文章编号:1671-4598(2023)07-0143-07 DOI:10.16526/j.cnki.11—4762/tp.2023.07.021 中图分类号:TP273 文献标识码:A

带状态约束的事件触发积分强化学习控制

田奋铭[1,2], 刘 飞[1,2]
(1. 江南大学 轻工过程先进控制教育部重点实验室, 江苏 无锡 214122;
2. 江南大学 自动化研究所, 江苏 无锡 214122)

摘要: 为克服全状态对称约束以及控制策略频繁更新的局限, 针对一类具有部分动力学未知的仿射非线性连续系统的最优控制问题, 提出一种带状态约束的事件触发积分强化学习的控制器设计方法; 该方法是一种基于数据的在线策略迭代方法; 引入系统转换将带有全状态约束的系统转化为不含约束的系统; 基于事件触发机制以及积分强化学习算法, 通过交替执行系统转换、策略评估、策略改进, 最终系统在满足全状态约束的情况下, 代价函数以及控制策略将分别收敛于最优值, 并能降低控制策略的更新频率; 此外, 利用李雅普诺夫函数对系统的稳定性进行严格的分析; 通过单连杆机械臂的仿真实验说明算法的可行性。

关键词: 仿射非线性系统; 最优控制; 事件触发控制; 积分强化学习; 神经网络

Event-triggered Integral Reinforcement Learning Control with State Constraints

TIAN Fenming[1,2], LIU Fei[1,2]
(1. Key Laboratory for Advanced Process Control of Light Industry of the Ministry of Education, Jiangnan University, Wuxi 214122, China;
2. Institute of Automation, Jiangnan University, Wuxi 214122, China)

Abstract: In order to overcome the limitations of full-state symmetry constraints and frequent update of control policy, aimed at an optimal control problem for a class of affine nonlinear continuous systems with partially unknown dynamics, a controller design method for the event-triggered integral reinforcement learning with state constraints is proposed. The method is a data-based online policy iteration approach. Firstly, the system transformation is introduced to transform the constrained system into the unconstrained system. Next, based on the event triggering mechanism and integral reinforcement learning algorithm, the system transformation, policy evaluation, and policy improvement are alternately operated. Finally, the system will satisfy that the full-state constraints, cost function and control policy will converge to the optimal values respectively, and it can reduce the update frequency of the control policy. In addition, the system stability is strictly analyzed by constructing the Lyapunov function. The simulation experiment of the single-link robotic arm verifies the effectiveness of the proposed algorithm.

Keywords: affine nonlinear system; optimal control; event-triggering control; integral reinforcement learning; neural network

0 引言

非线性连续系统的控制问题一直是现代控制理论的基本问题之一。针对非线性连续系统的控制问题有众多针对性的控制方法, 如: PID 控制[1-2]、自适应控制[3-5]、滑模控制[6-7]、以及多种方法综合应用[8-9]。然而, 对于大多数受控的系统, 在控制过程中必然要考虑状态约束, 以防止系统不稳定问题的发生。以路径跟踪任务中的车辆控制为例, 除了考虑跟踪性能外, 还必须将车辆的某些状态限制在稳定区域内。

针对带有状态约束系统的控制问题目前已经产生多种基本理论框架[10-15]。文献 [12] 针对带有时变状态约束的非线性纯反馈系统的跟踪控制问题, 利用矩阵变换以及反步法展开讨论, 最终实现轨迹跟踪控制并且系统的状态始终满足状态约束。文献 [13] 针对带有时不变对称状态约束系统的代价函数优化问题, 通过将状态约束转化为障碍函数并入代价函数, 使用神经网络逼近技术, 基于自适应动态规划算法, 在系统模型完全已知的情况下实现最优控制。文献 [14] 基于矩阵变换以及自适应评价设计算法, 利用 Critic-Actor 神经网络, 有效的解决了非线性纯反馈连续系统的“多人博弈”最优控制问题。模型预测控制 (MPC) 方法作为解决带有状态约束的优化控制问题最常用的方法, 实际上也是利用障碍函数法, 将状态约束并入代价函数中。尽管上述方法都能解决带有状态约束的优化控

收稿日期:2023-02-24; **修回日期**:2023-03-02。
基金项目:国家自然科学基金(61833007)。
作者简介:田奋铭(1997-),男,山西吕梁人,硕士研究生,主要从事强化学习控制方向的研究。
通讯作者:刘 飞(1965-),男,安徽宣城人,博士,教授,主要从事先进过程控制方向的研究。
引用格式:田奋铭,刘 飞.带状态约束的事件触发积分强化学习控制[J].计算机测量与控制,2023,31(7):143-149.

制问题，但都是基于系统动力学完全已知或者利用辨识手段获得动力学信息展开讨论。然而，如今的控制系统大多呈现强耦合、强非线性的特点，如航天航空等，精确的动力学大多难以获得，直接或间接地阻碍了带有状态约束系统的控制问题的研究。以机电伺服系统为例，机电伺服系统是一个多变量、强耦合的系统，系统的参数易受系统所处环境的影响，在考虑伺服系统跟踪控制问题的同时，也必须考虑状态约束问题[16]，因此考虑带状态约束且系统具有不确定性的最优控制问题十分必要。这里的不确定性主要指系统动力学部分未知、系统动力学全部未知、系统某些时变参数变化规律未知等。

近年来，积分强化学习（IRL）算法成为实现仿射非线性系统最优控制问题的重要方法之一[17-23]。该方法起源于动态规划，结合了强化学习理论以及伸进网络技术，利用系统的输入输出数据，结合在线策略迭代的思想，通过交替执行策略评估以及策略改进，最终在部分动力学未知的情况下实现最优控制，因此受到广泛学者的青睐。针对部分动力学未知的仿射非线性系统的最优控制问题，文献[18]提出积分强化学习算法。文献[19]在文献[18]的基础上考虑了输入受限的系统，并且在使用梯度下降法求解权重时采用了经验回放技术，进一步提高了算法的精度。针对系统动力学完全未知的情况，基于最小二乘法以及离线策略迭代技术，结合积分强化学习算法，成功实现最优控制[20]。考虑到积分强化学习算法是一种时间触发型算法，需要频繁进行策略评估以及策略更新运算，同时更新控制策略，为了降低控制策略的更新频率，将事件触发机制与积分强化学习算法结合起来，同时考虑稳态非零问题（当系统处于稳态时，控制策略与状态不为零），最终实现最优控制[23]。然而，据作者所知，利用积分强化学习算法解决带有状态约束的部分动力学未知系统的最优控制问题尚未得到广泛关注。

为了克服现存控制方法存在的局限性，最终实现最优控制。本文针对带有全状态约束且部分动力学未知系统的最优控制问题，基于 IRL 控制理论，提出一种带状态约束的事件触发积分强化学习算法。利用矩阵变换将带有约束的系统转化为无约束系统，基于转换之后系统的状态，利用 IRL 算法，通过交替执行策略评估以及策略改进，实现最优控制，从而避免对原系统未知动态的估计。此外，在控制过程中引入事件触发机制，以降低控制策略的更新频率，节约系统内存资源。

1 问题描述

考虑如下仿射非线性连续系统：

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

其中：$x = [x_1, \cdots, x_n]^T \in R^n$ 是系统可观测的状态，R^n 表示 n 维欧几里得空间，$u = [u_1, \cdots, u_m]^T \in R^m$ 是控制策略，$f(x) \in R^{n\times 1}$ 是未知的漂移动力学，$g(x) \in R^{n\times m}$ 是已知的输入动力学。假设控制系统（1）是稳定的。

定义系统（1）的代价函数，如下所述。

$$V(x_t) = \int_t^{\infty} e^{-\rho(\tau - t)} r(x,u) \mathrm{d}\tau \tag{2}$$

其中：为了使（2）收敛引入了衰减项 $e^{-\rho(\tau-t)}$，ρ 为折扣因子，是大于零的常数，$r(x,u) = x^TQx + u^TRu$ 是 t 时刻的奖励函数，Q 与 R 为正定矩阵，并且 $x^TQx \geqslant q\|x\|^2$，q 为正实数。

本文的控制目标是设计容许的控制策略 u 使得代价函数最优，即：

$$V^*(x_t) = \min_u \int_t^{\infty} e^{-\rho(\tau - t)} r(x,u) \mathrm{d}\tau \tag{3}$$

并且 u 是有界的（不为无穷大）。同时系统状态 $x_i (i = 1, \cdots, n)$ 始终是有界的，即：$|x_i| < a_i, a_i > 0$。

2 控制策略设计

控制策略设计主要包括五部分。首先利用矩阵变换技术将带有约束的仿射非线性连续系统转化为不含约束的仿射非线性连续系统，以克服状态约束控制系统的影响；其次介绍基本的积分强化学习算法；再次考虑到积分强化学习算法频繁策略更新，为减少计算量和提高控制效率，引入事件触发机制，基于李雅普诺夫稳定性定理，设计了事件触发条件，以减少控制策略的更新频率；然后利用神经网络逼近值函数的方法，准确地估计值函数；最后给出带状态约束的事件触发积分强化学习算法的流程。

2.1 系统转换

本节利用系统转换技术将带有状态约束的仿射非线性连续系统转化为不含约束的仿射非线性连续系统[12]。

在进行系统转换之前，首先，定义一组虚拟状态变量 $z = [z_1, \cdots, z_n]^T \subset R^n$，并且满足如下等式条件：

$$z_i(x_i) = arctanh\left(\frac{x_i}{a_i}\right) \tag{4}$$

其中：a_i 为 x_i 的边界值，$i = 1, 2 \cdots n$。注意到，$z_i(x_i)$ 具有如下性质：首先，$z_i(x_i)$ 是单调递增的函数；其次，$z_i(0) = 0$；最后，若 x_i 趋向于 $-a_i$ 时，z_i 趋向于负无穷，若 x_i 趋向于 a_i 时，z_i 趋向于正无穷。

引理 1[12]：对于任意初始状态，如果系统的初始状态满足状态约束，利用式（4）得到转换之后的系统，若设计控制策略使得转换之后系统的状态有界，并将控制策略作用于实际系统，则系统的实际状态满足状态约束。

对式（4）进行关于时间的导数求解，将得到一个虚拟系统，并且虚拟系统依然保持仿射非线性的形式。虚拟系统由下式给出：

$$\dot{z} = F(z) + G(z)u \tag{5}$$

其中：$F(z)$ 为关于 $f(x)$ 的函数，由于系统（1）中 $f(x)$ 未知，故 $F(z)$ 也未知。$G(z)$ 是关于 $g(x)$ 的函数，$G(z) = [G_1(z), G_2(z), \cdots, G_n(z)]^T$，$G_i(z) = \dfrac{g_i(x)}{(a_i - a_i tanh^2(z_i))}$，$i = 1, 2, \cdots, n$，$x = [a_1 tanh(z_1), \cdots, a_n tanh(z_n)]^T$。此外，假设系统（5）是稳定的，$G(z)$ 是有界的并且 $G(z)$ 满足利普希茨连续条件，即：

$\|G(z)\| \leqslant b_{OG}$，$\|G(z_1)-G(z_2)\| \leqslant b_G\|z_1-z_2\|$

其中：b_{OG} 与 b_G 是正实数。

通过将状态约束边界并入原始仿射非线性连续系统(1)，将得到一个新的无约束系统（5）。此外，如果转换之后的虚拟系统（5）的稳态值趋向于零，则系统的实际状态也趋向于零，那么，转换前后的控制系统具有相同的渐近稳定性。接下来，只需专注于对虚拟系统（5）设计控制策略使得代价函数最优即可。

2.2　积分强化学习算法

本节主要利用积分强化学习算法求解具有部分动力学未知的虚拟系统（5）的最优控制问题。定义虚拟系统的代价函数如下所示：

$$V(z_t)=\int_t^{\infty} e^{-\rho(\tau-t)}r(z,u)\mathrm{d}\tau \tag{6}$$

对于任意时间间隔 $\Delta t>0$，式（6）可以重写为：

$$V(z_t)=\int_t^{t+\Delta t} e^{-\rho(\tau-t)}r(z,u)\mathrm{d}\tau+e^{-\rho\Delta t}V(z_{t+\Delta t}) \tag{7}$$

上式也被称为积分强化学习－贝尔曼（IRL-Belleman）方程，是积分强化学习算法的核心。如果 $V(z_t)$ 是可微的，则：

$$\dot{V}(z_t)=\nabla V^T(z_t)(F(z)+G(z)u)-\rho V(z_t)+r(z,u) \tag{8}$$

其中：$\dot{V}(z_t)$ 为 $V(z_t)$ 关于时间的导数，$\nabla V(z_t)$ 为 $V(z_t)$ 在 t 时刻关于虚拟状态的导数。为了简化符号，下面用 $V(z)$ 表示 $V(z_t)$。

根据式（5）以及式（8），哈密顿函数定义为：

$$H(z,u,\nabla V(z))=\nabla V^T(z)(F(z)+G(z)u)-\rho V(z)+z^TQz+u^TRu \tag{9}$$

根据贝尔曼最优原理，对于最优的代价函数 $V^*(z)$，哈密顿函数满足：

$$\min_u H(z,u,\nabla V^*(z))=0 \tag{10}$$

令哈密顿函数关于控制策略的一阶偏导数为零，即可获得最优控制策略。最优控制策略如下所示：

$$u^*(z)=-0.5R^{-1}G^T(z)\nabla V^*(z) \tag{11}$$

结合式（7），此时最优代价函数 $V^*(z)$ 满足：

$$V^*(z_t)=\int_t^{t+\Delta t} e^{-\rho(\tau-t)}r(z,u)\mathrm{d}\tau+e^{-\rho\Delta t}V^*(z_{t+\Delta t}) \tag{12}$$

基于前面所述，积分强化学习中最关键的两步（策略评估以及策略改进）描述如下。

策略评估：

$$V^i(z_t)=\int_t^{t+\Delta t} e^{-\rho(\tau-t)}r(z,u)\mathrm{d}\tau+e^{-\rho\Delta t}V^i(z_{t+\Delta t}) \tag{13}$$

策略改进：

$$u^{i+1}(z)=-0.5R^{-1}G^T(z)\nabla V^i(z) \tag{14}$$

其中：i 为策略迭代指数。积分强化学习算法描述如下：首先给定初始可许的控制策略 u^0，通过交替执行策略评估(13) 以及策略改进 (14)，最终控制策略以及代价函数将收敛于最优值。

对于积分强化学习算法来说，控制器无需时刻更新控制策略，在 t 时刻采集系统状态信息，利用（13）以及(14) 分别进行策略评估以及策略改进，然后将改进的控制策略作用于系统，直至 $t+\Delta t$ 时刻，因此积分强化学习算法是一种时间触发型算法。对于 Δt 的选取，现有的文献一般都会选择固定值，每隔 Δt，进行一次策略改进。若系统处于稳态，仍然需要不断进行策略评估以及策略改进的计算。因此，下文将结合事件触发机制确定 Δt。

2.3　事件触发机制

本节主要利用李雅普诺夫函数确定事件触发条件，从而确定 Δt。在分析之前，给出如下条件。$u(z)$ 满足利普希茨连续条件，即：

$$\|u(z_1)-u(z_2)\| \leqslant \mathcal{L}_u\|(z_1)\|$$

其中：$\mathcal{L}_u$ 是正实数。

定义一个单调的触发序列 $\{t^i\}_{i=0}^{\infty}$，对于任意时刻 t^i，遵循 $t^i<t^{i+1}$。$[t^i,t^{i+1})$ 时间段内的系统状态误差表述如下：

$$e^i(t)=\bar{z}^i-z(t),t\in[t^i,t^{i+1}) \tag{15}$$

其中：$\bar{z}^i=z(t^i)$ 为 t^i 时刻的虚拟状态，$z(t)$ 是 t^i 时刻至 t^{i+1} 时刻之间的连续状态，$[t^i,t^{i+1})$ 被称为触发间隔。t^i 时刻产生新的控制策略作用于系统，同时计算系统稳定运行的状态误差边界，一旦违反边界即为 t^{i+1} 时刻，由此确定 $\Delta t=t^{i+1}-t^i$，即：

$$u(t)=\begin{cases}u(\bar{z}^i), & t\in(t^i,t^{i+1})\\ u(\bar{z}^{i+1}), & t=t^{i+1}\end{cases} \tag{16}$$

t^i 时刻，基于事件触发机制的控制策略重新表示为：$u(\bar{z}^i)=u(e^i(t)+z(t))$，则系统（5）重新描述为：$\dot{z}=F(z)+G(z)u(\bar{z}^i)$。

选取 $V(z)$ 作为李雅普诺夫函数，则：

$$\dot{V}(z)=\nabla V^T(z)(F(z)+G(z)u(\bar{z}^i))=\nabla V^T(z)(F(z)+G(z)u(z))+\nabla V^T(z)G(z)(u(\bar{z}^i)-u(z)) \tag{17}$$

结合式（10）以及式（14）可推导出：

$$\nabla V^T(z)(F(z)+G(z)u(z))=\rho V(z)-z^TQz-u(z)^TRu(z),\nabla V^T(z)G(z)=-2u^T(z)R$$

故，式（17）进一步推导为：

$$\dot{V}(z)=\rho V(z)-z^TQz-u(\bar{z}^i)Ru(\bar{z}^i)+(u(z)-u(\bar{z}^i))^TR(u(z)-u(\bar{z}^i))\leqslant -q\|z\|^2-\lambda_{\min}(R)\|u(\bar{z}^i)\|^2+\lambda_{\max}(R)\mathcal{L}_u^2\|e^i(t)\|^2+\rho V(z) \tag{18}$$

由文献［23］知：$V(z)$ 是有界的，即：$\|V(z)\|\leqslant\delta_V$，$\delta_V$ 是正实数，故：

$$\dot{V}(z)\leqslant -q\|z\|^2-\lambda_{\min}(R)\|u(\bar{z}^i)\|^2+\lambda_{\max}(R)\mathcal{L}_u^2|e^i(t)|^2+\rho\delta_V \tag{19}$$

综上，如果选择事件触发条件：

$$\|e^i(t)\|=\sqrt{\frac{q\|z\|^2+\lambda_{\min}(R)\|u(\bar{z}_i)\|^2-\rho\delta_V}{\lambda_{\max}(R)\mathcal{L}_u^2}} \tag{20}$$

则 $\dot{V}(z)\leqslant 0$，基于李雅普诺夫稳定性理论，系统（5）始终处于稳定状态。

2.4　神经网络实现

一般来说，直接求解 $V(z)$ 是不容易的。由逼近定理知，若 $V(z)$ 是连续的、平滑的以及可微的，则 $V(z)$ 及其

关于状态的导数 $\nabla V(z)$ 可以用神经网络近似，即：

$$V(z) = W^T\psi(z) + \varepsilon(z) \tag{21a}$$

$$\nabla V(z) = \nabla\psi^T(z)W + \nabla\varepsilon(z) \tag{21b}$$

上述网络也被称为评论神经网络，主要由三层组成：输入层、隐藏层以及输出层。简单起见，选择单隐藏层的神经网络结构，并将输入层到隐藏层的权重全部置为 1，这意味着隐藏层的输入即为输入层的输入。$\psi(z) \in R^{l\times 1}$ 是神经元的激活函数组成的向量，$\nabla\varphi(z)$ 为 $\varphi(z)$ 关于状态 z 的导数，l 为隐藏层神经元的数量。$W \in R^{l\times 1}$ 是隐藏层至输出层的常参数组成的权重向量。$\varepsilon(z)$ 为评论神经网络的近似误差，$\nabla\varepsilon(z)$ 为 $\varepsilon(z)$ 关于状态 z 的导数。

对于求解非线性程度很高的函数来说，现有的文献一般都会使用神经网络逼近定理来求解，但是如何设定神经元的数量以及选择合适的激活函数仍然是一个悬而未决的问题。针对上述情况，已经产生许多合适的激活函数，例如双曲正切函数和径向基函数。除此之外，虽然未知函数可以用神经网络来逼近，但结果未必满足未知函数的梯度，这主要是由初始权重决定的，以上只能依靠设计师的反复设计以及经验。由式（26）知，$\nabla V(z)$ 对于确定控制策略来说是必要的。

利用式（21a）逼近式（13）的解，则式（10）可以重写为：

$$\varepsilon_b = p(t) + W^{i,T}\Delta\psi(z_{t+\Delta t}) \tag{22}$$

其中：ε_b 是由于评论神经网络的逼近误差 $\varepsilon(z)$ 引起的，$p(t) = \int_t^{t+\Delta t} z^TQz + u^TRu\,\mathrm{d}\tau$ 暗含了该阶段内的奖励，$\Delta\psi(z_{t+\Delta t}) = e^{-t\alpha\Delta t}\psi(z_{t+\Delta t}) - \psi(z_t)$。

然而，在 $[t, t+\Delta t)$ 时间段内理想权重 W^i 是未知的。在忽略近似误差的情况下，式（21a）重写为：

$$\hat{V}(z) = \hat{W}^T\psi(z) \tag{23}$$

则，式（22）重写为：$0 = p(t) + \hat{W}^{i,T}\Delta\psi(z_{t+\Delta t})$。

利用梯度下降法，通过最小化残差方程 $E = \frac{e^2}{2}$，即可获得 $\hat{W}^i$ 的权重更新律。首先给定初始权重 $\hat{W}_0^i$，然后利用如下权重更新律在线调整 $\hat{W}_k^i$，权重更新律设计如下，

$$\hat{W}_k^i = \hat{W}_{k-1}^i - \alpha\frac{\partial E}{\partial e}\frac{\partial e}{\partial \hat{W}_{k-1}^i} = \hat{W}_{k-1}^i - \alpha\frac{\bar{\Delta\psi}}{m_\psi}e_{k-1}^i \tag{24}$$

其中：$\alpha > 0$ 为学习率，k 为权重迭代指标，$m_\psi = \Delta\psi^T(z_{t+\Delta t})\Delta\psi(z_{t+\Delta t}) + 1$，$\bar{\Delta\psi} = \frac{\Delta\psi(z_{t+\Delta t})}{m_\psi}$，$e_{k-1}^i = p(t) + \hat{W}_{k-1}^{i,T}\Delta\psi(z_{t+\Delta t})$。若存在一个常数 N，对于任意的 $k \geqslant N$，$\|\hat{W}_k^i - \hat{W}_{k-1}^i\| \leqslant \varepsilon_w$，其中 ε_w 为无限趋近于零的正常数，则 $W^i = \hat{W}^i = \hat{W}_k^i$。此外，权重误差动态 $\widetilde{W} = W - \hat{W}^i$ 为如下形式。

$$\dot{\widetilde{W}} = -\alpha\bar{\Delta\psi}\bar{\Delta\psi}^T\widetilde{W} + \alpha\bar{\Delta\psi}(e_{k-1}^i/m_\psi) \tag{25}$$

利用（14），则基于事件触发控制的策略更新调整为：

$$u^{i+1}(z) = -0.5R^{-1}G^T(z)\nabla\psi^T(z)\hat{W}^i \tag{26}$$

注意：由逼近定理知，当隐藏层神经元的数量无限趋近于无穷大时，$\varepsilon(z)$ 以及 $\nabla\varepsilon(z)$ 将无限趋近于 0，即：$\|\varepsilon(z)\| \leqslant \varepsilon_m$，$\|\nabla\varepsilon(z)\| \leqslant \varepsilon_{dm}$，$\varepsilon_m$ 与 ε_{dm} 为无限趋近于零的正常数。此外，理想权重 W 是范数有界的，$\|W\| \leqslant W_m$，W_m 为大于零的常数。对于任意神经元 $\varphi(z)$ 及 $\nabla\varphi(z)$ 是范数有界的，即：$\|\psi(z)\| \leqslant \psi_m$，$\|\nabla\psi(z)\| \leqslant \psi_{dm}$，$\psi_m > 0$，$\psi_{dm} > 0$；$\nabla\psi(z)$ 满足利普希茨连续条件，即：$\|\nabla\psi(z_1) - \nabla\psi(z_2)\| \leqslant \psi_{dmm}\|z_1 - z_2\|$，$\psi_{dmm} > 0$。此外，$e_{k-1}^i$ 是有界的，即：$\|e_{k-1}^i\| \leqslant \varepsilon_{bm}$，$\varepsilon_{bm} > 0$。

2.5 算法流程

带状态约束的事件触发积分强化学习算法归纳描述如下。

第一步：初始化，选择合适的初始控制策略 u^0、评论神经网络的初始权重 W_0、权重收敛误差 ε_w、权重学习率 α、神经元的数量以及各自的激活函数；

第二步：利用式（5）计算 $G(z)$；

第三步：$i = 0$；

第四步：结合式（20），确定事件触发条件 $e^i(t)$；

第五步：将 u^i 作用于控制系统，并且实时采集数据，并利用式（4）计算虚拟状态 z，直至满足事件触发条件；

第六步：利用式（25）以及式（27）分别获得权重 $\hat{W}^i$ 以及控制策略 u^{i+1}；

第七步：令 $i = i+1$；重复第四步。直至权重 $\hat{W}^i$ 不再发生变化，即获得最优权重 W^*。

3 稳定性分析

本节利用李雅普诺夫函数分析在事件触发条件下控制系统的稳定性。首先给出如下定理。

定理 1：考虑由非线性系统（1）转换之后的虚拟系统（5）、权重更新律以及策略更新律分别如式（24）和式（26）所示，如果选择事件触发条件为式（20），则权重误差动态是有界的，并且系统是稳定的。

证明：定义李雅普诺夫函数为：

$$L(t) = L_1(t) + L_2(t) + L_3(t) \tag{27}$$

其中：$L_1(t) = \int_t^{t+\Delta t} V^*(z(z))\,\mathrm{d}\tau$，$L_3(t) = 0.5\widetilde{W}^T\widetilde{W}$，$L_2(t) = \int_t^{t+\Delta t} V^*(\bar{z}^i)\,\mathrm{d}\tau$。

为了便于分析，下面分两种情况来讨论。

情形一：考虑事件不触发条件下系统的稳定性。首先，对式（27）进行求导，$\dot{L}(t) = \dot{L}_1(t) + \dot{L}_2(t) + \dot{L}_3(t)$，其中 $\dot{L}_2(t) = 0$。在分析其它项之前，先分析 $\dot{V}^*(z)$，

$$\dot{V}^*(z) = \nabla V^{*T}(z)(F(z) + G(z)u^*(\bar{z}^i)) = \rho V^*(z) - z^TQz - u^*(z)^TRu^*(z) - (\nabla\psi^T(\bar{z}^i)W + \nabla\varepsilon(z))^TG(z)(u^*(z) - u^*(\bar{z}^i)) \tag{28}$$

利用 Young 不等式和 Cauchy-schwarz 不等式，式（28）进一步推导为：

$$\dot{V}^*(z) \leqslant 0.5\|\nabla\psi^T(\bar{z}^i)W + \nabla\varepsilon(z)\|^2 + \rho V^*(z) - u^{*T}(z)Ru^*(z) + 0.5\|G(z)\theta_u\|^2 - z^TQz \leqslant \rho V^*(z) - z^TQz + \psi_{dm}{}^2W_m + \varepsilon_{dm}^2 + 0.5b_G^2\|\theta_u\|^2 \tag{29}$$

其中：$\theta_u = u^*(z) - u^*(\bar{z}^i)$，接下来析 θ_u。

$$\theta_u = 0.5R^{-1}[G^T(\bar{z}^i)\nabla\psi^T(\bar{z}^i) - G^T(z)\nabla\psi^T(z)]\hat{W} -$$

$$-0.5R^{-1}G^T(z)(\nabla\psi^T(z)\widetilde{W}+\nabla\varepsilon(z))=0.5R^{-1}H_1\hat{W}-0.5R^{-1}G^T(z)H_2 \tag{30}$$

其中：$H_1=[\nabla\psi^T(z)-\nabla\psi^T(\bar{z}^i)]G(\bar{z}^i)+\nabla\psi(z)^T[G(\bar{z}^i)-G(z)]$，$H_2=\nabla\psi^T(z)\widetilde{W}+\nabla\varepsilon(z)$。

进一步，式（30）推导为：

$$\|\theta_u\|^2\leqslant\|R^{-1}\|^2(\|H_{11}\|^2+\|H_{12}\|^2)\|\hat{W}\|^2+\|R^{-1}\|^2(\|H_{21}\widetilde{W}\|^2+\|H_{22}\|^2) \tag{31}$$

其中：$H_{11}=(\nabla\psi^T(\bar{z}^i)-\nabla\psi^T(z))G^T(\bar{z}^i)$，$H_{21}=G^T(z)\nabla\psi(z)$，$H_{22}=G^T(z)\nabla\varepsilon(z)$，$H_{12}=\nabla\psi^T(z)(G^T(\bar{z}^i)-G^T(z))$。由式(20)知，$e^i(t)$ 是有界的，即：$\|e^i(t)\|^2\leqslant b_e$，故，

$$\|\theta_u\|^2\leqslant\|R^{-1}\|^2(\psi_{\mathrm{dmm}}^2b_G^2+\psi_{dm}^2b_{GG}^2)b_e^2W_m^2+\|R^{-1}\|^2b_G^2\psi_{dm}^2\|\widetilde{W}\|^2+\|R^{-1}\|^2b_G^2\varepsilon_{dm}^2 \tag{32}$$

利用（29）以及（32），$\dot{V}^*(z)$ 简化为：

$$\dot{V}^*(z)\leqslant-z^TQz+\mathcal{L}_\sum\|\widetilde{W}\|^2+\mathcal{L}_{all} \tag{33}$$

其中：$\mathcal{L}_\sum=0.5b_G^2\|R^{-1}\|^2b_G^2\psi_{dm}^2$，$\mathcal{L}_{all}=\psi_{dm}^2W_m+\varepsilon_{dm}^2+0.5b_G^2(\|R^{-1}\|^2(\psi_{dmm}^2b_G^2+\psi_{dm}^2b_{GG}^2)b_e^2W_m^2+\|R^{-1}\|^2b_G^2\varepsilon_{dm}^2)+\rho\delta_V+\varepsilon_{dm}^2$。

接下来分析 $L_1(t)$，

$$\dot{L}_1(t)=\int_t^{t+\Delta t}\dot{V}^*(z(z))\mathrm{d}\tau\leqslant\int_t^{t+\Delta t}-z^TQz+\mathcal{L}_z\|\widetilde{W}\|^2+\mathcal{L}_{all}\mathrm{d}\tau \tag{34}$$

然后，讨论 $L_3(t)$，

$$\dot{L}_3(t)=-\alpha\widetilde{W}^T\bar{\Delta}\psi\bar{\Delta}\psi^T\widetilde{W}+\alpha\widetilde{W}^T\bar{\Delta}\psi(e_{k-1}^i/m_\psi)\leqslant-\alpha\widetilde{W}^T\bar{\Delta}\psi\bar{\Delta}\psi^T\widetilde{W}+0.5\alpha(\widetilde{W}^T\bar{\Delta}\psi\bar{\Delta}\psi^T\widetilde{W}+\varepsilon_{bm}^2)\leqslant-0.5\alpha\widetilde{W}^T\bar{\Delta}\psi\bar{\Delta}\psi^T\widetilde{W}+0.5\alpha\varepsilon_{bm}^2\leqslant-0.5\alpha\lambda_{min}(\bar{\Delta}\psi\bar{\Delta}\psi^T)\|\widetilde{W}\|^2+0.5\alpha\varepsilon_{bm}^2 \tag{35}$$

综上所述，

$$\dot{L}(t)\leqslant-0.5\alpha\lambda_{min}(\bar{\Delta}\psi\bar{\Delta}\psi^T)\|\widetilde{W}\|^2+0.5\alpha\varepsilon_{bm}^2+\int_t^{t+\Delta t}-z^TQz+\mathcal{L}_\sum\|\widetilde{W}\|^2+\mathcal{L}_{all}\mathrm{d}\tau\leqslant\int_t^{t+\Delta t}-z^TQz\mathrm{d}\tau-0.5\alpha\lambda_{min}(\bar{\Delta}\psi\bar{\Delta}\psi^T)\|\widetilde{W}\|^2+0.5\alpha\varepsilon_{bm}^2+\Delta t(\mathcal{L}_\sum\|\widetilde{W}\|^2+\mathcal{L}_{all}) \tag{36}$$

若权重误差满足：

$$\|\widetilde{W}\|>\sqrt{\frac{0.5\alpha\varepsilon_{bm}^2+\mathcal{L}_{all}\Delta t}{0.5\alpha\lambda_{min}(\bar{\Delta}\psi\bar{\Delta}\psi^T)-\Delta t\ \mathcal{L}_\sum}}$$

则 $\dot{L}(t)<0$，表明在事件不触发的情况下控制系统是稳定的，并且权重误差是有界的。

情形二：在事件触发的情况下，考虑间断点处的稳定性。

$$\Delta L=\int_t^{t+\Delta t}V^*(z^+)-V^*(\bar{z}^i)\mathrm{d}\tau+0.5\widetilde{W}^{+T}\widetilde{W}^+-0.5\widetilde{W}^T\widetilde{W}+\int_t^{t+\Delta t}V^*(\bar{z}^{i+1})-V^*(\bar{z}^i)\mathrm{d}\tau \tag{37}$$

其中：$V^*(z)$ 与 $\widetilde{W}$ 在触发时刻是连续函数，故上式中的前 3 项和等于 0。此外，注意到 $V^*(\bar{z}^{i+1})-V^*(\bar{z}^i)\leqslant-v(\|e_{i+1}\|)<0$。因此 $\Delta L<0$，故在触发时刻控制系统是稳定的。证毕

4　系统应用

为了验证带有状态约束的事件触发积分强化学习算法有效性，本节利用单连杆机械臂的仿射非线性连续系统进行仿真[12]，其动态系统描述如下：

$$\ddot{\theta}=-\frac{MgL}{\widetilde{G}}\sin(\theta(t))-\frac{\widetilde{D}}{\widetilde{G}}\dot{\theta}+\frac{1}{\widetilde{G}}u(t)$$

其中：$M=10$ kg 是机械臂的质量，$g=9.81$ m/s^2 是重力加速度，$L=0.5$ m 是机械臂的长度，$\widetilde{D}=2$ N 是粘性摩擦以及 $\widetilde{G}=10$ kg·m^2 是转动惯量。将机械臂的角度位置 θ 定义为状态 x_1，$\dot{\theta}$ 定义为状态 x_2，则系统可以重新描述为仿射非线性的形式：

$$\dot{x}=f(x)+g(x)u$$

其中：$x=[x_1,x_2]^T$，$g(x)=[0,0.1]^T$，$f(x)=[x_2\quad-4.905\sin(x_1)-0.2x_2]^T$。这里假设 $f(x)$ 是未知的，只用于收集系统数据，不用于控制律的设计。假设系统的初始状态为 $x=[0.5,0.5]^T$。

本实验的控制目标是设计控制策略 u 使得二次型代价函数最优，并且在控制过程中系统的状态满足约束，即：$|x_i|<1,i=1,2$。二次型代价函数如下所示。

$$V(x_t,u)=\int_t^\infty e^{-\rho(\tau-t)}r(x,u)\mathrm{d}\tau$$

其中：$\rho=0.9$ 为折扣因子，$r(z,u)=z^TQz+u^TRu$，$R=10$，$Q=diag(0.2,0.2)$。

为了克服状态约束，首先定义一组虚拟状态 $z=[z_1,z_2]^T$ 用于系统转换，转换之后的系统依然是仿射非线性连续系统，利用式（5），则 $G(z)$ 表述：为：

$$G^T(z)=\begin{bmatrix}0&\dfrac{0.1}{(1-(\tanh(z_2)))^2}\end{bmatrix}$$

此外，$F(z)$ 是未知的。转换之后的虚拟状态可以用（4）计算获得。定义转换之后系统的代价函数为：

$$V(z_t,u)=\int_t^\infty e^{-\rho(\tau-t)}r(z,u)\mathrm{d}\tau$$

选取 Critic 神经网络的结构为 2－8－1，其中：神经网络的输入变量的个数为 2，分别是系统经转换之后的虚拟状态 z_1 和 z_2。输入层至隐藏层的权重设置为 1。选择单隐藏层神经网络，并且隐藏层的神经元的数量为 8。输出层神经元的数量为 1，代表代价函数的值。隐藏层神经元代表的激活函数组成的向量用 $\psi(z)$ 表示，为：

$$\psi(z)=[z_1^2,z_1z_2,z_2^2,z_1^4,z_1z_2^3,z_1^2z_2^2,z_2z_1^3,z_2^4]^T$$

仿真过程中参数设置：初始控制策略 $u_0=-1$、评论神经网络权重收敛误差精度 $\varepsilon_w=0.005$、权重学习率为 $\alpha=0.9$。评论神经网络的初始权重：

$W_0=[8.67,-0.15,-5.87,6,0.8,-1.14,1.72,-2.23]^T$

仿真结果以及分析如下所示。

图 1 为虚拟状态的运行轨迹，其中，实线代表虚拟状态 z_1，虚线代表虚拟状态 z_2。由图所知，虚拟状态在整个控制过程中始终是有界的（不为无穷大），故系统的实际运行状态必然满足约束。

图 2 与图 3 为考虑状态约束与未考虑状态约束的对比图，虚线代表不考虑状态约束的运行轨迹，实线代表考虑状态约束的运行轨迹。两种情况都是在事件触发机

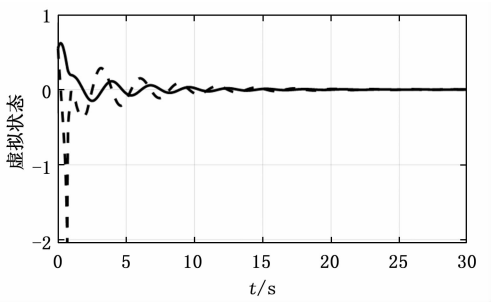

图 1 虚拟状态曲线

制下完成的，并且都选择相同初始参数，可以避免因参数不同而对系统状态轨迹的影响。由图知，相较于未考虑状态约束的情况，本文所提算法在整个控制过程中系统状态均未超过事先设置的状态约束，并且最终系统的状态收敛到稳态点附近，由此判定该算法能够解决带有状态约束的控制问题。结合图 1，虚拟状态以及实际状态都收敛到零点附近，因此转换前后的系统具备相同的渐近稳定性。此外，注意到由于考虑了状态约束，能使系统较快的收敛到稳态点附近。大约经过 5 s 之后，系统的状态全部收敛于零。

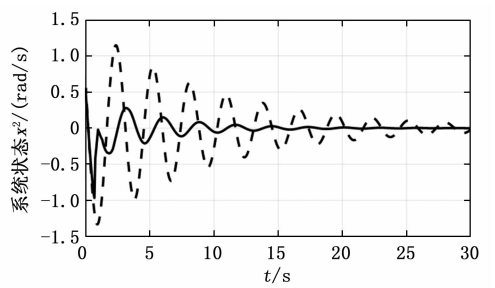

图 2 x_1 轨迹对比

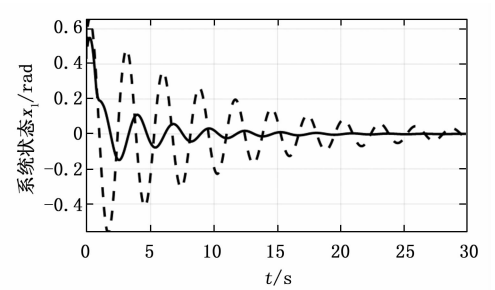

图 3 x_2 轨迹对比

图 4 为带状态约束的事件出发积分强化学习控制算法在整个控制过程中施加的控制策略。在经过大约 5 s 之后，控制策略也收敛于零。对于二次型代价函数，理想情况，最优代价函数的对应的稳态值为零。结合图 2 与图 3，5 s 之后代价函数的值一直稳定在 0 的较小邻域内，说明所提算法是可行的。此外，注意到图 4 中某个时刻控制策略显著增大是由于此时刻实际状态接近于边界但并未超过边界引起的。

事件触发时刻以及事件触发条件如图 5 所示，其中横

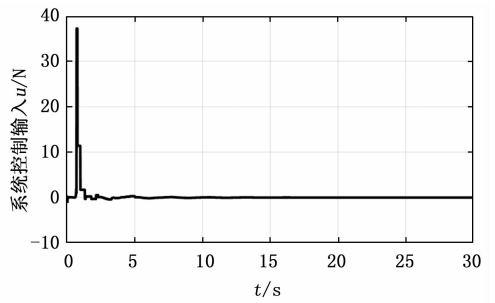

图 4 控制输入轨迹

轴代表触发时刻，纵轴代表触发条件误差，一旦超过这个误差，更新控制策略。由横轴触发时刻的间隔以及图 4 更新控制策略的时刻知，该算法并非是周期触发。图 6 是评论神经网络部分权重的收敛曲线。由图知，最终权重将收敛于某一值附近。

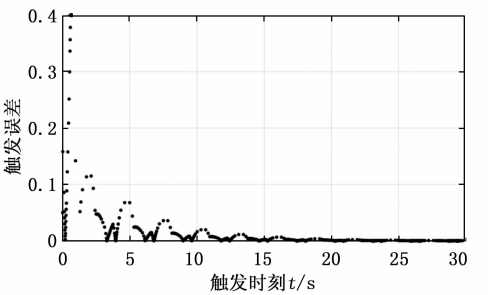

图 5 事件触发时刻以及触发误差

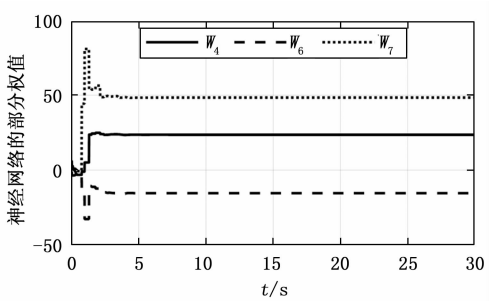

图 6 评论神经网络部分权值轨迹

5 结束语

本文基于事件触发机制的积分强化学习算法，设计仿射非线性连续系统的最优控制策略，将系统转换、事件触发机制、积分强化学习算法紧密地结合起来，利用李雅普诺夫函数给出满足系统稳定运行的事件触发条件。在实际工程系统中，由于系统的动力学大多难以获得并且受状态约束的影响，使本文算法更具普遍性。最后，针对单连杆机械臂的仿真结果表明所提方法的有效性。

参考文献：

[1] FU Y, CHAI T. Self-tuning control with a filter and a neural compensator for a class of nonlinear systems [J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24

(5)：837－843.
[2] THEORIN A，HAGGLUND T. Derivative backoff：the other saturation problem for PID controllers [J]. Journal of Process Control，2015，33：155－160.
[3] SHTESSEL Y B，MORENO J A，FRIDMAN L M. Twisting sliding mode control with adaptation：Lyapunov design，methodology and application [J]. Automatica，2017，75：229－235.
[4] LI Z，REN W，LIU X，et al. Distributed consensus of linear multi-agent systems with adaptive dynamic protocols [J]. Automatica，2013，49 (7)：1986－1995.
[5] 叶金鑫，谢丽蓉，王宏伟. 基于非均匀采样系统的无模型自适应控制 [J]. 电光与控制，2022，29 (5)：48－54.
[6] YANG J，LI S，YU X. Sliding-mode control for systems with mismatched uncertainties via a disturbance observer [J]. IEEE Transactions on Industrial Electronics，2013，60 (1)：160－169.
[7] LIU J，GAO Y，SU X，et al. Disturbance-observer based control for air management of PEM fuel cell systems via sliding mode technique [J]. IEEE Transactions on Control Systems Technology，2019，27 (3)：1129－1138.
[8] 侯明冬，王印松. 一类非线性大滞后系统自适应积分滑模控制 [J]. 控制理论与应用，2019，36 (7)：1182－1188.
[9] 姚雪莲，杨　艺. 可反馈线性化系统的鲁棒自适应容错控制设计 [J]. 控制理论与应用，2020，37 (9)：2001－2010.
[10] GREENE M L，DEPTULA P，NIVISON S. Sparse learning-based approximate dynamic programming with barrier constraints [J]. IEEE Control Systems Letters，2020，4 (3)：743－748.
[11] DEPTULA P，CHEN H，LICITRA R，et al. Approximate optimal motion planning to avoid unknown moving Avoidance regions [J]. IEEE Transactions on Robotics，2020，36 (2)：414－430.
[12] CAO Y，SONG Y，WEN C. Practical tracking control of perturbed uncertain nonaffine systems with full state constraints [J]. Automatica，2019，110 (C)：108608.
[13] XU J，WANG J，RAO J，et al. Adaptive dynamic programming for optimal control of discrete-time nonlinear system with state constraints based on control barrier function [J]. International Journal of Robust and Nonlinear Control，2022，32 (6)：3408－3424.
[14] YANG Y，VAMVOUDAKIS K Y，MODARES H. Safe reinforcement learning for dynamical games [J]. International Journal of Robust and Nonlinear Control，2020，30 (9)：3706－3726.
[15] NGUYEN H N，BOURDAIS R，GUTMAN P O，Fast model predictive control for linear periodic systems with state and control constraints [J]. International Journal of Robust and Nonlinear Control，2017，27 (17)：3703－3726.
[16] 吴跃飞，胡智琦，马大为，等. 具有状态约束的机电伺服系统自适应鲁棒控制 [J]. 上海交通大学学报，2014，48 (5)：707－712.
[17] VRABIE D，ABU-KHALAF M，PASTRAVANU O，et al. Adaptive optimal control for continuous-time linear systems based on policy iteration [J]. Automatica，2009，45 (2)：477－484.
[18] VRABIE D，LEWIS F. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems [J]. Neural Networks，2009，22 (3)：237－246.
[19] MODARES H，LEWIS F L，NAGHIBA-SISTANI M B. Integral reinforcement learning and experience replay for adaptive optimal control of partially unknown constrained-input continuous time systems [J]. Automatica，2014，50 (1)：193－202.
[20] LUO B，WU H，HUANG T，et al. Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design [J]. Automatica，2014，50 (12)：3281－3290.
[21] LI H，LIU D，WANG D. Integral Reinforcement Learning for Linear Continuous-Time Zero-Sum Games With Completely Unknown Dynamics [J]. IEEE ransactions on Automatica science and Engineering，2014，11 (3)：706－714.
[22] YANG X，LIU D，LUO，B，et al. Data-based robust adaptive control for a class of unknown nonlinear constrained-input systems via integral reinforcement learning [J]. Information Sciences，2016，369：731－747.
[23] YANG X，WEI Q. Adaptive critic learning for constrained optimal event-triggered control with discounted cost [J]. IEEE transactions on Neural Networks and Learning Systems，2021，32 (1)：91－104.

(上接第 142 页)

[16] ML Salfi，U Ahsun，HA Bhatti. Lateral and longitudinal guidance and control design of a UAV in auto landing phase [P]. International Bhurban Conference on Applied Sciences & Technology，2009.
[17] 习金程. 某型无人机天钩回收策略设计与软件开发 [D]. 南京：南京航空航天大学，2020.
[18] 李　凯. 小型舰载无人机弹射起飞与天钩回收技术研究 [D]. 南京：南京航空航天大学，2019.
[19] 张晗祺，郭志明，孙宗华，等. 高速无人机自主着陆纵向控制技术研究 [J]. 南京航空航天大学学报，2020，52 (6)：881－888.
[20] 邱潇颀，高长生，荆武兴. 变质心固定翼无人机动力学分析与抗扰控制 [J]. 北京航空航天大学学报，2022，48 (3)：430－437.
[21] HUANG H，ZHOU J. Solar sailing CubeSat attitude control method with satellite as moving mass [J]. Acta Astronautica，2019，159：331－341.
[22] HE L，CHEN X，KUMAR K D，et al. A novel three-axis attitude stabilization method using in-plane internal mass-shifting [J]. Aerospace Science and Technology，2019，92：489－500.