right\}=\\&\left(1+\frac{\underline{\delta}-1}{(1+\delta^{-1})^i}\right)\mathcal{J}^*(e(k))\end{aligned} \quad (17)$$

不等式 (15) 右边的证明过程与之类似, 这里不再详细展开. 接下来, 我们将证明随着迭代指标增加到无穷时代价函数的一致收敛性. 当 $i\to\infty$ 时, 对于 $0<\delta<\infty$, 可以推导出

$$\lim_{i\to\infty}\left\{\left(1+\frac{\underline{\delta}-1}{(1+\delta^{-1})^i}\right)\mathcal{J}^*(e(k))\right\}=\mathcal{J}^*(e(k)) \quad (18)$$

和

$$\lim_{i\to\infty}\left\{\left(1+\frac{\overline{\delta}-1}{(1+\delta^{-1})^i}\right)\mathcal{J}^*(e(k))\right\}=\mathcal{J}^*(e(k)) \quad (19)$$

定义 $V_\infty(e(k))=\lim_{i\to\infty}V_i(e(k))$, 进一步可以

得到 $V_\infty(e(k))=\mathcal{J}^*(e(k))$. 因为 Ω 是紧集, 因此可以得到代价函数序列一致收敛[18]. □

实际中值迭代算法的迭代指标不可能增大到无穷, 算法必须在有限的迭代步骤内停止. 通常值迭代过程的停止准则为 $|V_{i+1}(e(k))-V_i(e(k))|<\varsigma$, 其中 ς 是一个小的正数, 此时相应的跟踪控制律 $\nu_i(e(k))$ 可作用于受控系统. 然而, 满足条件 $|V_{i+1}(e(k))-V_i(e(k))|<\varsigma$ 的 $\nu_i(e(k))$ 可能不是可容许的跟踪控制律, 而只是一致最终有界的跟踪控制律[17]. 因此, 在有限的迭代次数内提出更合理的准则来判断系统稳定性和跟踪控制律的可容许性是必要的. 接下来, 我们着重分析跟踪控制律的可容许性.

定理 2. 定义迭代跟踪控制律 $\nu_i(e(k))$ 和迭代代价函数 $V_i(e(k))$ 如式 (10) 和式 (11) 所示, $V_0(e(k))=e^{\mathrm{T}}(k)\Lambda e(k)$. 对于任意的 $e(k)\neq 0$, 如果跟踪控制律 $\nu_i(e(k))$ 使得下式成立

$$V_{i+1}(e(k))-\gamma V_i(e(k))<U(e(k),\nu_i(e(k))) \tag{20}$$

则迭代指标为 i 时的跟踪控制律是可容许的.

证明. 根据式 (20), 一定存在一个常数 $-\infty<\varrho<1$ 满足

$$V_{i+1}(e(k))-\gamma V_i(e(k))<\varrho U(e(k),\nu_i(e(k))) \tag{21}$$

将式 (11) 代入式 (21), 可得

$$V_i(e(k+1))-V_i(e(k))<\frac{1}{\gamma}(\varrho-1)U(e(k),\nu_i(e(k))) \tag{22}$$

不等式 (22) 的右半部分是一个负数, 于是可得 $V_i(e(k+1))-V_i(e(k))<0$, 这意味着 $\nu_i(e(k))$ 是一个稳定控制律. 此外, 通过扩展不等式 (22) 可以得到

$$\begin{aligned}
&V_i(e(k+1))-V_i(e(k))<\frac{1}{\gamma}(\varrho-1)U(e(k),\nu_i(e(k)))\\
&V_i(e(k+2))-V_i(e(k+1))<\frac{1}{\gamma}(\varrho-1)U(e(k+1),\nu_i(e(k+1)))\\
&\qquad\vdots\\
&V_i(e(k+N))-V_i(e(k+N-1))<\frac{1}{\gamma}(\varrho-1)\times U(e(k+N-1),\nu_i(e(k+N-1)))
\end{aligned} \tag{23}$$

因为 $\nu_i(e(k))$ 是一个稳定控制律, 当 $N\to\infty$, 可以得到 $\lim_{N\to\infty}V_i(e(k+N))=0$. 于是, 式 (23) 可将进一步归纳为

$$\frac{1}{\gamma}(1-\varrho)\sum_{j=0}^{\infty}U(e(k+j),\nu_i(e(k+j)))<V_i(e(k)) \tag{24}$$

对于 $-\infty<\varrho<1$ 和有界的 $e(k)$ 而言, $V_i(e(k))$ 是有界的. 由此可以得到 $\sum_{j=0}^{\infty}U(e(k+j),\nu_i(e(k+j)))$ 是有界的. 由于折扣因子的取值范围为 $\gamma\in(0,1]$, 进一步地, 可以得到 $\sum_{j=0}^{\infty}\gamma^jU(e(k+j),\nu_i(e(k+j)))$ 是有界的, 这满足了可容许性的条件. □

定理 2 中给出了迭代跟踪控制律的可容许性判别条件. 需要注意的是, 可容许的 $\nu_i(e(k))$ 并不能保证跟踪控制律 $\nu_{i+\eta}(e(k))$ 也是可容许的, $\eta\in\mathbf{N}^+$. 此外, $\nu_i(e(k))$ 也不一定是近似最优控制律. 我们希望如果当前迭代步的跟踪控制律 $\nu_i(e(k))$ 为可容许控制律, 则该迭代步之后的所有跟踪控制律 $\nu_{i+\eta}(e(k))$ 都是可容许的.

在无折扣广义值迭代算法框架下, 当 $V_0(e(k))>V_1(e(k))$ 时, 迭代代价函数将以单调递减的形式收敛, 即

$$V_{i+1}(e(k))=U(e(k),\nu_i(e(k)))+V_i(e(k+1))<V_i(e(k)),\ i\in\mathbf{N} \tag{25}$$

根据式 (25), 可以得到

$$V_i(e(k+1))-V_i(e(k))<0 \tag{26}$$

这表明每一个迭代步的跟踪控制律都能够镇定被控系统. 这不仅克服了传统值迭代中控制律无法确保系统稳定的困难, 也避免了在策略迭代中求取初始可容许控制律. 值得一提的是, 代价函数单调递减的条件 $V_0(e(k))>V_1(e(k))$ 是容易实现的, 例如增大初始代价函数中矩阵 Λ 的元素值. 然而, 式 (25) 中引入折扣因子后, $V_{i+1}(e(k))<V_i(e(k))$ 成立并不能保证 $V_i(e(k+1))-V_i(e(k))<0$ 成立. 接下来, 利用单调递减代价函数具有的显著优势, 我们进一步将上述结论推广到具有折扣因子的广义值迭代算法. 因此, 后续的学习和分析过程都是在 $V_0(e(k))>V_1(e(k))$ 的前提条件下进行.

定理 3. 定义迭代跟踪控制律 $\nu_i(e(k))$ 和迭代代价函数 $V_i(e(k))$ 如式 (10) 和式 (11) 所示, $V_0(e(k))=e^{\mathrm{T}}(k)\Lambda e(k)$. 对于任意的 $e(k)\neq 0$, 如果折扣因子 γ 满足

$$\gamma>1-\frac{Q(e(k))}{V_0(e(k))},\ 0<\gamma\leq 1 \tag{27}$$

则 $\nu_i(e(k))$, $i\in\mathbf{N}$, 是可容许的跟踪控制律.

证明. 当 $V_0(e(k))>V_1(e(k))$ 时, 可以得到

$$V_{i+1}(e(k))-V_i(e(k))=U(e(k),\nu_i(e(k)))+\gamma V_i(e(k+1))-V_i(e(k))<0 \tag{28}$$

根据式 (28), 可以得到

$$\gamma V_i(e(k+1))-\gamma V_i(e(k))< -U(e(k),\nu_i(e(k)))+(1-\gamma)V_i(e(k)) \quad (29)$$

为了实现 $V_i(e(k+1))-V_i(e(k))<0$, 折扣因子需要满足以下不等式

$$0<1-\frac{U(e(k),\nu_i(e(k)))}{V_i(e(k))}<\gamma\le 1,\ e(k)\neq 0 \quad (30)$$

即当式 (30) 成立时, $\nu_i(e(k))$ 是一个稳定的跟踪控制律. 接下来, 我们证明 $\nu_i(e(k))$ 是一个可容许的跟踪控制律. 当 $V_i(e(k+1))-V_i(e(k))<0$ 时, 存在一个常数 $-\infty<\varrho<1$, 使得

$$\begin{aligned}&V_i(e(k+1))-V_i(e(k))<(\varrho-1)U(e(k),\nu_i(e(k)))\\&V_i(e(k+2))-V_i(e(k+1))<(\varrho-1)\times\\&\quad U(e(k+1),\nu_i(e(k+1)))\\&\qquad\vdots\\&V_i(e(k+N))-V_i(e(k+N-1))<(\varrho-1)\times\\&\quad U(e(k+N-1),\nu_i(e(k+N-1)))\end{aligned} \quad (31)$$

进而可得

$$(1-\varrho)\sum_{j=0}^{\infty}U(e(k+j),\nu_i(e(k+j)))<V_i(e(k)) \quad (32)$$

由于 $V_i(e(k))$ 是有界的, 结合式 (32) 的左边, 进一步可以得到 $\sum_{j=0}^{\infty}\gamma^j U(e(k+j),\ \nu_i(e(k+j)))$ 有界, 这意味着 $\nu_i(e(k))$ 是一个可容许的跟踪控制律. 由于 $U(e(k),\nu_i(e(k)))$ 不具备单调特性, 因此式 (30) 的成立只能表明 $\nu_i(e(k))$ 可以使得误差系统稳定, 不能作为通用的判别准则. 考虑到 $Q(e(k))\le U(e(k),\ \nu_i(e(k)))$, 可以得到

$$1-\frac{U(e(k),\nu_i(e(k)))}{V_i(e(k))}<1-\frac{Q(e(k)))}{V_i(e(k))},\ e(k)\neq 0 \quad (33)$$

即当折扣因子大于式 (33) 右半部分时, 即可保证跟踪控制律 $\nu_i(e(k))$ 的可容许性. 式 (33) 右侧的条件比左侧更加严格, 但其优点显著, 能够保证此后所有迭代控制律的可容许性. 为了方便, 定义 $\Psi_i(e(k))=1-Q(e(k))/V_i(e(k))$. 由于 $\{V_i(e(k))\}$ 是一个单调递减的序列, 可以得到 $\{\Psi_i(e(k))\}$ 也是一个单调递减的序列. 当条件 $\gamma>\Psi_i(e(k))$ 成立时, 我们可以得到 $\gamma>\Psi_{i+\eta}(e(k)),\ \eta\in\mathbf{N}^+$, 这意味着 $\nu_i(e(k))$ 及以后所有的迭代跟踪控制律 $\nu_{i+j}(e(k)))$ 都是可容许的. 也就是说, 条件 $\gamma>\Psi_i(e(k))$ 既保证 $V_{i+\eta}(e(k+1))-V_{i+\eta}(e(k))<0$, 同时使得 $\sum_{j=0}^{\infty}\gamma^j U(e(k+j),\nu_{i+\eta}\times(e(k+j)))$ 有界. 根据代价函数的单调性, 有

$$V_{i+\eta}(e(k))<V_i(e(k))<\cdots<V_0(e(k)),\ \eta\in\mathbf{N}^+$$

由此可以推出

$$\Psi_{i+\eta}(e(k))<\Psi_i(e(k))<\cdots<\Psi_0(e(k)),\ \eta\in\mathbf{N}^+ \quad (34)$$

因此, 我们最终可以得到, 当 $\gamma>\Psi_0(e(k))=1-Q(e(k))/V_0(e(k))$ 时, 每一个迭代步的跟踪控制律都是可容许的. □

值得一提的是, 在代价函数单调递减的情况下, $\gamma=1$ 能够满足定理 3 中的所有判别条件, 具有显著的优势. 折扣因子不为 1 时, 迭代控制律的可容许性得不到保证. 在下文中, 为了验证一般折扣因子的作用, 折扣因子不再取 $\gamma=1$. 事实上, 式 (27) 提出的可容许判别准则相对比较严格, 要求接近于 1 的折扣因子. 于是, 为了更易实现算法, 我们使用 $\gamma>\Psi_i(e(k))$ 作为实际的判别准则. 总而言之, 本文提出的迭代算法的停止准则为 $|V_{i+1}(e(k))-V_i(e(k))|<\varsigma$ 和 $\gamma>\Psi_i(e(k))$, 其中第 1 项用于保证跟踪控制律的近似最优性, 而第 2 项用于保证跟踪控制律的可容许性. 值得一提的是, 本文提出的稳定性条件是一个充分条件, 折扣因子较大时容易满足该条件从而使得控制律稳定, 而折扣因子较小时不能满足该稳定条件, 其稳定性无法确定.

3 基于神经网络的算法实现

由于系统 (1) 是非仿射的, 稳定控制和 $x(k+1)$ 相对于 $u(x(k))$ 的偏导数难以求解. 在本文中, 我们建立一个模型网络来辨识系统以求解稳定控制和上述偏导数. 此外, 分别构造评判网络和执行网络来近似代价函数和跟踪控制律. 接下来, 我们给出基于折扣广义值迭代算法的神经网络实现方案.

构造一个模型网络以学习非线性系统动态, 从而避免对系统精确数学模型的要求. 通过输入状态和控制律, 模型网络的输出表达式为

$$\hat{x}(k+1)=\omega_{m2}^{\mathrm{T}}\Theta_m(\omega_{m1}^{\mathrm{T}}x_m(k)+b_{m1})+b_{m2} \quad (35)$$

其中, $x_m(k)=[x^{\mathrm{T}}(k),u^{\mathrm{T}}(x(k))]^{\mathrm{T}}$, ω_{m2} 和 ω_{m1} 是权值矩阵, b_{m2} 和 b_{m1} 是阈值向量, Θ_m 是激活函数. 不失一般性, 定义模型网络的训练性能指标为

$$E_m=\frac{1}{2}[\hat{x}(k+1)-x(k+1)]^{\mathrm{T}}[\hat{x}(k+1)-x(k+1)] \quad (36)$$

本文中, 我们使用 MATLAB 神经网络工具箱来训练模型网络. 值得一提的是, 模型网络在算法的迭代过程开始前已经完成训练. 对于仿射系统, 稳定控制的求解依赖于原始系统的状态矩阵和控制矩阵. 然而, 本文的原始系统函数是非仿射的, 这就导致稳定控制的求解变得困难. 因此, 我们使用训

练好的模型网络表达式来求解稳定控制, 即

$$r(k+1)=\omega_{m2}^{\mathrm{T}}\Theta_m(\omega_{m1}^{\mathrm{T}}r_m(k)+b_{m1})+b_{m2} \tag{37}$$

其中, $r_m(k)=[r^{\mathrm{T}}(k),u^{\mathrm{T}}(r(k))]^{\mathrm{T}}$. 由于式 (37) 中除了 $u(r(k))$ 以外都是已知变量, 我们可以通过数值方法来计算稳定控制 $u(r(k))$.

在这里, 我们利用评判网络来近似代价函数 $V_i(e(k))$. 对于输入 $e(k)$, 评判网络的近似值为

$$\hat{V}_i(e(k))=\omega_{c2}^{\mathrm{T}}\Theta_c(\omega_{c1}^{\mathrm{T}}e(k)) \tag{38}$$

其中, ω_{c2} 和 ω_{c1} 是相应的权值矩阵, Θ_c 是激活函数. 结合式 (11) 和式 (38), 定义评判网络的训练性能指标为

$$E_i^c=\frac{1}{2}[\hat{V}_i(e(k))-V_i(e(k))]^{\mathrm{T}}[\hat{V}_i(e(k))-V_i(e(k))] \tag{39}$$

其中, 训练性能指标 E_i^c 随着迭代指标 i 不断变化.

通过权值矩阵 ω_{a2} 和 ω_{a1}, 我们使用执行网络来近似迭代跟踪控制律

$$\hat{\nu}_i(e(k))=\omega_{a2}^{\mathrm{T}}\Theta_a(\omega_{a1}^{\mathrm{T}}e(k)) \tag{40}$$

其中, Θ_a 是执行网络的激活函数. 类似地, 执行网络的训练性能指标定义为

$$E_i^a=\frac{1}{2}[\hat{\nu}_i(e(k))-\nu_i(e(k))]^{\mathrm{T}}[\hat{\nu}_i(e(k))-\nu_i(e(k))] \tag{41}$$

其中, $\nu_i(e(k))$ 可根据下式获得

$$\nu_i(e(k))=-\frac{\gamma}{2}R^{-1}\left(\frac{\partial x(k+1)}{\partial u(x(k))}\right)^{\mathrm{T}}\frac{\partial\hat{V}_i(e(k+1))}{\partial e(k+1)} \tag{42}$$

采用梯度下降算法, 评判网络和执行网络的权值矩阵更新规则为

$$\begin{aligned}&\omega_{c\tau}:=\omega_{c\tau}-\alpha_c\frac{\partial E_i^c}{\partial\omega_{c\tau}}\\&\omega_{a\tau}:=\omega_{a\tau}-\alpha_a\frac{\partial E_i^a}{\partial\omega_{a\tau}},\ \tau=1,2\end{aligned} \tag{43}$$

其中, $\alpha_c,\alpha_a\in(0,1)$ 分别为评判网络和执行网络的学习率.

4 仿真实验

本节开展两个仿真实验用于体现算法的控制性能, 首先针对一个非仿射的倒立摆装置, 其次考虑污水处理应用.

4.1 倒立摆装置

考虑一个具有双曲切线输入的倒立摆装置[27], 其离散时间状态方程为

$$x(k+1)=\begin{bmatrix}x_1(k)+0.1x_2(k)\\-0.6125\sin(x_1(k))+0.975x_2(k)\end{bmatrix}+\begin{bmatrix}0\\0.125\big(\tanh(u(x(k)))+u(x(k))\big)\end{bmatrix} \tag{44}$$

其中, $x(k)=[x_1(k),x_2(k)]^{\mathrm{T}}$ 是状态变量, $u(x(k))$ 是控制律, $x(0)=[-0.2,0.8]^{\mathrm{T}}$. 令代价函数如式 (6) 所示. 根据自适应评判领域常用的准则, 学习参数在表 1 中给出. 其选取原则是使得代价函数序列收敛.

表 1 基于广义值迭代算法的跟踪控制参数值

Table 1 Parameter values of tracking control based on generalized value iterative algorithm

符号	Q	R	Λ	γ
例1	I_2	$0.5I_2$	$40I_2$	0.97
例2	$0.01I_2$	$0.01I_2$	I_2	0.98

在开展迭代算法之前, 需要提前对三层结构的模型网络进行训练. 选取 1000 组样本数据并设定学习率 $\alpha_m=0.02$, 我们使用 MATLAB 神经网络工具箱来训练模型网络, 其中训练误差为 10^{-8}, 训练步数为 500. 当训练结束后, 模型网络的权值保持不变. 根据式 (36) 所示的性能指标, 模型网络的训练效果如图 1 所示.

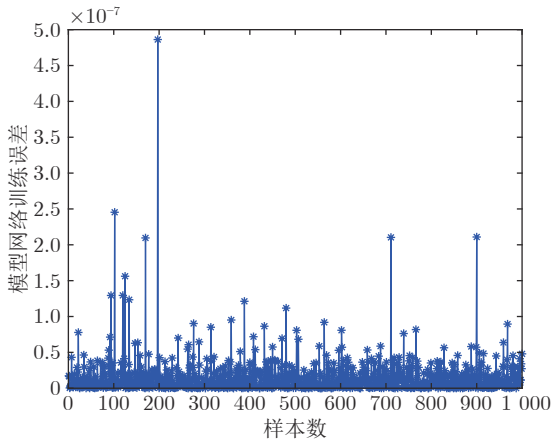

图 1 模型网络的训练误差

Fig. 1 The training errors of the model network

接下来, 给出需要跟踪的参考轨迹方程为

$$r(k+1)=\begin{bmatrix}r_1(k)+0.1r_2(k)\\-0.2492r_1(k)+0.9888r_2(k)\end{bmatrix} \tag{45}$$

其中, $r(k)=[r_1(k),r_2(k)]^{\mathrm{T}}$, $r(0)=[-0.1,0.2]^{\mathrm{T}}$. 根据式 (37), 我们使用 MATLAB 中的 “fsolve”来求

解稳定控制. 为了执行迭代算法, 我们建立结构同为 2–8–1 的评判网络和执行网络. 在神经网络的更新中, 两个网络的初始权值范围为 $[-0.2, 0.2]$, 激活函数选为 $\tanh(\cdot)$, 学习率为 $\alpha_c = \alpha_a = 0.05$. 基于选定的参数, 我们执行具有折扣因子的广义值迭代算法来获得近似最优的跟踪控制律. 值得一提的是, 当停止准则中两个条件满足时, 即 $|V_{i+1}(e(k)) - V_i(e(k))| < \varsigma$ 和 $\gamma > \Psi_i$, 其中 $\varsigma = 10^{-5}$, 我们停止算法的迭代. 在每一次迭代时, 我们训练评判网络和执行网络直到性能指标 E_i^c 和 E_i^a 小于 10^{-8} 或者达到最大训练步 500 .

执行迭代算法后, 迭代代价函数的收敛曲线如图 2 所示, 折扣因子和 Ψ_i 在图 3 中给出, 评判网络和执行网络的权值矩阵范数收敛效果在图 4 中给出. 当 $i = 13$ 时, 条件 $\gamma > \Psi_i$ 成立. 即在 13 次迭代之后的所有跟踪控制律都为可容许控制律. 而条件 $|V_{i+1}(e(k)) - V_i(e(k))| < \varsigma$ 成立时的迭代指标为 $i = 233$. 上述收敛效果验证了所提算法的有效性且

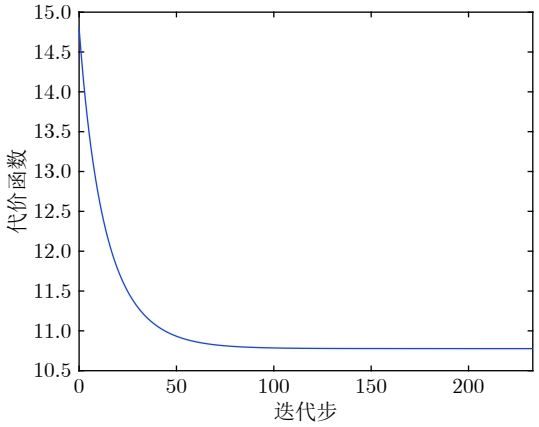

图 2　代价函数收敛过程

Fig. 2　The convergence process of the cost function

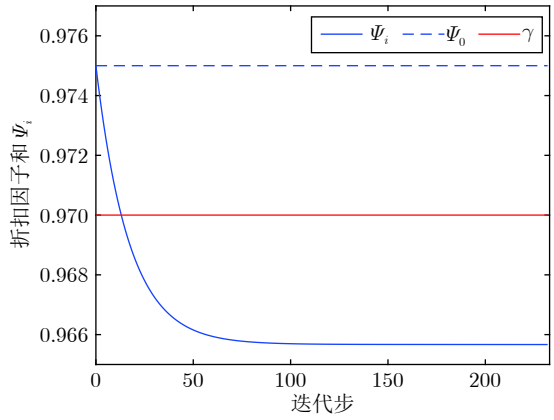

图 3　折扣因子和 Ψ_i 曲线

Fig. 3　The curves of the discount factor and Ψ_i

此时的迭代跟踪控制律具有可容许性和近似最优性. 接下来, 对于给定的初始状态 $x(0)$ 和 $r(0)$, 我们使用训练好的执行网络产生近似最优跟踪控制律. 值得注意的是, 原始系统的控制律是稳定控制和跟踪控制律的和, 即 $u(x(k)) = u(e(k)) + u(r(k))$. 在运行 120 个时间步之后, 系统状态, 参考轨迹和控制律曲线如图 5 所示. 此外, 跟踪误差和跟踪控制律的曲线如图 6 所示. 可以看到, 本文提出的跟踪控制方法能够使得原始系统快速地跟踪上参考轨迹, 进一步验证了所提跟踪技术的可行性和有效性.

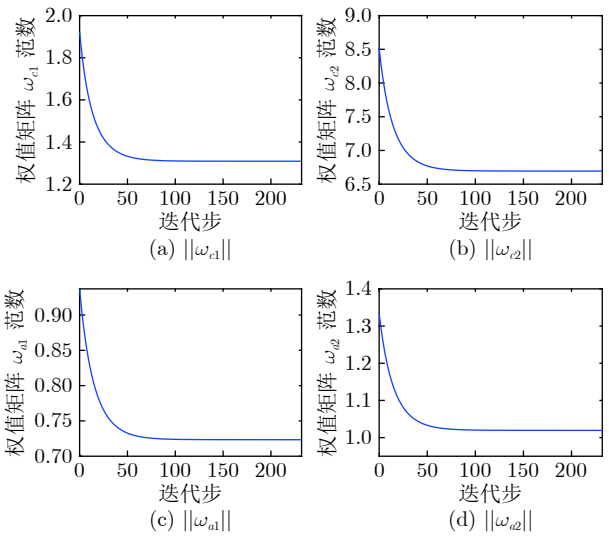

图 4　权值矩阵范数收敛过程

Fig. 4　The convergence process of the norm of weight matrices

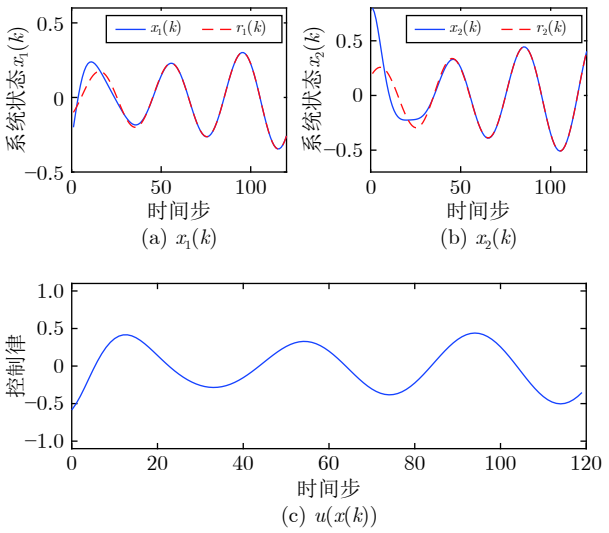

图 5　系统状态和控制律轨迹

Fig. 5　Trajectories of the state and the control law

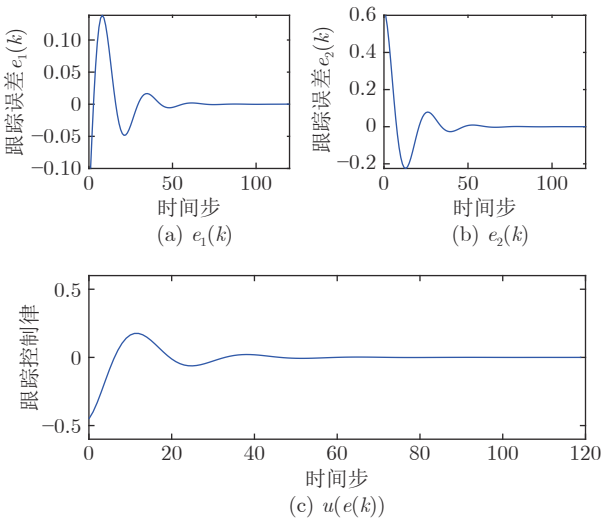

图 6 跟踪误差和跟踪控制律轨迹

Fig. 6 Trajectories of the error and the tracking control law

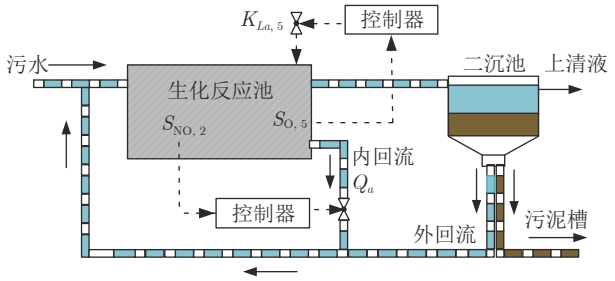

图 7 污水处理过程示意图

Fig. 7 The simple structure of the wastewater treatment process

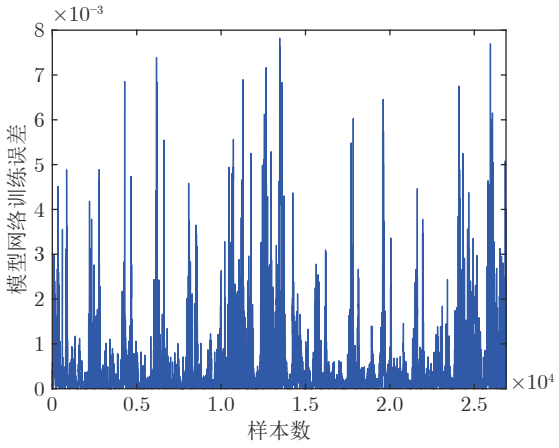

图 8 模型网络的训练误差

Fig. 8 The training errors of the model network

4.2 污水处理应用

污水处理是实现水资源循环利用的一个重要途径. 大多数污水处理厂采用活性污泥工艺来处理污水, 其中脱氮除磷是主要的实现目标. 以污水处理国际标准模型 (Benchmark simulation model No.1, BSM1) 为平台, 我们将提出的值迭代跟踪算法应用于污水处理中溶解氧浓度和硝态氮浓度的控制设计. 在污水处理反应过程中, 通常要求溶解氧浓度 ($S_{O,5}$) 和硝态氮浓度 ($S_{NO,2}$) 维持在合理的水平, 即 2 mg/l 和 1 mg/l[28–29]. 此外, 氧传递系数 $K_{La,5}$ 和内回流量 Q_a 是对应的控制变量. 在这里, 定义系统状态为 $x(k)=[S_{O,5}, S_{NO,2}]^{\mathrm{T}}$, 参考轨迹为 $r(k)=[2,1]^{\mathrm{T}}$, 控制输入为 $u(x(k))=[K_{La,5}, Q_a]^{\mathrm{T}}$. 图 7 给出了污水处理过程的简单结构图. 污水处理过程具有的非线性和不确定性使其难以建立精确的数学模型. 因此, 我们使用一个结构为 4–12–2 的模型网络来学习系统的复杂动态. 利用晴天情况下的 26 880 组输入输出数据来训练模型网络, 其中学习率为 0.02, 训练步为 800, 训练精度为 10^{-4}. 训练结束后, 模型网络权值不再变化且训练误差如图 8 所示. 然后, 我们使用 MATLAB 中的 “fsolve” 函数来求解稳定控制. 由于跟踪的参考轨迹 $r=[2,1]^{\mathrm{T}}$ 是常数, 于是得到的稳定控制也为常数, 即 $u(r(k))=[206, 29\,166]^{\mathrm{T}}$.

接下来, 我们实现数据驱动的折扣广义值迭代算法. 效用函数中的正定矩阵 和初始代价函数中的矩阵以及其他学习参数在表 1 中给出. 从实际平台中, 我们可以观测到溶解氧浓度和硝态氮浓度的初始值 $x(0)=[0.5, 3.7]^{\mathrm{T}}$. 我们构造结构为 2–20–1 的评判网络和 2–20–2 的执行网络来近似代价函数和跟踪控制律. 在每个迭代步内, 设置学习率 $\alpha_c=\alpha_a=0.05$, 我们使用 1 000 个训练步来训练评判网络和执行网络直到误差小于 10^{-8}. 在 771 次迭代后, 代价函数, Ψ_i 和权值矩阵范数收敛结果分别展示在图 9 ~ 11 中. 可以看出, 代价函数是单调递减的且在第 124 次迭代时跟踪控制律的可容许条件得到满足.

对于给定的零初始值 $x(0)$, 我们将得到的近似最优跟踪控制律作用于受控系统. 在运行 600 个时间步后, 系统的状态响应曲线和控制律曲线如图 12 所示, 而跟踪误差和跟踪控制律的曲线在图 13 中给出. 可以清楚地看到, 溶解氧浓度和硝态氮浓度维持在理想值. 这验证了所提折扣广义值迭代算法的有效性以及停止准则的可用性.

为了验证算法的自适应能力, 我们对系统控制阶段的前 200 个时间步施加一个大的干扰量. 具体为在氧传递系数中增加一个取值为 $[-25, 25]$ 的扰动分量, 同时在内回流量中增加一个取值为 $[-150, 150]$

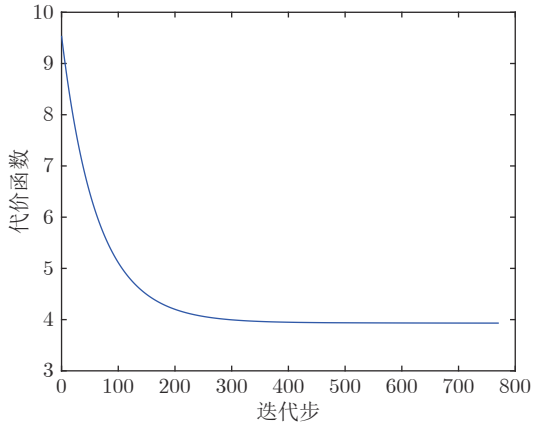

图 9　代价函数收敛过程

Fig. 9　The convergence process of the cost function

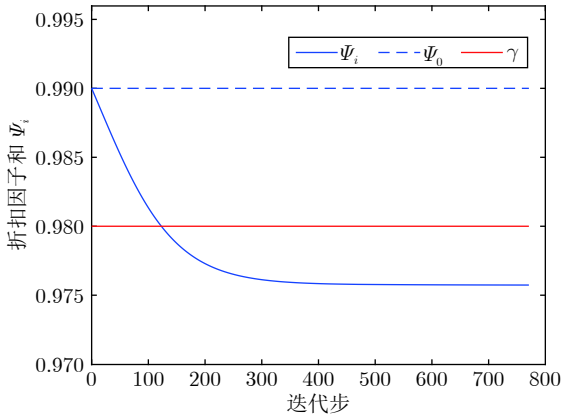

图 10　折扣因子和 Ψ_i 曲线

Fig. 10　The curves of the discount factor and Ψ_i

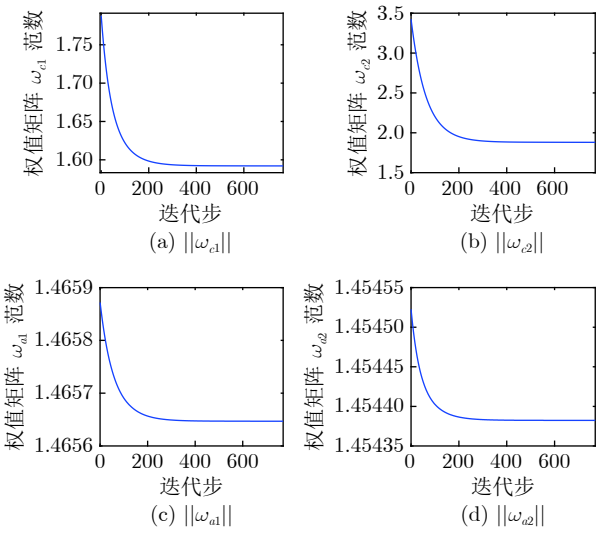

图 11　权值矩阵范数收敛过程

Fig. 11　The convergence process of the norm of weight matrices

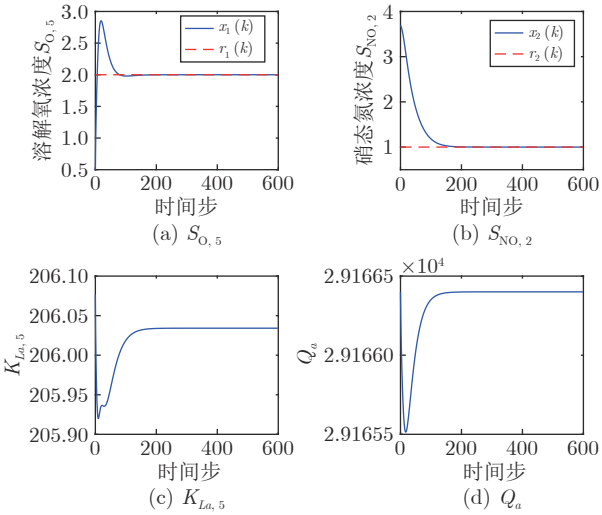

图 12　系统状态和控制律轨迹

Fig. 12　Trajectories of the state and the control law

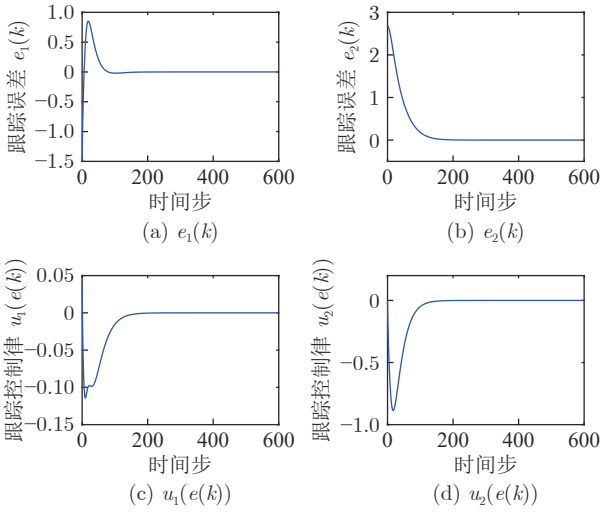

图 13　跟踪误差和跟踪控制律轨迹

Fig. 13　Trajectories of the error and the tracking control law

的扰动分量. 这时系统状态和控制输入的变化曲线如图 14 所示. 在干扰的作用下, 系统仍能跟踪上期望的设定值, 这反映了本文设计的算法具有自适应性和鲁棒性.

5　结论

针对非仿射系统的跟踪设计问题, 我们提出了一种基于折扣广义值迭代的自适应控制方法. 首先, 利用系统的输入输出数据, 建立模型网络来获得稳定控制和提供下一时刻状态相对于控制律的偏导数, 这个过程不要求精确的数学模型或系统动态矩阵. 然后, 基于折扣广义值迭代的性质, 通过使迭代

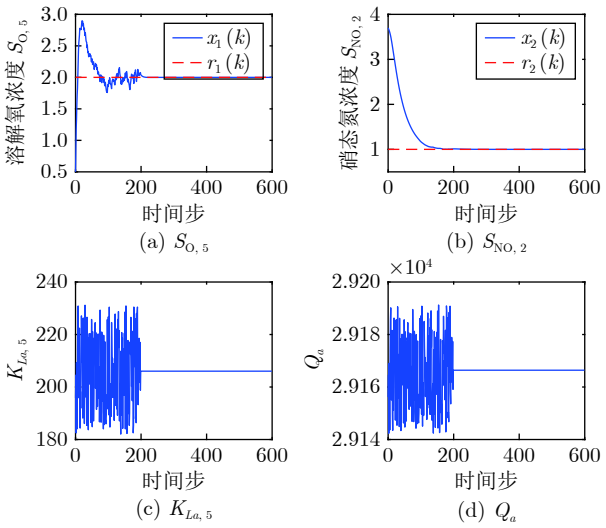

图 14 带有干扰的系统状态和控制律轨迹

Fig. 14 Trajectories of the state and the control law with the disturbance input

中的代价函数单调递减从而给出迭代跟踪控制律的可容许性判别准则. 在两个停止条件的作用下, 本文获得的跟踪控制律具有可容许性和近似最优性. 最后, 通过两个仿真实例验证了所提轨迹跟踪策略的有效性. 目前的研究是基于离线迭代开展的, 未来我们将致力于扩展该方法到在线控制领域以及实际场景应用.

References

1 Liu Y J, Zeng Q, Tong S C, Chen C L P, Liu L. Actuator failure compensation-based adaptive control of active suspension systems with prescribed performance. *IEEE Transactions on Industrial Electronics*, 2020, **67**(8): 7044−7053

2 Wang T C, Li Y M. Neural-network adaptive output-feedback saturation control for uncertain active suspension systems. *IEEE Transactions on Cybernetics*, 2020. DOI: 10.1109/TCYB.2020.3001581

3 Wang Ding. Research progress on learning-based robust adaptive critic control. *Acta Automatica Sinica*, 2019, **45**(6): 1031−1043
(王鼎. 基于学习的鲁棒自适应评判控制研究进展. 自动化学报, 2019, **45**(6): 1031−1043)

4 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
(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)

5 Song R Z, Zhu L. Optimal fixed point tracking control for discrete-time nonlinear systems via ADP. *IEEE/CAA Journal of Automatica Sinica*, 2019, **6**(3): 657−666

6 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

7 Wang D, Liu D R, Wei Q L. Finite-horizon neuro-optimal tracking control for a class of discrete-time nonlinear systems using adaptive dynamic programming approach. *Neurocomputing*, 2012, **78**: 14−22

8 Kiumarsi B, Lewis F L. Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(1): 140−151

9 Wang D, He H B, Liu D R. Adaptive critic nonlinear robust control: A survey. *IEEE Transactions on Cybernetics*, 2017, **47**(10): 3429−3451

10 Li J N, Ding J L, Chai T Y, Lewis F L, Sarangapani J. Adaptive interleaved reinforcement learning: Robust stability of affine nonlinear systems with unknown uncertainty. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. DOI: 10.1109/TNNLS.2020.3027653

11 Zhang Q C, Zhao D B. Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics. *IEEE Transactions on Cybernetics*, 2019, **49**(8): 2874−2885

12 Ha M M, Wang D, Liu D R. Event-triggered adaptive critic control design for discrete-time constrained nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(9): 3158−3168

13 Dong L, Zhong X N, Sun C Y, He H B. Adaptive event-triggered control based on heuristic dynamic programming for nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(7): 1594−1605

14 Wang D, Ha M M, Qiao J F. Self-learning optimal regulation for discrete-time nonlinear systems under event-driven formulation. *IEEE Transactions on Automatic Control*, 2020, **65**(3): 1272−1279

15 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

16 Liu D, 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

17 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

18 Li H L, Liu D R. Optimal control for discrete-time affine nonlinear systems using general value iteration. *IET Control Theory and Applications*, 2012, **6**(18): 2725−2736

19 Wei Q L, Lewis F L, Liu D R, Song R Z, Lin H Q. Discrete-time local value iteration adaptive dynamic programming: Convergence analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018, **48**(6): 875−891

20 Ha M M, Wang D, Liu D R. Generalized value iteration for discounted optimal control with stability analysis. *Systems and Control Letters*, 2021, **147**: 104847

21 Song R Z, Xiao W D, Sun C Y. Optimal tracking control for a class of unknown discrete-time systems with actuator saturation via data-based ADP algorithm. *Acta Automatica Sinica*, 2013, **39**(9): 1413−1420

22 Ha M M, Wang D, Liu D R. Data-based nonaffine optimal tracking control using iterative DHP approach. *IFAC-PapersOnLine*, 2020, **53**(2): 4246−4251

23 Wang D, Ha M M, Qiao J F. Data-driven iterative adaptive critic control toward an urban wastewater treatment plant. *IEEE Transactions on Industrial Electronics*, 2021, **68**(8): 7362−7369

24 Wang D, Zhao M M, Ha M M, Ren J. Neural optimal tracking control of constrained nonaffine systems with a wastewater

treatment application. *Neural Networks*, 2021, **143**: 121−132

25 Wang D, Zhao M M, Qiao J F. Intelligent optimal tracking with asymmetric constraints of a nonlinear wastewater treatment system. *International Journal of Robust and Nonlinear Control*, 2021, **31**(14): 6773−6787

26 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

27 Wang D, Qiao J F. Approximate neural optimal control with reinforcement learning for a torsional pendulum device. *Neural Networks*, 2019, **117**: 1−7

28 Bo Y C, Qiao J F. Heuristic dynamic programming using echo state network for multivariable tracking control of wastewater treatment process. *Asian Journal of Control*, 2015, **17**(5): 1654−1666

29 Han Hong-Gui, Zhang Lin-Lin, Wu Xiao-Long, Qiao Jun-Fei. Dataknowledge driven multiobjective optimal control for municipal wastewater treatment process. *Acta Automatica Sinica*, 2021, **47**(11): 2538−2546
(韩红桂, 张琳琳, 伍小龙, 乔俊飞. 数据和知识驱动的城市污水处理过程多目标优化控制. 自动化学报, 2021, **47**(11): 2538−2546)

王 鼎 北京工业大学信息学部教授. 2009 年获得东北大学理学硕士学位, 2012 年获得中国科学院自动化研究所工学博士学位. 主要研究方向为强化学习与智能控制. 本文通信作者.
E-mail: dingwang@bjut.edu.cn
(**WANG Ding** Professor at the Faculty of Information Technology, Beijing University of Technology. He received his master degree in operations research and cybernetics from Northeastern University, and his Ph.D. degree in control theory and control engineering from Institute of Automation, Chinese Academy of Sciences, in 2009 and 2012, respectively. His research interest covers reinforcement learning and intelligent control. Corresponding author of this paper.)

赵明明 北京工业大学硕士研究生. 主要研究方向为强化学习和智能控制.
E-mail: zhaomm@emails.bjut.edu.cn
(**ZHAO Ming-Ming** Master student at the Faculty of Information Technology, Beijing University of Technology. His research interest covers reinforcement learning and intelligent control.)

哈明鸣 北京科技大学博士研究生. 2016 年获得北京科技大学学士学位, 2019 年获得北京科技大学硕士学位. 主要研究方向为最优控制, 自适应动态规划, 强化学习.
E-mail: hamingming_0705@foxmail.com
(**HA Ming-Ming** Ph.D. candidate at the School of Automation and Electrical Engineering, University of Science and Technology Beijing. He received his master and bachelor degrees from the School of Automation and Electrical Engineering, University of Science and Technology Beijing, in 2016 and 2019, respectively. His research interest covers optimal control, adaptive dynamic programming, and reinforcement learning.)

乔俊飞 北京工业大学信息学部教授. 主要研究方向为污水处理过程智能控制, 神经网络结构设计与优化.
E-mail: junfeq@bjut.edu.cn
(**QIAO Jun-Fei** Professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers intelligent control of wastewater treatment processes, structure design and optimization of neural networks.)