

面向非线性系统的知识驱动高效安全评判学习控制

王鼎^{1, 2, 3, 4} 李鑫^{1, 2, 3, 4} 赵明明^{1, 2, 3, 4} 乔俊飞^{1, 2, 3, 4}

摘要 针对具有状态约束的离散时间非线性系统, 开发一种知识驱动高效安全评判学习控制算法. 首先, 基于知识迁移技术, 利用历史任务中的先验知识来改善评判学习控制算法在当前目标任务中的学习效率. 同时, 构造一类单调递减的衰减函数以避免先验知识影响控制策略的最优性. 其次, 通过设计合适的障碍函数, 成功将具有状态约束的最优调节问题转化为无约束最优调节问题, 进而建立一种安全评判学习控制框架. 利用所提控制算法, 可高效求解一类具有状态约束的非线性系统最优控制问题. 此外, 为提供严谨的理论依据, 给出所提控制算法的收敛性证明以及控制策略的稳定性判别准则. 最后, 利用两个数值仿真实例验证了所提算法的有效性.

关键词 自适应动态规划; 控制障碍函数; 安全评判学习控制; 知识迁移; 非线性系统

引用格式 王鼎, 李鑫, 赵明明, 乔俊飞. 面向非线性系统的知识驱动高效安全评判学习控制. 自动化学报, 2026, 52(8): 1757–1768

DOI 10.16383/j.aas.c250724 **CSTR** 32138.14.j.aas.c250724

Knowledge-driven Efficient Safe Critic Learning Control for Nonlinear Systems

WANG Ding^{1, 2, 3, 4} LI Xin^{1, 2, 3, 4} ZHAO Ming-Ming^{1, 2, 3, 4} QIAO Jun-Fei^{1, 2, 3, 4}

Abstract In this paper, a knowledge-driven efficient safe critic learning control algorithm is developed for discrete-time nonlinear systems subject to state constraints. First, leveraging knowledge transfer techniques, prior knowledge from historical tasks is utilized to enhance the learning efficiency of the critic learning control algorithm in the current target task. Meanwhile, a class of monotonically decreasing decay functions is constructed to avoid the impact of prior knowledge on the optimality of the control strategy. Secondly, by designing an appropriate barrier function, the original state-constrained optimal regulation problem is transformed into an unconstrained one, thereby enabling the development of a safe critic learning control framework. By leveraging the proposed control algorithm, the optimal control problem for a class of nonlinear systems with state constraints can be efficiently solved. Furthermore, to provide a rigorous theoretical basis, the convergence proof of the proposed control algorithm and the stability criterion of the control strategy are presented. Finally, the effectiveness of the proposed algorithm is verified through two numerical simulation examples.

Keywords adaptive dynamic programming; control barrier functions; safe critic learning control; knowledge transfer; nonlinear systems

Citation Wang Ding, Li Xin, Zhao Ming-Ming, Qiao Jun-Fei. Knowledge-driven efficient safe critic learning control for nonlinear systems. *Acta Automatica Sinica*, 2026, 52(8): 1757–1768

随着强化学习与自适应动态规划的发展, 评判

学习控制 (critic learning control, CLC)^[1–3] 成为非线性最优控制设计的重要方法之一. CLC 的核心思想是通过近似求解 HJB (Hamilton-Jacobi-Bellman) 方程^[4–5], 以迭代方式逼近最优值函数和最优控制策略, 在系统模型未知或动态特性复杂的场景中展现出良好适用性^[6–8]. 然而, 在实际控制系统中, 状态变量与控制输入常受物理边界、执行器饱和等约束条件限制^[9–10]. 控制输入约束^[11–12] 可通过投影映射、饱和函数或非线性变换等方式直接进行处理, 而状态约束因依赖系统动态演化过程, 难以在控制器设计中显式地对其进行处理, 成为非线性控制领域更具挑战性的研究难题^[13–14].

为解决状态约束下的非线性最优控制问题, 相关学者提出多种方法. 例如, 模型预测控制^[15–17] 通

收稿日期 2025-12-15 录用日期 2026-03-04
Manuscript received December 15, 2025; accepted March 4, 2026

国家自然科学基金 (62473012, 62222301), 北京市自然科学基金 (F251019), 新一代人工智能国家科技重大专项 (2021ZD0112302, 2021ZD0112301) 资助

Supported by National Natural Science Foundation of China (62473012, 62222301), Beijing Natural Science Foundation (F251019), and National Science and Technology Major Project (2021ZD0112302, 2021ZD0112301)

本文责任编辑 杨涛

Recommended by Associate Editor YANG Tao

1. 北京工业大学信息科学技术学院 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. 智慧环保北京实验室 北京 100124 4. 北京人工智能研究院 北京 100124

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

过在线滚动优化并显式嵌入状态约束,可有效保障安全性,但实时计算负担较重,难以满足快速动态响应需求. 基于风险度量的安全强化学习方法^[18-20]通过引入条件风险价值,在策略优化中显式考虑违反约束的代价,提升了在不确定环境下的系统安全性. 然而,此类方法通常仅提供统计意义上的高概率安全保证,难以严格确保状态轨迹始终满足硬性约束. 传统控制障碍函数(control barrier function, CBF)方法^[21-23]虽能满足状态约束,却往往以牺牲控制策略性能为代价,难以在安全性与最优性之间实现有效平衡.

近年来,有学者将 CBF 与 CLC 相结合,构建了安全 CLC (safe CLC, SCLC) 框架^[24-25]. 在该类方法中,通过将 CBF 项嵌入效用函数,将带状态约束的最优调节问题转化为无约束形式,从而适配 CLC 的迭代求解机制并提供严格安全保证^[26]. 总的来说,此类算法在状态约束下的非线性最优控制问题取得良好应用效果. 然而,现有 SCLC 算法的迭代过程通常从零初始值函数出发,缺乏有效先验信息引导,导致其在复杂非线性系统中的收敛速度较慢,限制了 SCLC 算法的实际应用^[27].

为提升 CLC 算法的学习效率,部分学者尝试设计加速学习机制以改善其收敛性能^[28-29]. Luo 等^[30]提出一种多步策略评估机制,通过在值函数更新中引入更多精确信息,有效提升了值迭代算法的计算效率. 然而,多步策略评估机制的引入,导致迭代计算中状态向量经常超出 CBF 的定义域,使得算法的迭代过程崩溃. 因此,该加速学习机制与融合 CBF 的 SCLC 算法的适配性较差^[31]. Ha 等^[32]提出一种融合松弛因子的可调值迭代算法,通过调整松弛因子可灵活调节收敛速度,但当松弛因子取值较大时,算法的收敛性无法得到严格保证. Gao 等^[33]提出一种融合 knowledge shaping 的强化学习控制方法,通过嵌入先验知识实现迭代加速,为高效 SCLC 算法的设计提供了新思路. 遗憾的是,该工作中关于算法收敛性的分析过程有待完善,且未能为状态约束场景提供有效的性能保障. 在此背景下,本文围绕非线性系统在状态约束下的最优安全控制问题,开展以下研究工作:

1) 受迁移学习启发,本文开发一种知识驱动高效 SCLC (knowledge-driven efficient SCLC, KD-ESCLC) 算法. 该算法利用在历史任务中的先验值函数信息,通过设计单调递减的衰减函数,将先验知识以可控方式注入当前任务学习过程. 在提升值函数收敛速度的同时,确保其最终收敛至最优值函数,避免先验知识对当前任务中控制策略最优性的

干扰.

2) 针对非对称状态约束,通过构造一种含安全系数的 CBF 并将其融入效用函数,建立一个可权衡安全裕度与控制成本的 SCLC 框架. 该框架适用于模型未知的非线性系统,能够通过迭代学习获得近似最优安全控制策略.

3) 从理论与实验两方面验证所提算法的有效性. 在理论方面,严格证明了 KD-ESCLC 算法的收敛性,并给出控制策略稳定性的判别准则,为算法终止迭代提供理论依据;在实验方面,基于两个离散时间系统仿真实例,结合超参数分析验证了所提算法在加速收敛和安全保障方面的有效性.

符号说明. $\mathbf{R}^n$ 和 $\mathbf{R}^{n \times m}$ 分别表示 n 维欧氏空间和 $n \times m$ 维实矩阵空间; $\|\cdot\|$ 表示向量或者矩阵的范数; $\mathbf{T}$ 是转置符号; $\mathbf{I}_n$ 表示 n 维单位矩阵; $\mathbf{R}$ 表示实数集合; $\mathbf{N}$ 表示非负整数集合.

1 问题描述

为描述研究问题,本节先构建在无约束场景下的非线性最优控制问题,明确最优控制策略的核心求解目标. 在此基础上,引入状态约束条件,建立符合实际需求的控制问题框架.

1.1 无约束非线性最优控制问题

考虑如下一类离散时间非线性系统:

$$s_{k+1} = F(s_k, a_k), \quad k = 0, 1, 2, \dots \quad (1)$$

其中, $s_k \in \Omega_s \subset \mathbf{R}^n$ 表示系统状态向量, $a_k \in \Omega_a \subset \mathbf{R}^m$ 表示控制输入向量, $F(\cdot, \cdot)$ 表示未知的系统函数, Ω_s 和 Ω_a 是两个紧集.

假设 1. 系统函数 $F(s, a)$ 在紧集 Ω_s 上是可微的,且满足 $F(0, 0) = 0$, 即状态变量 $s = 0$ 为控制输入 $a = 0$ 作用下的一个平衡点.

假设 2. 系统 (1) 是可镇定的,即至少存在一个反馈控制策略 $\pi(s_k)$, 使得闭环系统渐近稳定.

为量化控制策略的性能,定义值函数如下:

$$V(s_k) = \sum_{h=k}^{\infty} U(s_h, a_h) = U(s_k, a_k) + V(s_{k+1}) \quad (2)$$

其中, $U(s_k, a_k)$ 表示 k 时刻的效用函数,其形式为

$$U(s_k, a_k) = s_k^T Q s_k + a_k^T R a_k \quad (3)$$

其中, $Q \in \mathbf{R}^{n \times n}$ 和 $R \in \mathbf{R}^{m \times m}$ 表示正定有界的权重矩阵. $\|Q\|$ 的大小反映了对系统状态调节速度与精度的重视程度, $\|R\|$ 则对应控制输入能量消耗在效用评估中的权重占比. 系统对状态收敛速度的要求越高、对状态偏差的约束越严格, $\|Q\|$ 相对 $\|R\|$ 取值越大.

CLC 算法的核心目标是获取最优控制策略 $\pi^*(s_k)$, 使得从任意状态 $s_k \in \Omega_s$ 到达平衡点的累积效用最小. 为此, 最优值函数 $V^*(s_k)$ 和最优控制策略 $\pi^*(s_k)$ 分别为

$$V^*(s_k) = \min_a \sum_{h=k}^{\infty} U(s_h, a_h) = \min_{a_k} \{U(s_k, a_k) + V^*(s_{k+1})\} \quad (4)$$

和

$$\pi^*(s_k) = \arg \min_a \sum_{h=k}^{\infty} U(s_h, a_h) = \arg \min_{a_k} \{U(s_k, a_k) + V^*(s_{k+1})\} \quad (5)$$

上述模型构成无约束非线性最优控制问题框架, 求解出最优值函数 $V^*(s_k)$ 后即可获得对应的最优控制策略 $\pi^*(s_k)$. 然而, 该模型未考虑实际系统中普遍存在的状态约束, 忽略此类约束将严重降低控制策略的实际适用性. 为此, 下文进一步构建具有状态约束的非线性最优控制问题模型.

1.2 状态约束下的非线性最优控制问题

针对系统 (1) 的状态约束, 定义满足所有约束条件的状态集合为安全集 Θ . 不失一般性, 该安全集可表示为

$$\Theta = \{s_k | \phi(s_k^{[j]}) \geq 0, \forall j \in \Phi\} \quad (6)$$

其中, $\phi(s_k^{[j]}) = [b_j - s_k^{[j]}, s_k^{[j]} - c_j]^T$ 是关于第 j 维状态 $s_k^{[j]}$ 的连续可微函数, 常数 $c_j < 0$ 和 $b_j > 0$ 分别表示该维状态的下界与上界; $\Phi = \{d_1, d_2, \dots, d_z, \dots, d_Z\}$ 表示受约束状态的维度指标集合, 满足 $d_z \in \{1, 2, \dots, n\}$ 且 $Z \leq n$. 进一步定义安全集 Θ 的内部与边界分别为

$$\text{Int}\Theta = \{s_k | \phi(s_k^{[j]}) > 0, \forall j \in \Phi\} \quad (7)$$

和

$$\partial\Theta = \{s_k | \phi(s_k^{[j]}) = 0, \exists j \in \Phi\} \quad (8)$$

其中, $\text{Int}\Theta$ 表示严格满足所有状态约束的内点安全区域, 而 $\partial\Theta$ 为至少一个状态分量处于约束边界的临界状态集合. 引入状态约束后, 最优控制策略 $\pi^*(s_k)$ 需在满足约束条件的前提下最小化累积效用, 可表示为

$$\begin{aligned} \pi^*(s_k) &= \arg \min_a \sum_{h=k}^{\infty} U(s_h, a_h) \\ \text{s.t. } (1), \quad s_k &\in \Theta \end{aligned} \quad (9)$$

由于存在约束条件, 无法直接利用 CLC 框架

迭代求解式 (9) 的数值解. 在文献 [10–11] 中, 控制输入约束通常可通过执行器饱和、投影映射或非线性变换等方式直接处理. 相比之下, 状态约束则因依赖于未知系统动态演化, 难以通过显式设计来实现, 构成更具挑战性的问题. 为解决这一难题, 引入 CBF 作为约束处理工具.

定义 1. κ 类函数^[24]. 若连续函数 $\chi: [0, x) \rightarrow [0, \infty)$ 是严格递增的, 且 $\chi(0) = 0$, 则称其为 κ 类函数.

定义 2. CBF^[25]. 若函数 $B(s^{[j]})$ 满足以下条件, 则称其为 CBF:

$$\begin{cases} \frac{1}{\chi_1(\phi(s^{[j]}))} \leq B(s^{[j]}) \leq \frac{1}{\chi_2(\phi(s^{[j]}))} \\ \dot{B}(s^{[j]}) \leq \chi_3(\phi(s^{[j]})) \end{cases} \quad (10)$$

其中, $\chi_1(\cdot)$ 、 $\chi_2(\cdot)$ 和 $\chi_3(\cdot)$ 为 κ 类函数, $\forall s \in \text{Int}\Theta$.

为适配 CLC 算法的迭代求解机制, 所设计的 CBF 需满足以下兼容条件:

$$\begin{cases} B(s^{[j]}) \geq 0, & s^{[j]} \in (c_j, b_j) \\ B(0) = 0, & s^{[j]} = 0 \\ B(s^{[j]}) \rightarrow \infty, & s^{[j]} \rightarrow c_j \text{ 或 } s^{[j]} \rightarrow b_j \end{cases} \quad (11)$$

受相关研究^[24–25, 29]的启发, 设计含安全系数的 CBF 如下:

$$B(s^{[j]}) = \alpha \left(\ln \left(\frac{b_j}{b_j - s^{[j]}} \right) + \ln \left(\frac{c_j}{c_j - s^{[j]}} \right) - \frac{s^{[j]}}{b_j} - \frac{s^{[j]}}{c_j} \right) \quad (12)$$

其中, $\alpha > 0$ 表示安全系数. 当 $c_j = -5$ 、 $b_j = 2$ 时, $B(s^{[j]})$ 的函数图像如图 1 所示. 可以发现, 其严格满足条件 (11). 与文献 [24–25] 中的 CBF 相比, $B(s^{[j]})$ 不仅适用于非对称状态约束场景, 而且可通过调节安全系数 α 权衡控制策略的安全裕度与控制成本. 通过融合 $B(s^{[j]})$, 定义一个安全效用函数

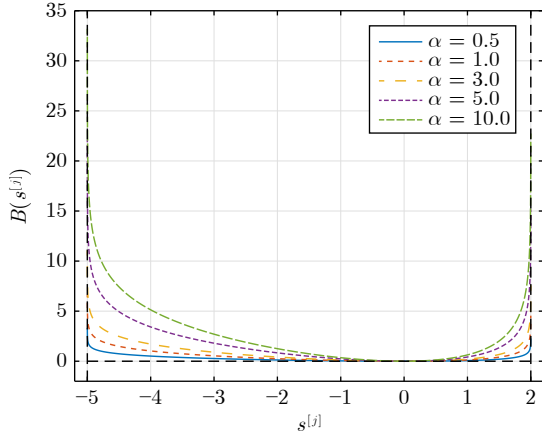
$$U_B(s_k, a_k) = s_k^T Q s_k + a_k^T R a_k + \sum_{z=1}^Z B(s_k^{[d_z]}) \quad (13)$$

对应的安全值函数为

$$V_B(s_k) = \sum_{h=k}^{\infty} U_B(s_h, a_h) = U_B(s_k, a_k) + V_B(s_{k+1}) \quad (14)$$

基于上述定义, 最优安全值函数 $V_B^*(s_k)$ 和最优安全控制策略 $\pi_B^*(s_k)$ 分别表示为

$$V_B^*(s_k) = \min_{a_k} \{U_B(s_k, a_k) + V_B^*(s_{k+1})\} \quad (15)$$

图 1 $B(s^{[j]})$ 在不同安全系数下的轨迹Fig.1 Trajectories of $B(s^{[j]})$ under different safety coefficients

和

$$\pi_B^*(s_k) = \arg \min_{a_k} \{U_B(s_k, a_k) + V_B^*(s_{k+1})\} \quad (16)$$

由式 (16) 可知, 获取最优安全控制策略 $\pi_B^*(s_k)$ 的关键是求解式 (15) 对应的最优安全值函数, 但该方程难以通过解析方法直接求解. 为此, 本文构建 KD-ESCLC 算法, 用于迭代逼近式 (15) 的数值解.

2 控制算法设计

在本节中, 首先对常规 SCLC 算法进行了回顾. 其次, 通过引入知识迁移技术, 将先验知识嵌入 SCLC 算法框架, 进而完成 KD-ESCLC 算法的开发. 为确保所提算法在提升收敛速度的同时不破坏原有算法的收敛性, 本文严格证明了 KD-ESCLC 算法的收敛性. 最后, 针对 KD-ESCLC 算法的迭代控制策略, 提出一种稳定性判别准则, 为算法迭代终止提供理论依据与科学指导.

2.1 SCLC 算法

构造迭代值函数序列 $\{V_B^{(l)}(s_k)\}$ 和迭代控制策略序列 $\{\pi_B^{(l)}(s_k)\}$, 其中 $l \in \mathbf{N}$ 表示迭代指标. 设定初始值函数 $V_B^{(0)}(\cdot) = 0$. 在迭代过程中, 策略提升的具体形式为

$$\begin{aligned} \pi_B^{(l)}(s_k) = \arg \min_{a_k} \{ & U_B(s_k, a_k) + V_B^{(l)}(s_{k+1}) \} = \\ & - \frac{1}{2} R^{-1} \left(\frac{\partial F(s_k, a_k)}{\partial a_k} \right)^T \frac{\partial V_B^{(l)}(s_{k+1})}{\partial s_{k+1}} \end{aligned} \quad (17)$$

值函数更新的规则可表示为

$$\begin{aligned} V_B^{(l+1)}(s_k) = \min_{a_k} \{ & U_B(s_k, a_k) + V_B^{(l)}(s_{k+1}) \} = \\ & U_B(s_k, \pi_B^{(l)}(s_k)) + V_B^{(l)}(s_{k+1}) \end{aligned} \quad (18)$$

当迭代指标 $l \rightarrow \infty$ 时, 迭代值函数 $V_B^{(l)}(s_k)$ 将收敛至最优安全值函数 $V_B^*(s_k)$, 迭代控制策略 $\pi_B^{(l)}(s_k)$ 也将收敛至最优安全控制策略 $\pi_B^*(s_k)$. 借助以下引理 [31] 的核心结论, 说明了 SCLC 算法的收敛性及相关特性.

引理 1. 对于任意 $s_k \in \text{Int}\Theta$, 当以下条件成立时:

$$0 \leq V_B^*(s_{k+1}) \leq \eta U_B(s_k, a_k) \quad (19)$$

$$\tau_1 V_B^*(s_k) \leq V_B^{(0)}(s_k) \leq \tau_2 V_B^*(s_k) \quad (20)$$

其中, η , τ_1 和 τ_2 为常数, 且 $\eta < \infty$, $0 \leq \tau_1 \leq 1 < \tau_2 < \infty$, 则迭代值函数 $V_B^{(l)}(s_k)$ 满足以下不等式:

$$\begin{aligned} \left(1 - \frac{1 - \tau_1}{(1 + \eta^{-1})^l} \right) V_B^*(s_k) \leq V_B^{(l)}(s_k) \leq \\ \left(1 - \frac{1 - \tau_2}{(1 + \eta^{-1})^l} \right) V_B^*(s_k) \end{aligned} \quad (21)$$

当 $l \rightarrow \infty$ 时, 有 $\lim_{l \rightarrow \infty} (1 - (1 - \tau_1)/(1 + \eta^{-1})^l) V_B^*(s_k) = V_B^*(s_k) = \lim_{l \rightarrow \infty} (1 - (1 - \tau_2)/(1 + \eta^{-1})^l) V_B^*(s_k)$. 根据式 (21), 可进一步得出 $\lim_{l \rightarrow \infty} V_B^{(l)}(s_k) = V_B^*(s_k)$, 从而完成收敛性的证明.

2.2 KD-ESCLC 算法

随着非线性系统规模的持续扩大, 其动态特性与控制复杂度也随之显著提升. SCLC 算法虽能通过迭代求解的方式获得式 (15) 的数值解, 进而得到近似最优安全值函数与最优安全控制策略, 但该算法采用零初始化的值函数, 即初始值函数不含任何有效信息, 导致迭代收敛缓慢. 为解决这一问题, 本节设计了一种能够利用历史任务的先验知识来改善收敛速度的 KD-ESCLC 算法. 在不显著增加计算成本的前提下, 该算法能够有效提升迭代计算的效率, 同时保障控制策略的最优性与安全性.

2.2.1 先验知识的获取与处理

先验知识来源于与当前目标控制任务具有一定相似性的历史控制任务. 对于该历史任务, 其受控系统的状态演化方程可表示为

$$s'_{k+1} = F'(s'_k, a'_k) \quad (22)$$

其中, s'_k 为历史系统的状态向量, a'_k 为对应的控制输入, $F'(\cdot, \cdot)$ 为历史系统的未知动态函数. 设 $\pi_P(s'_k)$ 为能使闭环系统 $s'_{k+1} = F'(s'_k, \pi_P(s'_k))$ 实现

渐近稳定的已知控制策略, 基于此, 可定义历史任务的先验值函数为

$$V_P(s'_k) = \sum_{h=k}^{\infty} U_P(s'_h, \pi_P(s'_h)) \quad (23)$$

其中, $U_P(\cdot, \cdot)$ 为历史任务中的效用函数. 需注意的, 目标任务与历史任务的系统动态、优化目标和约束条件等可能存在差异. 若直接将先验值函数 $V_P(s'_k)$ 用于目标任务的迭代学习, 易导致迭代过程不稳定, 甚至偏离最优控制策略. 因此, 需要从先验值函数中提取更具适配性的先验效用函数, 具体形式为

$$U_P(s') = V_P(s') - V_P(F'(s', \pi_P(s'))) \quad (24)$$

该先验效用函数保留了在历史任务中与控制性能相关的核心信息, 可用于引导常规 SCLC 算法的迭代过程, 实现先验知识的有效迁移.

2.2.2 KD-ESCLC 算法的构建

通过融合先验知识与常规 SCLC 算法, 设计高效效用函数 $U_E(s_k, a_k)$, 其表达式为

$$U_E(s_k, a_k) = U_B(s_k, a_k) + \beta(l)U_P(s_k) \quad (25)$$

其中, $\beta(l) = Ae^{-ql}$ 为单调递减的衰减函数, 用于调整先验知识的影响权重; $A > 0$ 为衰减函数的初始幅值; $q > 0$ 为衰减速率. 随着迭代指标 l 增加, $\beta(l)$ 逐渐趋近于 0, 可避免先验知识对目标任务最优控制策略的持续干扰, 确保最终收敛至目标任务的最优解.

基于高效效用函数 $U_E(s_k, a_k)$, 重新构造 KD-ESCLC 算法的迭代值函数序列 $\{V_E^{(l)}(s_k)\}$ 与迭代控制策略序列 $\{\pi_E^{(l)}(s_k)\}$. 初始值函数仍设为 $V_E^{(0)}(s_k) = 0$. $\{\pi_E^{(l)}(s_k)\}$ 和 $\{V_E^{(l)}(s_k)\}$ 的迭代更新方式分别表示为

$$\begin{aligned} \pi_E^{(l)}(s_k) = \arg \min_{a_k} \left\{ U_E(s_k, a_k) + V_E^{(l)}(s_{k+1}) \right\} = \\ - \frac{1}{2} R^{-1} \left(\frac{\partial F(s_k, a_k)}{\partial a_k} \right)^T \frac{\partial V_E^{(l)}(s_{k+1})}{\partial s_{k+1}} \end{aligned} \quad (26)$$

和

$$\begin{aligned} V_E^{(l+1)}(s_k) = \min_{a_k} \left\{ U_E(s_k, a_k) + V_E^{(l)}(s_{k+1}) \right\} = \\ U_E(s_k, \pi_E^{(l)}(s_k)) + V_E^{(l)}(s_{k+1}) \end{aligned} \quad (27)$$

由于高效效用函数 $U_E(s_k, a_k)$ 集成了先验效用函数 $U_P(s_k)$, 迭代值函数序列 $\{V_E^{(l)}(s_k)\}$ 可从先验知识中获取初始引导信息, 从而加快向最优安全值

函数 $V_B^*(s_k)$ 逼近的速度. 同时, 衰减函数 $\beta(l)$ 的单调性确保了当 $l \rightarrow \infty$ 时, $\beta(l) \rightarrow 0$, 此时高效效用函数 $U_E(s_k, a_k)$ 逐渐趋近于安全效用函数 $U_B(s_k, a_k)$, 进而保证 $V_E^{(l)}(s_k) \rightarrow V_B^*(s_k)$ 和 $\pi_E^{(l)}(s_k) \rightarrow \pi_B^*(s_k)$. 下述定理 1 将对该收敛性结论进行严格证明.

定理 1. 给定初始值函数 $V_E^{(0)}(s_k) = 0$. 若迭代控制策略序列 $\{\pi_E^{(l)}(s_k)\}$ 与迭代值函数序列 $\{V_E^{(l)}(s_k)\}$ 分别根据式 (26) 与 (27) 更新, 则 $\lim_{l \rightarrow \infty} V_E^{(l)}(s_k) = V_B^*(s_k)$ 和 $\lim_{l \rightarrow \infty} \pi_E^{(l)}(s_k) = \pi_B^*(s_k)$.

证明. 首先定义高效效用函数的上界函数

$$\bar{U}_E(s_k, a_k) = U_B(s_k, a_k) + AU_P(s_k) \quad (28)$$

然后, 构造一个上界值函数序列 $\{\bar{V}_E^{(l)}(s_k)\}$ 和上界控制策略序列 $\{\bar{\pi}_E^{(l)}(s_k)\}$. 初始上界值函数设为 $\bar{V}_E^{(0)}(s_k) = 0$. 此外, 定义这两个序列迭代更新的方式为

$$\bar{\pi}_E^{(l)}(s_k) = \arg \min_{a_k} \left\{ \bar{U}_E(s_k, a_k) + \bar{V}_E^{(l)}(s_{k+1}) \right\} \quad (29)$$

和

$$\bar{V}_E^{(l+1)}(s_k) = \min_{a_k} \left\{ \bar{U}_E(s_k, a_k) + \bar{V}_E^{(l)}(s_{k+1}) \right\} \quad (30)$$

根据式 (25), 可得

$$U_B(s_k, a_k) \leq U_E(s_k, a_k) \leq \bar{U}_E(s_k, a_k) \quad (31)$$

基于式 (18) 和 (27), 当 $l = 0$ 时, 可得

$$\begin{aligned} V_E^{(1)}(s_k) = \min_{a_k} \left\{ U_E(s_k, a_k) + V_E^{(0)}(s_{k+1}) \right\} = \\ \min_{a_k} \left\{ U_B(s_k, a_k) + \beta(l)U_P(s_k) \right\} \geq \\ \min_{a_k} U_B(s_k, a_k) = \\ \min_{a_k} \left\{ U_B(s_k, a_k) + V_B^{(0)}(s_{k+1}) \right\} = \\ V_B^{(1)}(s_k) \end{aligned} \quad (32)$$

假设对于一个特定 $l \in \mathbf{N}$, $V_E^{(l)}(s_k) \geq V_B^{(l)}(s_k)$ 成立. 然后, 可推导出

$$\begin{aligned} V_E^{(l+1)}(s_k) = \min_{a_k} \left\{ U_E(s_k, a_k) + V_E^{(l)}(s_{k+1}) \right\} \geq \\ \min_{a_k} \left\{ U_B(s_k, a_k) + V_E^{(l)}(s_{k+1}) \right\} \geq \\ \min_{a_k} \left\{ U_B(s_k, a_k) + V_B^{(l)}(s_{k+1}) \right\} = \\ V_B^{(l+1)}(s_k) \end{aligned} \quad (33)$$

根据数学归纳法, 对于任意 $l \in \mathbf{N}$, $V_E^{(l)}(s_k) \geq V_B^{(l)}(s_k)$ 成立.

根据式 (27) 和 (30), 当 $l = 0$ 时, 可得

$$\begin{aligned}
V_E^{(1)}(s_k) &= \min_{a_k} \left\{ U_E(s_k, a_k) + V_E^{(0)}(s_{k+1}) \right\} \leq \\
&\min_{a_k} \left\{ \bar{U}_E(s_k, a_k) + V_E^{(0)}(s_{k+1}) \right\} = \\
&\min_{a_k} \bar{U}_E(s_k, a_k) = \\
&\min_{a_k} \left\{ \bar{U}_E(s_k, a_k) + \bar{V}_E^{(0)}(s_{k+1}) \right\} = \\
&\bar{V}_E^{(1)}(s_k)
\end{aligned} \quad (34)$$

假设对于一个特定 $l \in \mathbf{N}$, $V_E^{(l)}(s_k) \leq \bar{V}_E^{(l)}(s_k)$ 成立. 然后, 可推导出

$$\begin{aligned}
V_E^{(l+1)}(s_k) &= \min_{a_k} \left\{ U_E(s_k, a_k) + V_E^{(l)}(s_{k+1}) \right\} \leq \\
&\min_{a_k} \left\{ \bar{U}_E(s_k, a_k) + V_E^{(l)}(s_{k+1}) \right\} \leq \\
&\min_{a_k} \left\{ \bar{U}_E(s_k, a_k) + \bar{V}_E^{(l)}(s_{k+1}) \right\} = \\
&\bar{V}_E^{(l+1)}(s_k)
\end{aligned} \quad (35)$$

根据数学归纳法, 对于任意 $l \in \mathbf{N}$, $V_E^{(l)}(s_k) \leq \bar{V}_E^{(l)}(s_k)$ 成立.

综上所述, 对于任意 $l \in \mathbf{N}$, $V_B^{(l)}(s_k) \leq V_E^{(l)}(s_k) \leq \bar{V}_E^{(l)}(s_k)$ 成立. 另外, 由引理 1 可知 $\lim_{l \rightarrow \infty} V_B^{(l)}(s_k) = V_B^*(s_k)$ 和 $\lim_{l \rightarrow \infty} \bar{V}_E^{(l)}(s_k) = \bar{V}_E^*(s_k)$, 因此, 可以得出值函数序列 $\{V_E^{(l)}(s_k)\}$ 是有界的. 同时, 考虑到衰减函数 $\beta(l)$ 随迭代指标 l 单调递减且 $\lim_{l \rightarrow \infty} \beta(l) = 0$, 结合引理 1 可推导出, $\lim_{l \rightarrow \infty} V_E^{(l)}(s_k) = V_B^*(s_k)$. 相应地, $\lim_{l \rightarrow \infty} \pi_E^{(l)}(s_k) = \pi_B^*(s_k)$. $\square$

由定理 1 可知, 当 $l \rightarrow \infty$ 时, KD-ESCLC 算法的迭代控制策略 $\pi_E^{(l)}(s_k)$ 将收敛至最优安全控制策略 $\pi_B^*(s_k)$. 但在实际工程中无法实现无限次迭代, 且零初始化值函数导致迭代控制策略稳定性无法保证, 这使得无法确认迭代应该何时终止. 因此, 需设计稳定性判别准则, 判断当前迭代控制策略是否满足稳定性要求, 为算法迭代终止提供科学依据. 在如下的定理 2 中, 给出了一个有效的判断条件.

定理 2. 对于迭代值函数 $V_E^{(l)}(s_k)$ 与迭代控制策略 $\pi_E^{(l)}(s_k)$, 若对任意 $s_k \in \text{Int}\Theta$, 以下不等式均成立:

$$V_E^{(l+1)}(s_k) - V_E^{(l)}(s_k) \leq \psi U_E(s_k, \pi_E^{(l)}(s_k)) \quad (36)$$

其中 $0 < \psi < 1$, 则控制策略 $\pi_E^{(l)}(s_k)$ 可确保当 $k \rightarrow \infty$ 时, $s_k \rightarrow 0$.

证明. 基于式 (27) 与式 (36), 可推导出

$$V_E^{(l)}(s_k) - V_E^{(l)}(s_{k+1}) \geq \bar{\psi} U_E(s_k, \pi_E^{(l)}(s_k)) \quad (37)$$

其中, $\bar{\psi} = 1 - \psi$. 对于 $k = 0, 1, \dots, N$, 有:

$$\begin{aligned}
V_E^{(l)}(s_0) - V_E^{(l)}(s_1) &\geq \bar{\psi} U_E(s_0, \pi_E^{(l)}(s_0)) \\
V_E^{(l)}(s_1) - V_E^{(l)}(s_2) &\geq \bar{\psi} U_E(s_1, \pi_E^{(l)}(s_1)) \\
&\vdots \\
V_E^{(l)}(s_N) - V_E^{(l)}(s_{N+1}) &\geq \bar{\psi} U_E(s_N, \pi_E^{(l)}(s_N))
\end{aligned} \quad (38)$$

基于式 (38), 可得到以下不等式:

$$V_E^{(l)}(s_0) - V_E^{(l)}(s_{N+1}) \geq \bar{\psi} \sum_{h=0}^N U_E(s_h, \pi_E^{(l)}(s_h)) \geq 0 \quad (39)$$

由定理 1 可知, 迭代值函数序列 $\{V_E^{(l)}(s_k)\}$ 是有界的, 因此式 (39) 左侧 $V_E^{(l)}(s_0) - V_E^{(l)}(s_{N+1})$ 的取值有界, 进而求和项 $\bar{\psi} \sum_{h=0}^N U_E(s_h, \pi_E^{(l)}(s_h))$ 有界. 此外, 序列 $\{\bar{\psi} \sum_{h=0}^N U_E(s_h, \pi_E^{(l)}(s_h))\}$ 关于 N 单调非减. 因此, 当 $N \rightarrow \infty$ 时, $\bar{\psi} \sum_{h=0}^N U_E(s_h, \pi_E^{(l)}(s_h))$ 必收敛至有限值, 这意味着当 $k \rightarrow \infty$ 时, $U_E(s_k, \pi_E^{(l)}(s_k)) \rightarrow 0$. 由于 $U_E(s_k, \pi_E^{(l)}(s_k))$ 为正定函数, 故当 $k \rightarrow \infty$ 时, $s_k \rightarrow 0$. $\square$

在定理 2 中, 参数 ψ 是稳定性判别准则的核心阈值, 需严格满足 $0 < \psi < 1$. ψ 越接近 0, 对控制策略的性能要求越严苛, 越能保障系统状态快速收敛至平衡点. 在实际应用中, 低维简单且约束严格的系统可选取较小 ψ 以强化稳定性, 高维复杂系统则取适中值, 平衡稳定性与迭代效率.

3 神经网络实现方案

在所提算法的实现过程中, 需引入神经网络作为函数逼近器, 从而构建模型网络 (model network, MN)、评判网络 (critic network, CN) 与执行网络 (action network, AN), 三者分别用于逼近未知系统动态、迭代值函数与迭代控制策略. MN 的输出表达式为

$$\hat{s}_{k+1} = \varpi_{m2}^T \delta \left(\varpi_{m1}^T [s_k^T, a_k^T]^T + \beta_{m1} \right) + \beta_{m2} \quad (40)$$

其中, $\varpi_{m1} \in \mathbf{R}^{(n+m) \times h_m}$ 和 $\varpi_{m2} \in \mathbf{R}^{h_m \times n}$ 表示 MN 的权重矩阵, h_m 为 MN 隐含层神经元数量; $\beta_{m1} \in \mathbf{R}^{h_m}$ 和 $\beta_{m2} \in \mathbf{R}^n$ 表示 MN 的阈值向量; $\delta(\cdot)$ 表示双曲正切激活函数. MN 的训练数据集 D_1 由被控系统的输入输出实测数据构建, 其形式为 $D_1 = \{[s_k, t, a_k, t, s_{k+1}, t] \mid t = 1, 2, \dots, M_1\}$, 其中 M_1 为数据集 D_1 的样本总量. 定义 MN 的逼近误差为

$$e_m(s_k, t, a_k, t) = \varpi_{m2}^T \delta \left(\varpi_{m1}^T [s_k^T, t, a_k^T, t]^T + \beta_{m1} \right) + \beta_{m2} - s_{k+1, t} \quad (41)$$

进而构建 MN 的性能函数为

$$E_m(s_k, a_k) = \frac{1}{2} \sum_{t=1}^{M_1} \|e_m(s_k, t, a_k, t)\|^2 \quad (42)$$

由于高效效用函数 $U_E(s_k, a_k)$ 中集成了 CBF, 其定义域特性要求 CN 的训练样本必须严格取自安全集的内部区域 $\text{Int}\Theta$. 因此, 用于训练 CN 和 AN 的数据集 D_2 可表示为 $D_2 = \{s_k, t \mid t = 1, 2, \dots, M_2, \forall s_k, t \in \text{Int}\Theta\}$, 其中 M_2 表示数据集 D_2 的样本总数. CN 和 AN 的输出可分别表示为

$$\hat{V}_E^{(l)}(s_k) = (\varpi_{c2}^{(l)})^T \delta \left((\varpi_{c1}^{(l)})^T s_k + \beta_{c1}^{(l)} \right) + \beta_{c2}^{(l)} \quad (43)$$

和

$$\hat{\pi}_E^{(l)}(s_k) = (\varpi_{a2}^{(l)})^T \delta \left((\varpi_{a1}^{(l)})^T s_k + \beta_{a1}^{(l)} \right) + \beta_{a2}^{(l)} \quad (44)$$

其中, $\varpi_{c1}^{(l)} \in \mathbf{R}^{n \times h_c}$ 、 $\varpi_{c2}^{(l)} \in \mathbf{R}^{h_c}$ 、 $\varpi_{a1}^{(l)} \in \mathbf{R}^{n \times h_a}$ 和 $\varpi_{a2}^{(l)} \in \mathbf{R}^{h_a \times m}$ 表示神经网络权重; $\beta_{c1}^{(l)} \in \mathbf{R}^{h_c}$ 、 $\beta_{c2}^{(l)} \in \mathbf{R}$ 、 $\beta_{a1}^{(l)} \in \mathbf{R}^{h_a}$ 和 $\beta_{a2}^{(l)} \in \mathbf{R}^m$ 表示神经网络阈值, h_c 和 h_a 分别表示 CN 和 AN 隐含层的神经元数量. CN 和 AN 的逼近误差分别定义为

$$e_c^{(l)}(s_k, t) = \hat{V}_E^{(l+1)}(s_k, t) - \left[U_E \left(s_k, t, \pi_E^{(l)}(s_k, t) \right) + \hat{V}_E^{(l)}(\hat{s}_{k+1}, t) \right] \quad (45)$$

和

$$e_a^{(l)}(s_k, t) = \hat{\pi}_E^{(l)}(s_k, t) + \frac{1}{2} R^{-1} \left(\frac{\partial \hat{s}_{k+1}, t}{\partial \hat{\pi}_E^{(l)}(s_k, t)} \right)^T \times \frac{\partial V_E^{(l)}(\hat{s}_{k+1}, t)}{\partial \hat{s}_{k+1}, t} \quad (46)$$

对应地, CN 与 AN 的性能指标函数分别为

$$E_c^{(l)}(s) = \frac{1}{2} \sum_{t=1}^{M_2} \|e_c^{(l)}(s_k, t)\|^2 \quad (47)$$

和

$$E_a^{(l)}(s) = \frac{1}{2} \sum_{t=1}^{M_2} \|e_a^{(l)}(s_k, t)\|^2 \quad (48)$$

在 MN、CN 和 AN 的训练中, 采用 Levenberg-Marquardt 算法调整权重矩阵和阈值向量, 以找到使 $E_m(s_k, a_k)$ 、 $E_c^{(l)}(s)$ 和 $E_a^{(l)}(s)$ 最小化的最优权值参数. 需要注意的是, CN 和 AN 需要在每个迭代步都进行训练, 而 MN 只需在迭代开始前完成一次训练即可. 另外, 本节所有神经网络的构建与训练均基于 MATLAB 神经网络工具箱. 在 MN、CN 和 AN 的训练过程中, 需要配置的参数包括隐藏层

神经元 h_m 、 h_c 、 h_a , 最大训练回合数 N_m 、 N_c 、 N_a , 训练目标值 G_m 、 G_c 、 G_a . 为直观体现 KD-ESCLC 算法的逻辑, 通过算法 1 给出其整体执行流程.

算法 1. KD-ESCLC 算法

输入. 初始值函数 $V_E^{(0)}(\cdot)$ 、先验效用函数 $U_P(\cdot)$ 、衰减函数 $\beta(l)$ 、权重矩阵 Q 和 R 、安全系数 α 、最大迭代次数 $l_{\max}$ 、数据集 D_1 和 D_2 .

输出. $\hat{\pi}_B^*(s_k)$ 和 $\hat{V}_B^*(s_k)$.

1. 优化 MN 参数使式 (42) 的性能函数 E_m 最小, 完成 MN 的训练.
2. 令迭代指标 $l = 0$.
3. **for** l **do**
4. 优化 CN 参数使式 (47) 的性能函数 $E_c^{(l)}(s)$ 最小, 完成 CN 的更新.
5. 优化 AN 参数使式 (48) 的性能函数 $E_a^{(l)}(s)$ 最小, 完成 AN 的更新.
6. **if** $\max_{s_k \in D_2} |\hat{V}_E^{(l+1)}(s_k) - \hat{V}_E^{(l)}(s_k)| \leq 10^{-6}$ **then**
7. 获得 $\hat{\pi}_B^*(s_k)$ 和 $\hat{V}_B^*(s_k)$, 停止迭代.
8. **end**
9. **if** $l \geq l_{\max}$ **then**
10. 停止迭代.
11. **end**
12. $l \leftarrow l + 1$
13. **end**

4 仿真实验

为验证所提 KD-ESCLC 算法的迭代学习效率与安全控制性能, 同时探究关键设计参数对算法控制性能的影响, 本节通过两个典型离散时间仿真实例开展系统性验证.

例 1. 考虑如下离散时间倒立摆系统^[31]:

$$s_{k+1} = \begin{bmatrix} s_k^{[1]} + 0.1 s_k^{[2]} \\ -0.612 \, 5 \sin(s_k^{[1]}) + 0.975 \, 0 s_k^{[2]} \end{bmatrix} + \begin{bmatrix} 0 \\ 0.125 \, 0(a_k + \tanh(a_k)) \end{bmatrix} \quad (49)$$

其中, 系统状态 $s_k = [s_k^{[1]}, s_k^{[2]}]^T$, 且 $s_k^{[1]}$ 和 $s_k^{[2]}$ 分别表示当前的角度和角速度; 控制输入为 a_k . 此次仿真旨在利用 KD-ESCLC 算法高效获得近似最优安全控制策略, 在保证约束条件 $s_k^{[2]} \in (-0.45, 0.50)$ 成立的前提下, 以最小代价实现系统状态收敛至平衡点.

算法设计参数与神经网络训练参数分别如表 1、表 2 所示, 且所有神经网络的最大训练回合数均设定为 200. 依据算法 1 的执行流程, 首先需完成

表 1 KD-ESCLC 算法的设计参数

Table 1 Design parameters of the KD-ESCLC algorithm

参数值	Q	R	α	A	q	$l_{\max}$	ψ
例 1	I_2	1	1	2	0.2	40	0.8
例 2	I_4	I_2	2	10	0.1	400	0.6

表 2 神经网络的训练参数

Table 2 Training parameters of neural networks

参数值	h_m	h_c	h_a	G_m	G_c	G_a
例 1	20	30	15	10^{-6}	10^{-8}	10^{-6}
例 2	30	40	20	10^{-7}	10^{-9}	10^{-7}

MN 的训练. 通过 MATLAB 的 rand 函数构建数据集 D_1 , 样本总数设为 $M_1 = 1\,000$, 采样区域限定为 $\{|s_k^{[1]}| \leq 0.6, |s_k^{[2]}| \leq 0.6, |a_k| \leq 0.6\}$, 并按 6 : 4 的比例划分训练集与测试集. MN 训练完成后, 其测试结果如图 2 所示, 逼近误差绝对值可稳定限制在 1.5×10^{-4} 以内. 该结果验证了 MN 对倒立摆系统动态特性的高精度逼近能力, 为后续算法迭代提供了可靠的系统模型支撑.

随后, 继续利用 MATLAB 的 rand 函数建立数据集 D_2 来完成所提算法的迭代训练. 其中样本总数 $M_2 = 500$, 采样区域设为 $\{|s_k^{[1]}| \leq 0.60, -0.43 < s_k^{[2]} < 0.48, |a_k| \leq 0.60\}$. 注意, 数据集 D_2 的采样区域必须严格限定于安全集的内部区域 $\text{Int}\Theta$. 先验效用函数设为 $U_P(s) = V^*(s) - V^*(F(s, \pi^*(s)))$, 其中 $V^*(s)$ 与 $\pi^*(s)$ 为系统 (49) 在无约束、 $Q = I_2$ 且 $R = 1$ 条件下的最优值函数与最优控制策略, 对应式 (4) 与式 (5). 最后, 按照算法 1 完成迭代计算. 为了比

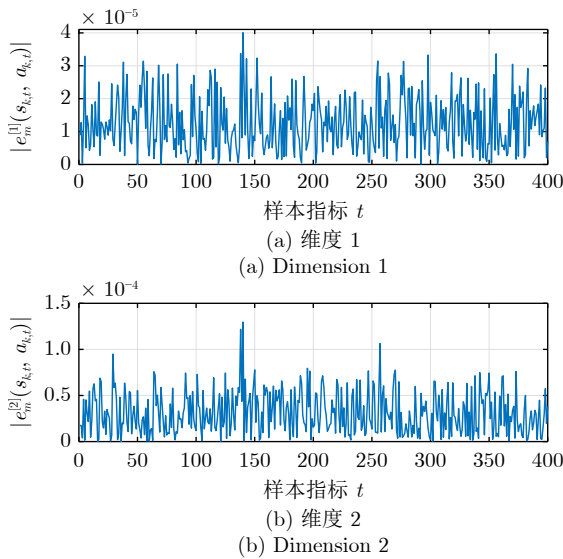
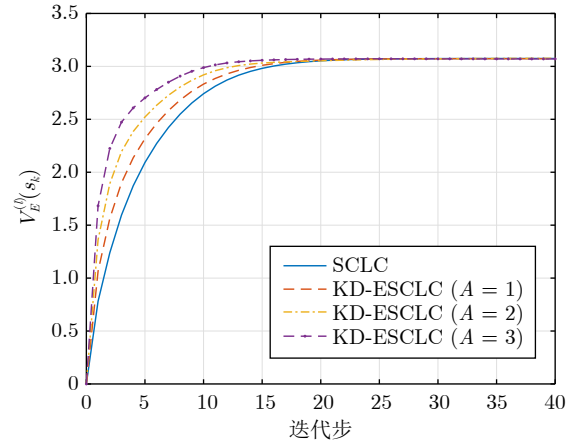


图 2 MN 的测试误差 (例 1)

Fig. 2 The test error of the MN (Example 1)

较 KD-ESCLC 算法与 SCLC 算法的迭代效率, 同时分析衰减函数初始幅值 A 对 KD-ESCLC 算法迭代效率的影响, 本节在不同 A 值下进行了多次迭代计算并对迭代值函数序列 $\{V_E^{(l)}(s_k)\}$ 的收敛过程进行记录. 如图 3 所示, KD-ESCLC 算法收敛速度显著快于常规 SCLC, 且 A 越大收敛越快, 所有曲线最终均收敛至相同值. 实验结果说明增大 A 可增强先验知识初始引导、加快值函数的收敛且不影响其最优性.

图 3 不同 A 值下的迭代值函数收敛轨迹 (例 1)Fig. 3 Convergence trajectories of the iterative value function under different A values (Example 1)

在获取近似最优安全值函数 $\hat{V}_B^*(s_k)$ 后, 由式 (17) 可推导得到最优安全控制策略 $\hat{\pi}_B^*(s_k)$. 为验证其控制性能, 构建包含 50 个不同初始状态的集合 $S_0 = \{s_0, r | r = 1, 2, \dots, 50\} \subset \text{Int}\Theta$, 在 $\hat{\pi}_B^*(s_k)$ 作用下, 系统状态演化轨迹如图 4 所示. 从图 4 中可观察到所有状态轨迹均严格位于安全区域内且最终收敛到平衡点, 充分验证了 KD-ESCLC 算法的安全控制性能. 为凸显 KD-ESCLC 算法在安全性方面的优势, 图 5 给出在常规 CLC 算法下的状态轨迹作为对比. 从图 5 中可清晰地看到, 部分轨迹超出角速度约束边界. 这一结果说明常规 CLC 算法无法满足状态约束, 进一步显示了 KD-ESCLC 算法在安全控制方面的必要性与有效性.

为研究安全系数 α 对算法控制性能的影响, 分别求解了不同 α 值对应的近似最优安全控制策略 $\hat{\pi}_B^*(s_k)$, 并将其作用于系统 (49). 图 6 和图 7 分别展示了不同安全系数的状态轨迹与控制输入轨迹. 可以发现, 随着安全系数 α 增大, 系统状态轨迹会远离安全边界, 同时控制输入幅值也相应增大. 这一现象表明增大 α 可提升系统状态的安全裕度, 但也是需要消耗更多的控制成本.

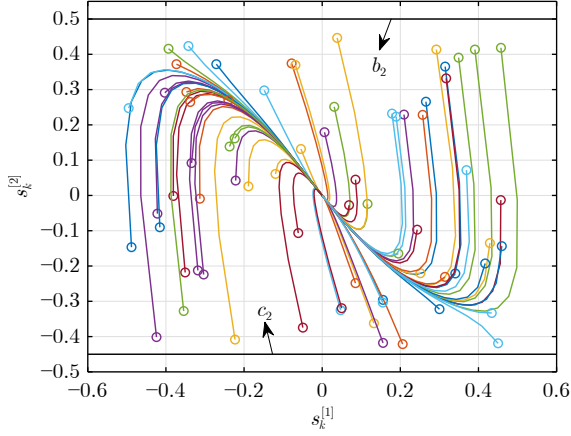


图 4 KD-ESCLC 算法下的系统状态轨迹 (例 1)

Fig.4 Trajectories of the system states under the KD-ESCLC algorithm (Example 1)

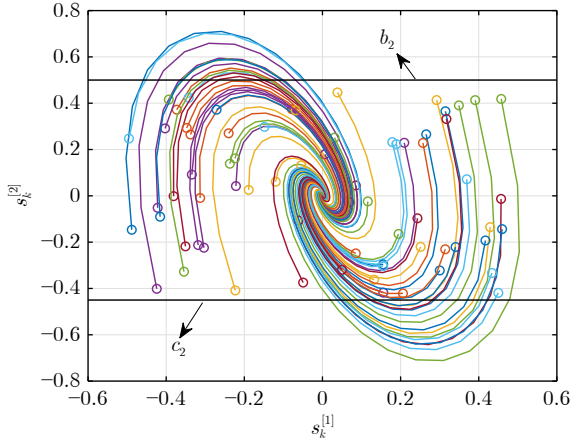


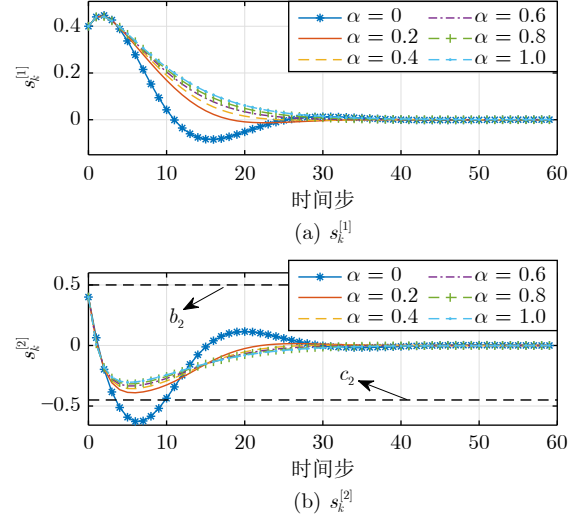
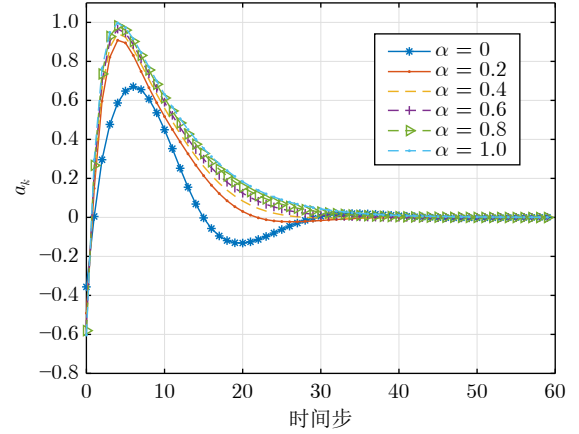
图 5 常规 CLC 算法下的系统状态轨迹 (例 1)

Fig.5 Trajectories of the system states under the conventional CLC algorithm (Example 1)

例 2. 考虑如下离散时间质量-弹簧系统^[20]:

$$s_{k+1} = \begin{bmatrix} 0.9998 & 0.00016 & 0.02 & 0 \\ 0.0004 & 0.9996 & 0 & 0.02 \\ -0.016 & 0.016 & 0.9998 & 0.00016 \\ 0.04 & -0.03999 & 0.0004 & 0.9996 \end{bmatrix} s_k + \begin{bmatrix} 0 & 0 \\ 0 & 0.00019999 \\ 0.0079996 & 0 \\ 0 & 0.019997 \end{bmatrix} a_k \quad (50)$$

其中, 系统状态 $s_k = [s_k^{[1]}, s_k^{[2]}, s_k^{[3]}, s_k^{[4]}]^T$, 控制输入 $a_k = [a_k^{[1]}, a_k^{[2]}]^T$, 设定状态约束为 $s_k^{[2]} \in (-0.80, 0.90)$. 首先利用 MATLAB 的 rand 函数构建 MN 的训练数据集 D_1 , 其样本总量 $M_1 = 3000$, 采样区域为 $\{|s_k^{[1]}| \leq 1, |s_k^{[2]}| \leq 1, |s_k^{[3]}| \leq 1, |s_k^{[4]}| \leq 1, |a_k^{[1]}| \leq 1,$

图 6 不同安全系数 α 下的系统状态轨迹 (例 1)Fig.6 Trajectories of the system states under different safety coefficients α (Example 1)图 7 不同安全系数 α 下的控制输入轨迹 (例 1)Fig.7 Trajectories of the control input under different safety coefficients α (Example 1)

$|a_k^{[2]}| \leq 1\}$. 当 MN 训练完成后, 其测试结果如图 8 所示, 逼近误差绝对值可稳定限制在 2.5×10^{-5} 以内, 这表明 MN 可精准拟合系统 (50) 的动态特性. 继而建立数据集 D_2 并完成 KD-ESCLC 算法的迭代计算, 其中样本总数 $M_2 = 15000$, 采样区域为 $\{|s_k^{[1]}| \leq 1.00, -0.78 < s_k^{[2]} < 0.88, |s_k^{[3]}| \leq 1.00, |s_k^{[4]}| \leq 1.00, |a_k^{[1]}| \leq 1.00, |a_k^{[2]}| \leq 1.00\}$. 在例 2 中, 设置了先验效用函数 $U_P(s) = V^*(s) - V^*(F(s, \pi^*(s)))$, 其中 $V^*(s)$ 和 $\pi^*(s)$ 分别表示系统 (50) 在无约束条件、 $Q = I_4$ 且 $R = I_2$ 条件下的最优值函数和最优控制策略.

为验证衰减函数初始幅值 A 对高维系统算法收敛效率的影响, 在不同 A 值下开展迭代实验. 值函数序列 $\{V_E^{(l)}(s_k)\}$ 的收敛过程如图 9 所示, 随着

A 值增大, 迭代值函数逼近最优值函数的速度显著加快. 还可发现, 引入先验知识后虽然无法保证迭代值函数序列的单调性, 但并不破坏值函数序列的收敛性, 这一结果与定理 1 的结论一致.

通过 KD-ESCLC 算法迭代得到近似最优安全控制策略 $\hat{\pi}_B^*(s_k)$ 后, 将其与常规 CLC 算法的近似最优控制策略 $\hat{\pi}^*(s_k)$ 进行对比. 设定统一初始状态 $s_0 = [-0.70, -0.75, -0.60, -0.70]^T$, 该状态位于安全集内且靠近约束边界, 更能体现控制策略的性能差异. 分别记录两种策略下的系统状态与控制输入轨迹, 结果如图 10、图 11 所示. 从图 10 中可以发现, KD-ESCLC 算法作用下的 $s_k^{[2]}$ 始终严格位于

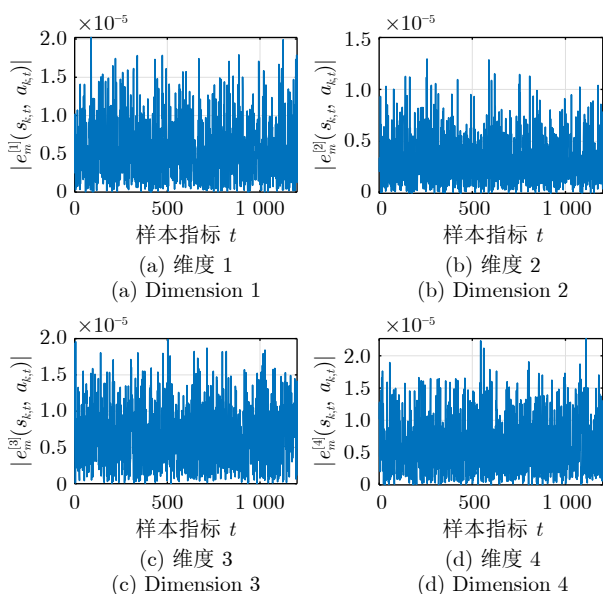


图 8 MN 的测试误差 (例 2)

Fig. 8 The test error of the MN (Example 2)

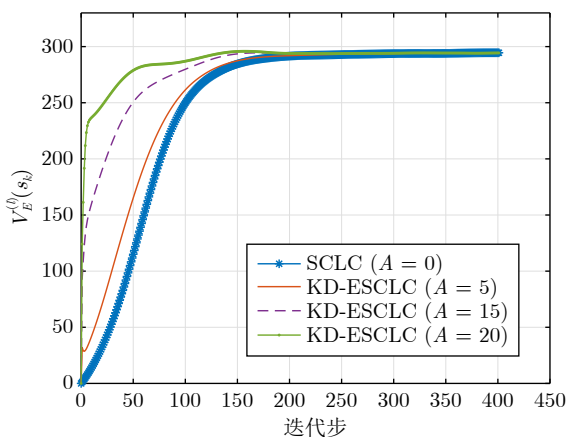


图 9 不同 A 值下的迭代值函数收敛轨迹 (例 2)

Fig. 9 Convergence trajectories of the iterative value function under different A values (Example 2)

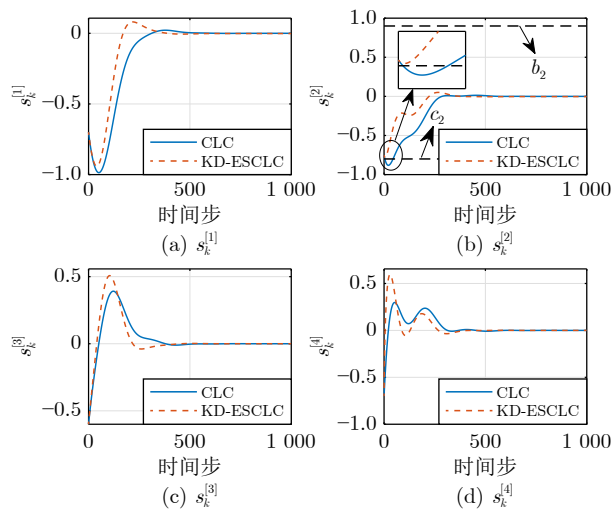


图 10 不同控制算法下的系统状态轨迹 (例 2)

Fig. 10 Trajectories of the system states under different control algorithms (Example 2)

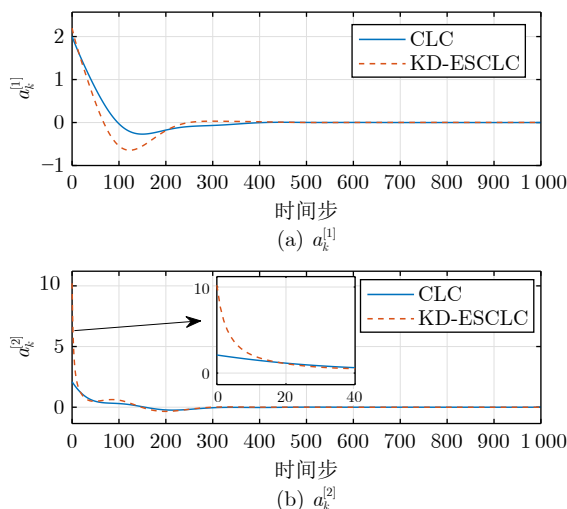


图 11 不同控制算法下的控制输入轨迹 (例 2)

Fig. 11 Trajectories of the control inputs under different control algorithms (Example 2)

$(-0.8, 0.9)$ 的约束范围内, 且随时间步收敛至平衡点; 而常规 CLC 算法的 $s_k^{[2]}$ 轨迹超出了约束边界. 根据图 11 可以发现, 相较于常规近似最优控制策略 $\hat{\pi}^*(s_k)$, 近似最优安全控制策略 $\hat{\pi}_B^*(s_k)$ 会产生幅值更大的控制输入.

例 1 与例 2 的仿真结果表明, KD-ESCLC 算法在未知复杂系统最优控制中性能优异, 无论是倒立摆还是质量-弹簧系统, 均能兼顾迭代效率、策略安全性与最优性.

5 结束语

面向一类具有状态约束的离散时间未知系统,

本文开发了一种 KD-ESCLC 算法. 首先, 该算法引入知识迁移技术, 通过设计单调递减的衰减函数充分利用历史任务的先验知识协助完成当前任务. 在提升算法迭代效率的同时, 避免先验知识对控制策略最优性造成干扰. 其次, 通过构造含安全系数的 CBF, 将带状态约束的最优调节问题转化为无约束形式, 构建了可权衡安全裕度与控制成本的 SCLC 框架, 有效求解满足状态约束的近似最优安全控制策略. 再次, 严格证明了 KD-ESCLC 算法的收敛性, 并给出控制策略的稳定性判别准则, 以提供理论支撑. 最后, 通过倒立摆与质量-弹簧系统的仿真进一步验证了算法的有效性. 相较于常规 CLC 算法, 该算法在迭代收敛速度与状态约束保障方面均表现更优.

参考文献

- Wang D, Gao N, Liu D, Li J, Lewis F. Recent progress in reinforcement learning and adaptive dynamic programming for advanced control applications. *IEEE/CAA Journal of Automatica Sinica*, 2024, **11**(1): 18–36
- Zhao M, Wang D, Qiao J, Ha M, Ren J. Advanced value iteration for discrete-time intelligent critic control: A survey. *Artificial Intelligence Review*, 2023, **56**: 12315–12346
- Wang Ding, Wang Jiang-Yu, Qiao Jun-Fei. Data-driven policy optimization for stochastic systems involving adaptive critic. *Acta Automatica Sinica*, 2024, **50**(5): 980–990
(王鼎, 王将宇, 乔俊飞. 融合自适应评判的随机系统数据驱动策略优化. 自动化学报, 2024, **50**(5): 980–990)
- Heydari A. Stability analysis of optimal adaptive control using value iteration with approximation errors. *IEEE Transactions on Automatic Control*, 2018, **63**(9): 3119–3126
- Wang Ding, Liu Ao, Qiao Jun-Fei. Robust adaptive critic control with hybrid iteration for continuous-time systems. *Acta Automatica Sinica*, 2026, **52**(1): 137–147
(王鼎, 刘奥, 乔俊飞. 连续时间系统混合迭代鲁棒自适应评判控制. 自动化学报, 2026, **52**(1): 137–147)
- Wang D, Hu L, Zhao M, Qiao J. Adaptive critic for event-triggered unknown nonlinear optimal tracking design with wastewater treatment applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(9): 6276–6288
- Liu D, Xue S, Zhao B, Luo B, Wei Q. Adaptive dynamic programming for control: A survey and recent advances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, **51**(1): 142–160
- Wang D, Zhao M, Ha M, Qiao J. Stability and admissibility analysis for zero-sum games under general value iteration formulation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(11): 8707–8718
- Fu J, Wen G, Yu X. Safe consensus tracking with guaranteed full state and input constraints: A control barrier function-based approach. *IEEE Transactions on Automatic Control*, 2023, **68**(12): 8075–8081
- Wang Ding, Zhao Ming-Ming, Liu De-Rong, Qiao Jun-Fei, Song Shi-Jie. Research advances on data-driven adaptive critic control. *Acta Automatica Sinica*, 2025, **51**(6): 1170–1190
(王鼎, 赵明明, 刘德荣, 乔俊飞, 宋世杰. 数据驱动自适应评判控制研究进展. 自动化学报, 2025, **51**(6): 1170–1190)
- Wang Ding, Zhao Hui-Ling, Li Xin. Adaptive critic control for wastewater treatment systems based on multiobjective particle swarm optimization. *Chinese Journal of Engineering*, 2024, **46**(5): 908–917
(王鼎, 赵慧玲, 李鑫. 基于多目标粒子群优化的污水处理系统自适应评判控制. 工程科学学报, 2024, **46**(5): 908–917)
- Shen H, Yu X, Yan H, Park J H, Wang J. Robust fixed-time sliding mode attitude control for a 2-DOF helicopter subject to input saturation and prescribed performance. *IEEE Transactions on Transportation Electrification*, 2025, **11**(1): 1223–1233
- Liang L, Zhao B, Zhou J, Zhang Z. Impact angle controlled integrated guidance and control with input and state constraints. *International Journal of Control*, 2024, **97**(4): 796–810
- Zhu L, Huang D, Li X, Wang Q. Cooperative operation control of virtual coupling high-speed trains with input saturation and full-state constraints. *IEEE Transactions on Automation Science and Engineering*, 2024, **21**(3): 3497–3510
- Castroviejo-Fernandez M, Leung J, Kolmanovsky I. Robust reference governor for input-constrained model predictive control to enforce state constraints at low computational cost. *International Journal of Control*, 2025, **98**(5): 1111–1124
- Wiltz A, Chen F, Dimarogonas D. Parallelized robust distributed model predictive control in the presence of coupled state constraints. *Automatica*, 2025, **171**: Article No. 111952
- Keusch R, Loeliger H, Geyer T. Long-horizon direct model predictive control for power converters with state constraints. *IEEE Transactions on Control Systems Technology*, 2024, **32**(2): 340–350
- Zhang H, Lin Y, Han S, Wang S, Lv K. Off-policy conservative distributional reinforcement learning with safety constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025, **55**(3): 2033–2045
- Kim D, Oh S. Efficient off-policy safe reinforcement learning using trust region conditional value at risk. *IEEE Robotics and Automation Letters*, 2022, **7**(3): 7644–7651
- Kim D, Oh S. TRC: Trust region conditional value at risk for safe reinforcement learning. *IEEE Robotics and Automation Letters*, 2022, **7**(2): 2621–2628
- Li B, Wen S, Yan Z, Wen G, Huang T. A survey on the control lyapunov function and control barrier function for nonlinear-affine control systems. *IEEE/CAA Journal of Automatica Sinica*, 2023, **10**(3): 584–602
- Farzan S, Azimi V, Hu A P, Rogers J. Adaptive control of wireborne underactuated brachiating robots using control Lyapunov and barrier functions. *IEEE Transactions on Control Systems Technology*, 2022, **30**(6): 2598–2614
- Zeng D, Jiang Y, Wang Y, Zhang H, Feng Y. Robust adaptive control barrier functions for input-affine systems: Application to uncertain manipulator safety constraints. *IEEE Control Systems Letters*, 2024, **8**: 279–284
- Marvi Z, Kiumarsi B. Safe reinforcement learning: A control barrier function optimization approach. *International Journal of Robust and Nonlinear Control*, 2021, **31**(6): 1923–1940
- Roshanravan S, Shamaghdari S. Integrated fault detection and fault-tolerant optimal tracking control for unknown nonlinear systems with state and input constraints using safe reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2026, **56**(1): 46–58
- Zhang L, Xie L, Jiang Y, Li Z, Liu X, Su H. Optimal control for constrained discrete-time nonlinear systems based on safe reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(1): 854–865
- Yazdani N, Moghaddam R, Kiumarsi B, Modares H. A safety-certified policy iteration algorithm for control of constrained nonlinear systems. *IEEE Control Systems Letters*, 2020, **4**(3): 686–691

- 28 Wang D, Wang J, Zhao M, Xin P, Qiao J. Adaptive multi-step evaluation design with stability guarantee for discrete-time optimal learning control. *IEEE/CAA Journal of Automatica Sinica*, 2023, **10**(9): 1797–1809
- 29 Zhao M, Wang D, Song S, Qiao J. Accelerated value iteration-based safe Q-learning for data-driven optimal tracking control. *IEEE Transactions on Cybernetics*, 2025, **55**(7): 3511–3524
- 30 Luo B, Liu D, Huang T, Liu J. Output tracking control based on adaptive dynamic programming with multistep policy evaluation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(10): 2155–2165
- 31 Wang D, Li X, Wang H, Qiao J. Safe optimal tracking control via multi-step critic learning for unknown nonlinear systems. *IEEE Transactions on Automation Science and Engineering*, 2025, **22**: 23379–23392
- 32 Ha M, Wang D, Liu D. Novel discounted adaptive critic control designs with accelerated learning formulation. *IEEE Transactions on Cybernetics*, 2024, **54**(5): 3003–3016
- 33 Gao X, Si J, Huang H. Reinforcement learning control with knowledge shaping. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(3): 3156–3167



王 鼎 北京工业大学信息科学技术学院教授。2009 年获得东北大学硕士学位, 2012 年获得中国科学院自动化研究所博士学位。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

School of Information Science and Technology, Beijing University of Technology. He received his master degree from Northeastern University in 2009 and Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2012. His research interests include reinforcement learning and intelligent control. Corresponding author of this paper.)



李 鑫 北京工业大学信息科学技术学院博士研究生。主要研究方向为强化学习和智能控制。

E-mail: lixin229038@emails.bjut.edu.cn

(**LI Xin** Ph.D. candidate at the School of Information Science and Technology, Beijing University of Technology. His research interests include reinforcement learning and intelligent control.)



赵明明 北京工业大学信息科学技术学院博士研究生。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** Ph.D. candidate at the School of Information Science and Technology, Beijing University of Technology. His research interests include reinforcement learning and intelligent control.)



乔俊飞 北京工业大学信息科学技术学院教授。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: adqiao@bjut.edu.cn

(**QIAO Jun-Fei** Professor at the School of Information Science and Technology, Beijing University of Technology. His research interests include intelligent control of wastewater treatment processes, structure design and optimization of neural networks.)