

一类非仿射系统的执行依赖启发式在线跟踪控制

赵慧玲^{1,2,3,4}, 王鼎^{1,2,3,4}, 任进^{1,2,3,4}

- (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)

摘要: 针对非仿射系统的跟踪控制问题, 提出一种基于执行依赖启发式动态规划 (ADHDP) 结构的在线设计方法。首先, 考虑一类未知非仿射系统, 进行从跟踪控制问题到误差调节问题的转换; 然后, 设计基于 ADHDP 结构的跟踪控制器, 并采用在线学习的方式, 实现系统控制与执行网络和评判网络的训练同步, 使得系统状态能够跟踪期望轨迹; 最后, 通过一个仿真实例验证所提方法的有效性。

关键词: 跟踪控制; 在线学习; 执行依赖设计

中图分类号: TP13

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202144

Action dependent heuristic online tracking control for a class of nonaffine systems

ZHAO Huiling^{1,2,3,4}, WANG Ding^{1,2,3,4}, REN Jin^{1,2,3,4}

1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China
3. Beijing Institute of Artificial Intelligence, Beijing 100124, China
4. Beijing Laboratory of Smart Environmental Protection, Beijing 100124, China

Abstract: To solve the tracking control problem for nonaffine systems, an online design method was developed by using the action dependent heuristic dynamic programming (ADHDP) structure. Firstly, the tracking control problem for the unknown nonaffine system was transformed into the error regulation problem. Then, the ADHDP tracking controller was designed and the online learning method was adopted to synchronize the system control with the training of action networks and critic networks, so that the desired trajectory could be tracked by the system state. Finally, a simulation example was given to verify the effectiveness of the proposed method.

Key words: tracking control, online learning, action dependent design

1 引言

自适应动态规划 (adaptive dynamic programming, ADP) 是多个领域交叉形成的一种近似最优设

计方法。该方法融合了动态规划 (dynamic programming, DP)、强化学习 (reinforcement learning, RL) 以及神经网络 (neural network, NN) 的思想。其中, DP 是一种求解决策过程最优化的数学方法,

收稿日期: 2021-11-01; 修回日期: 2021-11-20

通信作者: 王鼎, dingwang@bjut.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61773373, No.61890930-5, No.62021003); 国家重点研发计划资助项目 (No.2021ZD0112300-2, No.2018YFC1900800-5); 北京市自然科学基金资助项目 (No.JQ19013)

Foundation Items: The National Natural Science Foundation of China (No.61773373, No.61890930-5, No.62021003), The National Key Research and Development Project of China (No.2021ZD0112300-2, No.2018YFC1900800-5), Beijing Natural Science Foundation (No.JQ19013)

其中的最优性原理为 ADP 提供了理论依据, 但 DP 存在明显的“维数灾难”问题。后来, 人工智能领域兴起了一种 RL 算法, 它能够在智能体与环境交互的过程中通过学习策略实现回报最大化。为了解决复杂非线性系统的最优控制问题, 美国著名学者韦伯斯 (Werbos) 结合 DP 原理以及 RL 思想, 将 NN 等逼近工具作为实现手段, 率先提出了 ADP 这一概念。ADP 在求解非线性系统的哈密顿-雅可比-贝尔曼 (Hamilton-Jacobi-Bellman, HJB) 方程^[1-4]方面展现了巨大潜力。因此, 学者们开始广泛关注 ADP, 并对其开展了深入的研究。近些年来, 大量学者从离散时间、连续时间、状态约束^[5-6]、跟踪控制^[7-10]等不同方向对 ADP 展开了深入挖掘。更有一些学者将其应用到具体的工业控制^[11-12]中。RL^[13-15]作为当前的研究热点之一, 很多学者利用 RL 的思想设计自适应控制器, 例如 Luo B 等人^[16]从实际系统的数据样本中学习最优跟踪控制策略, 设计了一种无模型最优跟踪控制器。此外, 还有一些学者将在线学习的方式应用到 ADP 中, Li S 等人^[17]提出了基于自适应动态规划的在线跟踪控制方法, 以解决部分具有时延的不确定非线性系统的跟踪问题。

1992 年, Werbos 又进一步引入了 ADP 两种基本结构的执行依赖形式, 其中, 执行依赖启发式动态规划 (action dependent heuristic dynamic programming, ADHDP) 与基本结构中的启发式动态规划 (heuristic dynamic programming, HDP) 原理大致相同, 都采用了 RL 中的执行-评判结构, 评判网络用于近似系统的代价函数, 起着评估系统性能的作用; 执行网络则根据控制方法输出近似最优的控制信号, 评判网络和执行网络通常采用反向传播 (back propagation, BP) 神经网络。但与 HDP 不同的是, ADHDP 可以不需要模型网络, 执行网络可以与评判网络直接相连, 且用于训练执行网络的反向传播信号不再是控制律, 而是代价函数, 这样的特点使那些建模困难的控制对象更容易控制且能够有效减少建模带来的误差。另外, 评判网络的输入不仅需要状态变量, 还需要控制变量。

本文针对一类非仿射系统, 将该系统的跟踪控制问题转换为误差调节问题, 在 ADHDP 的框架结构下, 采用在线学习的方式取代求解 HJB 方程的迭代过程, 实现控制器训练与系统控制的同步在线进行。最后通过一个仿真例子进行对比实验, 验证本文所提方法的有效性。

2 问题描述

考虑一类非仿射系统:

$$\mathbf{x}(t+1) = \Phi(\mathbf{x}(t), \mathbf{u}(t)), t \in \mathbf{N} \quad (1)$$

其中, $\mathbf{x}(t) \in \mathbb{R}^n$ 是 n 维的状态向量, $\mathbf{u}(t) \in \mathbb{R}^m$ 是 m 维的控制向量, $t \in \mathbf{N}$ ($\mathbf{N}$ 是非负整数集)。系统函数 $\Phi(\cdot, \cdot)$ 在 $\Omega_x \subseteq \mathbb{R}^n$ 上是 Lipschitz 连续的, 并且 $\Phi(0, 0) = 0$ 。

定义期望轨迹如下:

$$\mathbf{r}(t+1) = \mathbf{F}(\mathbf{r}(t)) \quad (2)$$

其中, $\mathbf{r}(t) \in \mathbb{R}^n$ 表示 t 时刻的期望值, $\mathbf{F}(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}^n$ 是可微的。则跟踪误差可以表示为:

$$\mathbf{e}(t) = \mathbf{x}(t) - \mathbf{r}(t) \quad (3)$$

假设对于任意的 $\mathbf{r}(t)$, 存在一个参考控制输入 $\mathbf{u}(\mathbf{r}(t))$, 并满足:

$$\mathbf{r}(t+1) = \Phi(\mathbf{r}(t), \mathbf{u}(\mathbf{r}(t))) \quad (4)$$

可以通过求解式 (4) 得到 $\mathbf{u}(\mathbf{r}(t))$, 定义:

$$\mathbf{u}(\mathbf{r}(t)) = \boldsymbol{\sigma}(\mathbf{r}(t)) \quad (5)$$

其中, $\boldsymbol{\sigma}(\mathbf{r}(t))$ 是式 (4) 的反解形式。这里设计一个反馈控制 $\mathbf{u}(\mathbf{x}(t))$ 满足 $\mathbf{x}(t+1) = \Phi(\mathbf{x}(t), \mathbf{u}(\mathbf{x}(t)))$, 目的是使状态向量跟踪上期望轨迹。则跟踪控制律可以表示为:

$$\mathbf{u}(\mathbf{e}(t)) = \mathbf{u}(\mathbf{x}(t)) - \boldsymbol{\sigma}(\mathbf{r}(t)) \quad (6)$$

根据式 (1) ~ 式 (6), 可以构造一个新的误差系统:

$$\mathbf{e}(t+1) = \Phi(\mathbf{e}(t) + \mathbf{r}(t), \mathbf{u}(\mathbf{e}(t)) + \boldsymbol{\sigma}(\mathbf{r}(t)) - \mathbf{F}(\mathbf{r}(t))) \quad (7)$$

求解最优跟踪控制问题的目的是找到一组控制序列使得代价函数最小, 定义代价函数如下:

$$J(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t))) = \sum_{j=t}^{\infty} U(\mathbf{e}(j), \mathbf{u}(\mathbf{e}(j))) \quad (8)$$

其中, $U(\mathbf{e}(j), \mathbf{u}(\mathbf{e}(j)))$ 为效用函数, 将其设定为:

$$U(\mathbf{e}(j), \mathbf{u}(\mathbf{e}(j))) = \mathbf{e}^T(j) \mathbf{Q} \mathbf{e}(j) + \mathbf{u}^T(\mathbf{e}(j)) \mathbf{R} \mathbf{u}(\mathbf{e}(j)) \quad (9)$$

其中, $\mathbf{Q}$ 和 $\mathbf{R}$ 是正定矩阵, 且 $U(0, 0) = 0$ 。

方便起见, 将 $J(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t)))$ 简写为 $J(\mathbf{e}(t))$ 。将式 (8) 展开得到如下形式:

$$\begin{aligned} J(\mathbf{e}(t)) &= U(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t))) + \\ &\sum_{j=t+1}^{\infty} \{ \mathbf{e}^T(j) \mathbf{Q} \mathbf{e}(j) + \mathbf{u}^T(\mathbf{e}(j)) \mathbf{R} \mathbf{u}(\mathbf{e}(j)) \} = \\ &U(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t))) + J(\mathbf{e}(t+1)) \end{aligned} \quad (10)$$

接下来,可以定义最优控制问题的 Hamiltonian 函数:

$$H(J, \mathbf{e}, \mathbf{u}) = J(\mathbf{e}(t+1)) - J(\mathbf{e}(t)) + U(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t))) \quad (11)$$

根据 Bellman 最优性原理, 最优代价函数 $J^*(\mathbf{e}(t))$ 满足以下离散时间 HJB 方程:

$$J^*(\mathbf{e}(t)) = \min_{\mathbf{u}(\mathbf{e}(t))} \{U(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t))) + J^*(\mathbf{e}(t+1))\} \quad (12)$$

最优控制策略 $\mathbf{u}^*(\mathbf{e}(t))$ 满足:

$$\mathbf{u}^*(\mathbf{e}(t)) = \arg \min_{\mathbf{u}(\mathbf{e}(t))} \{U(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t))) + J^*(\mathbf{e}(t+1))\} \quad (13)$$

最后, 原始系统的最优控制策略 $\mathbf{u}^*(\mathbf{x}(t))$ 可以表示为:

$$\mathbf{u}^*(\mathbf{x}(t)) = \mathbf{u}(\mathbf{r}(t)) + \mathbf{u}^*(\mathbf{e}(t)) \quad (14)$$

3 基于 ADHDP 结构的在线控制方法的实现

ADHDP 结构由被控系统、评判网络和执行网络 3 个部分组成。首先构造一个模型网络对未知的非仿射系统进行辨识, 进而求解稳定控制输入; 然后展示评判网络的训练过程; 最后给出执行网络的结构及更新准则。该方法的框架如图 1 所示。

3.1 系统辨识

对于未知的非仿射非线性系统, 可以采用一个 3 层的 BP 神经网络来逼近系统动态, 表达式如下:

$$\hat{\mathbf{x}}(t+1) = \mathbf{w}_{m2}^\top(t) \delta_m(\mathbf{w}_{m1}^\top(t) \mathbf{X}(t) + \boldsymbol{\rho}(t)) + \boldsymbol{\rho}_2(t) \quad (15)$$

其中, $\mathbf{X}(t) = [\mathbf{x}^\top(t), \mathbf{u}^\top(\mathbf{x}(t))]^\top$ 为该模型网络的输入, 输出 $\hat{\mathbf{x}}(t+1)$ 为下一时刻的状态预测值, $\mathbf{w}_{m1}(t)$ 和 $\boldsymbol{\rho}(t)$ 分别为输入层与隐含层之间的网络权值和阈值向量, $\mathbf{w}_{m2}(t)$ 和 $\boldsymbol{\rho}_2(t)$ 分别为隐含层与输出层之

间的网络权值和阈值向量, $\delta_m(\cdot)$ 为该模型网络的激活函数, 这里选取双曲正切函数:

$$\delta_m(\tau) = \frac{1 - e^{-\tau}}{1 + e^{-\tau}} \quad (16)$$

其中, $\tau = \mathbf{w}_{m1}^\top(t) \mathbf{X}(t) + \boldsymbol{\rho}(t)$ 。

定义模型网络的预测误差为输出的近似值与实际值之间的差值, 即:

$$\boldsymbol{\varepsilon}_m(t) = \hat{\mathbf{x}}(t+1) - \mathbf{x}(t+1) \quad (17)$$

训练的性能度量函数为:

$$E_m(t) = \frac{1}{2} \boldsymbol{\varepsilon}_m^\top(t) \boldsymbol{\varepsilon}_m(t) \quad (18)$$

模型网络各层之间的网络权值与阈值向量的更新规则均采用如下梯度下降算法:

$$\mathbf{w}_{m1}(t+1) = \mathbf{w}_{m1}(t) + \lambda_m \left[-\frac{\partial E_m(t)}{\partial \mathbf{w}_{m1}(t)} \right] \quad (19)$$

$$\boldsymbol{\rho}(t+1) = \boldsymbol{\rho}(t) + \lambda_m \left[-\frac{\partial E_m(t)}{\partial \boldsymbol{\rho}(t)} \right] \quad (20)$$

$$\mathbf{w}_{m2}(t+1) = \mathbf{w}_{m2}(t) + \lambda_m \left[-\frac{\partial E_m(t)}{\partial \mathbf{w}_{m2}(t)} \right] \quad (21)$$

$$\boldsymbol{\rho}_2(t+1) = \boldsymbol{\rho}_2(t) + \lambda_m \left[-\frac{\partial E_m(t)}{\partial \boldsymbol{\rho}_2(t)} \right] \quad (22)$$

其中, λ_m 为神经网络训练过程的学习率。

通过式 (4) 可以采用上述训练好的模型来表示期望轨迹:

$$\mathbf{r}(t+1) = \mathbf{w}_{m2}^\top(t) \delta_m(\mathbf{w}_{m1}^\top(t) [\mathbf{r}^\top(t), \mathbf{u}^\top(\mathbf{r}(t))]^\top + \boldsymbol{\rho}(t)) + \boldsymbol{\rho}_2(t) \quad (23)$$

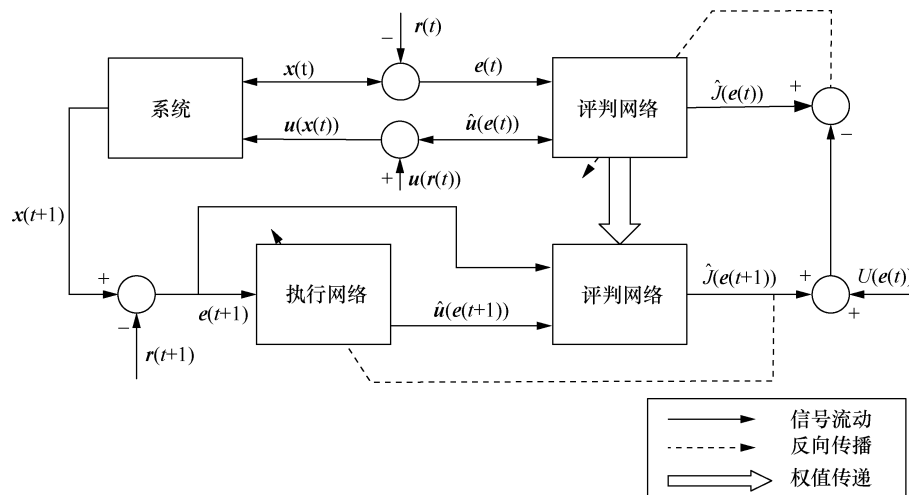


图 1 基于 ADHDP 结构的在线控制方法框架

接下来便可以通过数学求解的方法得到 $u(r(t))$ 。

3.2 评判网络

在 ADHDP 框架中, 评判网络用于近似系统的代价函数, 与传统 HDP 结构不同的是, ADHDP 的评判网络的输入包括跟踪误差 $e(t)$ 和跟踪控制律 $u(e(t))$ 两个部分; 相同的是输出都为当前时刻的代价函数。采用一个 3 层的 BP 神经网络来构造评判网络, 其训练过程包括正向计算和误差反向传播两个过程。其中, 正向计算过程如下:

$$\hat{J}(e(t)) = w_{c2}^T(t) \delta_c(w_{c1}^T(t)[e^T(t), u^T(e(t))]^T) \quad (24)$$

其中, $w_{c1}(t)$ 和 $w_{c2}(t)$ 分别为输入层与隐含层之间和隐含层与输出层之间的网络权值, $\delta_c(\cdot)$ 为评判网络的激活函数, 这里同样选取双曲正切函数。

方便起见, 将经过评判网络激活函数生成的序列表示为如下形式:

$$\phi_c(e(t), u(e(t))) = \delta_c(w_{c1}^T(t)[e^T(t), u^T(e(t))]^T) \quad (25)$$

接下来是评判网络的误差反向传播过程, 训练评判网络的目标是逼近代价函数, 定义误差函数为:

$$\begin{aligned} \varepsilon_c(t) = & \hat{J}(e(t)) - [U(e(t), u(e(t))) + \hat{J}(e(t+1))] = \\ & -U(e(t), u(e(t))) - w_{c2}^T(t) \times \\ & [\phi_c(e(t+1), u(e(t+1))) - \phi_c(e(t), u(e(t)))] \end{aligned} \quad (26)$$

令

$$D_c = \phi_c(e(t+1), u(e(t+1))) - \phi_c(e(t), u(e(t))) \quad (27)$$

训练评判网络的目标函数为:

$$E_c(t) = \frac{1}{2} \varepsilon_c^T(t) \varepsilon_c(t) \quad (28)$$

采用梯度下降法对评判网络的权值进行更新, 输入层与隐含层之间的权值保持不变, 只更新隐含层与输出层之间的权值, 其更新规则为:

$$w_{c2}(t+1) = w_{c2}(t) + \lambda_c \left[-\frac{\partial E_c(t)}{\partial w_{c2}(t)} \right] \quad (29)$$

$$\frac{\partial E_c(t)}{\partial w_{c2}(t)} = \varepsilon_c(t) \frac{D_c}{1 + D_c^T D_c} \quad (30)$$

其中, λ_c 为评判网络的学习率。

3.3 执行网络

执行网络同样采用一个 3 层的 BP 神经网络, 其输入为跟踪误差 $e(t)$, 输出为控制策略 $\hat{u}(e(t))$ 。定义其正向计算的表达式如下:

$$\hat{u}(e(t)) = w_{a2}^T(t) \delta_a(w_{a1}^T(t)e(t)) \quad (31)$$

其中, $w_{a1}(t)$ 和 $w_{a2}(t)$ 分别为执行网络输入层与隐含层之间和隐含层与输出层之间的网络权值, $\delta_a(\cdot)$ 为执行网络的激活函数, 这里同样选取为双曲正切函数。

在 ADHDP 框架下, 定义执行网络的预测误差为:

$$\varepsilon_a(t) = \hat{J}(e(t+1)) \quad (32)$$

定义执行网络的目标函数为:

$$E_a(t) = \frac{1}{2} \varepsilon_a^T(t) \varepsilon_a(t) \quad (33)$$

与上面提到的评判网络相似, 执行网络输入层与隐含层之间的权值保持不变, 隐含层与输出层之间的权值更新规则同样采用梯度下降算法。

$$w_{a2}(t+1) = w_{a2}(t) + \lambda_a \left[-\frac{\partial E_a(t)}{\partial w_{a2}(t)} \right] \quad (34)$$

其中, λ_a 为执行网络的学习率。

算法步骤如下。

步骤 1: 初始化评判网络和执行网络各层之间的权值参数 w_{c1} 、 w_{c2} 、 w_{a1} 、 w_{a2} , 以及它们的学习率 λ_c 和 λ_a , 设置效用函数的参数, 初始化时刻 $t=0$ 。

步骤 2: 输入系统当前时刻的状态 $x(t)$, 根据式 (3) 计算出状态向量与期望轨迹之间的误差向量 $e(t)$, 将 $e(t)$ 输入执行网络, 通过式 (31) 正向计算出控制策略 $\hat{u}(e(t))$ 。

步骤 3: 根据当前时刻的跟踪误差 $e(t)$ 及控制策略 $\hat{u}(e(t))$, 通过式 (9) 和式 (24) 计算出效用函数 $U(e(t), \hat{u}(e(t)))$ 和代价函数 $\hat{J}(e(t))$ 。

步骤 4: 根据式 (6) 得到 $u(x(t))$, 并作用于被控系统上, 得到下一时刻的状态向量 $x(t+1)$, 重复步骤 2, 得到下一时刻的控制策略 $\hat{u}(e(t+1))$ 。

步骤 5: 重复步骤 3, 得到下一时刻的效用函数 $U(e(t+1), \hat{u}(e(t+1)))$ 和代价函数 $\hat{J}(e(t+1))$ 。

步骤 6: 通过式 (26) 计算评判网络的误差函数 $\varepsilon_c(t)$, 并根据式 (29) 自适应地调整评判网络的权值。

步骤 7: 判断评判网络的预测误差 $\varepsilon_c(t)$ 是否达到控制精度或最大训练次数, 若是, 则执行步骤 8; 否则, 跳转到步骤 3。

步骤 8: 通过式 (32) 计算执行网络的误差函数 $\varepsilon_a(t)$, 并根据式 (34) 自适应地修正执行网络的权值。

步骤 9: 判断执行网络的预测误差 $\varepsilon_a(t)$ 是否达

到控制精度或最大训练次数,若是,则继续执行步骤 10; 否则,跳转到步骤 8。

步骤 10: 令 $t=t+1$, 若 t 达到最大时间步, 则执行步骤 11; 否则, 跳转到步骤 2。

步骤 11: 结束训练。

4 仿真实验

本节通过一个仿真例子来验证所提方法的有效性, 考虑一个非仿射的动态系统如下:

$$\mathbf{x}(t+1) = \begin{bmatrix} \tanh(\mathbf{x}_1(t)) + 0.05 \tanh(\mathbf{x}_2(t)) \\ -0.3 \tanh(\mathbf{x}_1(t)) + \tanh(\mathbf{x}_2(t)) \end{bmatrix} + \begin{bmatrix} 0 \\ \sin(\mathbf{u}(t)) \end{bmatrix} \quad (35)$$

选取初值为 $\mathbf{x}(0)=[5,3]^T$ 。

首先通过 MATLAB 的神经网络工具箱对该系统进行建模, 从而求取参考控制输入。选取一个 3 层的 BP 神经网络对系统进行辨识, 其结构为 3-8-2(分别对应输入层、隐含层、输出层的神经元个数)。隐含层神经元个数的选择至关重要, 经过不断调试, 最终将隐含层神经元个数确定为 8。学习率的选取反映了神经网络调整权值的快慢, 经过调试, 选取系统辨识神经网络的学习率为 $\lambda_m = 0.02$, 随机生成 1 000 组样本作为模型网络的输入, 将非仿射系统生成的 1 000 组样本数据作为模型网络的输出, 对模型网络进行 500 次迭代训练。最终获得模型网络的训练误差如图 2 所示。基于以上训练好的模型网络, 可以直接使用 MATLAB 中的 fsolve 函数求解稳态控制 $\mathbf{u}(\mathbf{r}(t))$ 。

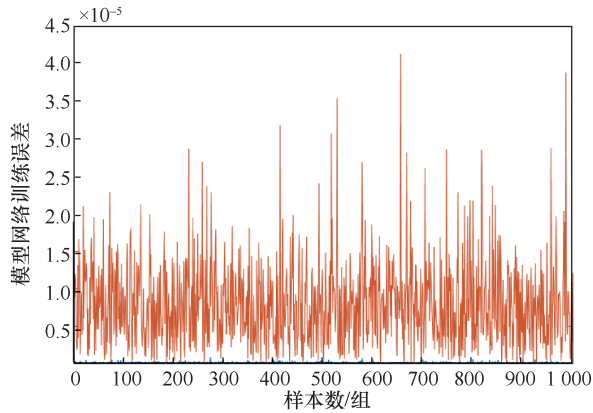


图2 模型网络训练误差

接下来, 定义期望轨迹如下:

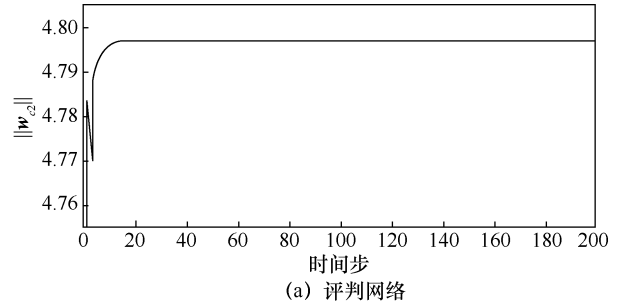
$$\mathbf{r}(t+1) = \begin{bmatrix} 0.9963\mathbf{r}_1(t) + 0.0498\mathbf{r}_2(t) \\ -0.2492\mathbf{r}_1(t) + 0.9888\mathbf{r}_2(t) \end{bmatrix} \quad (36)$$

选取参考初值为 $\mathbf{r}(0)=[0.06,0.4]^T$ 。

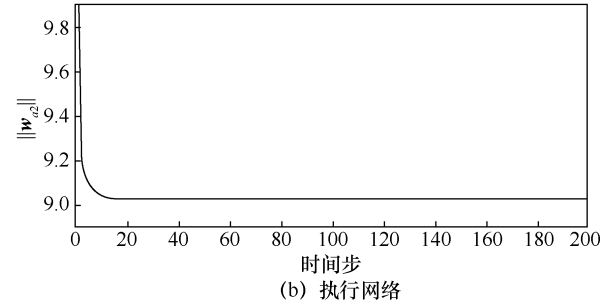
在本控制方法中, 评判网络和执行网络的结构为 3-8-1、2-8-1, 学习率 λ_c 、 λ_a , 内部最大循环次数 N_c 、 N_a , 训练目标精度 ρ_c 、 ρ_a , 效用函数 $U(\mathbf{e}(t), \mathbf{u}(\mathbf{e}(t)))$ 中的参数 $\mathbf{Q}$ 和 R 的取值均在表 1 中给出, 确定系统运行最大时间步 $t_{\max} = 200$ 。评判网络和执行网络的权值调整采用基于梯度下降算法的在线学习的方式进行更新, 如式 (29) 和式 (34) 所示。图 3 展示的是评判网络和执行网络隐含层与输出层之间的权值范数收敛曲线。由图 3 可以看到, 随着时间步的前进, 评判网络和执行网络的权值范数逐渐收敛至一个定值。

表1 基于 ADHDP 结构的在线控制方法的参数值

$\mathbf{Q}$	R	λ_c 、 λ_a	N_c 、 N_a	ρ_c 、 ρ_a
$\begin{bmatrix} 0.01 & 0 \\ 0 & 0.5 \end{bmatrix}$	0.01	0.05	1 000	1×10^{-12}



(a) 评判网络



(b) 执行网络

图3 权值范数收敛曲线

为了验证本文所提方法的有效性, 将该方法与比例积分微分 (proportional-integral-derivative, PID) 控制算法做了一个简单的对比实验, 其中, PID 控制算法的参数经过试错法设定, 变量一的控制参数为 5、3、0.5, 变量二的控制器参数为 1、0.2、0.1。采用两种方案分别对同一个系统进行控制后, 系统状态与期望轨迹跟踪效果如图 4 所示。由图 4 可以看到, 采用基于 ADHDP 结构的在线控制方法的系统状态在 20 个时间步之后基本跟踪上了期望

轨迹,而PID控制算法的跟踪效果要差一些,且本文所提方法的超调量较PID控制算法的超调量要小。两种方法的系统跟踪误差如图5所示,本文所提方法的跟踪误差在20个时间步之后基本趋近于零或在零点附近轻微波动,PID控制算法的控制效果相对较差。基于ADHDP结构的在线控制方法的误差控制策略收敛曲线如图6所示。结合几个仿真结果图可以看出,当误差控制律 $u(e(t))$ 逐渐趋近于零时,跟踪误差也逐渐趋近于零,系统状态最终跟踪上了期望轨迹。经过对比仿真实验,可以证明本文所提方法能够有效地解决非仿射非线性动态系统的跟踪控制问题。

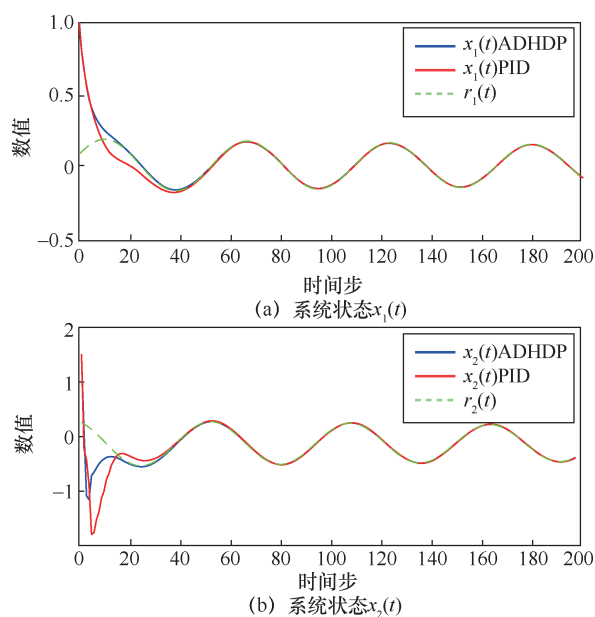


图4 系统状态与期望轨迹跟踪效果

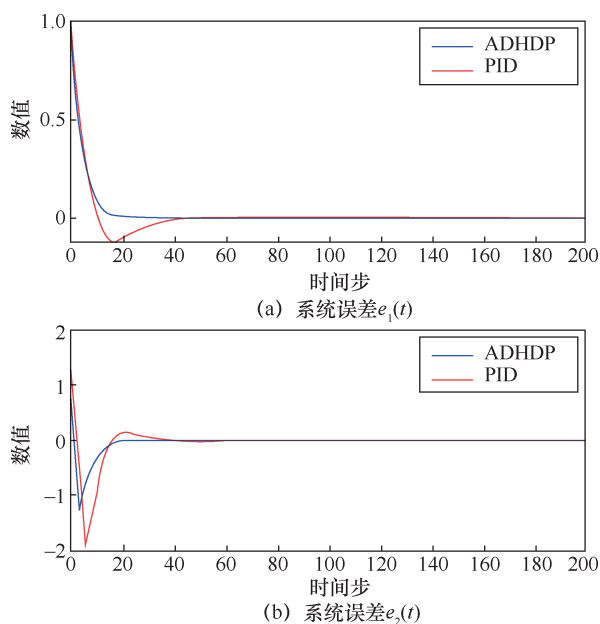


图5 系统跟踪误差

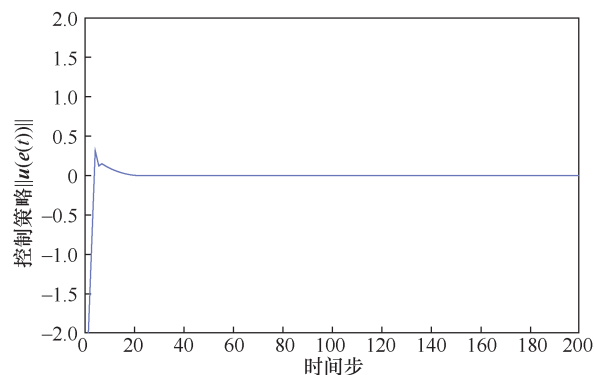


图6 基于ADHDP结构的在线控制方法的误差控制收敛曲线

5 结束语

本文提出了一种基于ADHDP结构的在线控制方法来求解未知非仿射系统的最优跟踪控制问题。首先,将最优跟踪控制问题转化为误差调节问题。然后,建立ADHDP框架,并采用基于梯度下降算法的在线学习方案进行控制器设计。最后,利用一个非仿射系统的仿真实例证明了所提出的最优跟踪控制方法的有效性。

参考文献:

- [1] LIU D R, WEI Q L. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(3): 621-634.
- [2] LIU D R, WEI Q L, WANG D, et al. Adaptive dynamic programming for optimal residential energy management[M]//Adaptive dynamic programming with applications in optimal control. Cham: Springer International Publishing, 2017: 483-535.
- [3] WANG D, HA M M, QIAO J F. Self-learning optimal regulation for discrete-time nonlinear systems under event-driven formulation[J]. IEEE Transactions on Automatic Control, 2020, 65(3): 1272-1279.
- [4] WANG D, QIAO J F, CHENG L. An approximate neuro-optimal solution of discounted guaranteed cost control design[J]. IEEE Transactions on Cybernetics, 2020: 1-10.
- [5] LI Y M, LIU Y J, TONG S C. Observer-based neuro-adaptive optimized control of strict-feedback nonlinear systems with state constraints[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021: 1-15.
- [6] YANG X, HE H B, ZHONG X N. Approximate dynamic programming for nonlinear-constrained optimizations[J]. IEEE Transactions on Cybernetics, 2021, 51(5): 2419-2432.
- [7] WANG D, HA M M, CHENG L. Neuro-optimal trajectory tracking with value iteration of discrete-time nonlinear dynamics[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021: 1-12.
- [8] 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[J]. Neurocomputing, 2012, 78(1): 14-22.
- [9] KIUMARSI B, LEWIS F L. Actor-critic-based optimal tracking for

- partially unknown nonlinear discrete-time systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(1): 140-151.
- [10] HA M M, WANG D, LIU D R. Data-based nonaffine optimal tracking control using iterative DHP approach[J]. IFAC-PapersOnLine, 2020, 53(2): 4246-4251.
- [11] WANG D, HA M M, QIAO J F. Data-driven iterative adaptive critic control toward an urban wastewater treatment plant[J]. IEEE Transactions on Industrial Electronics, 2021, 68(8): 7362-7369.
- [12] WEI Q L, LIAO Z H, SONG R Z, et al. Self-learning optimal control for ice-storage air conditioning systems via data-based adaptive dynamic programming[J]. IEEE Transactions on Industrial Electronics, 2021, 68(4): 3599-3608.
- [13] 李金娜, 程薇燃. 基于强化学习的数据驱动多智能体系统最优一致性综述[J]. 智能科学与技术学报, 2020, 2(4): 327-340.
- LI J N, CHENG W R. An overview of optimal consensus for data driven multi-agent system based on reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 327-340.
- [14] 李涛, 魏庆来. 基于深度强化学习的智能暖气温度控制系统[J]. 智能科学与技术学报, 2020, 2(4): 348-353.
- LI T, WEI Q L. Intelligent heating temperature control system based on deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 348-353.
- [15] 王金予, 魏欣然, 石文磊, 等. 强化学习在资源优化领域的应用[J]. 大数据, 2021, 7(5): 131-149.
- WANG J Y, WEI X R, SHI W L, et al. Applications of reinforcement learning in the field of resource optimization[J]. Big Data Research, 2021, 7(5): 131-149.
- [16] LUO B, LIU D R, HUANG T W, et al. Model-free optimal tracking control via critic-only Q-learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(10): 2134-2144.
- [17] LI S, DING L, GAO H B, et al. ADP-based online tracking control of partially uncertain time-delayed nonlinear system and application to wheeled mobile robots[J]. IEEE Transactions on Cybernetics, 2020, 50(7): 3182-3194.

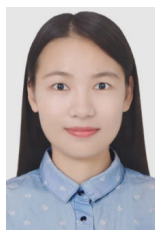
[作者简介]



赵慧玲(1998—), 女, 北京工业大学信息学部硕士生, 主要研究方向为强化学习与智能控制。



王鼎(1984—), 男, 博士, 北京工业大学信息学部教授、博士生导师, 主要研究方向为强化学习与智能控制。



任进(1999—), 女, 北京工业大学信息学部硕士生, 主要研究方向为强化学习与智能控制。