基于强化学习的一类具有输入约束非线性系统最优控制

罗 傲, 肖文彬, 周 琪†, 鲁仁全

(广东工业大学 广东省智能决策与协同控制重点实验室, 广东 广州 510006)

摘要: 针对部分系统存在输入约束和不可测状态的最优控制问题, 本文将强化学习中基于执行–评价结构的近似最优算法与反步法相结合, 提出了一种最优跟踪控制策略. 首先, 利用神经网络构造非线性观测器估计系统的不可测状态. 然后, 设计一种非二次型效用函数解决系统的输入约束问题. 相比现有的最优方法, 本文提出的最优跟踪控制方法不仅具有反步法在处理n阶系统跟踪问题上的优势, 而且保证了所有虚拟控制器均为最优, 同时, 该方法可以简化控制器设计过程. 最后, 基于李雅普诺夫稳定性理论, 证明了闭环系统中的所有信号一致最终有界. 通过仿真结果验证该方法的有效性.

关键词: 输入约束; 不可测状态; 最优控制; 强化学习; 反步法

引用格式: 罗傲, 肖文彬, 周琪, 等. 基于强化学习的一类具有输入约束非线性系统最优控制. 控制理论与应用, 2022, 39(1): 154 – 164

DOI: 10.7641/CTA.2021.00898

Optimal control for a class of nonlinear systems with input constraints based on reinforcement learning

LUO Ao, XIAO Wen-bin, ZHOU Qi†, LU Ren-quan

(Guangdong Province Key Laboratory of Intelligent Decision and Cooperative Control, Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: In this paper, by incorporating the approximate optimization algorithm, which is derived from actor-critic structure in reinforcement learning, into the backstepping, an optimal tracking control strategy is proposed for a class of nonlinear systems with immeasurable states and input constraints. First, a nonlinear observer is constructed with neural network to estimate the immeasurable states. Then, a non-quadratic cost function is designed to solve the problem of controller constraints. Compared with the existing optimization methods, the optimal tracking control method proposed in this paper not only has the advantage of backstepping technique in addressing the n-order system tracking problem, but also ensures that all virtual controllers are optimal. And this method simplifies the controller design. Finally, according to Lyapunov stability theory, it is proven that all signals in the closed-loop system are uniformly ultimately bounded. The effectiveness of the proposed method is verified by the simulation results.

Key words: input constraints; immeasurable states; optimal control; reinforcement learning; backstepping

Citation: LUO Ao, XIAO Wenbin, ZHOU Qi, et al. Optimal control for a class of nonlinear systems with input constraints based on reinforcement learning. *Control Theory & Applications*, 2022, 39(1): 154 – 164

1 引言

随着世界能源危机的日益严峻, 最优控制已经成为实际生产生活中一项重要设计方案[1–3]. 对于被控对象效用函数为二次型的线性系统主要通过求解黎卡提(Riccati)方程获得其最优解. 然而, 实际中大部分被控系统具有复杂非线性的特点, 需要求解哈密顿–雅可比–贝尔曼(Hamilton–Jacobi–Bellman, HJB)方程. 由于HJB方程具有强非线性, 从而导致难以获得方程解析解, 同时, 随着系统维数的增加, 求解过程中会产生“维数灾”问题. 因此, 如何通过近似算法获得HJB方程解, 是当前优化控制领域研究的热点问题[4–6]. 近年来, 基于强化学习执行–评价网络的自适应评判控制策略被广泛应用于非线性系统的最优控制问题. 该方法利用神经网络或模糊逻辑系统的近似

收稿日期: 2020−12−14; 录用日期: 2021−07−08.

†通信作者. E-mail: zhouqi2009@gmail.com; Tel.: +86 13342816257.

本文责任编委: 岳东.

国家自然科学基金项目(62121004, 61973091), “广东特支计划”本土创新创业团队项目(2019BT02X353), 广东省重点领域研发计划项目(2021B0101410005)资助.

Supported by the National Natural Science Foundation of China (62121004, 61973091), the Local Innovative and Research Teams Project of Guangdong Special Support Program (2019BT02X353) and the Key Area Research and Development Program of Guangdong Province (2021B0101410005).

特性得到HJB方程解, 为非线性系统的最优控制问题的求解提供了新途径[7–11].

反步法作为一种系统化的控制设计方法, 具有结构简单, 鲁棒性强的优势, 能有效解决复杂非线性系统控制问题, 极大地促进了非线性系统控制理论的研究[12–27]. 综合反步法在解决n阶不确定非线性系统跟踪问题上的优势, 许多专家学者尝试将强化学习与反步法相结合并提出了相应的控制算法[28–32]. 其中, 文献[28–29]为实现跟踪控制, 将控制器设计分为两个部分: 基于反步法的前馈控制器和基于强化学习的最优控制器, 但该方法无法保证所有虚拟控制器为最优, 并且设计过程复杂. 文献[30]第1次提出最优反步法的概念, 在反步法的每一步应用强化学习中执行–评价结构, 利用执行网络和评价网络分别产生控制律和评价控制性能, 得到各子系统的最优虚拟控制器, 从而使整个系统控制性能达到最优. 文献[31–32]均采用最优反步法的控制设计思想, 分别解决了船舶系统和含有未知动态的非线性系统的跟踪控制问题. 然而, 上述文献所提出的最优控制算法都是在假设被控系统状态完全已知的前提下设计的, 但是, 实际被控系统的部分状态往往难以通过测量直接获取. 此外, 现有的基于最优反步法的文献都忽略了对虚拟控制器反复求导所引起的“复杂度爆炸”问题.

值得注意的是, 在实际工程中, 由于系统部件的物理局限性, 控制输入往往会受到约束, 如果所设计的控制算法没有充分考虑这一问题, 则无法获得理想的控制性能, 甚至无法使得被控系统稳定. 基于最优控制的优点, 研究人员试图对系统中存在输入约束的优化问题进行处理, 使系统在约束条件下达到控制目的, 同时优化相应的性能指标[33–37]. 文献[35]针对具有输入约束的矿物研磨过程控制问题, 通过引入参考调节器使得输入处于约束的范围内, 同时应用强化学习策略, 得到具有约束的最优调节器. 文献[36–37]通过设计一种非二次型的效用函数, 分别解决了具有输入约束的连续时间下系统模型自由的优化镇定问题和部分系统动态未知的事件触发优化问题. 但以上方法需要假设系统状态完全可测, 且未考虑高阶系统的跟踪问题.

为解决实际系统存在输入约束和不可测状态而导致难以实现最优控制的问题, 本文基于强化学习方法提出一种最优跟踪控制策略. 本文的主要贡献总结如下: 1) 针对高阶非线性系统, 本文提出基于强化学习执行–评价结构的最优反步法控制策略, 不仅能使系统输出快速跟踪理想信号, 还能优化性能指标, 降低控制资源的占用率. 与文献[28]所提出的最优方法相比, 本文所采用的设计方法更为简单, 并能使虚拟控制序列均为最优. 2) 与文献[30]相比, 本文所提出的最优跟踪控制方法不需要系统状态完全可测, 放宽系统状态可测这一约束条件, 更加符合实际工程情况. 3) 利用一种非二次型效用函数解决控制器约束问题. 同时引入动态面技术, 避免设计过程中“复杂度爆炸”的现象, 简化控制器设计过程.

2　预备知识与问题阐述

2.1　系统描述

考虑一类严格反馈非线性系统

$$\begin{cases} \dot{x}_i = x_{i+1} + f_i(\underline{x}_i),\ 1 \leqslant i \leqslant n-1, \\ \dot{x}_n = u + f_n(\underline{x}_n), \\ y = x_1, \end{cases} \tag{1}$$

其中: $\underline{x}_i = [x_1\ \ x_2\ \ \cdots\ \ x_i]^{\mathrm{T}} \in \mathbb{R}^i$为系统状态向量, $y \in \mathbb{R}$表示系统输出, $f_i(\underline{x}_i)$是未知光滑非线性函数, u表示受约束的系统输入, 约束范围为$\pm\beta$, β为已知正常数. 系统只有输入和输出可直接测量.

假设 1[24]　非线性函数$f(x)$在紧集Ω_{x}内具有利普希茨特性, 即存在一个已知的常数l_i, 对于$\forall X_1, X_2 \in \mathbb{R}^i$使得以下不等式成立:

$$|f_i(X_1) - f_i(X_2)| \leqslant l_i \|X_1 - X_2\|,\ i = 1, \cdots, n,$$

其中$\|X\|$表示向量X的2–范数.

引理 1[17]　根据神经网络的近似特性, 任何定义在紧集Ω_{x}上的连续非线性函数$f(x)$可以用神经网络逼近, 具体形式如下:

$$f(x) = W^{\mathrm{T}} S(x),$$

其中: $W = [W_1\ \ W_2\ \ \cdots\ \ W_p]^{\mathrm{T}} \in \mathbb{R}^p$表示权值向量, $S(x) = [s_1(x)\ \ \cdots\ \ s_p(x)]^{\mathrm{T}}$为基函数向量, p表示神经元个数,

$$s_i(x) = \exp[-\frac{(x-u_i)^{\mathrm{T}}(x-u_i)}{\phi_i^2}],$$

$x \in \Omega_{\mathrm{x}} \subset \mathbb{R}^p$是输入向量, $u_i = [u_{i1}\ \ \cdots\ \ u_{ip}]^{\mathrm{T}}$是高斯函数的中心, ϕ_i是高斯函数宽度. 基于神经网络逼近特性, 存在理想权值向量W^*, 使以下方程成立:

$$f(x) = W^{*\mathrm{T}} S(x) + \varepsilon(x),\ \forall x \in \Omega_{\mathrm{x}} \subset \mathbb{R}^p,$$
$$W^* := \arg\min_{W \in \mathbb{R}^p} \{ \sup_{x \in \Omega_{\mathrm{x}}} |f(x) - W^{\mathrm{T}} S(x)| \},$$

其中: $W^* = [W_1^*\ \ W_2^*\ \ \cdots\ \ W_p^*]^{\mathrm{T}} \in \mathbb{R}^p$, 逼近误差$\varepsilon(x)$, 满足$|\varepsilon(x)| \leqslant \delta^*$, δ^*为正常数.

2.2　观测器设计

基于神经网络近似特性, 可得

$$f_i(\hat{\underline{x}}_i) = W_{\mathrm{f}i}^{*\mathrm{T}} S_{\mathrm{f}i}(\hat{\underline{x}}_i) + \varepsilon_{\mathrm{f}i},$$

其中: $W_{\mathrm{f}i}^*$为理想权值, $\varepsilon_{\mathrm{f}i}$为逼近误差, $\hat{\underline{x}}_i$为$\underline{x}_i$的估计. 系统(1)可变形为

$$\begin{cases} \dot{x} = A_0 x + F^* + \varepsilon - \Delta f + E_n u, \\ y = x_1, \end{cases} \tag{2}$$

其中:

$$A_0=\begin{bmatrix}\mathbf{0}_{(n-1)\times 1} & I_{n-1}\\ 0 & \mathbf{0}_{1\times(n-1)}\end{bmatrix},$$

$x=[x_1\ \ x_2\ \ \cdots\ \ x_n]^{\mathrm{T}}$, $\varepsilon=[\varepsilon_{\mathrm{f}1}\ \ \varepsilon_{\mathrm{f}2}\ \ \cdots\ \ \varepsilon_{\mathrm{f}n}]^{\mathrm{T}}$, 且满足$||\varepsilon||\leqslant\delta_{\mathrm{f}}^*$, δ_{f}^*为正常数.

$$E_n=[0\ \cdots\ 0\ \ 1]^{\mathrm{T}}\in\mathbb{R}^n,$$

$$\Delta f=[\Delta f_1\ \ \Delta f_2\ \ \cdots\ \ \Delta f_n]^{\mathrm{T}},$$

$$\Delta f_i=f_i(\hat{\underline{x}}_i)-f_i(\underline{x}_i),$$

$$F^*=[W_{\mathrm{f}1}^{*\mathrm{T}}S_{\mathrm{f}1}\ \ W_{\mathrm{f}2}^{*\mathrm{T}}S_{\mathrm{f}2}\ \ \cdots\ \ W_{\mathrm{f}n}^{*\mathrm{T}}S_{\mathrm{f}n}]^{\mathrm{T}},$$

$$S_{\mathrm{f}i}=S_{\mathrm{f}i}(\hat{\underline{x}}_i).$$

由于系统(1)中状态不可测, 为了完成控制器设计, 本文将设计如下状态观测器来估计系统状态:

$$\begin{cases}\dot{\hat{x}}=A\hat{x}+Ky+\hat{F}+E_nu,\\ \hat{y}=\hat{x}_1,\end{cases}\tag{3}$$

其中:

$$A=\begin{bmatrix}-k_1 & I_{n-1}\\ -\underline{k}_i & \mathbf{0}_{1\times(n-1)}\end{bmatrix},\ \underline{k}_i=[k_2\ \ k_3\ \ \cdots\ \ k_n]^{\mathrm{T}},$$

$$\hat{F}=[\hat{W}_{\mathrm{f}1}^{\mathrm{T}}S_{\mathrm{f}1}\ \ \hat{W}_{\mathrm{f}2}^{\mathrm{T}}S_{\mathrm{f}2}\ \ \cdots\ \ \hat{W}_{\mathrm{f}n}^{\mathrm{T}}S_{\mathrm{f}n}]^{\mathrm{T}},$$

$\hat{W}_{\mathrm{f}i}$为$W_{\mathrm{f}i}^*$的估计值, $K=[k_1\ \ k_2\ \ \cdots\ \ k_n]^{\mathrm{T}}$, 系数$k_i$需使得多项式$p(s)=s^n+k_1s^{n-1}+\cdots+k_n$为赫尔维茨多项式. 因此, 对任意给定的$Q^{\mathrm{T}}=Q>0$, 存在正定矩阵$P^{\mathrm{T}}=P>0$, 使得$A^{\mathrm{T}}P+P^{\mathrm{T}}A=-Q$. 令$\tilde{x}=\hat{x}-x$, $\tilde{x}=[\tilde{x}_1\ \ \tilde{x}_2\ \ \cdots\ \ \tilde{x}_n]^{\mathrm{T}}$, 观测误差的状态空间形式为

$$\dot{\tilde{x}}=A\tilde{x}-\varepsilon+\Delta f+\tilde{F},\tag{4}$$

其中:

$$\tilde{F}=[\tilde{W}_{\mathrm{f}1}^{\mathrm{T}}S_{\mathrm{f}1}\ \ \tilde{W}_{\mathrm{f}2}^{\mathrm{T}}S_{\mathrm{f}2}\ \ \cdots\ \ \tilde{W}_{\mathrm{f}n}^{\mathrm{T}}S_{\mathrm{f}n}]^{\mathrm{T}},$$

$$\tilde{W}_{\mathrm{f}}=\hat{W}_{\mathrm{f}}-W_{\mathrm{f}}^*,\ \tilde{W}_{\mathrm{f}}=[\tilde{W}_{\mathrm{f}1}\ \ \tilde{W}_{\mathrm{f}2}\ \ \cdots\ \ \tilde{W}_{\mathrm{f}n}]^{\mathrm{T}},$$

$$W_{\mathrm{f}}^*=[W_{\mathrm{f}1}^*\ \ W_{\mathrm{f}2}^*\ \ \cdots\ \ W_{\mathrm{f}n}^*]^{\mathrm{T}}.$$

定理 1 利用神经网络观测器(3), 若针对神经网络权值$\hat{W}_{\mathrm{f}}$设计如下自适应律:

$$\dot{\hat{W}}_{\mathrm{f}}=-\eta\tilde{x}_1S_{\mathrm{f}}-\sigma\hat{W}_{\mathrm{f}},\tag{5}$$

其中: $S_{\mathrm{f}}=[S_{\mathrm{f}1}\ \ S_{\mathrm{f}2}\ \ \cdots\ \ S_{\mathrm{f}n}]^{\mathrm{T}}$, σ与η为正设计参数, 则观测误差$\tilde{x}$与权值误差$\tilde{W}_{\mathrm{f}}$最终一致有界.

证 选择如下李雅普诺夫函数:

$$L_0=\tilde{x}^{\mathrm{T}}P\tilde{x}+\frac{1}{2}\tilde{W}_{\mathrm{f}}^{\mathrm{T}}\tilde{W}_{\mathrm{f}},$$

其导数为

$$\dot{L}_0=-\tilde{x}^{\mathrm{T}}Q\tilde{x}+2\tilde{x}^{\mathrm{T}}P(-\varepsilon+\Delta f+\tilde{F})+\tilde{W}_{\mathrm{f}}^{\mathrm{T}}\dot{\tilde{W}}_{\mathrm{f}}.\tag{6}$$

根据式(5), 并应用杨氏不等式, 可得

$$\dot{L}_0\leqslant-a_{01}||\tilde{x}||^2-a_{02}||\tilde{W}_{\mathrm{f}}||^2+d_0,\tag{7}$$

其中:

$$a_{01}=\lambda_{\min}(Q)-r_0-2||P||^2-\frac{1}{2}\eta^2,$$

$$a_{02}=\frac{1}{2}\sigma-\frac{3}{2}\lambda_{\max}(S_{\mathrm{f}}S_{\mathrm{f}}^{\mathrm{T}}),$$

$$d_0=\delta_{\mathrm{f}}^{*2}+\frac{1}{2}\sigma||W_{\mathrm{f}}^{*\mathrm{T}}||^2,\ r_0=1+||P||^2\sum_{i=1}^{n}l_i^2,$$

$\lambda_{\max}(S_{\mathrm{f}}S_{\mathrm{f}}^{\mathrm{T}})$为$S_{\mathrm{f}}S_{\mathrm{f}}^{\mathrm{T}}$最大特征值.

由式(7)可知, 当以下不等式满足时, 观测误差$||\tilde{x}||$和权值误差$||\tilde{W}_{\mathrm{f}}||$最终一致有界

$$||\tilde{x}||>\sqrt{\frac{d_0}{a_{01}}},\tag{8}$$

$$||\tilde{W}_{\mathrm{f}}||>\sqrt{\frac{d_0}{a_{02}}}.\tag{9}$$

注 1 系统(1)中的控制器u是在集合Ω上的容许控制器, 即u在集合Ω上连续, $u(0)=0$, 并且u可以使得系统(1)在Ω上稳定. 定义为$u\in\Psi(\Omega)$. 本文所研究的最优问题是寻找一个满足约束条件的控制器$u\in\Psi(\Omega)$, 使得所设计的代价函数最小.

3 主要结果

3.1 最优控制器设计和稳定性分析

本节中, 基于强化学习的最优反步法技术将应用于一类n阶非线性系统最优跟踪控制问题研究. 不同于单一的反步法, 最优反步法设计的控制器不仅能保证系统稳定, 而且可以优化控制性能, 设计算法结构如图1所示.

定义如下坐标变换:

$$\begin{cases}z_1^*=x_1-y_{\mathrm{r}},\\ z_1=\hat{x}_1-y_{\mathrm{r}},\\ z_i=\hat{x}_i-q_i,\\ m_i=q_i-\hat{\alpha}_{i-1},\ i=2,3,\cdots,n,\end{cases}\tag{10}$$

其中: z_1^*表示系统跟踪误差, z_1表示观测器跟踪误差, z_i代表虚拟误差面, q_i代表一阶滤波器的输出信号, m_i为滤波误差, y_{r}为理想跟踪信号, $\hat{\alpha}_{i-1}$为近似最优虚拟控制器.

注 2 由式(8)–(9)可知观测器估计误差$\tilde{x}$最终一致有界, 因此, $\tilde{z}_1=z_1-z_1^*$同样最终一致有界. 在随后的设计中将用z_1取代z_1^*进行最优设计.

步骤 1 跟踪误差z_1对时间t的微分为

$$\dot{z}_1=\hat{x}_2+g_1-\dot{y}_{\mathrm{r}},\tag{11}$$

其中$g_1=\hat{W}_{\mathrm{f}1}^{\mathrm{T}}S_{\mathrm{f}1}+k_1(y-\hat{x}_1)$.

定义z_1子系统代价函数为

$$V_1(z_1)=\int_t^{\infty} r_1(z_1(s),\alpha_1(z_1))\mathrm{d}s,$$

其中: 效用函数为$r_1(z_1(s),\alpha_1(z_1))=z_1^2+\alpha_1^2$, α_1表示虚拟控制器. 将$\hat{x}_2$看作最优虚拟控制α_1^*, 即$\hat{x}_2\triangleq\alpha_1^*$, 则最优代价函数为

$$V_1^*(z_1)=\min_{\alpha_1\in\Psi(\Omega_{z1})}(\int_t^{\infty} r_1(z_1(s),\alpha_1(z_1))\mathrm{d}s)=\int_t^{\infty} r_1(z_1(s),\alpha_1^*(z_1))\mathrm{d}s,$$

其中: Ω_{z1}为包含原点的紧集. 梯度$\dfrac{\partial V_1^*}{\partial z_1}$可写为

$$\frac{\partial V_1^*}{\partial z_1}=2\beta_1 z_1+\frac{\partial V_1^{\mathrm{o}}}{\partial z_1}-2\dot{y}_{\mathrm{r}}+2g_1,\tag{12}$$

其中:

$$\frac{\partial V_1^{\mathrm{o}}}{\partial z_1}=-2\beta_1 z_1+\frac{\partial V_1^*}{\partial z_1}+2\dot{y}_{\mathrm{r}}-2g_1,$$

β_1为设计正常数, $\dfrac{\partial V_1^{\mathrm{o}}}{\partial z_1}$为连续函数. 根据贝尔曼最优性原理, z_1子系统的HJB方程为

$$H_1(z_1,\alpha_1^*,\frac{\partial V_1^*}{\partial z_1})=z_1^2+\alpha_1^{*2}+(g_1-\dot{y}_{\mathrm{r}}+\alpha_1^*)\times(2\beta_1 z_1-2\dot{y}_{\mathrm{r}}+2g_1+\frac{\partial V_1^{\mathrm{o}}}{\partial z_1})=0.\tag{13}$$

基于最优控制理论, 由$\dfrac{\partial H_1}{\partial\alpha_1^*}=0$得最优虚拟控制器$\alpha_1^*$为

$$\alpha_1^*=-\beta_1 z_1-\frac{1}{2}\frac{\partial V_1^{\mathrm{o}}}{\partial z_1}+\dot{y}_{\mathrm{r}}-g_1.\tag{14}$$

由于$\dfrac{\partial V_1^{\mathrm{o}}}{\partial z_1}$在集合$\Omega_{z_1}$上连续, 因此可以通过神经网络逼近, 由此可得以下等式:

$$\frac{\partial V_1^{\mathrm{o}}}{\partial z_1}=W_1^{*\mathrm{T}}S_1+\varepsilon_1,\tag{15}$$

$$\frac{\partial V_1^*}{\partial z_1}=2\beta_1 z_1+W_1^{*\mathrm{T}}S_1+\varepsilon_1-2\dot{y}_{\mathrm{r}}+2g_1,\tag{16}$$

$$\alpha_1^*=-\beta_1 z_1-\frac{1}{2}(W_1^{*\mathrm{T}}S_1+\varepsilon_1)+\dot{y}_{\mathrm{r}}-g_1,\tag{17}$$

其中: 逼近误差ε_1满足$|\varepsilon_1|\leqslant\delta_1^*$, δ_1^*为正常数. 令$S_1=S_1(z_1)$.

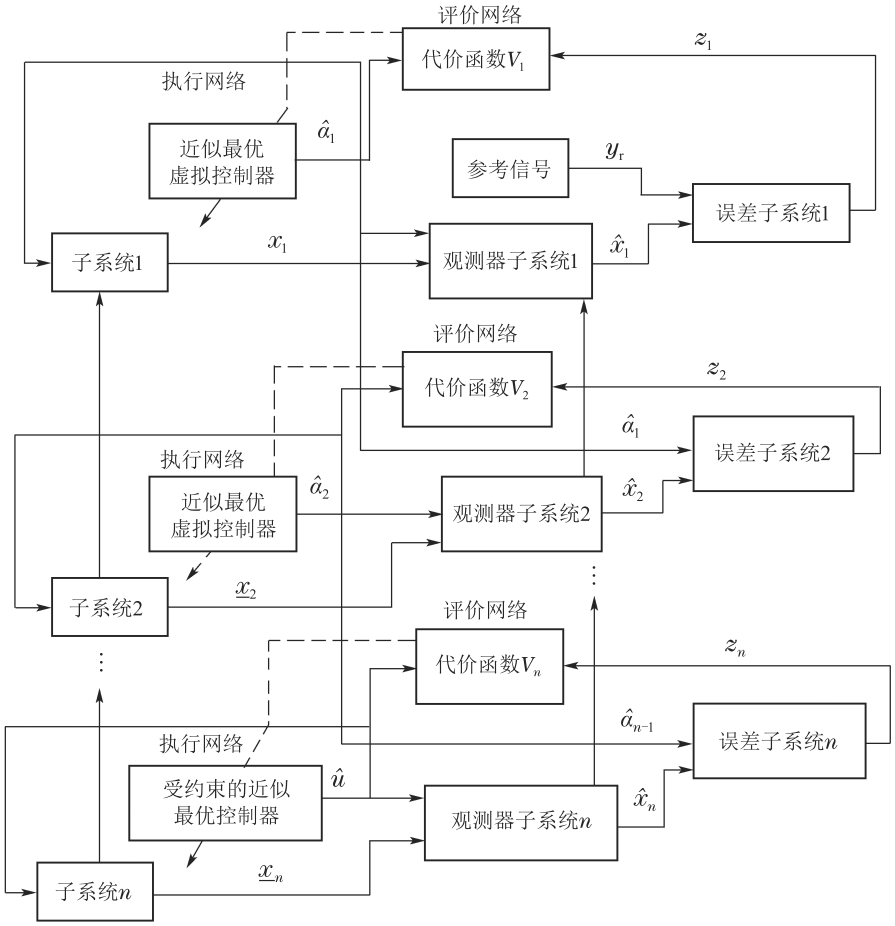

图 1 基于强化学习的最优控制设计框图

Fig. 1 Block diagram of the optimal control design based on reinforcement learning

将式(15)(17)代入式(13)得到

$$H_1(z_1,\alpha_1^*,W_1^*)=-(\beta_1^2-1)z_1^2-\beta_1 z_1 W_1^{*\mathrm{T}}S_1-2\dot{y}_{\mathrm{r}}g_1-\frac{1}{4}W_1^{*\mathrm{T}}S_1S_1^{\mathrm{T}}W_1^*+\dot{y}_{\mathrm{r}}^2+g_1^2+\epsilon_1, \tag{18}$$

其中: $\epsilon_1=\varepsilon_1(g_1-\dot{y}_{\mathrm{r}}+\alpha_1^*)+\frac{1}{4}\varepsilon_1^2$, 且满足$|\epsilon_1|\leqslant\epsilon_1^*$. 由于理想权值$W_1^*$未知, 因此最优虚拟控制器(17)不可用. 基于强化学习思想, 利用评价网络和执行网络, 得到以下式子:

$$\frac{\partial\hat{V}_1^{\mathrm{o}}}{\partial z_1}=\hat{W}_{\mathrm{c1}}^{\mathrm{T}}S_1, \tag{19}$$

$$\hat{\alpha}_1=-\beta_1 z_1-\frac{1}{2}\hat{W}_{\mathrm{a1}}^{\mathrm{T}}S_1+\dot{y}_{\mathrm{r}}-g_1, \tag{20}$$

其中: $\hat{V}_1^{\mathrm{o}}$和$\hat{\alpha}_1$分别为V_1^{o}和α_1^*的估计, $\hat{W}_{\mathrm{c1}}$和$\hat{W}_{\mathrm{a1}}$分别是评价和执行神经网络的权值. 将式(19)–(20)代入式(13), 得到近似HJB方程

$$H_1(z_1,\hat{\alpha}_1,\hat{W}_{\mathrm{c1}})=z_1^2+\hat{\alpha}_1^2+(2\beta_1 z_1+\hat{W}_{\mathrm{c1}}^{\mathrm{T}}S_1-2\dot{y}_{\mathrm{r}}+2g_1)(g_1-\dot{y}_{\mathrm{r}}+\hat{\alpha}_1). \tag{21}$$

基于式(13)(21), 贝尔曼余差为

$$\zeta_1=H_1(z_1,\hat{\alpha}_1,\hat{W}_{\mathrm{c1}})-H_1(z_1,\alpha_1^*,W_1^*)=H_1(z_1,\hat{\alpha}_1,\hat{W}_{\mathrm{c1}}).$$

定义目标函数, $E_1=\frac{1}{2}\zeta_1^2$. 为了使得贝尔曼余差最小, 针对目标函数采用梯度下降法设计如下评价网络自适应律:

$$\dot{\hat{W}}_{\mathrm{c1}}=-\frac{\gamma_{\mathrm{c1}}}{||w_1||^2+1}\frac{\partial E_1}{\partial\hat{W}_{\mathrm{c1}}}=-\frac{\gamma_{\mathrm{c1}}}{||w_1||^2+1}w_1(w_1^{\mathrm{T}}\hat{W}_{\mathrm{c1}}-(\beta_1^2-1)z_1^2+g_1^2+\frac{1}{4}\hat{W}_{\mathrm{a1}}^{\mathrm{T}}S_1S_1^{\mathrm{T}}\hat{W}_{\mathrm{a1}}+\dot{y}_{\mathrm{r}}^2-2\dot{y}_{\mathrm{r}}g_1), \tag{22}$$

其中: $w_1=S_1(g_1+\hat{\alpha}_1-\dot{y}_{\mathrm{r}})$, 学习率$\gamma_{\mathrm{c1}}>0$. 相应的执行网络自适应律为

$$\dot{\hat{W}}_{\mathrm{a1}}=\frac{\gamma_{\mathrm{c1}}}{4(||w_1||^2+1)}S_1S_1^{\mathrm{T}}\hat{W}_{\mathrm{a1}}w_1^{\mathrm{T}}\hat{W}_{\mathrm{c1}}+\frac{1}{2}S_1z_1-\gamma_{\mathrm{a1}}S_1S_1^{\mathrm{T}}\hat{W}_{\mathrm{a1}}, \tag{23}$$

其中学习率$\gamma_{\mathrm{a1}}>\gamma_{\mathrm{c1}}^2$.

假设 2[6] 信号$w_iw_i^{\mathrm{T}}$, $i=1,2,\cdots,n$, 在区间$[t,t+\bar{t}_i]$上满足以下持续激励条件:

$$\varrho_iI_{p_i}\leqslant w_iw_i^{\mathrm{T}}\leqslant\xi_iI_{p_i}, \tag{24}$$

其中: $\varrho_i>0,\xi_i>0,\bar{t}_i>0$, 且$I_{p_i}\in\mathbb{R}^{p_i\times p_i}$为单位矩阵.

注 3 假设2中, 矩阵$w_iw_i^{\mathrm{T}}$每一元素都应满足小于等于关系, 即, 每一元素都在区间$[\varrho_i,\xi_i]$内. 该假设可以使得评价网络的估计权值$\hat{W}_{\mathrm{c}i}$更好地收敛于理想权值W_i^*.

由式(10), 误差动态式(11)可改写为

$$\dot{z}_1=z_2+\hat{\alpha}_1+g_1-\dot{y}_{\mathrm{r}}+m_2. \tag{25}$$

对z_1子系统选择如下李雅普诺夫函数:

$$L_1=\frac{1}{2}z_1^2+\frac{1}{2}\tilde{W}_{\mathrm{a1}}^{\mathrm{T}}\tilde{W}_{\mathrm{a1}}+\frac{1}{2}\tilde{W}_{\mathrm{c1}}^{\mathrm{T}}\tilde{W}_{\mathrm{c1}},$$

其中: $\tilde{W}_{\mathrm{a1}}=\hat{W}_{\mathrm{a1}}-W_1^*$, $\tilde{W}_{\mathrm{c1}}=\hat{W}_{\mathrm{c1}}-W_1^*$.

将式(20)(22)–(23)(25)代入L_1对时间t的微分得

$$\dot{L}_1=z_1z_2-\beta_1z_1^2-\frac{1}{2}\hat{W}_{\mathrm{a1}}^{\mathrm{T}}S_1z_1+\frac{1}{2}\tilde{W}_{\mathrm{a1}}^{\mathrm{T}}S_1z_1-\gamma_{\mathrm{a1}}\tilde{W}_{\mathrm{a1}}^{\mathrm{T}}S_1S_1^{\mathrm{T}}\hat{W}_{\mathrm{a1}}+z_1m_2+\frac{\gamma_{\mathrm{c1}}}{4(||w_1||^2+1)}\times\tilde{W}_{\mathrm{a1}}^{\mathrm{T}}S_1S_1^{\mathrm{T}}\hat{W}_{\mathrm{a1}}w_1^{\mathrm{T}}\hat{W}_{\mathrm{c1}}-\frac{\gamma_{\mathrm{c1}}}{||w_1||^2+1}\tilde{W}_{\mathrm{c1}}^{\mathrm{T}}w_1\times(w_1^{\mathrm{T}}\hat{W}_{\mathrm{c1}}-(\beta_1^2-1)z_1^2+\frac{1}{4}\hat{W}_{\mathrm{a1}}^{\mathrm{T}}S_1S_1^{\mathrm{T}}\hat{W}_{\mathrm{a1}}+g_1^2-2\dot{y}_{\mathrm{r}}g_1+\dot{y}_{\mathrm{r}}^2). \tag{26}$$

应用杨氏不等式和假设2, 并结合式(18)(22), 式(26)可写为

$$\dot{L}_1\leqslant\frac{1}{2}z_2^2-a_{11}z_1^2-a_{12}\tilde{W}_{\mathrm{a1}}^{\mathrm{T}}\tilde{W}_{\mathrm{a1}}-a_{13}\tilde{W}_{\mathrm{c1}}^{\mathrm{T}}\tilde{W}_{\mathrm{c1}}+\bar{d}_1+\frac{1}{2}m_2^2, \tag{27}$$

其中:

$$d_1=\frac{2\gamma_{\mathrm{a1}}+1}{4}(W_1^{*\mathrm{T}}S_1)^2+\frac{\gamma_{\mathrm{c1}}}{2(||w_1||^2+1)}\epsilon_1^2.$$

由于d_1中均为有界量, 因此存在一个常数$\bar{d}_1$使得$d_1\leqslant\bar{d}_1$. $a_{11}=\beta_1-\frac{5}{4}$,

$$a_{12}=(\frac{1}{2}\gamma_{\mathrm{a1}}-\frac{1}{8}\gamma_{\mathrm{c1}}^2-\frac{\xi_1}{8}W_1^{*\mathrm{T}}W_1^*)\times\lambda_{\min}(S_1S_1^{\mathrm{T}}),$$

$$a_{13}=\frac{\varrho_1}{2(||w_1||^2+1)}\gamma_{\mathrm{c1}}-\frac{\xi_1}{8(||w_1||^2+1)^2}\times W_1^{*\mathrm{T}}S_1S_1^{\mathrm{T}}W_1^*.$$

步骤 2($i=2,3,\cdots,n-1$) 基于动态面技术[25], 选取如下一阶滤波器用于避免反步法中“复杂度爆炸”问题

$$\tau_i\dot{q}_i+q_i=\hat{\alpha}_{i-1},\ q_i(0)=\hat{\alpha}_{i-1}(0), \tag{28}$$

其中: τ_i是正设计参数, 由$m_i=q_i-\hat{\alpha}_{i-1}$可得$\dot{q}_i=-\frac{m_i}{\tau_i}$, 则有

$$\dot{m}_i=-\frac{m_i}{\tau_i}+M_i, \tag{29}$$

其中$M_i=-\dot{\hat{\alpha}}_{i-1}$为有界连续函数.

注 4 每一步设计出相对应的$\hat{\alpha}_i$后, 在下一步的最优

控制的设计中需对其进行求导, 并且之后的每一步都会对其进行反复求导, 从而产生“复杂度爆炸”问题. 引入动态面技术将虚拟控制信号通过一阶低通滤波器得到估计值q_i, 在下一步设计过程中, 用估计值代替虚拟控制信号可以避免对其进行求导, 简化了控制器设计.

第i个子系统的误差动态方程为

$$\dot{z}_i = \hat{x}_{i+1} + g_i - \dot{q}_i, \tag{30}$$

其中$g_i = \hat{W}_{\mathrm{f}i}^{\mathrm{T}} S_{\mathrm{f}i} + k_i(y - \hat{x}_1)$.

令第i个子系统的最优虚拟控制器为$\hat{x}_{i+1} = \alpha_i^*$, 可得最优代价函数

$$V_i^*(z_i) = \min_{\alpha_i \in \Psi(\Omega_{\mathrm{z}i})} (\int_t^\infty r_i(z_i(s), \alpha_i(z_i))\mathrm{d}s) = \int_t^\infty r_i(z_i(s), \alpha_i^*(z_i))\mathrm{d}s,$$

其中: 效用函数$r_i(z_i(s), \alpha_i^*(z_i)) = z_i^2 + \alpha_i^{*2}$, α_i^{*2}表示最优虚拟控制器, $\Omega_{\mathrm{z}i}$表示包含原点的紧集.

与第1步类似, 梯度$\dfrac{\partial V_i^*}{\partial z_i}$可写为

$$\frac{\partial V_i^*}{\partial z_i} = 2\beta_i z_i + \frac{\partial V_i^{\mathrm{o}}}{\partial z_i} - 2\dot{q}_i + 2g_i,$$

β_i为正设计参数, 对应的HJB方程为

$$H_i(z_i, \alpha_i^*, \frac{\partial V_i^*}{\partial z_i}) = z_i^2 + \alpha_i^{*2} + (g_i - \dot{q}_i + \alpha_i^*) \times (2\beta_i z_i - 2\dot{q}_i + 2g_i + \frac{\partial V_i^{\mathrm{o}}}{\partial z_i}) = 0. \tag{31}$$

根据最优控制原理, 最优虚拟控制器为

$$\alpha_i^* = -\beta_i z_i - \frac{1}{2}\frac{\partial V_i^{\mathrm{o}}}{\partial z_i} + \dot{q}_i - g_i. \tag{32}$$

利用神经网络逼近性质

$$\frac{\partial V_i^{\mathrm{o}}}{\partial z_i} = W_i^{*\mathrm{T}} S_i + \varepsilon_i, \tag{33}$$

$$\alpha_i^* = -\beta_i z_i - \frac{1}{2}(W_i^{*\mathrm{T}} S_i + \varepsilon_i) + \dot{q}_i - g_i, \tag{34}$$

其中逼近误差ε_i满足$|\varepsilon_i| \leqslant \delta_i^*$. 令$S_i = S_i(z_i)$.

由式(31)(33)–(34)可知以下式子成立:

$$H_i(z_i, \alpha_i^*, W_i^*) = -(\beta_i^2 - 1)z_i^2 - \beta_i z_i W_i^{*\mathrm{T}} S_i - 2\dot{q}_i g_i + \dot{q}_i^2 - \frac{1}{4} W_i^{*\mathrm{T}} S_i S_i^{\mathrm{T}} W_i^* + \epsilon_i + g_i^2 - \varepsilon_i \dot{q}_i, \tag{35}$$

其中$\epsilon_i = \varepsilon_i(g_i + \alpha_i^*) + \dfrac{1}{4}\varepsilon_i^2$.

与第1步类似, 式(33)–(34)可写为

$$\frac{\partial \hat{V}_i^{\mathrm{o}}}{\partial z_i} = \hat{W}_{\mathrm{c}i}^{\mathrm{T}} S_i, \tag{36}$$

$$\hat{\alpha}_i = -\beta_i z_i - \frac{1}{2}\hat{W}_{\mathrm{a}i}^{\mathrm{T}} S_i + \dot{q}_i - g_i. \tag{37}$$

第i阶子系统的贝尔曼余差为

$$\zeta_i = H_i(z_i, \hat{\alpha}_i, \hat{W}_{\mathrm{c}i}) - H_i(z_i, \alpha_i^*, W_i^*) = H_i(z_i, \hat{\alpha}_i, \hat{W}_{\mathrm{c}i}).$$

令$E_i = \dfrac{1}{2}\zeta_i^2$, 根据梯度下降法, 设计评价网络的自适应律为

$$\dot{\hat{W}}_{\mathrm{c}i} = -\frac{\gamma_{\mathrm{c}i}}{||w_i||^2 + 1}\frac{\partial E_i}{\partial \hat{W}_{\mathrm{c}i}} = -\frac{\gamma_{\mathrm{c}i}}{||w_i||^2 + 1} w_i (w_i^{\mathrm{T}} \hat{W}_{\mathrm{c}i} - (\beta_i^2 - 1)z_i^2 + g_i^2 + \dot{q}_i^2 + \frac{1}{4}\hat{W}_{\mathrm{a}i}^{\mathrm{T}} S_i S_i^{\mathrm{T}} \hat{W}_{\mathrm{a}i} - 2\dot{q}_i g_i), \tag{38}$$

其中: 学习率$\gamma_{\mathrm{c}i} > 0$, $w_i = S_i(-\beta_i z_i - \dfrac{1}{2}\hat{W}_{\mathrm{a}i}^{\mathrm{T}} S_i)$. 相应执行网络自适应律为

$$\dot{\hat{W}}_{\mathrm{a}i} = \frac{1}{2} S_1 z_i - \gamma_{\mathrm{a}i} S_i S_i^{\mathrm{T}} \hat{W}_{\mathrm{a}i} + \frac{\gamma_{\mathrm{c}i}}{4(||w_i||^2 + 1)} \times S_i S_i^{\mathrm{T}} \hat{W}_{\mathrm{a}i} w_i^{\mathrm{T}} \hat{W}_{\mathrm{c}i}, \tag{39}$$

其中学习率$\gamma_{\mathrm{a}i} > \gamma_{\mathrm{c}i}^2$.

由式(10)可知第i阶系统误差动态为

$$\dot{z}_i = z_{i+1} + m_{i+1} + \hat{\alpha}_i + g_i - \dot{q}_i. \tag{40}$$

针对第i阶子系统选取如下李雅普诺夫函数:

$$L_i = \sum_{k=1}^{i-1} L_k + \frac{1}{2}z_i^2 + \frac{1}{2}\tilde{W}_{\mathrm{a}i}^{\mathrm{T}}\tilde{W}_{\mathrm{a}i} + \frac{1}{2}\tilde{W}_{\mathrm{c}i}^{\mathrm{T}}\tilde{W}_{\mathrm{c}i} + \frac{1}{2}m_i^2.$$

与第1步化简相似, L_i对时间t求微分后可化简为

$$\dot{L}_i \leqslant \sum_{k=1}^{i-1} \dot{L}_k + \frac{1}{2}z_{i+1}^2 - a_{i1}z_i^2 - a_{i2}\tilde{W}_{\mathrm{a}i}^{\mathrm{T}}\tilde{W}_{\mathrm{a}i} - a_{i3}\tilde{W}_{\mathrm{c}i}^{\mathrm{T}}\tilde{W}_{\mathrm{c}i} + \bar{d}_i + \frac{1}{2}m_{i+1}^2 + m_i\dot{m}_i + \frac{\gamma_{\mathrm{c}i}^2\delta_i^{*2}}{2c_i}m_i^2, \tag{41}$$

其中:

$$d_i = \frac{2\gamma_{\mathrm{a}i} + 1}{4}(W_i^{*\mathrm{T}} S_i)^2 + \frac{\gamma_{\mathrm{c}i}}{2(||w_i||^2 + 1)}\epsilon_i^2,$$

由于d_i中均为有界量, 因此存在一个常数$\bar{d}_i$使得$d_i \leqslant \bar{d}_i$. $a_{i1} = \beta_i - \dfrac{5}{4}$,

$$a_{i2} = \frac{\gamma_{\mathrm{a}i}}{2}\lambda_{\min}(S_i S_i^{\mathrm{T}}) - \frac{\gamma_{\mathrm{c}i}}{8(||w_i||^2 + 1)} \times \lambda_{\max}(S_i S_i^{\mathrm{T}} W_i^* W_i^{*\mathrm{T}} S_i S_i^{\mathrm{T}}) - \frac{\gamma_{\mathrm{c}i}}{8(||w_i||^2 + 1)^2}\lambda_{\max}(S_i W_i^{*\mathrm{T}} w_i w_i^{\mathrm{T}} W_i^* S_i^{\mathrm{T}}),$$

$$a_{i3} = \lambda_{\min}(w_i w_i^{\mathrm{T}})(\frac{3\gamma_{\mathrm{c}i}}{8(||w_i||^2 + 1)} - \frac{c_i}{2\tau_i^2(||w_i||^2 + 1)^2}),$$

c_i为正参数.

步骤 3 根据式(10), 第n阶系统误差动态方程为

$$\dot{z}_n = g_n + u - \dot{q}_n, \tag{42}$$

其中$g_n = \hat{W}_{\rm fn}^{\rm T} S_{\rm fn} + k_n(y - \hat{x}_1)$.

z_n子系统最优代价函数为

$$V_n^*(z_n) = \min_{u\in\Psi(\Omega_{zn})}(\int_t^\infty r_n(z_n(s), u(z_n))\mathrm{d}s) = \int_t^\infty r_n(z_n(s), u^*(z_n))\mathrm{d}s,$$

其中: $r_n(z_n(s), u^*(z_n)) = \varsigma z_n^2 + R(u^*)$, u^*表示最优实际控制器, ς为正数, Ω_{zn}表示包含原点的紧集.

$$R(u) = 2\int_0^u \beta\psi^{-1}(\frac{s}{\beta})\mathrm{d}s,$$

β为正常数. $\psi(\cdot)$是一个有界且一阶导数为正常数的单调奇函数, 满足$|\psi(\cdot)| \leqslant 1$. 当$\psi(\cdot)$是单调奇函数时$R(u)$是正定的. 不失一般性, 本文选取$\psi(\cdot) = \tanh(\cdot)$.

由最优代价函数$V_n^* = \beta_n z_n^2 + V_n^{\rm o}$, β_n为正设计参数, 得第n个子系统HJB方程为

$$H_n(z_n, u^*, \frac{\partial V_n^{\rm o}}{\partial z_n}) = \varsigma z_n^2 + R(u^*) + 2\beta_n z_n(g_n - \dot{q}_n + u^*) + \frac{\partial V_n^{\rm o}}{\partial z_n}(g_n - \dot{q}_n + u^*) = 0. \tag{43}$$

根据最优性原理, 可得

$$u^* = -\beta\tanh(\frac{1}{\beta}\beta_n z_n + \frac{1}{2\beta}\frac{\partial V_n^{\rm o}}{\partial z_n}). \tag{44}$$

利用神经网络逼近性质

$$\frac{\partial V_n^{\rm o}}{\partial z_n} = W_n^{*\rm T} S_n + \varepsilon_n, \tag{45}$$

其中: 逼近误差ε_n满足$|\varepsilon_n| \leqslant \delta_n^*$. 令$S_n = S_n(z_n)$.

与第i步相似, 式(44)–(45)可写为

$$\hat{u} = -\beta\tanh(\frac{1}{\beta}\beta_n z_n + \frac{1}{2\beta}\hat{W}_{\rm an}^{\rm T} S_n), \tag{46}$$

$$\frac{\partial \hat{V}_n^{\rm o}}{\partial z_n} = \hat{W}_{\rm cn}^{\rm T} S_n. \tag{47}$$

令

$$N_1 = \frac{1}{\beta}\beta_n z_n + \frac{1}{2\beta}\hat{W}_{\rm an}^{\rm T} S_n,$$

$$N_2 = \frac{1}{\beta}\beta_n z_n + \frac{1}{2\beta}W_n^{*\rm T} S_n.$$

由式(43)(46)–(47), 可得第n阶子系统的贝尔曼余差为

$$\zeta_n = H_n(z_n, \hat{u}, \hat{W}_{\rm cn}) - H_n(z_n, u^*, W_n^*) = \varsigma z_n^2 + R(\hat{u}) + (2\beta_n z_n + \hat{W}_{\rm cn}^{\rm T} S_n)\times(g_n + \hat{u} - \dot{q}_n).$$

令$E_n = \frac{1}{2}\zeta_n^2$. 采用梯度下降法, 得到评价网络自适应律

$$\dot{\hat{W}}_{\rm cn} = -\frac{\gamma_{cn}}{1 + ||w_n||^2} w_n(2\beta_n z_n(g_n + \hat{u} - \dot{q}_n) + \hat{W}_{\rm cn}^{\rm T} w_n + \varsigma z_n^2 + R(\hat{u})), \tag{48}$$

其中: $w_n = S_n(g_n + \hat{u} - \dot{q}_n)$, 学习率$\gamma_{cn} > 0$. 设计执行网络自适应律如下:

$$\dot{\hat{W}}_{\rm an} = -\gamma_{\rm an} S_n S_n^{\rm T}\hat{W}_{\rm an} + \frac{\beta\gamma_{\rm cn}}{1 + ||w_n||^2} S_n \hat{W}_{\rm cn}^{\rm T}\hat{w}_n \times (\tanh\ N_1 - \mathrm{sgn}(N_1)), \tag{49}$$

其中$\gamma_{\rm an} > 0$.

选取如下李雅普诺夫函数:

$$L_n = L_{n1} + L_{n2},$$

其中:

$$L_{n1} = V_n^*,\ L_{n2} = \frac{1}{2}\tilde{W}_{\rm an}^{\rm T}\tilde{W}_{\rm an} + \frac{1}{2}\tilde{W}_{\rm cn}^{\rm T}\tilde{W}_{\rm cn} + \frac{1}{2}m_n^2.$$

对L_{n1}微分可得

$$\dot{L}_{n1} = (2\beta_n z_n + W_n^{*\rm T} S_n)(g_n - \dot{q}_n) + \varepsilon_n(g_n + \hat{u}) - \beta(2\beta_n z_n + W_n^{*\rm T} S_n)\tanh\ N_1 - \varepsilon_n\dot{q}_n. \tag{50}$$

由式(43)(45)可得

$$\varepsilon_{\rm HJB} + \varepsilon_n\dot{q}_n = (2\beta_n z_n + W_n^{*\rm T} S_n)(g_n - \dot{q}_n) - \beta\times(2\beta_n z_n + W_n^{*\rm T} S_n) + \varsigma z_n^2 + 2\beta\times\int_0^{-\beta\tanh\ N_2}\tanh^{-1}(\frac{s}{\beta})\mathrm{d}s, \tag{51}$$

其中: $\varepsilon_{\rm HJB} + \varepsilon_n\dot{q}_n$为神经网络余差, 且$\varepsilon_{\rm HJB}$为有界量, 满足$\varepsilon_{\rm HJB} \leqslant \epsilon_{\rm HJB}$, $\epsilon_{\rm HJB}$为正常数.

将式(51)代入式(50), 且$\tanh\ N_i \leqslant \sqrt{h}$, 通过杨氏不等式化简可得

$$\dot{L}_{n1} \leqslant -(\varsigma - \beta_n^2)z_n^2 + d_{n1},$$

其中:

$$d_{n1} = \beta W_n^{*\rm T} S_n(\tanh\ N_2 - \tanh\ N_1) + 4\beta^2 h + \epsilon_{\rm HJB} + \varepsilon_n(g_n + \hat{u}),$$

h为正常数.

由杨氏不等式可得

$$m_i M_i \leqslant \frac{m_i^2 M_i^2}{2} + \frac{1}{2},$$

并且存在一个正常数$\bar{M}$满足$|M_i| \leqslant \bar{M}$.

令

$$\Gamma(N_1) = 2\beta\int_0^{-\beta\tanh\ N_1}\tanh^{-1}(\frac{s}{\beta})\mathrm{d}s,$$

$$\Gamma(N_2)=2\beta\int_0^{-\beta\tanh N_2}\tanh^{-1}(\frac{s}{\beta})\mathrm{d}s.$$

由文献[36]推导可知

$$\begin{aligned}\Gamma(N_1)-\Gamma(N_2)=&\Theta+2\beta^2(N_1\tanh N_1-N_2\times\\&\tanh N_2)+2\beta^2(N_2\mathrm{sgn}(N_2)-\\&N_1\mathrm{sgn}(N_1)),\end{aligned}\tag{52}$$

其中:

$$\Theta=2\ln\frac{1+\exp(-2N_2\mathrm{sgn}(N_2))}{1+\exp(-2N_1\mathrm{sgn}(N_1))},$$

Θ为有界量.

对L_n求微分,并将式(48)–(49)代入, 化简得

$$\begin{aligned}\dot{L}_n\leqslant&-a_{n1}z_n^2-a_{n2}\tilde{W}_{\mathrm{a}n}^{\mathrm{T}}\tilde{W}_{\mathrm{a}n}-a_{n3}\tilde{W}_{\mathrm{c}n}^{\mathrm{T}}\tilde{W}_{\mathrm{c}n}+\\&\frac{1}{2}-(\frac{1}{\tau_n}-\frac{\gamma_{\mathrm{c}n}^2\delta_n^{*2}}{2c_n}-\frac{\bar{M}^2}{2})m_n^2+d_n,\end{aligned}\tag{53}$$

其中:

$$\begin{aligned}&a_{n1}=\varsigma-\beta_n^2-\frac{8\gamma_{\mathrm{c}n}}{1+||w_n||^2}\beta^2\beta_n^2,\\&a_{n2}=(\frac{\gamma_{\mathrm{a}n}}{2}-\frac{1}{2})\lambda_{\min}(S_nS_n^{\mathrm{T}}),\\&a_{n3}=\frac{15\varrho_n\gamma_{\mathrm{c}n}}{32(1+||w_n||^2)}-\frac{c_n\xi_n}{2\tau_n^2(1+||w_n||^2)^2},\end{aligned}$$

c_n为正参数, 且

$$\begin{aligned}d_n=&\frac{8\gamma_{\mathrm{c}n}}{1+||w_n||^2}(\epsilon_{\mathrm{HJB}}+\Theta+\beta W_n^{*\mathrm{T}}S_n(\mathrm{sgn}(N_2)-\\&\mathrm{sgn}(N_1)))^2+\frac{\gamma_{\mathrm{a}n}}{2}W_n^{*\mathrm{T}}S_nS_n^{\mathrm{T}}W_n^*+d_{n1}+\frac{1}{2}\times\\&(\frac{\beta\gamma_{\mathrm{c}n}W_n^{*\mathrm{T}}w_n}{1+||w_n||^2}(\tanh N_1-\mathrm{sgn}(N_1)))^2.\end{aligned}$$

由上式可知d_n中均为有界量, 因此存在一个常数$\bar{d}_n$使得$d_n\leqslant\bar{d}_n$.

选取李雅普诺夫函数

$$L=L_{n-1}+L_n+L_0.\tag{54}$$

由式(7)(27)(41)(53)–(54)可知

$$\begin{aligned}\dot{L}\leqslant&-a_1||z||^2-a_2||\tilde{W}_{\mathrm{a}}||^2-a_3||\tilde{W}_{\mathrm{c}}||^2+\bar{d}-\\&a_{01}||\tilde{x}||^2-a_{02}||\tilde{W}_{\mathrm{f}}||^2-a_4||m||^2,\end{aligned}\tag{55}$$

其中:

$$\begin{aligned}&a_1=\min\{a_{11},a_{i1}-\frac{1}{2}\},\ i=2,3,\cdots,n,\\&a_2=\min\{a_{i2}\},a_3=\min\{a_{i3}\},\ i=1,2,\cdots,n,\\&a_4=\min\{\frac{1}{\tau_i}-\frac{\gamma_{\mathrm{c}i}^2\delta_i^{*2}}{2c_i}-\frac{\bar{M}^2+1}{2}\},\\&i=2,3,\cdots,n,\\&\bar{d}=d_0+\bar{d}_1+\bar{d}_2+\cdots+\bar{d}_n+\frac{n-1}{2},\end{aligned}$$

$$m=[m_2\ \ m_3\ \ \cdots\ \ m_n]^{\mathrm{T}}.$$

由此可得

$$\begin{cases}||z||>\sqrt{\dfrac{\bar{d}}{a_1}},\ ||\tilde{W}_{\mathrm{a}}||>\sqrt{\dfrac{\bar{d}}{a_2}},\\||\tilde{W}_{\mathrm{c}}||>\sqrt{\dfrac{\bar{d}}{a_3}},\ ||\tilde{x}||>\sqrt{\dfrac{\bar{d}}{a_{01}}},\\||\tilde{W}_{\mathrm{f}}||>\sqrt{\dfrac{\bar{d}}{a_{02}}},\ ||m||>\sqrt{\dfrac{\bar{d}}{a_4}}.\end{cases}\tag{56}$$

选取适当参数, 使得$a_1,a_2,a_3,a_4,a_{01},a_{02}$为正, 且当上述不等式满足时, 系统所有信号为一致最终有界.

3.2　数值仿真

本节将通过数值仿真验证所提出的基于强化学习最优控制方法的有效性. 考虑如下系统:

$$\begin{cases}\dot{x}_1=\sin(2x_1)+x_2,\\\dot{x}_2=2\sin\ x_1+\cos\ x_2-1+u,\\y=x_1.\end{cases}$$

选取如下设计参数:

$$\begin{aligned}&k_1=60,\ k_2=60,\ \eta=1.6,\ \sigma=5,\ \beta=5,\\&\varsigma=1300,\ \beta_1=11,\ \beta_2=20,\ \gamma_{\mathrm{a}1}=5,\ \gamma_{\mathrm{c}1}=0.5,\\&\gamma_{\mathrm{a}2}=5,\ \gamma_{\mathrm{c}2}=0.015.\end{aligned}$$

系统初始状态为

$$x_1(0)=0.1,\ x_2(0)=1,\ \hat{x}_1(0)=0.15,\ \hat{x}_2(0)=0.1.$$

神经网络初始值选择为

$$\begin{aligned}\hat{W}_{\mathrm{f}1}(0)&=[0.1\ \ \cdots\ \ 0.1]^{\mathrm{T}}\in\mathbb{R}^{5\times1},\\\hat{W}_{\mathrm{f}2}(0)&=[0.2\ \ \cdots\ \ 0.2]^{\mathrm{T}}\in\mathbb{R}^{5\times1},\\\hat{W}_{\mathrm{a}1}(0)&=[1.2\ \ \cdots\ \ 1.2]^{\mathrm{T}}\in\mathbb{R}^{6\times1},\\\hat{W}_{\mathrm{c}1}(0)&=[0.4\ \ \cdots\ \ 0.4]^{\mathrm{T}}\in\mathbb{R}^{6\times1},\\\hat{W}_{\mathrm{a}2}(0)&=[1.3\ \ \cdots\ \ 1.3]^{\mathrm{T}}\in\mathbb{R}^{6\times1},\\\hat{W}_{\mathrm{c}2}(0)&=[0.6\ \ \cdots\ \ 0.6]^{\mathrm{T}}\in\mathbb{R}^{6\times1}.\end{aligned}$$

在学习过程的前5 s, 给控制端施加一个噪声信号

$$n(t)=0.5\sin(2t)\cos(2t)+0.2\sin(20t)\cos(7t)^3$$

改善学习效果, 理想跟踪信号$y_{\mathrm{r}}=\sin\ t$, 图2–7为仿真结果.

图2表示参考信号y_{r}和系统的输出信号y的响应曲线. 图3表示观测器输出跟踪理想轨迹的误差z_1以及观测误差$\tilde{x}_1$, 可以看出观测器的跟踪误差z_1以及观测误差$\tilde{x}_1$收敛到0附近, 因此在优化设计中$\hat{x}_1$可以替代x_1, z_1可以替代z_1^*. 图4表示系统状态x_1, x_2

和估计值$\hat{x}_1, \hat{x}_2$的运行轨迹, 可以看出所设计的观测器有很好的估计效果. 图5表示自适应参数$||\hat{W}_{a1}||$, $||\hat{W}_{a2}||$的曲线. 图6表示自适应参数$||\hat{W}_{c1}||$, $||\hat{W}_{c2}||$的曲线. 图7表示系统的输入轨迹.

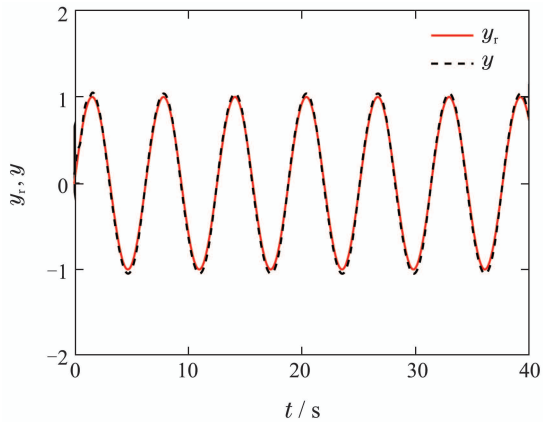

图 2 系统输出y和参考信号y_r

Fig. 2 System output y and reference signal y_r

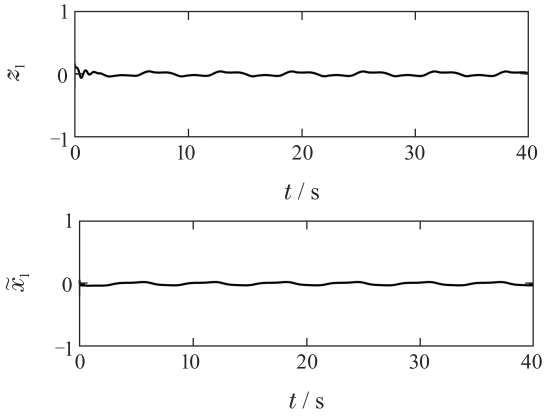

图 3 观测器跟踪误差z_1和观测误差$\tilde{x}_1$

Fig. 3 The observer tracking error z_1 and the observer error $\tilde{x}_1$

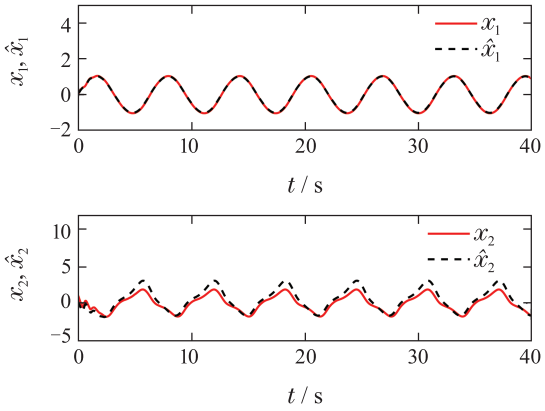

图 4 状态$x_1, \hat{x}_1$和状态$x_2, \hat{x}_2$

Fig. 4 States $x_1, \hat{x}_1$ and $x_2, \hat{x}_2$

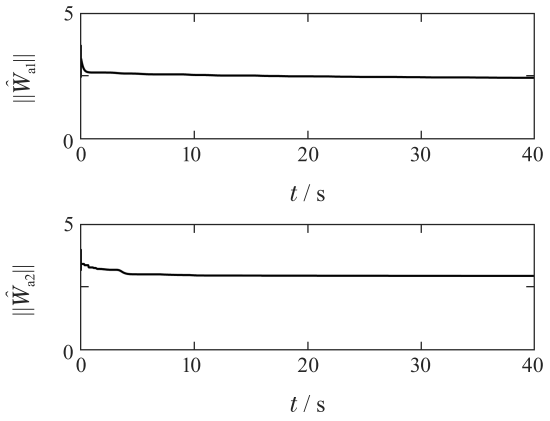

图 5 自适应律$||\hat{W}_{a1}||$和$||\hat{W}_{a2}||$

Fig. 5 Adaptive laws $||\hat{W}_{a1}||$ and $||\hat{W}_{a2}||$

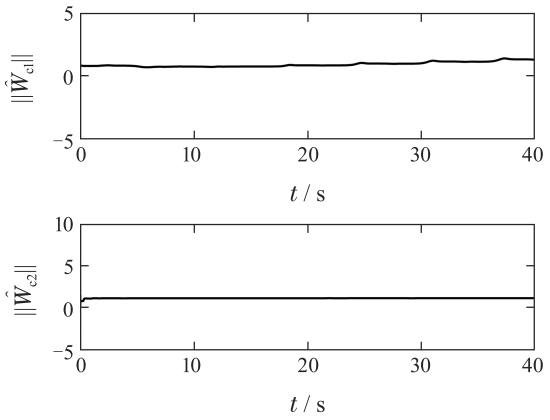

图 6 自适应律$||\hat{W}_{c1}||$和$||\hat{W}_{c2}||$

Fig. 6 Adaptive laws $||\hat{W}_{c1}||$ and $||\hat{W}_{c2}||$

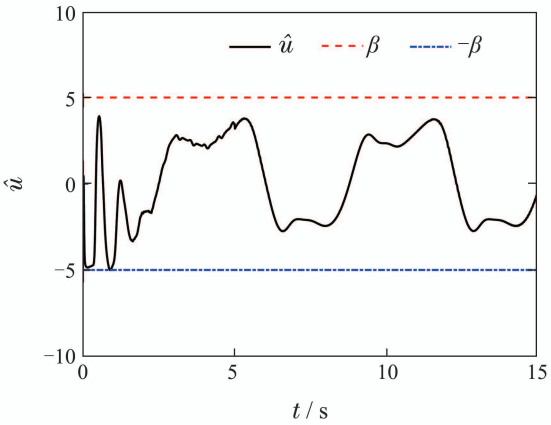

图 7 系统输入$\hat{u}$

Fig. 7 System input $\hat{u}$

4 总结

本文针对存在不可测状态和输入约束系统的最优控制问题, 基于反步法和强化学习的思想, 提出了一种最优跟踪控制策略. 首先, 构建非线性观测器估计系统不可测状态, 然后, 在反步法中的每一步引入执行–评价网络结构的强化学习方法, 得到相应的最优控制器, 并通过引入动态面技术避免反

步法中出现“复杂度爆炸”问题. 最后, 采用李雅普诺夫稳定性理论验证了闭环系统的稳定性. 仿真结果表明了本文设计方法的有效性. 在将来的研究工作中, 我们将利用优化的思想解决多智能体系统最优编队控制问题.

参考文献:

[1] LIU Q, LENG J W, YAN D X, et al. Digital twin-based designing of the configuration, motion, control, and optimization model of a flow-type smart manufacturing system. *Journal of Manufacturing Systems*, 2021, 58: 52 – 64.

[2] REN H R, LU R Q, XIONG J L, et al. Optimal filtered and smoothed estimators for discrete-time linear systems with multiple packet dropouts under markovian communication constraints. *IEEE Transactions on Cybernetics*, 2020, 50(9): 4169 – 4181.

[3] LI H Y, WU Y, CHEN M. Adaptive fault-tolerant tracking control for discrete-time multiagent systems via reinforcement learning algorithm. *IEEE Transactions on Cybernetics*, 2020, DOI: 10.1109/TCYB.2020.2982168.

[4] BEARD R, SARIDIS G, WEN J. Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation. *Automatica*, 1997, 33(12): 2159 – 2177.

[5] 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.

[6] 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.

[7] ZHAO B, LIU D R, LUO C M. Reinforcement learning-based optimal stabilization for unknown nonlinear systems subject to inputs with uncertain constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(10): 4330 – 4340.

[8] FU Z J, XIE W F, RAKHEJA S, et al. Observer-based adaptive optimal control for unknown singularly perturbed nonlinear systems with input constraints. *IEEE Journal of Automatica Sinica*, 2017, 4(1): 48 – 57.

[9] WANG D, HE H B, LIU D R. Improving the critic learning for event-based nonlinear H_∞ control design. *IEEE Transactions on Cybernetics*, 2017, 47(10): 3417 – 3428.

[10] HAMIDREZA M, FRANK L L, MOHAMMAD B N S. Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, 24(10): 1513 – 1525.

[11] ZHAO B, LIU D R, ALIPPI C. Sliding-mode surface-based approximate optimal control for uncertain nonlinear systems with asymptotically stable critic structure. *IEEE Transactions on Cybernetics*, 2019, DOI: 10.1109/TCYB.2019.2962011.

[12] ZHOU Q, WANG L J, WU C W, et al. Adaptive fuzzy tracking control for a class of pure-feedback nonlinear systems with time-varying delay and unknown dead zone. *Fuzzy Sets and Systems*, 2017, 329: 36 – 60.

[13] DONG G W, CAO L, YAO D Y, et al. Adaptive attitude control for multi-MUAV systems with output dead-zone and actuator fault. *IEEE/CAA Journal of Automatica Sinica*, 2020, DOI: 10.1109/JAS.2020.1003605.

[14] LIN G H, LI H Y, MA H, et al. Human-in-the-loop consensus control for nonlinear multi-agent systems with actuator faults. *IEEE/CAA Journal of Automatica Sinica*, 2020, DOI: 10.1109/JAS.2020.1003596.

[15] MA H, LIANG H J, MA H J, et al. Adaptive event-triggered control for a class of nonlinear systems with periodic disturbances. *Science China Information Sciences*, 2020, 63(5): 150212.

[16] ZHOU Qi, CHEN Guangdeng, LU Renquan, et al. Disturbance-observer-based event-triggered control for multi-agent systems with input saturation. *Scientia Sinica Informationis*, 2019, 49(11): 1502 – 1516.
(周琪, 陈广登, 鲁仁全, 等. 基于扰动观测器的输入饱和多智能体系统事件触发控制. 中国科学: 信息科学, 2019, 49(11): 1502 – 1516.)

[17] LIANG H J, LIU G L, ZHANG H G, et al. Neural-network-based event-triggered adaptive control of nonaffine nonlinear multiagent systems with dynamic uncertainties. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, DOI: 10.1109/TNNLS.2020.3003950.

[18] LIANG H J, GUO X Y, PAN Y N, et al. Event-triggered fuzzy bipartite tracking control for network systems based on distributed reduced-order observers. *IEEE Transactions on Fuzzy Systems*, 2020, DOI: 10.1109/TFUZZ.2020.2982618.

[19] PAN Y N, DU P H, XUE H, et al. Singularity-free fixed-time fuzzy control for robotic systems with user-defined performance. *IEEE Transactions on Fuzzy Systems*, 2020, DOI: 10.1109/TFUZZ.2020.2999746.

[20] MA H, REN H R, ZHOU Q, et al. Approximation-based nussbaum gain adaptive control of nonlinear systems with periodic disturbances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, DOI: 10.1109/TSMC.2021.3050993.

[21] ZHENG Xiaohong, DONG Guowei, ZHOU Qi, et al. Fault-tolerant control for stochastic multi-agent systems with output constraints. *Control Theory & Applications*, 2020, 37(5): 961 – 968.
(郑晓宏, 董国伟, 周琪, 等. 带有输出约束条件的随机多智能体系统容错控制. 控制理论与应用, 2020, 37(5): 961 – 968.)

[22] YANG Bin, ZHOU Qi, CAO Liang, et al. Event-triggered control for multi-agent systems with prescribed performance and full state constraints. *Acta Automatica Sinica*, 2019, 45(8): 1527 – 1535.
(杨彬, 周琪, 曹亮, 等. 具有指定性能和全状态约束的多智能体系统事件触发控制. 自动化学报, 2019, 45(8): 1527 – 1535.)

[23] LIU Y, LIU X P, JING Y W, et al. A novel finite-time adaptive fuzzy tracking control scheme for nonstrict feedback systems. *IEEE Transactions on Fuzzy Systems*, 2019, 27(4): 646 – 658.

[24] XIAO W B, CAO L, LI H Y, et al. Observer-based adaptive consensus control for nonlinear multi-agent systems with time-delay. *Science China Information Sciences*, 2020, 63(132202): 1 – 17.

[25] LI T S, WANG D, FENG G, et al. A DSC approach to robust adaptive NN tracking control for strict-feedback nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2010, 40(3): 915 – 927.

[26] LIAN S K, MENG W, LIN Z M, et al. Adaptive attitude control of a quadrotor using fast non-singular terminal sliding mode. *IEEE Transactions on Industrial Electronics*, 2021, DOI: 10.1109/TIE.2021.3057015.

[27] CHEN M, SHAO S Y, JIANG B. Adaptive neural control of uncertain nonlinear systems using disturbance observer. *IEEE Transactions on Cybernetics*, 2017, 47(10): 3110 – 3123.

[28] LI Y M, SUN K K, TONG S C. Observer-based adaptive fuzzy fault-tolerant optimal control for SISO nonlinear systems. *IEEE Transactions on Cybernetics*, 2019, 49(2): 649 – 661.

[29] GUO Zijie, BAI Weiwei, ZHOU Qi, et al. Adaptive optimal control for a class of nonlinear systems with dead zone input and prescribed performance. *Acta Automatica Sinica*, 2019, 45(11): 2128 – 2136.
(郭子杰, 白伟伟, 周琪, 等. 基于性能指标约束的一类输入死区非线性系统最优控制. 自动化学报, 2019, 45(11): 2128 – 2136.)

[30] WEN G X, GE S S, TU F W. Optimized backstepping for tracking control of strict-feedback systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(8): 3850 – 3862.

[31] WEN G X, GE S S, CHEN C L P, et al. Adaptive tracking control of surface vessel using optimized backstepping technique. *IEEE Transactions on Cybernetics*, 2019, 49(9): 3420 – 3431.

[32] WEN G X, CHEN C L P, GE S S. Simplified optimized backstepping control for a class of nonlinear strict-feedback systems with unknown dynamic functions. *IEEE Transactions on Cybernetics*, 2020, DOI: 10.1109/TCYB.2020.3002108.

[33] JIANG Y, FAN J L, CHAI T Y, et al. Data-driven flotation industrial process operational optimal control based on reinforcement learning. *IEEE Transactions on Industrial Informatics*, 2018, 14(5): 1974 – 1989.

[34] JIANG Y, FAN J L, CHAI T Y, et al. Dual-rate operational optimal control for flotation industrial process with unknown operational model. *IEEE Transactions on Industrial Electronics*, 2019, 66(6): 4587 – 4599.

[35] LU X L, KIUMARSI B, CHAI T Y, et al. Operational control of mineral grinding processes using adaptive dynamic programming and reference governor. *IEEE Transactions on Industrial Informatics*, 2019, 15(4): 2210 – 2221.

[36] YANG X, LIU D R, WANG D. Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints. *International Journal of Control*, 2014, 87(3): 553 – 566.

[37] ZHU Y H, ZHAO D B, HE H B, et al. Event-triggered optimal control for partially unknown constrained-input systems via adaptive dynamic programming. *IEEE Transactions on Industrial Electronics*, 2017, 64(5): 4101 – 4109.

作者简介:

罗 傲 硕士研究生, 目前研究方向为强化学习、多智能体系统和自适应控制, E-mail: luoao2019@163.com;

肖文彬 博士研究生, 目前研究方向为非线性系统自适应控制、多智能体协同控制和自适应动态规划, E-mail: xiaowb992@163.com;

周 琪 教授, 博士生导师, 目前研究方向为非线性系统的模糊控制、神经控制、鲁棒控制及其应用, E-mail: zhouqi2009@gmail.com;

鲁仁全 教授, 博士生导师, 目前研究方向为复杂系统、网络控制系统和非线性系统, E-mail: rqlu@gdut.edu.cn.