DOI: 10.13382/j.jemi.B2206002

基于 actor-critic 框架的在线积分强化学习算法研究*

蔡 军 苟文耀 刘 颜
(重庆邮电大学自动化学院 重庆 400065)

摘 要:针对轮式移动机器人动力学系统难以实现无模型的最优跟踪控制问题,提出了一种基于 actor-critic 框架的在线积分强化学习控制算法。首先,构建 RBF 评价神经网络并基于近似贝尔曼误差设计该网络的权值更新律,以拟合二次型跟踪控制性能指标函数。其次,构建 RBF 行为神经网络并以最小化性能指标函数为目标设计权值更新律,补偿动力学系统中的未知项。最后,通过 Lyapunov 理论证明了所提出的积分强化学习控制算法可以使得价值函数,行为神经网络权值误差与评价神经网络权值误差一致最终有界。仿真和实验结果表明,该算法不仅可以实现对恒定速度以及时变速度的跟踪,还可以在嵌入式平台上进行实现。

关键词:积分强化学习;RBF 神经网络;非线性仿射系统;跟踪控制
中图分类号:TP13;TN911.4 **文献标识码**:A **国家标准学科分类代码**:410.15

Research on online integral reinforcement learning algorithm based on actor-critic framework

Cai Jun Gou Wenyao Liu Yan
(School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: For the problem that it is difficult to achieve model-free optimal tracking control in the dynamic system of wheeled mobile robot, a new online integral reinforcement learning control algorithm based on actor-critic framework is proposed in this paper. Firstly, the critic neural network based on RBF is constructed to fit the quadratic tracking control performance index function and the weight updating law of the network is designed based on the approximate Behrman error. Secondly, the RBF actor neural network is constructed to compensate the unknown terms in the dynamic system and the weight updating law is designed to minimize the performance index function. Finally, it is proved by Lyapunov theory that the proposed integral reinforcement learning control algorithm can make the value function, the critic and actor neural network weights error uniformly and finally bounded. Simulation and experimental results show that the algorithm not only realizes the tracking of constant or time-varying velocity, but also can be implemented on the embedded platform.

Keywords: integral reinforcement learning; RBF neural network; nonlinear affine system; tracking control

0 引 言

轮式移动机器人作为机器人领域的一个重要分支,由于其在军事,农业,工业,家居等领域的广泛应用,近几年来受到了不少学者与工程师的青睐[1-3]。然而,大多数的研究都仅仅关注于轮式移动机器人的运动学模型[4-5]。相比而言,运动学模型不需要考虑复杂的动态特性以及模型参数,如小车质量、电机电阻、转动惯量等。但在高移速或者高负载的情况下,为了更好的实现跟踪轨迹,动力学特性是需要考虑的[6-7]。

在早期基于动力学模型的研究中,由动力学控制器产生的控制信号通常是电机的力矩,电压或者电流[8-10]。然而,一般的机器人通常接收速度指令。因此,Martins 等[7]建立的以期望线速度和期望角速度为控制量的动力学模型是有实际意义的。此外,在基于动力学模型的研究中,大部分的研究工作都只有理论的分析与仿真的结果,只有少部分的学者在实时平台上对他们所提出的算

收稿日期: 2022-11-14 Received Date: 2022-11-14
*基金项目:重庆市教委科学技术研究项目(KJZD-M202200603)、重庆市自然科学基金项目(CSTB2022NSCQ-MSX0380)资助

法进行了验证[6,11]。

在运动学与动力学模型的基础上,学者们针对轮式移动机器人的轨迹跟踪控制问题提出了各种控制方案。Obaid 等[12]提出了一种时变的反步跟踪控制器,可以在有限时间间隔内产生平滑有界的速度输出。Huang 等[13]提出了一种反馈线性化时滞控制器,并构建了高增益预测器对反馈线性化控制器中的预测状态进行了估计。Zhang 等[14]提出了一种基于非奇异递归结构滑模的有限时间跟踪控制方法,并利用光滑双曲先切函数设计扰动观测器,保证了跟踪误差在有限时间内收敛到原点的小范围内。Mehrez 等[15]将不带稳定约束或代价的模型预测控制用于轮式移动机器人的设定点跟踪,并通过利用成本可控性的概念来保证闭环的渐近稳定性。然而,上述控制方案大多是基于模型的,这将难以在实际的轮式移动机器人上实现,并且上诉研究方案没有考虑轮式移动机器人的最优跟控制问题。

自适应动态规划[16-18]在动态规划理论的基础上,结合了 actor-critic 框架和神经网络,避免了传统动态规划在求解最优解时需要依赖精准的数学模型和计算量过大的问题,是求解非线性最优控制问题的有效手段。一般而言,actor 网络依据当前状态计算控制量,而 critic 网络则用于评估当前控制量的价值,即对价值函数进行拟合。基于对价值函数的不同处理,critic 网络的优化方法通常分为一下两种:一种是对价值函数求导得到哈密顿-雅克比-贝尔曼方程[19-21],另一种则是对价值函数进行时序差分处理[22-24],得到贝尔曼方程。基于后者的 critic 网络更新方法也被称为积分强化学习,在该方法中仅利用了系统的输入输出数据,从而摆脱了对模型的依赖。本文也将在积分强化学习[25-26]的基础之上做进一步的研究。

本文的主要贡献如下:1)针对一类具有多输入多输出特性的一阶非线性仿射系统,通过未知项补偿的方法,设计了无模型的在线积分强化学习控制律,实现了最优跟踪控制。2)在嵌入式微处理器上对所设计的在线积分强化学习算法进行了实现,并应用于轮式移动机器人的内环动力学控制器中,实现了轮式移动机器人的轨迹跟踪控制。

1　问题描述

考虑如下一阶非线性仿射系统:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{u}(t) \tag{1}$$

式中:$\boldsymbol{x} \in R^n$ 表示状态,$\boldsymbol{u} \in R^m$ 表示控制量 $\boldsymbol{f}(\boldsymbol{x}) \in R^n$ 表示系统函数。定义跟踪误差为:

$$\boldsymbol{e} = \boldsymbol{x}_d - \boldsymbol{x} \tag{2}$$

式中:$\boldsymbol{x}_d$ 表示期望轨迹。

定义价值函数为:

$$V(\boldsymbol{e}(t)) = \int_t^{\infty} e^{-\gamma(\tau - t)}(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{e} + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u})\,\mathrm{d}\tau \tag{3}$$

其中,$0 \leqslant \gamma < 1$ 表示折扣因子。价值函数满足如下 Bellman 方程:

$$p(t) + e^{-\gamma T}V(\boldsymbol{e}(t)) - V(\boldsymbol{e}(t - T)) = 0 \tag{4}$$

式中:T 表示控制周期,$p(t)$ 为积分强化项。

$$p(t) = \int_{t-T}^{t} e^{-\gamma(\tau - t + T)}(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{e} + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u})\,\mathrm{d}\tau \tag{5}$$

本文将设计控制律 $\boldsymbol{u}(t)$ 使得由式(3)定义的价值函数最小化,且控制律中不包含模型信息 $\boldsymbol{f}(\boldsymbol{x})$。

2　积分强化学习控制器设计

本节将提出基于 actor-critic 框架的控制结构。由于 $V(\boldsymbol{e}(t))$ 包含了系统在未来时刻的信息,在当前时刻是不可得到的,因此设计评价网络拟合价值函数,并设计行为网络拟合系统中的未知项 $\boldsymbol{f}(\boldsymbol{x})$。

2.1　RBF 神经网络

RBF 神经网络结构简单,能在一定精度下逼近任何非线性函数。RBF 神经网络有 3 层:输入层,隐含层和输出层。隐含层的神经元激活函数为高斯径向基函数:

$$h_j = \exp\left(-\frac{\|\boldsymbol{s} - \boldsymbol{c}_j\|^2}{2b^2}\right), j = 1, \cdots, M \tag{6}$$

式中:$\boldsymbol{s} \in R^N$ 为输入向量,$\boldsymbol{c}_j \in R^N$ 为隐含层节点所包含的中心向量,且 $\boldsymbol{c} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_M]$,$b$ 为高斯基函数的宽度,M 是隐含层节点数量。网络的输出为:

$$\boldsymbol{y} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{s}) \tag{7}$$

式中:$\boldsymbol{h}(\boldsymbol{s}) = [h_1, \cdots, h_M]$ 为隐含层的输出,$\boldsymbol{W} \in R^{M \times O}$ 为输出层的权值,$\boldsymbol{y} \in R^O$ 为网络的输出。则 $\boldsymbol{y}$ 对 $\boldsymbol{s}$ 的梯度为:

$$\nabla \boldsymbol{y}_s = \nabla \boldsymbol{h}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W} \tag{8}$$

$\nabla \boldsymbol{h}(\boldsymbol{s})$ 可由如下雅克比矩阵表示:

$$\nabla \boldsymbol{h}(\boldsymbol{s}) = \begin{bmatrix} \dfrac{\partial h_1}{\partial s_1} & \cdots & \dfrac{\partial h_1}{\partial s_N} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_M}{\partial s_1} & \cdots & \dfrac{\partial h_M}{\partial s_N} \end{bmatrix} \tag{9}$$

其中:

$$\frac{\partial h_j}{\partial s_i} = \frac{sign(c_{ij} - s_i)\exp\left(-\dfrac{\|\boldsymbol{s} - \boldsymbol{c}_j\|^2}{2b^2}\right) \mid c_{ij} - s_i \mid}{b^2} \tag{10}$$

2.2　评价网络设计

由 RBF 神经网络拟合的价值函数可表示为:

$$V(\boldsymbol{e}) = \boldsymbol{W}_c^{\mathrm{T}}\boldsymbol{h}_c(\boldsymbol{e}) + \varepsilon_c \tag{11}$$

式中:$\boldsymbol{W}_c \in R^p$ 为评价网络的理想权值,ε_c 为评价网络有

界的逼近误差。则价值函数 $V(\boldsymbol{e}(t))$ 对误差 $\boldsymbol{e}(t)$ 的梯度可为：

$$\nabla \boldsymbol{V}_e = \nabla \boldsymbol{h}_c^{\mathrm{T}} \boldsymbol{W}_c + \nabla \boldsymbol{\varepsilon}_c \tag{12}$$

式中：$\nabla \boldsymbol{\varepsilon}_c$ 为评价网络逼近误差的梯度，满足 $\| \nabla \boldsymbol{\varepsilon}_c \| \leqslant b_c$。评价网络的实际输出为：

$$\hat{V} = \hat{\boldsymbol{W}}_c^{\mathrm{T}} \boldsymbol{h}_c(\boldsymbol{e}) \tag{13}$$

式中：$\hat{\boldsymbol{W}}_c$ 为评价网络的实际权，$\hat{V}$ 是对价值函数的估计值。定义评价网络理想权值实际权值的误差如下：

$$\tilde{\boldsymbol{W}}_c = \boldsymbol{W}_c - \hat{\boldsymbol{W}}_c \tag{14}$$

将式（11）代入 Bellman 方程（4）中可得 Bellman 误差：

$$\boldsymbol{\varepsilon}_B \triangleq p(t) + \boldsymbol{W}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c(t) \tag{15}$$

式中：$\Delta \boldsymbol{h}_c(t) = e^{-\gamma T} \boldsymbol{h}_c(\boldsymbol{e}(t)) - \boldsymbol{h}_c(\boldsymbol{e}(t - T))$。该 Bellman 误差也可由网络逼近误差表示。定义近似 Bellman 误差为：

$$\hat{\boldsymbol{\varepsilon}}_B \triangleq p(t) + \hat{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c(t) \tag{16}$$

基于梯度下降算法设计 $\hat{\boldsymbol{W}}_c$ 的更新律使得目标函数 $E_c = \hat{\boldsymbol{\varepsilon}}_B^2 / 2$ 最小化：

$$\dot{\hat{\boldsymbol{W}}}_c = -\alpha_c (\Delta \bar{\boldsymbol{h}}_c \hat{\boldsymbol{\varepsilon}}_B - \nabla \boldsymbol{h}_c \dot{\boldsymbol{e}}) \tag{17}$$

式中：$\Delta \bar{\boldsymbol{h}}_c = m_s \Delta \boldsymbol{h}_c$，$m_s = 1/(1 + \Delta \boldsymbol{h}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c)^2$，$\alpha_c$ 为学习率。与一般梯度下降的设计不同，式中 $\nabla \boldsymbol{h}_c \dot{\boldsymbol{e}}$ 的加入可保证系统的稳定性。

2.3 行为网络设计

构建基于 RBF 的行为神经网络拟合系统方程中的未知项 $\boldsymbol{f}(\boldsymbol{x})$：

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}_a^{\mathrm{T}} \boldsymbol{h}_a(\boldsymbol{x}) + \boldsymbol{\varepsilon}_a \tag{18}$$

式中：$\boldsymbol{W}_a \in R^{q \times n}$ 为行为网络的理想权值，$\boldsymbol{\varepsilon}_a$ 为行为网络的逼近误差，满足 $\| \boldsymbol{\varepsilon}_a \| \leqslant b_a$。行为网络的实际输出为：

$$\hat{\boldsymbol{f}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}_a^{\mathrm{T}} \boldsymbol{h}_a(\boldsymbol{x}) \tag{19}$$

式中：$\hat{\boldsymbol{W}}_a$ 是行为神经网络的实际权值。定义行为神经网络权值误差为：

$$\tilde{\boldsymbol{W}}_a = \boldsymbol{W}_a - \hat{\boldsymbol{W}}_a \tag{20}$$

设计行为网络的权值自适应更新律为：

$$\dot{\hat{\boldsymbol{W}}}_a = -\alpha_a \boldsymbol{h}_a(\boldsymbol{x}) \hat{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c(\boldsymbol{e}) \tag{21}$$

式中：α_a 为行为神经网络的学习率。

2.4 控制律设计

设计如下基于 RBF 神经网络的控制律为：

$$\boldsymbol{u} = \boldsymbol{g}^{-1}(\boldsymbol{x})(-\hat{\boldsymbol{f}}(\boldsymbol{x}) + \dot{\boldsymbol{x}}_d + \eta \nabla \boldsymbol{h}_c^{T}(\boldsymbol{e}) \hat{\boldsymbol{W}}_c) \tag{22}$$

式中：η 为可调参数。

3 稳定性分析

定理 1 对于一阶非线性仿射系统（1），在控制律（22）以及评价网络权值更新律（17）和行为网络权值更新律（21）的作用下，价值函数 $V(\boldsymbol{e})$，RBF 网络权值误差 $\tilde{\boldsymbol{W}}_a$ 和 $\tilde{\boldsymbol{W}}_c$ 是有界的。

证明 选取如下 Lyapunov 函数：

$$L = V + \frac{1}{2} \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \alpha_c^{-1} \tilde{\boldsymbol{W}}_c + \frac{1}{2} tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \alpha_a^{-1} \tilde{\boldsymbol{W}}_a) \tag{23}$$

令 $L_1 = V - \hat{V} + \frac{1}{2} \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \alpha_c^{-1} \tilde{\boldsymbol{W}}_c$，$L_2 = \hat{V} + \frac{1}{2} tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \alpha_a^{-1} \tilde{\boldsymbol{W}}_a)$，即 $L = L_1 + L_2$。对 L_1 求导，有：

$$\dot{L}_1 = \dot{V} - \dot{\hat{V}} - \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \alpha_c^{-1} \dot{\hat{\boldsymbol{W}}}_c =$$

$$(\nabla \boldsymbol{h}_c^{\mathrm{T}} \boldsymbol{W}_c + \nabla \boldsymbol{\varepsilon}_c)^{\mathrm{T}} \dot{\boldsymbol{e}} - (\nabla \boldsymbol{h}_c^{\mathrm{T}} \hat{\boldsymbol{W}}_c)^{\mathrm{T}} \dot{\boldsymbol{e}} - \tilde{\boldsymbol{W}}\mathrm{T}_c \alpha_c^{-1} \dot{\hat{\boldsymbol{W}}}_c =$$

$$\tilde{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c \dot{\boldsymbol{e}} + \nabla \boldsymbol{\varepsilon}_c^{\mathrm{T}} \dot{\boldsymbol{e}} + \tilde{\boldsymbol{W}}_c^{\mathrm{T}} (\Delta \bar{\boldsymbol{h}}_c \hat{\boldsymbol{\varepsilon}}_B - \nabla \boldsymbol{h}_c \dot{\boldsymbol{e}}) =$$

$$\nabla \boldsymbol{\varepsilon}_c^{\mathrm{T}} \dot{\boldsymbol{e}} + \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \bar{\boldsymbol{h}}_c (p(t) + \hat{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c) =$$

$$\nabla \boldsymbol{\varepsilon}_c^{\mathrm{T}} \dot{\boldsymbol{e}} + \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \bar{\boldsymbol{h}}_c (p(t) + \boldsymbol{W}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c - \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c) =$$

$$-\tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \bar{\boldsymbol{h}}_c \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c + \nabla \boldsymbol{\varepsilon}_c^{\mathrm{T}} \dot{\boldsymbol{e}} + \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \bar{\boldsymbol{h}}_c (1 - e^{\gamma T}) \varepsilon_c =$$

$$m_s \left(-\tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c + \frac{\nabla \boldsymbol{\varepsilon}_c^{T} \dot{\boldsymbol{e}}}{m_s} + \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c (1 - e^{\gamma T}) \varepsilon_c \right)$$

令 $\varepsilon_e = 2 \nabla \boldsymbol{\varepsilon}_c^{\mathrm{T}} \dot{\boldsymbol{e}}$ 再由杨氏不等式可知：

$$\tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c (1 - e^{\gamma T}) \varepsilon_c \leqslant \frac{1}{2} \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \bar{\boldsymbol{h}}_c + \frac{1}{2} (1 - e^{\gamma T})^2 \varepsilon_c^2$$

则有：

$$\dot{L}_1 \leqslant -m_s \left(\frac{1}{2} \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c - \frac{1}{2} \left((1 - e^{\gamma T})^2 \varepsilon_c^2 + \frac{\varepsilon_e}{m_s} \right) \right) \leqslant -\frac{1}{2} m_s \left(\| \tilde{\boldsymbol{W}}_c^{\mathrm{T}} \Delta \boldsymbol{h}_c \|^2 - (1 - e^{\gamma T})^2 \varepsilon_c^2 - \frac{\varepsilon_e}{m_s} \right)$$

于是，当下式成立时，有 $\dot{L}_1 \leqslant 0$。

$$\| \tilde{\boldsymbol{W}}_c \| \geqslant \sqrt{\frac{(1 - e^{\gamma T})^2 \varepsilon_c^2 + \varepsilon_e / m_s}{\| \Delta \boldsymbol{h}_c \|^2}} \tag{24}$$

再对 L_2 求导，有：

$$\dot{L}_2 = \dot{\hat{V}} - tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \alpha_a^{-1} \dot{\hat{\boldsymbol{W}}}_a) =$$

$$\hat{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c (\dot{\boldsymbol{x}}_d - \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{u}) - tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \alpha_a^{-1} \dot{\hat{\boldsymbol{W}}}_a) =$$

$$\hat{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c (-\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \boldsymbol{h}_a(\boldsymbol{x}) - \boldsymbol{\varepsilon}_a - \eta \nabla \boldsymbol{h}_c^{\mathrm{T}}(\boldsymbol{e}) \hat{\boldsymbol{W}}_c) - tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \alpha_a^{-1} \dot{\hat{\boldsymbol{W}}}_a) =$$

$$-\hat{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c \tilde{\boldsymbol{W}}_a^{\mathrm{T}} \boldsymbol{h}_a(\boldsymbol{x}) - tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}} \alpha_a^{-1} \dot{\hat{\boldsymbol{W}}}_a) - \hat{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c \boldsymbol{\varepsilon}_a - \eta \| \hat{\boldsymbol{W}}_c^{\mathrm{T}} \nabla \boldsymbol{h}_c \|^2$$

由于：

$$-\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\boldsymbol{\varepsilon}_a \leqslant \|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\| \cdot \|\boldsymbol{\varepsilon}_a\| \leqslant b_a\|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\|,$$

又由于：

$$\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\ \tilde{\boldsymbol{W}}_a^{\mathrm{T}}\boldsymbol{h}_a(\boldsymbol{x}) = tr(\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\ \tilde{\boldsymbol{W}}_a^{\mathrm{T}}\boldsymbol{h}_a(\boldsymbol{x})) = tr(\tilde{\boldsymbol{W}}_a^{\mathrm{T}}\boldsymbol{h}_a(\boldsymbol{x})\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c),$$

代入行为网络权值更新律，则有：

$$\dot{L}_2 \leqslant b_a\|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\| - \eta\|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\|^2 \leqslant -(\eta\|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\| - b_a)\|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\|$$

于是，当：

$$\eta > \frac{b_a}{\|\hat{\boldsymbol{W}}_c^{\mathrm{T}}\nabla\boldsymbol{h}_c\|} \tag{25}$$

有 $\dot{L}_2 \leqslant 0$。

综上，在满足条件（24），（25）的前提下，有 $L \geqslant 0$，$\dot{L} = \dot{L}_1 + \dot{L}_2 \leqslant 0$，于是当 $t \to \infty$ 时，V 和 $\tilde{\boldsymbol{W}}_a$ 与 $\tilde{\boldsymbol{W}}_c$ 有界。证明成立。

4　仿真与实验结果分析

本节中，将以轮式移动机器人为被控对象，如图 1 所示，并在仿真与实物上验证所提出算法的有效性。

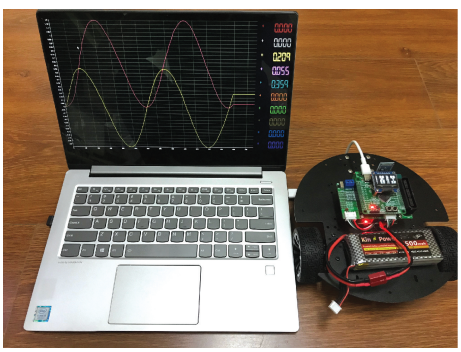

图 1　轮式移动机器人的实物图

Fig. 1　Physical picture of wheeled mobile robot

4.1　轮式移动机器人速度跟踪控制仿真

考虑如下轮式移动机器人的动力学模型：

$$\begin{bmatrix}\dot{v}\\ \dot{w}\end{bmatrix} = \begin{bmatrix}\dfrac{\theta_3}{\theta_1}w^2 - \dfrac{\theta_4}{\theta_1}v\\ -\dfrac{\theta_5}{\theta_2}uw - \dfrac{\theta_6}{\theta_2}w\end{bmatrix} + \begin{bmatrix}\dfrac{1}{\theta_1} & 0\\ 0 & \dfrac{1}{\theta_2}\end{bmatrix}\begin{bmatrix}v_{ref}\\ w_{ref}\end{bmatrix} \tag{26}$$

该模型符合式（1）中一阶非线性仿射系统的定义，其中状态 $\boldsymbol{x} = [v, w]^{\mathrm{T}}$，$v$ 和 w 表示轮式移动机器人的线速度与角速度，控制量 $\boldsymbol{u} = [v_{ref}, w_{ref}]^{\mathrm{T}}$，$v_{ref}$ 和 w_{ref} 表示线速度与角速度指令。$\boldsymbol{\theta} = [\theta_1, \cdots, \theta_6]^{\mathrm{T}}$ 为模型参数，其值如下：$\theta_1 = 0.0228$，$\theta_2 = 0.0568$，$\theta_3 = -0.0001$，$\theta_4 = 1.0030$，$\theta_5 = 0.0732$，$\theta_6 = 0.9981$。

取期望速度 $v_d = 0.3$，$w_d = 0.5$. 行为网络中取 $\boldsymbol{c}_a = \begin{bmatrix}-1,-0.8,-0.6,-0.4,-0.2,0,0.2,0.4,0.6,0.8,1\\ -1,-0.8,-0.6,-0.4,-0.2,0,0.2,0.4,0.6,0.8,1\end{bmatrix}$，$b_a = 3$，$\alpha_a = 0.03$。评价网络中取 $\boldsymbol{c}_c = \begin{bmatrix}-1,-0.5,0,0.5,1\\ -1,-0.5,0,0.5,1\end{bmatrix}$，$b_c = 2.5$，$\alpha_c = 0.08$。评价网络与行为网络的权值初始化为 0。控制周期 $T = 0.01$。控制器中余下参数的选取如下：$\gamma = 0.01$，$\boldsymbol{Q} = 5I_{2\times2}$，$\boldsymbol{R} = I_{2\times2}$，$\eta = 5$。

本文所提出算法的跟踪效果如图 2 所示。可以看出，轮式移动机器人的线速度和角速度都达到了设定的期望值。图 3 和 4 分别展示了 critic 和 actor 的网络参数，经过一定次数的更新后，权值最终收敛（注：图 4 中虚线部分表示 $\boldsymbol{W}_{a1}$，实线部分表示 $\boldsymbol{W}_{a2}$，都是 11 维的向量，且 $\boldsymbol{W}_a = [\boldsymbol{W}_{a1}, \boldsymbol{W}_{a2}]$）。在设定的期望速度下，轮式移动机器人的实际运动轨迹是半径为 0.6 m 的圆，如图 5 所示。

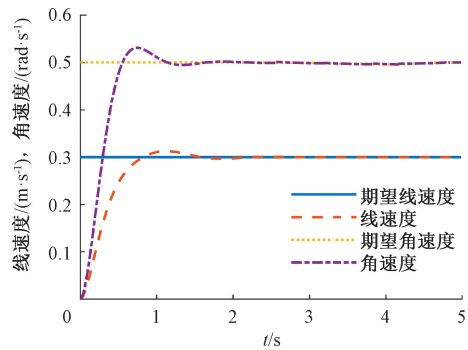

图 2　线速度和角速度跟踪效果

Fig. 2　Linear and angular velocity tracking effects

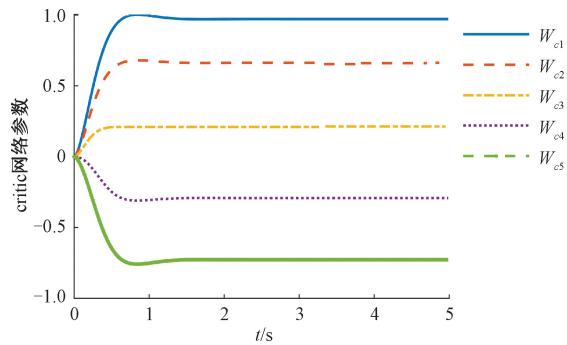

图 3　critic 网络参数更新

Fig. 3　Parameter updating of critic network

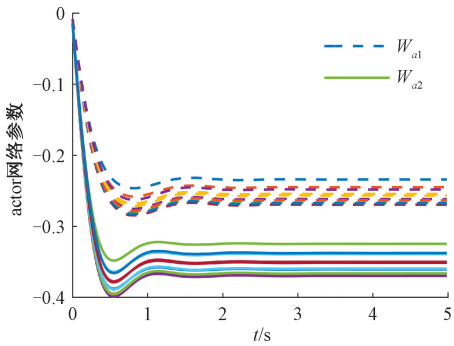

图 4　actor 网络参数更新

Fig. 4　Parameter updating of actor network

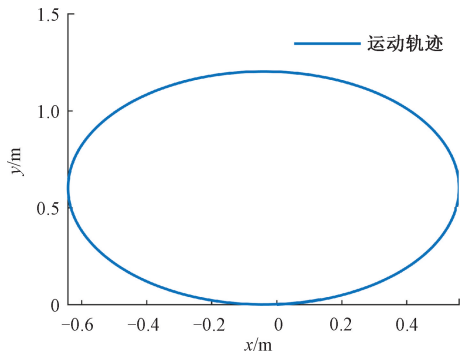

图 5　轮式移动机器人实际运动轨迹(上位机数据)

Fig. 5　Actual motion trajectory of wheeled mobile robot (data from upper computer)

4.2　轮式移动机器人轨迹跟踪控制实验

在本次实验中,将考虑完整的轮式移动机器人模型,即运动学模型和动力学模型。给出轮式移动机器人的运动学模型如下:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{\psi}\end{bmatrix}=\begin{bmatrix}\cos\psi & -a\sin\psi\\ \sin\psi & a\cos\psi\\ 0 & 1\end{bmatrix}\begin{bmatrix}v\\ w\end{bmatrix} \tag{27}$$

式中:x,y 表示位置,ψ 为轮式移动机器人的偏航角,参考点与两轮中心点的距离 $a=0.04$。本文所提出的控制算法仍然作为轮式移动机器人的动力学控制器,期望速度由如下运动学控制器给出:

$$\begin{bmatrix}v_d\\ w_d\end{bmatrix}=\begin{bmatrix}\cos\psi & \sin\psi\\ -\dfrac{1}{a}\sin\psi & \dfrac{1}{a}\cos\psi\end{bmatrix}\begin{bmatrix}\dot{x}_d+l_x\tanh\left(\dfrac{k_x}{l_x}\tilde{x}\right)\\ \dot{y}_d+l_y\tanh\left(\dfrac{k_y}{l_y}\tilde{y}\right)\end{bmatrix} \tag{28}$$

式中,x_d 与 y_d 表示期望轨迹,$\tilde{x}=x_d-x$,$\tilde{y}=y_d-y$ 表示位置误差,k_x,k_y,l_x,l_y 为控制器中的参数,均取值为 0.1。设定“8”字形期望轨迹为:$x_d=\sin(0.5t)$,$y_d=\cos(0.25t)-1$。

综上,轮式移动机器人轨迹跟踪控制的控制结构如图 6 所示,其中动力学控制器为本文所提出的基于 RBF 的改进积分强化学习算法,控制器中的参数与 4.1 节的仿真一致。

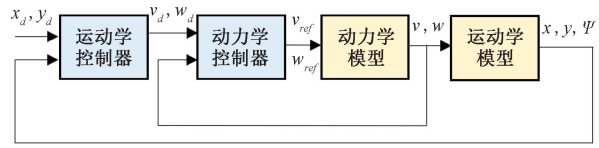

图 6　轮式移动机器人控制结构

Fig. 6　Wheeled mobile robot control structure

本实验中的数据均为在实物上采集而得,并由 MATLAB 制图。图 7 展示了对“8”字形轨迹的跟踪效果,可见本文所提出的控制算法不仅适用于恒定速度的跟踪,也适用于对时变速度的跟踪。图 8 和 9 分别展示了 x 轴与 y 轴方向上的跟踪效果,图 10 和 11 分别展示了对期望线速度和角速度的跟踪效果,该期望线速度与角速度由外环运动学控制器提供。由 actor 网络观测的系统中的未知项 $\boldsymbol{f}(\boldsymbol{x})$ 如图 12 所示,值得注意,该未知项中还隐含了未知的扰动项。

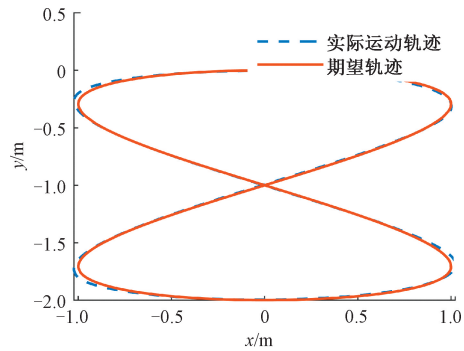

图 7　“8”字形轨迹跟踪

Fig. 7　“8” shape trajectory tracking

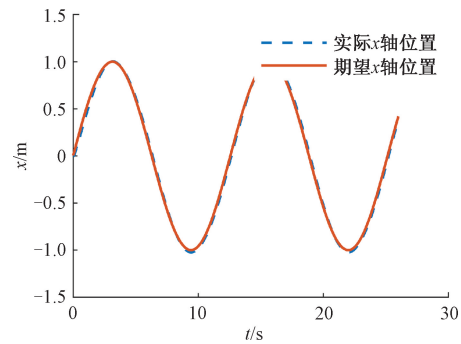

图 8　x 轴位置跟踪效果

Fig. 8　x-axis position tracking effect

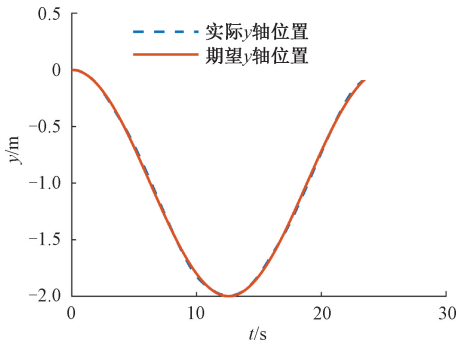

图 9　y 轴位置跟踪效果

Fig. 9　y-axis position tracking effect

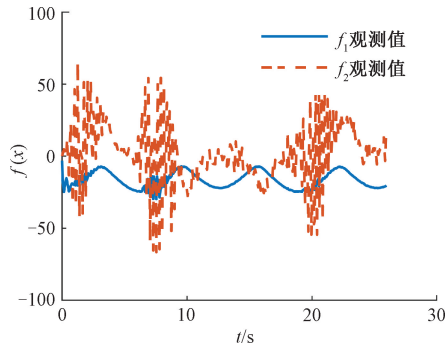

图 12　actor 网络对 $\boldsymbol{f}(\boldsymbol{x})$ 的观测

Fig. 12　Observation of $\boldsymbol{f}(\boldsymbol{x})$ by actor network

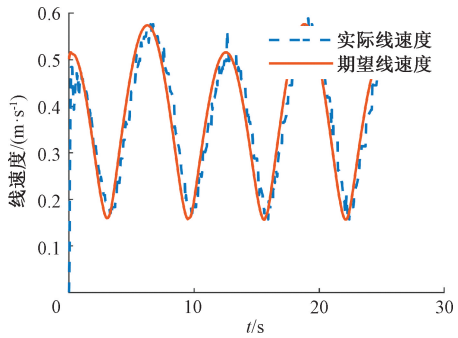

图 10　线速度跟踪效果

Fig. 10　Angular velocity tracking effect

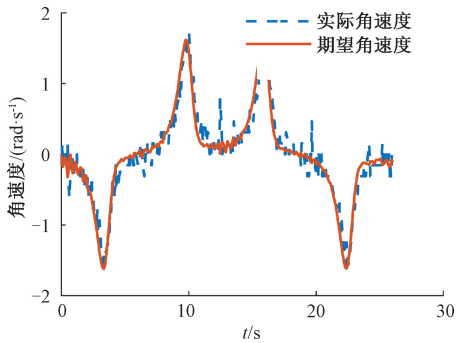

图 11　角速度跟踪效果

Fig. 11　Angular velocity tracking effect

5　结　论

本文针对连续一阶非线性仿射系统的跟踪控制问题，提出了基于 actor-critic 框架的在线积分强化学习算法。该算法通过行为 RBF 神经网络补偿系统中的未知项，通过评价 RBF 神经网络逼近设定的二次型性能指标函数，并根据 Lyapunov 理论证明了本算法可以实现最优跟踪控制。

参考文献

[1] CHANG S, WANG Y, ZUO Z, et al. Fixed-time formation control for wheeled mobile robots with prescribed performance [J]. IEEE Transactions on Control Systems Technology, 2022, 30(2): 844-851.

[2] 杨傲雷，金宏宙，陈灵，等. 融合深度学习与粒子滤波的移动机器人重定位方法[J]. 仪器仪表学报，2021，42(7)：226-233.
YANG AO L, JIN H ZH, CHEN L, et al. Mobile robot relocation method combining deep learning and particle filtering[J]. Chinese Journal of Scientific Instrument, 2021, 42(7): 226-233.

[3] ZHAO L, LIA J, LIC H, et al. Double-loop tracking control for a wheeled mobile robot with unmodeled dynamics along right angle roads [J]. ISA Transactions, 2022.

[4] SONG Z, REN H, ZHANG J, et al. Kinematic analysis and motion control of wheeled mobile robots in cylindrical workspaces [J]. IEEE Transactions on Automation Science and Engineering, 2016, 13(2): 1207-1214.

[5] LI W, DING L, LIU Z, et al. Kinematic bilateral tele-driving of wheeled mobile robots coupled with slippage[J]. IEEE Transactions on Industrial Electronics, 2017, 64(3): 2147-2157.

[6] GHEISARNEJAD M, KHOOBAN M H. An intelligent non-integer pid controller-based deep reinforcement learning: Implementation and experimental results[J]. IEEE Transactions on Industrial Electronics, 2021, 68(4): 3609-3618.

[7] MARTINS F N, SARCINELLI-FILHO M, CARELLI R. A velocity-based dynamic model and its properties for differential drive mobile robots[J]. Journal of Intelligent & Robotic Systems, 2017, 85(2): 277-292.

[8] HE W, DAVID A O, YIN Z, et al. Neural network

control of a robotic manipulator with input deadzone and output constraint[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 46(6): 759-770.

[9] ONAT A, OZKAN M. Dynamic adaptive trajectory tracking control of nonholonomic mobile robots using multiple models approach[J]. Advanced Robotics, 2015, 29(14): 913-928.

[10] HE W, DONG Y, SUN C. Adaptive neural impedance control of a robotic manipulator with input saturation[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 46(3): 334-344.

[11] CHWA D. Robust distance-based tracking control of wheeled mobile robots using vision sensors in the presence of kinematic disturbances[J]. IEEE Transactions on Industrial Electronics, 2016, 63(10): 6172-6183.

[12] OBAID M A M, HUSAIN A R. Time varying backstepping control for trajectory tracking of mobile robot[J]. International Journal of Computational Vision and Robotics, 2017, 7(1-2): 172-181.

[13] HUANG Y M, LEI J. Predictive feedback linearization tracking control for WMR systems with control delay[J]. Integrated Ferroelectrics, 2021, 217(1): 279-288.

[14] ZHANG H, LI B, XIAO B, et al. Nonsingular recursive-structure sliding mode control for high-order nonlinear systems and an application in a wheeled mobile robot[J]. ISA Transactions, 2022, 130: 553-564.

[15] MEHREZ M W, WORTHMANN K, CENERINI J P V, et al. Model predictive control without terminal constraints or costs for holonomic mobile robots[J]. Robotics and Autonomous Systems, 2020, 127: 103468.

[16] LUO B, LIU D, WU H N, et al. Policy gradient adaptive dynamic programming for data-based optimal control[J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3341-3354.

[17] WANG D, HE H, LIU D. Adaptive critic nonlinear robust control: A survey[J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3429-3451.

[18] WEI Q, LIU D, LIN Q, et al. Adaptive dynamic programming for discrete-time zero-sum games[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4): 957-969.

[19] CUI L, XIE X, WANG X, et al. Event-triggered single-network ADP method for constrained optimal tracking control of continuous-time non-linear systems[J]. Applied Mathematics and Computation, 2019, 352: 220-234.

[20] YANG X, ZHAO B. Optimal neuro-control strategy for nonlinear systems with asymmetric input constraints[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(2): 575-583.

[21] ZHANG K, ZHANG H, XIAO G, et al. Tracking control optimization scheme of continuous-time nonlinear system via online single network adaptive critic design method[J]. Neurocomputing, 2017, 251: 127-135.

[22] VAMVOUDAKIS K G, VRABIE D, LEWIS F L. Online adaptive algorithm for optimal control with integral reinforcement learning: Online adaptive algorithm for optimal control[J]. International Journal of Robust and Nonlinear Control, 2014, 24(17): 2686-2710.

[23] XUE S, LUO B, LIU D, et al. Neural network-based event-triggered integral reinforcement learning for constrained H∞ tracking control with experience replay[J]. Neurocomputing, 2022, 513: 25-35.

[24] ZHONG W, WANG M, WEI Q, et al. A new neuro-optimal nonlinear tracking control method via integral reinforcement learning with applications to nuclear systems[J]. Neurocomputing, 2022, 483: 361-369.

[25] GUO X, YAN W, CUI R. Integral reinforcement learning-based adaptive NN control for continuous-time nonlinear MIMO systems with unknown control directions[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(11): 4068-4077.

[26] WANG N, GAO Y, ZHANG X. Data-driven performance-prescribed reinforcement learning control of an unmanned surface vehicle[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5456-5467.

作者简介

蔡军,2000年于电子科技大学获得学士学位,2004年于重庆邮电大学获得硕士学位,现为重庆邮电大学教授,主要研究方向为模式识别、机器人应用。

E-mail: 1073230317@qq.com

Cai Jun received his B. Sc. degree from University of Electronic Science and Technology of China in 2000 and M. Sc. degree from Chongqing University of Posts and Telecommunications in 2004, respectively. Now he is a professor in Chongqing University of Posts and Telecommunications. His main research interests include pattern recognition and robotics and its applications.

苟文耀(通信作者),2016 年于山东建筑大学获得学士学位,现为重庆邮电大学在读研究生,主要研究方向为强化学习和自适应动态规划。

E-mail:501421956@qq.com

Gou Wenyao (Corresponding author) received his B. Sc. degree from Shandong Jianzhu University in 2016. Now he is a M. Sc. candidate at Chongqing University of Posts and Telecommunications. His main research interests include reinforcement learning and adaptive dynamic programming.

刘颜,2016 年于山东科理工大学获得学士学位,现为重庆邮电大学在读研究生,主要研究方向为强化学习和自适应动态规划。

Liu Yan received her B. Sc. degree from Shandong University of Science and Technology in 2016. Now she is a M. Sc. candidate at Chongqing University of Posts and Telecommunications. Her main research interests include reinforcement learning and adaptive dynamic programming.