引用格式:蒲俊,马清亮,李远冬,等. 基于数据驱动自适应动态规划的输入约束的非线性系统 H_∞ 控制[J]. 电光与控制,2019,26(7):40-45. PU J, MA Q L, LI Y D, et al. H_∞ control of nonlinear systems with input constraints based on data-driven adaptive dynamic programming[J]. Electronics Optics & Control, 2019, 26(7):40-45.

基于数据驱动自适应动态规划的输入约束的非线性系统 H_∞ 控制

蒲 俊, 马清亮, 李远冬, 顾 凡
(火箭军工程大学,西安 710025)

摘 要: 提出了一种包含在线采样、离线学习两个阶段的基于数据驱动的迭代自适应动态规划(ADP)算法,仅通过在线数据,解决输入约束的连续未知模型的非线性系统的 H_∞ 控制问题。通过策略迭代(PI)和迭代强化学习(IRL)方法推导出无模型(HJI)方程。构建3个神经网络,在线采集系统数据结束后,利用离线学习方法,近似求解无模型 HJI 方程,进而得到值函数、控制策略和扰动策略,神经网络的未知参数通过最小二乘方法求解。仿真结果验证了算法的可行性。

关键词: 自适应动态规划; H_∞ 控制; 输入约束; 最优控制; 神经网络
中图分类号: TP13 **文献标志码**: A **doi**:10.3969/j.issn.1671-637X.2019.07.008

H_∞ Control of Nonlinear Systems with Input Constraints Based on Data-Driven Adaptive Dynamic Programming

PU Jun, MA Qing-liang, LI Yuan-dong, GU Fan
(Rocket Force University of Engineering, Xi'an 710025, China)

Abstract: A data-driven iterative Adaptive Dynamic Programming (ADP) algorithm including online sampling and offline learning is proposed. The H_∞ control problem of the nonlinear system with input constraints and unknown models is solved only by online data. The model-free Hamiton-Jacobi-Isaacs (HJI) equation is derived by using the methods of Policy Iteration (PI) and Iteration Reinforcement Learning (IRL). Three neural networks are constructed. After the online acquisition of system data is completed, the off-line learning method is used to approximately solve the model-free HJI equation, and then the value function, control strategy and disturbance strategy are obtained. The parameter of the neural network is solved by the least squares method. Simulation results verify the feasibility of the algorithm.

Key words: adaptive dynamic programming; H_∞ control; input constraints; optimal control; neural network

0 引言

最优控制一直是控制理论非常活跃的分支,其目的就是要找到一个控制策略,使得所定义的性能指标最小[1],而在各类控制应用中,系统中存在大量扰动,给系统带来负面影响,H_∞ 控制提供了一个有力的工具减少扰动的影响[2]。根据博弈论的思想,H_∞ 控制器的设计相当于一个双玩家的零和博弈游戏(Zero-Sum Game, ZSG)[3],控制器在最坏扰动下最小化性能指标,达到最优控制的目的。对于连续的非线性系统,可以通过求解 Hamiton-Jacobi-Isaacs (HJI) 方程获得 ZSG 解,但是非线性系统 HJI 方程是非线性偏微分方程,求解几乎是不可能的。

迭代自适应动态规划(Adaptive Dynamic Programming, ADP)技术融合了最优控制、自适应控制、强化学习理论,利用函数近似结构估计值函数,这就为近似求解 HJI 方程提供了有效途径[4]。WU 等[5]利用强化学

收稿日期:2018-07-11 修回日期:2018-08-29
基金项目:国家自然科学基金(61773387)
作者简介:蒲 俊(1994—),男,四川绵阳人,硕士生,研究方向为先进控制理论、非线性系统控制。

习的思想，推导出不依赖系统模型的有效计算同步策略更新算法，设计 H_∞ 反馈控制器；YASINI 等[6]融合策略迭代（Policy Iteration，PI）和迭代强化学习（Iteration Reinforcement Learning，IRL）算法解决部分系统模型未知的双玩家 ZSG 问题；ZHANG 等[3]提出迭代自适应动态规划方法解决一类非线性系统零和博弈问题；LUO 等[7]设计基于数据驱动的 H_∞ 控制器，其设计方法不再需要系统的模型信息，但是 PI 算法在近似求解 HJI 方程时需要初始的容许控制策略，其对于 HJI 方程往往是次优；YANG 等[8]用 ADP 算法设计输入约束的非线性系统 H_∞ 控制器，但依赖系统的模型信息，而实际系统不可避免地会受到多种因素的影响，很大程度上造成其模型未知或部分未知；ZHU 等[9]提出迭代 ADP 算法求解无模型 H_∞ 控制问题，但没有考虑输入受约束情况，系统由于工作环境和本身性质，其控制输入受到约束的情况非常常见。

受此启发，本文针对一类输入受约束的非线性系统，基于 ADP 技术，并结合博弈论和数据驱动思想，融合 PI 和 IRL 算法推导出无模型 HJI 方程，求解非线性系统 H_∞ 控制问题。

1　问题描述

考虑如下一类连续非线性系统

$$\begin{cases}\dot{\boldsymbol{x}}=\boldsymbol{f}(\boldsymbol{x})+\boldsymbol{g}(\boldsymbol{x})\boldsymbol{u}(t)+\boldsymbol{k}(\boldsymbol{x})\boldsymbol{w}(t)\\ \boldsymbol{z}=\boldsymbol{h}(\boldsymbol{x})\end{cases}\tag{1}$$

式中：$\boldsymbol{x}\in\mathbf{R}^n$ 是系统状态；$\boldsymbol{u}(t)\in\mathbf{R}^m$ 是控制输入；$\boldsymbol{w}(t)\in\mathbf{R}^q$ 是扰动输入，满足 $\boldsymbol{w}(t)\in L_2[0,\infty)$；$\boldsymbol{z}\in\mathbf{R}^p$ 是虚拟输出；$\boldsymbol{f}(\boldsymbol{x})\in\mathbf{R}^n$，$\boldsymbol{g}(\boldsymbol{x})\in\mathbf{R}^{n\times m}$ 和 $\boldsymbol{k}(\boldsymbol{x})\in\mathbf{R}^{n\times q}$ 是系统内部模型，假设系统模型都是未知的，$\boldsymbol{f}(\boldsymbol{x})$，$\boldsymbol{g}(\boldsymbol{x})$ 和 $\boldsymbol{k}(\boldsymbol{x})$ 为 Lipschitz 连续且 $\boldsymbol{f}(0)=0$，因此 $\boldsymbol{x}=0$ 是式（1）系统的一个平衡点。在后续的计算式中为了方便表示，用 $\boldsymbol{u}$ 和 $\boldsymbol{w}$ 分别表示 $\boldsymbol{u}(t)$ 和 $\boldsymbol{w}(t)$。

H_∞ 控制就是要找到一个控制策略使得如下性能指标，对所有的 $\boldsymbol{w}\in L_2[0,\infty)$ 和 $\boldsymbol{x}(0)=0$ 都是非正的，即

$$J_1(\boldsymbol{x}(0),\boldsymbol{u},\boldsymbol{w})=\int_0^\infty(\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x})+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}-\gamma^2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})\mathrm{d}t\tag{2}$$

$\boldsymbol{R}>0$，$\gamma\geqslant\gamma^*\geqslant0$，如果这个控制策略存在，就称系统具有 $L_2-g_{\mathrm{gain}}\leqslant\gamma$，$\gamma^*$ 表示使该问题有解的最小值。

受文献[10]启发，对于控制输入 $\boldsymbol{u}$ 受约束情况，引用广义的非二次函数 $Y(\boldsymbol{u})=2\int_0^{\boldsymbol{u}}\lambda\tanh^{-\mathrm{T}}(s/\lambda)\boldsymbol{R}\mathrm{d}s$，$\boldsymbol{R}>0$ 为对角矩阵。只关注反馈策略和完整的状态信息，给出一个控制策略 $\boldsymbol{u}(t)=\boldsymbol{u}(\boldsymbol{x}(t))$ 和一个扰动策略 $\boldsymbol{w}(t)=\boldsymbol{w}(\boldsymbol{x}(t))$，定义值函数为

$$V(\boldsymbol{x}(0))=\int_0^\infty(\boldsymbol{h}(x)^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x})+Y(\boldsymbol{u})-\gamma^2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})\mathrm{d}\tau\equiv\int_0^\infty r(\boldsymbol{x}(t),\boldsymbol{u},\boldsymbol{w})\mathrm{d}\tau\ 。\tag{3}$$

定义 Lyapunov 方程为

$$r(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})+\nabla\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}+\boldsymbol{k}\boldsymbol{w})=0\quad V(0)=0\tag{4}$$

式中，符号 ∇ 为求偏导数，即 $\nabla\boldsymbol{V}=\partial\boldsymbol{V}/\partial\boldsymbol{x}$。

定义 Hamiltonian 函数为

$$H(\boldsymbol{x},\nabla\boldsymbol{V},\boldsymbol{u},\boldsymbol{w})\equiv r(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})+\nabla\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}+\boldsymbol{k}\boldsymbol{w})\ 。\tag{5}$$

根据博弈理论，求解 H_∞ 控制问题相当于求解一个双玩家 ZSG 问题。控制策略使性能指标最大化，而扰动策略让性能指标最小化。对于连续非线性 ZSG 就是要找到最优的反馈控制策略和扰动策略，即

$$\boldsymbol{V}^*=\min_{\boldsymbol{u}}\max_{\boldsymbol{w}}V(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})\tag{6}$$

假如满足下列纳什均衡条件

$$\min_{\boldsymbol{u}}\max_{\boldsymbol{w}}V(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})=\min_{\boldsymbol{w}}\max_{\boldsymbol{u}}V(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w})\tag{7}$$

ZSG 问题有唯一解，即存在鞍点 $(\boldsymbol{u}^*,\boldsymbol{w}^*)$，$\boldsymbol{u}^*$ 即为所求系统式（1）的 H_∞ 最优控制策略，$\boldsymbol{w}^*$ 为最坏扰动策略，进一步根据稳定条件，$\boldsymbol{u}^*$ 和 $\boldsymbol{w}^*$ 分别表示为

$$\boldsymbol{u}^*=-\lambda\tanh\left(\frac{1}{2\lambda}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\boldsymbol{V}^*\right)\tag{8}$$

$$\boldsymbol{w}^*=\frac{1}{2}\gamma^{-2}\boldsymbol{k}^{\mathrm{T}}\nabla\boldsymbol{V}^*\ 。\tag{9}$$

将式（8）、式（9）代入式（4），得 HJI 方程为

$$(\nabla\boldsymbol{V}^*)^{\mathrm{T}}\boldsymbol{f}+\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x})+\lambda^2\bar{\boldsymbol{R}}\ln(\underline{1}-\tanh^2(\boldsymbol{D}^*))+\frac{1}{4\gamma^2}(\nabla\boldsymbol{V}^*)^{\mathrm{T}}\boldsymbol{k}(\boldsymbol{x})\boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}\nabla\boldsymbol{V}^*=0\quad\boldsymbol{V}^*(0)=0\tag{10}$$

式中：$\bar{\boldsymbol{R}}=(s_1\ \cdots\ s_m)\in\mathbf{R}^m$；$\boldsymbol{D}^*=\frac{1}{2\lambda}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\boldsymbol{V}^*$；$\underline{1}$ 是一个所有元素都等于 1 的列向量。

假设 1　选择 $\gamma>0$ 且式（1）系统是零状态可观的。在集合 $\Omega\in\mathbf{R}^n$ 上存在一个控制策略 $\boldsymbol{u}(\boldsymbol{x})$ 使系统渐近稳定且有 $L_2-g_{\mathrm{gain}}\leqslant\gamma$，则式（10）HJI 方程在 Ω 上存在一个光滑解。

此处，假设 1 保证了非线性系统 H_∞ 控制问题的有解性。

2　基于 ADP 的无模型策略迭代算法

非线性系统的 HJI 方程是一个非线性的偏微分方程，其解析解不一定存在，可用策略迭代方法得到其近似解，算法步骤如下。

1）分别给定初始稳定的控制策略 $\boldsymbol{u}_1$ 和扰动策略 $\boldsymbol{w}_1$，此时 $i=1$。

2）策略评价。依据已知 $\boldsymbol{u}_1$ 和 $\boldsymbol{w}_1$，利用如下 Lyapunov 方程求解 $\boldsymbol{V}_i$，即

$$r(\boldsymbol{x},\boldsymbol{u}_i,\boldsymbol{w}_i)+\nabla \boldsymbol{V}_i^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}_i+\boldsymbol{k}\boldsymbol{w}_i)=0 \quad \boldsymbol{V}_i(0)=0 \text{ 。} \tag{11}$$

3) 策略提高。$i=i+1$,更新 $\boldsymbol{u}_i$ 和 $\boldsymbol{w}_i$,即

$$\begin{cases}\boldsymbol{u}_{i+1}(\boldsymbol{x})=-\lambda\tanh(\boldsymbol{D}_{i+1})\\ \boldsymbol{w}_{i+1}(\boldsymbol{x})=\dfrac{1}{2\gamma^2}\boldsymbol{R}\boldsymbol{k}^{\mathrm{T}}\nabla \boldsymbol{V}_i\end{cases} \tag{12}$$

式中:$\boldsymbol{D}_{i+1}=\dfrac{1}{2\lambda}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla \boldsymbol{V}_i$。

4) 如果 $\|\boldsymbol{V}_{i+1}-\boldsymbol{V}_i\|\leqslant\varepsilon,\varepsilon>0$,则停止计算,输出 $\boldsymbol{V}_i$,否则 $i=i+1$,并转至步骤2)。

定理1 假设1成立且满足 Kantorovich's 收敛条件[11],在式(11)和式(12)迭代相当于牛顿方法求解式(10)的 HJI 方程,当 $i\to\infty$ 时,$\boldsymbol{u}_i\to\boldsymbol{u}^*$,$\boldsymbol{w}_i\to\boldsymbol{w}^*$,$\boldsymbol{V}_i\to\boldsymbol{V}^*$。$\boldsymbol{V}^*$为值函数最优解,$(\boldsymbol{u}^*,\boldsymbol{w}^*)$为最优纳什均衡策略。

证明 文献[10]已进行详细证明,本文不赘述。

式(4)是 $\boldsymbol{V}_i$ 的线性偏微分方程。迭代求解式(4)相对于直接求解 HJI 方程更为可行,但是迭代方法需要系统完整的模型信息,有的采用 IRL 算法,可以不依赖 $\boldsymbol{f}$,但其他信息仍必要。但是在非线性系统考虑输入约束时具体模型往往是难以获得的。

通过模型依赖 PI 算法和 IRL 算法推导无模型的 ADP 算法,给出两个随机的控制策略 $\boldsymbol{u}$ 和扰动策略 $\boldsymbol{w}$ 使式(1)系统在一个闭区间上稳定,$\boldsymbol{u}_i,\boldsymbol{w}_i$ 表示式(11)和式(12)第 i 次迭代的结果,进一步计算 $\boldsymbol{V}_i,\boldsymbol{u}_{i+1}$ 和 $\boldsymbol{w}_{i+1}$。$\boldsymbol{V}_i$ 对时间的导数 $\dot{\boldsymbol{V}}_i=\nabla \boldsymbol{V}_i^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}+\boldsymbol{k}\boldsymbol{w})$,根据式(11)和式(12)有

$$\dot{\boldsymbol{V}}_i=\nabla \boldsymbol{V}_i^{\mathrm{T}}(\boldsymbol{g}(\boldsymbol{u}-\boldsymbol{u}_i)+\boldsymbol{k}(\boldsymbol{w}-\boldsymbol{w}_i))-r(\boldsymbol{x},\boldsymbol{u}_i,\boldsymbol{w}_i)=$$
$$2\lambda(\boldsymbol{D}_i)^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{u}-\boldsymbol{u}_i)+2\gamma^2\boldsymbol{w}_{i+1}^{\mathrm{T}}(\boldsymbol{w}-\boldsymbol{w}_i)-r(\boldsymbol{x},\boldsymbol{u}_i,\boldsymbol{w}_i) \text{ 。} \tag{13}$$

根据 IRL 算法,同时在区间 $[t,t+\Delta t]$ 对式(13)两边积分得到

$$\boldsymbol{V}_i(t+\Delta t)-\boldsymbol{V}_i(t)-\int_t^{t+\Delta t}2\lambda(\boldsymbol{D}_{i+1})^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{u}-\boldsymbol{u}_i)\mathrm{d}\tau-$$
$$\int_t^{t+\Delta t}2\gamma^2\boldsymbol{w}_{i+1}^{\mathrm{T}}(\boldsymbol{w}-\boldsymbol{w}_i)\mathrm{d}\tau+\int_t^{t+\Delta t}r(\boldsymbol{x},\boldsymbol{u}_i,\boldsymbol{w}_i)\mathrm{d}\tau=0 \tag{14}$$

式中,$\boldsymbol{V}_i,\boldsymbol{D}_{i+1}$ 和 $\boldsymbol{w}_{i+1}$ 是需要求解的未知函数或函数向量。

无模型迭代 ADP 算法主要思想是求解式(14)无模型方程来代替求解式(11)和式(12)的基于模型迭代方程。式(14)方程中不再包含系统的模型信息 $\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{g}(\boldsymbol{x})$ 和 $\boldsymbol{k}(\boldsymbol{x})$,而是利用系统的实时数据 $\boldsymbol{u}_i$ 和 $\boldsymbol{w}_i$。事实上系统的模型信息包含于可用的系统数据中。所以,在用无模型策略学习方法近似求解式(14)方程前,需要在线采样收集所需要的可用系统数据。

3 算法实现

使用3个神经网络(评价网络、控制网络、扰动网络)分别近似值函数 $\boldsymbol{V}_i$、控制策略 $\boldsymbol{u}_{i+1}$ 和扰动策略 $\boldsymbol{w}_{i+1}$。

根据 Weirstrass 高阶近似理论[12],式(14)的解 $\boldsymbol{V}_i$ 用神经网络表示为

$$\boldsymbol{V}_1^i(\boldsymbol{x})=\boldsymbol{W}_{1,i+1}^{\mathrm{T}}\boldsymbol{\phi}_1(\boldsymbol{x})+\boldsymbol{\varepsilon}_{1,i+1}(\boldsymbol{x}) \tag{15}$$
$$\boldsymbol{D}_{i+1}(\boldsymbol{x})=\boldsymbol{W}_{2,i+1}^{\mathrm{T}}\boldsymbol{\phi}_2(\boldsymbol{x})+\boldsymbol{\varepsilon}_{2,i+1}(\boldsymbol{x}) \tag{16}$$
$$\boldsymbol{w}_{i+1}(\boldsymbol{x})=\boldsymbol{W}_{3,i+1}^{\mathrm{T}}\boldsymbol{\phi}_3(\boldsymbol{x})+\boldsymbol{\varepsilon}_{3,i+1}(\boldsymbol{x}) \tag{17}$$
$$\boldsymbol{D}_i(\boldsymbol{x})=\boldsymbol{W}_{2,i}^{\mathrm{T}}\boldsymbol{\phi}_2(\boldsymbol{x})+\boldsymbol{\varepsilon}_{2,i}(\boldsymbol{x}) \tag{18}$$
$$\boldsymbol{w}_i(\boldsymbol{x})=\boldsymbol{W}_{3,i}^{\mathrm{T}}\phi_3(\boldsymbol{x})+\boldsymbol{\varepsilon}_{3,i}(\boldsymbol{x}) \tag{19}$$

其中:$\boldsymbol{\phi}_1:\mathbf{R}^n\to\mathbf{R}^{K_1}$,$\boldsymbol{\phi}_2:\mathbf{R}^n\to\mathbf{R}^{K_2}$,$\boldsymbol{\phi}_3:\mathbf{R}^n\to\mathbf{R}^{K_3}$,为线性独立基础函数向量;$\boldsymbol{W}_{1,\cdot}\in\boldsymbol{R}^{k_1}$,$\boldsymbol{W}_{2,\cdot}\in\boldsymbol{R}^{k_2\times m}$,$\boldsymbol{W}_{3,\cdot}\in\mathbf{R}^{k_3\times q}$为权值矩阵;$\boldsymbol{\varepsilon}_{1,\cdot}$,$\boldsymbol{\varepsilon}_{2,\cdot}$ 和 $\boldsymbol{\varepsilon}_{3,\cdot}$ 是神经网络重建误差;K_1,K_2 和 K_3 是隐藏层神经元数。假设基函数参数和重建误差在 Ω 上,当 $K_1\to\infty$,$K_2\to\infty$,$K_3\to\infty$ 时,$\boldsymbol{\varepsilon}_{1,\cdot}\to0,\boldsymbol{\varepsilon}_{2,\cdot}\to0,\boldsymbol{\varepsilon}_{3,\cdot}\to0$。

为确定 $\boldsymbol{W}_{1,i+1},\boldsymbol{W}_{2,i+1},\boldsymbol{W}_{3,i+1}$ 的理想参数,用 $\hat{\boldsymbol{W}}_{1,i+1}$,$\hat{\boldsymbol{W}}_{2,i+1}$,$\hat{\boldsymbol{W}}_{3,i+1}$ 表示理想权值矩阵,神经网络表示为

$$\hat{\boldsymbol{V}}_i(\boldsymbol{x})=\hat{\boldsymbol{W}}_{1,i+1}^{\mathrm{T}}\boldsymbol{\phi}_1(\boldsymbol{x}) \tag{20}$$
$$\hat{\boldsymbol{D}}_{i+1}(\boldsymbol{x})=\hat{\boldsymbol{W}}_{2,i+1}^{\mathrm{T}}\boldsymbol{\phi}_2(\boldsymbol{x}) \tag{21}$$
$$\hat{\boldsymbol{w}}_{i+1}(\boldsymbol{x})=\hat{\boldsymbol{W}}_{3,i+1}^{\mathrm{T}}\boldsymbol{\phi}_3(\boldsymbol{x}) \tag{22}$$

式中,$\hat{\boldsymbol{W}}_{1,i+1}$,$\hat{\boldsymbol{W}}_{2,i+1}$ 和 $\hat{\boldsymbol{W}}_{3,i+1}$ 是当前的估计值。

为多次迭代,定义一个严格增的时间序列 $\{t_k\}_{k=0}^{l}$,$l>0$ 为采集数据点的数量。由于 $\hat{\boldsymbol{V}}_i(\boldsymbol{x})$,$\hat{\boldsymbol{D}}_{i+1}(\boldsymbol{x})$,$\hat{\boldsymbol{w}}_{i+1}(\boldsymbol{x})$ 替代 $\boldsymbol{V}_i(\boldsymbol{x})$,$\boldsymbol{D}_{i+1}(\boldsymbol{x})$,$\boldsymbol{w}_{i+1}(\boldsymbol{x})$ 代入方程(14),因此存在残差表示为

$$e_k=\hat{\boldsymbol{V}}_i(t_{k+1})-\hat{\boldsymbol{V}}_i(t_k)+\int_{t_k}^{t_{k+1}}r(\boldsymbol{x},\boldsymbol{u}_i,\boldsymbol{w}_i)\mathrm{d}\tau-$$
$$\int_{t_k}^{t_{k+1}}2\lambda(\hat{\boldsymbol{D}}_{i+1})^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{u}-\boldsymbol{u}_i)\mathrm{d}\tau-\int_{t_k}^{t_{k+1}}2\gamma^2\boldsymbol{w}_{i+1}^{\mathrm{T}}(\boldsymbol{w}-\boldsymbol{w}_i)\mathrm{d}\tau=$$
$$(\boldsymbol{\phi}_1(\boldsymbol{x}(t_{k+1}))-\boldsymbol{\phi}_1(\boldsymbol{x}(t_k)))^{\mathrm{T}}\hat{\boldsymbol{W}}_{1,i+1}-2\lambda\int_{t_k}^{t_{k+1}}[\boldsymbol{\phi}_2(\boldsymbol{x})^{\mathrm{T}}\cdot$$
$$\hat{\boldsymbol{W}}_{2,i+1}\boldsymbol{R}(\boldsymbol{u}+\lambda\tanh(\hat{\boldsymbol{W}}_{2,i}^{\mathrm{T}}\boldsymbol{\phi}_2(\boldsymbol{x})))]-\int_{t_k}^{t_{k+1}}2\gamma^2\boldsymbol{\phi}_3(\boldsymbol{x})^{\mathrm{T}}\cdot$$
$$\hat{\boldsymbol{W}}_{3,i+1}(\boldsymbol{w}-\hat{\boldsymbol{W}}_{3,i+1}^{\mathrm{T}}\boldsymbol{\phi}_3(\boldsymbol{x}))\mathrm{d}\tau+\int_{t_k}^{t_{k+1}}[\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x})-$$
$$\boldsymbol{\phi}_3(\boldsymbol{x})^{\mathrm{T}}\hat{\boldsymbol{W}}_{3,i+1}\hat{\boldsymbol{W}}_{3,i+1}^{\mathrm{T}}\boldsymbol{\phi}_3(\boldsymbol{x})+$$
$$\int_0^{-\lambda\tanh(\boldsymbol{W}_{2,i}^{\mathrm{T}}\boldsymbol{\phi}_2(\boldsymbol{x}))}(\lambda\tanh^{-1}(s/\lambda))^{\mathrm{T}}\boldsymbol{R}\mathrm{d}s]\mathrm{d}\tau \text{ 。} \tag{23}$$

根据 Kronecker 积⊗,式(23)可改写为

$$e_k=\boldsymbol{\theta}_k(\bar{\boldsymbol{W}}_i)^{\mathrm{T}}\bar{\boldsymbol{W}}_{i+1}+\xi(\bar{\boldsymbol{W}}_i) \tag{24}$$

式中：$\bar{\boldsymbol{W}}_{i+1}=(\hat{\boldsymbol{W}}_{1,i+1}^{\mathrm{T}},vec\,\hat{\boldsymbol{W}}_{2,i+1}^{\mathrm{T}},vec\,\hat{\boldsymbol{W}}_{3,i+1}^{\mathrm{T}})^{\mathrm{T}}\in\mathbf{R}^{\bar{K}}$，$\bar{K}=K_1+mK_2+qK_3$，$\bar{\boldsymbol{W}}_k$可用相同方式表达，符号 $vec(\cdot)$ 表示矩阵向量化，迭代标志 $i\in\{0,1,\cdots\}$，时间序列标志 $k\in\{0,\cdots,l\}$，$\boldsymbol{\theta}_k(\bar{\boldsymbol{W}}_i)$ 和 $\xi_k(\bar{\boldsymbol{W}}_i)$ 可以分别定义为

$$\boldsymbol{\theta}_k(\bar{\boldsymbol{W}}_i)=\begin{pmatrix}\boldsymbol{\phi}_1(\boldsymbol{x}(t_{k+1}))-\boldsymbol{\phi}_1(\boldsymbol{x}(t_k))\\-\int_{t_k}^{t_{k+1}}2\lambda\boldsymbol{R}(\boldsymbol{u}+\lambda\tanh(\hat{\boldsymbol{W}}_{2,i}^{\mathrm{T}}\boldsymbol{\phi}_2))\otimes\boldsymbol{\phi}_2\mathrm{d}\tau\\-\int_{t_k}^{t_{k+1}}2\gamma^2(\boldsymbol{w}-\hat{\boldsymbol{W}}_{3,i}^{\mathrm{T}}\boldsymbol{\phi}_3)\otimes\boldsymbol{\phi}_3\mathrm{d}\tau\end{pmatrix}\tag{25}$$

$$\xi_k(\bar{\boldsymbol{W}}_i)=\int_{t_k}^{t_{k+1}}[\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{\phi}_3(\boldsymbol{x})^{\mathrm{T}}\hat{\boldsymbol{W}}_{3,i+1}\hat{\boldsymbol{W}}_{3,i+1}^{\mathrm{T}}\boldsymbol{\phi}_3(\boldsymbol{x})+\int_0^{-\lambda\tanh(\boldsymbol{W}_{2,i}^{\mathrm{T}}\boldsymbol{\phi}_2(\boldsymbol{x}))}(\lambda\tanh^{-1}(s/\lambda))^{\mathrm{T}}\boldsymbol{R}\mathrm{d}s]\mathrm{d}\tau\ 。\tag{26}$$

根据最小二乘（Least-Squares，LS）原理找到一组权值使残差最小，即

$$\min_{\bar{\boldsymbol{W}}_{i+1}}\sum_{k=0}^{l-1}e_k^2\ 。\tag{27}$$

假设 2　持续激励（Persistency of Excitation，PE）对于 $i\geqslant0$ 存在 $l_0>0$ 和 $\delta>0$，因此对所有的 $l\geqslant l_0$ 有

$$\frac{1}{l}\sum_{k=0}^{l-1}\boldsymbol{\theta}_k(\bar{\boldsymbol{W}}_i)\boldsymbol{\theta}_k(\bar{\boldsymbol{W}}_i)^{\mathrm{T}}\geqslant\delta\boldsymbol{I}_{\bar{K}}\tag{28}$$

式中，$\boldsymbol{I}_{\bar{K}}$ 是适当维数的单位矩阵。

根据 $\boldsymbol{\theta}_k(\bar{\boldsymbol{W}}_i)$ 的定义，为了保证 PE 条件，$\boldsymbol{u}$ 和 $\boldsymbol{\omega}$ 需要与 $\hat{\boldsymbol{u}}_{i+1}$ 和 $\hat{\boldsymbol{\omega}}_{i+1}$ 大不相同而且状态需要被持续激励，所以 $\boldsymbol{u}$ 和 $\boldsymbol{\omega}$ 设计为深度探测信号。LS 表达如下

$$\bar{\boldsymbol{W}}_{i+1}=-(\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i)^{\mathrm{T}}\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i))^{-1}\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i)^{\mathrm{T}}\boldsymbol{\Xi}(\bar{\boldsymbol{W}}_i)\tag{29}$$

其中，

$$\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i)=(\boldsymbol{\theta}_0(\bar{\boldsymbol{W}}_i)\quad\cdots\quad\boldsymbol{\theta}_{l-1}(\bar{\boldsymbol{W}}_i))^{\mathrm{T}};\tag{30}$$

$$\boldsymbol{\Xi}(\bar{\boldsymbol{W}}_i)=(\xi_0(\bar{\boldsymbol{W}}_i)\quad\cdots\quad\xi_{l-1}(\bar{\boldsymbol{W}}_i))^{\mathrm{T}}\ 。\tag{31}$$

当计算得到 $\bar{\boldsymbol{W}}_{i+1}$，它就替换 $\bar{\boldsymbol{W}}_i$ 开始下面的迭代。

所提出的 ADP 算法实际上是无模型策略学习方法，通过给定适当的初始策略权值和在线采集系统数据计算出 $\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_k)$ 和 $\boldsymbol{\Xi}(\bar{\boldsymbol{W}}_k)$，而后用式（25）迭代求解 $\bar{\boldsymbol{W}}_{k+1}$，最后利用式（20）～式（22）近似求解 $\hat{V}_i(\boldsymbol{x})$，$\hat{\boldsymbol{D}}_{i+1}(\boldsymbol{x})$ 和 $\hat{\boldsymbol{w}}_{i+1}(\boldsymbol{x})$，进而得到 H_∞ 控制器。

为了式（29）中矩阵 $\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i)^{\mathrm{T}}\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i)$ 的逆存在，矩阵 $\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i)$ 是列满秩的，在实际计算中，采样次数 l 需要满足 $l\geqslant\mathrm{rank}(\boldsymbol{\Theta}(\bar{\boldsymbol{W}}_i))$，即 $l\geqslant K_1+mK_2+qK_3$。

图 1 为 ADP 算法的流程图，包括两个阶段，在采集阶段时，输入深度控制信号和扰动信号，在线采集系统数据。充分时间后，算法转到学习阶段，训练神经网络权值，采用离线迭代收敛方式使权值收敛到固定值，如果没有收敛，则需要重新返回第一阶段采集更多数据，最终，控制网络输出 H_∞ 控制器，其中 $\boldsymbol{u}'\in\Omega$ 为随机容许控制策略。

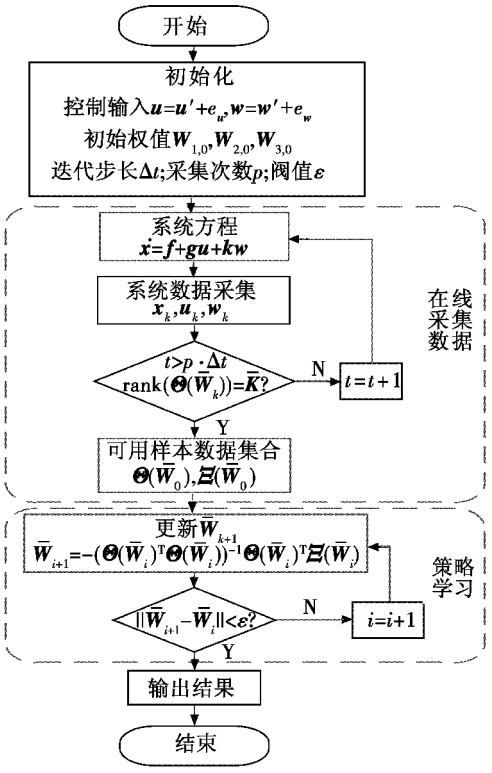

图 1　ADP 算法流程图

Fig. 1　Flow chart of ADP algorithm

定理 2　如果假设 2 成立，对 $\forall\varepsilon>0$ 存在 $i^*>0$，$K_1^*>0$，$K_2^*>0$ 和 $K_3^*>0$，那么如果 $i^*>i$，$K_1>K_1^*$，$K_2>K_2^*$ 和 $K_3>K_3^*$，则

$$\begin{cases}|\hat{V}_i(\boldsymbol{x})-V_i(\boldsymbol{x})|\leqslant\varepsilon\\\|\hat{\boldsymbol{D}}_i-\boldsymbol{D}_i\|\leqslant\varepsilon\\\|\hat{\boldsymbol{w}}_i-\boldsymbol{w}_i\|\leqslant\varepsilon\\|\hat{V}_i(\boldsymbol{x})-V^*(\boldsymbol{x})|\leqslant\varepsilon\\\|\hat{\boldsymbol{D}}_i-\boldsymbol{D}^*\|\leqslant\varepsilon\\\|\hat{\boldsymbol{w}}_i-\boldsymbol{w}^*\|\leqslant\varepsilon\end{cases}\tag{32}$$

对所有 $\boldsymbol{x}\in\Omega$ 成立，类似的结论证明详见文献[12]。

4　算例仿真

本章通过 Matlab 软件仿真结果说明本文算法的有效性。

考虑非线性系统

$$\dot{\boldsymbol{x}}=\boldsymbol{f}(\boldsymbol{x})+\boldsymbol{g}(\boldsymbol{x})\boldsymbol{u}+\boldsymbol{k}(\boldsymbol{x})\boldsymbol{w}\tag{33}$$

式中：$\boldsymbol{f}(\boldsymbol{x})=\begin{pmatrix}-0.25x_1\\0.5x_2^3-0.25\gamma^{-2}x_1^2x_2-0.25x_2\end{pmatrix}$；$\boldsymbol{g}(\boldsymbol{x})=(0\ \ x_2)^{\mathrm{T}}$；$k(x)=(0\ \ x_1)^{\mathrm{T}}$；$\boldsymbol{x}=(x_1\ \ x_2)^{\mathrm{T}}\in\mathbf{R}^2$ 为系统

状态，$\boldsymbol{u}\in\mathbf{R}^m$ 和 $\boldsymbol{w}\in\mathbf{R}^q$ 分别为控制输入和扰动输入。

选择 $\boldsymbol{h}(\boldsymbol{x})=(x_1\quad x_2)^{\mathrm{T}}$，$\boldsymbol{R}=\boldsymbol{I}$，$\gamma=4$，$\boldsymbol{I}$ 是单位矩阵。控制输入 $\boldsymbol{u}$ 被约束为 $\|\boldsymbol{u}\|\leqslant 1$，那么 $Y(\boldsymbol{u})$ 可以定义为

$$Y(\boldsymbol{u})=2\int_0^{\boldsymbol{u}}\tanh^{-\mathrm{T}}(s)\boldsymbol{R}\mathrm{d}s\quad 。\tag{34}$$

根据近似式(21)和式(22)，选择 3 个 3 层的前馈神经网络，其基函数向量定义为 $\boldsymbol{\phi}_1(1)=\boldsymbol{\phi}_2(1)=\boldsymbol{\phi}_3(1)=(x_1^2\quad x_1x_2\quad x_2^2\quad x_1^4\quad x_1^4)^{\mathrm{T}}$。

初始状态设置为 $\boldsymbol{x}_0=(0.5\quad -0.5)^{\mathrm{T}}$，选择深度控制输入 $\boldsymbol{u}$ 和扰动输入 $\boldsymbol{w}$，即

$$\boldsymbol{u}=-\tanh(\sin 1.1\pi t+\sin 1.4\pi t+\sin 1.2\pi t+\sin 2.9\pi t-3.2\sin 3.2\pi t+1)\quad ,\tag{35}$$

$$\boldsymbol{w}=1.1\sin\pi t-\sin 1.5\pi t-\sin 1.8\pi t-\sin 2\pi t+0.2\quad 。\tag{36}$$

收敛阈值为 $\varepsilon=10^{-6}$，积分计算步长为 0.1 s，采样次数 $l=50$，所以，无模型学习阶段在 5 s 时，通过采样数据获得最优控制策略。神经网络初始迭代权值 $\boldsymbol{W}_1^0=\boldsymbol{W}_2^0=\boldsymbol{W}_3^0=(0\quad 0\quad 0\quad 0\quad 0)^{\mathrm{T}}$。

图 2 为评价网络的权值 $\boldsymbol{W}_1$ 随迭代次数的变化；图 3 为控制网络的权值 $\boldsymbol{W}_2$ 随迭代次数的变化；图 4 为扰动网络的权值 $\boldsymbol{W}_3$ 随迭代次数的变化。5 s 后，最优控制策略和最优扰动策略代替深度控制信号加在被控系统，全局的系统状态曲线如图 5 所示，图 6 为系统的控制输入曲线，图 7 为系统的全部扰动输入曲线。

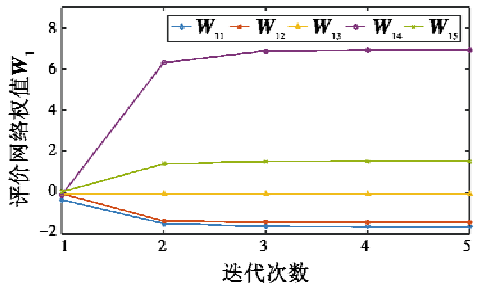

图 2 W_1 参数更新

Fig. 2 W_1 parameter updating

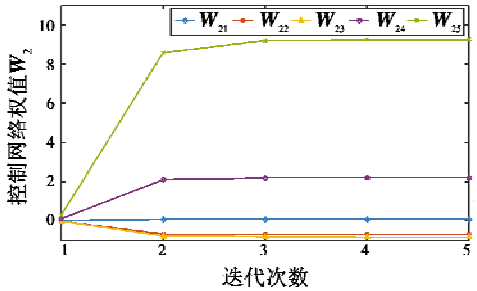

图 3 W_2 参数更新

Fig. 3 W_2 parameter updating

由图 2 ~ 图 4 可以看出，评价网络、控制网络和扰动网络的权值 $\boldsymbol{W}_1$，$\boldsymbol{W}_2$，$\boldsymbol{W}_3$ 在 4 次迭代后收敛情况分别为 $\boldsymbol{W}_1=(-1.632\quad -1.424\quad -0.065\quad 6.919\quad 1.537)$，$\boldsymbol{W}_2=(0.071\quad -0.692\quad -0.838\quad 2.197\quad 9.272)$，$\boldsymbol{W}_3=(-0.043\quad 0.005\quad 0.005\quad 0.018\quad -0.147)$。

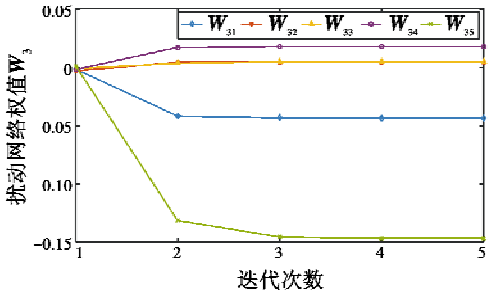

图 4 W_3 参数更新

Fig. 4 W_3 parameter updating

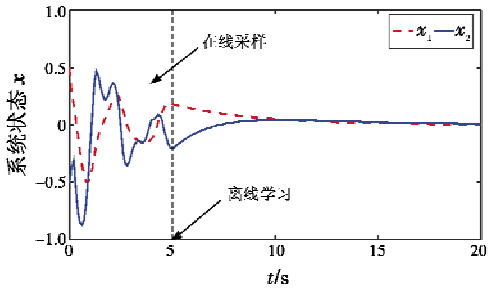

图 5 系统状态

Fig. 5 Curve of system state

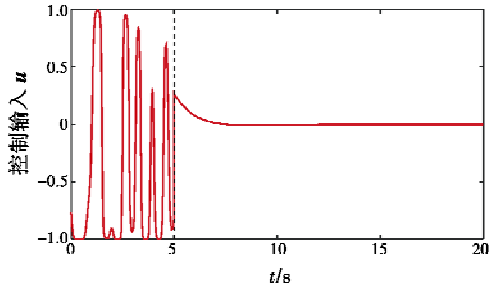

图 6 控制输入 $\boldsymbol{u}$

Fig. 6 Control input $\boldsymbol{u}$

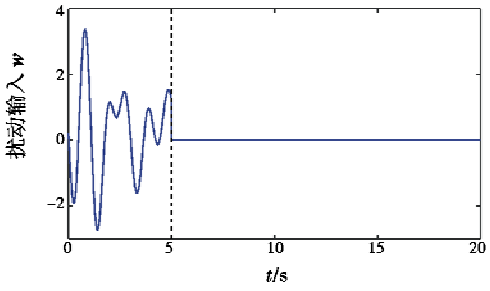

图 7 扰动输入 $\boldsymbol{w}$

Fig. 7 Disturbance input $\boldsymbol{w}$

进而可以得出鞍点为

$$\boldsymbol{u}^*(\boldsymbol{x})=0.071x_1^2-0.692x_1x_2-0.838x_2^2+2.197x_1^4+9.272x_1^4\quad ,\tag{37}$$

$$w^*(x) = -0.043x_1^2 + 0.005x_1x_2 + 0.005x_2^2 + 0.018x_1^4 - 0.147x_1^4 \text{ 。} \quad (38)$$

由图 5 可以看出，5 s 时，系统由采样阶段进入无模型策略学习阶段，5 s 后，系统在所设计的控制器和最坏扰动条件下，逐渐收敛到零，验证了所得控制策略的合理性。从图 6、图 7 可以看出，控制输入被约束在 $\|u\| \leqslant 1$，控制和扰动输入最终收敛到零。

5　结束语

本文针对一类内部动力模型未知的非线性系统，基于数据驱动的自适应动态规划思想，提出了一种通过在线采集系统数据，迭代求解输入约束 H_∞ 最优控制问题的无模型算法。该算法分为在线采集数据和离线学习两个部分，在离线学习时，构建 3 个神经网络，利用实时系统数据，近似求解无模型 HJI 方程，最后，算例仿真验证了算法的有效性。

参 考 文 献

[1] BERTSEKAS D P. Dynamic programming and optimal control[M]. Belmont: Athena Scientific, 1995.

[2] 顾凡，马清亮，岳瑞华，等. 基于遗传算法的多项式非线性系统静态输出反馈 H_∞ 控制[J]. 电光与控制，2017，24(10)：55-58，89.

[3] ZHANG H, WEI Q, LIU D. An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential game[J]. Automatica, 2011, 47(1): 207-214.

[4] LIU D R, YANG X, WANG D, et al. Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints[J]. IEEE Transactions on Cybernetics, 2015, 45(7): 1372-1385.

[5] LUO B, WU H N. Computationally efficient simultaneous policy update algorithm for nonlinear H_∞ state feedback control with Galerkin's method[J]. International Journal of Robust and Nonlinear Control, 2013, 23(7): 991-1012.

[6] YASINI S, KARIMPOUR A, SISTANI M B, et al. Online concurrent reinforcement learning algorithm to solve two-player zero-sum games for partially unknown nonlinear continuous-time systems[J]. International Journal of Adaptive Control and Signal Processing, 2015, 29(4): 473-493.

[7] LUO B, HUANG T, WU H N, et al. Data-driven H_∞ control for nonlinear distributed parameter systems[J]. IEEE Transactions on Neural Networks & Learning Systems, 2015, 26(11): 2949-2961.

[8] YANG X, LIU D, WEI Q, et al. Adaptive dynamic programming for H_∞ control of constrained-input nonlinear systems[C]//Proceedings of the 34th Chinese Control Conference, IEEE, 2015: 3027-3032.

[9] ZHU Y H, ZHAO D B, LI X J, et al. Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data[J]. IEEE Transactions on Neural Networks & Learning Systems, 2017, 28(3): 714-725.

[10] ZHANG Q C, ZHAO D B, ZHU Y H. Data-driven adaptive dynamic programming for continuous-time fully cooperative games with partially constrained inputs[J]. Neurocomputing, 2017, 238: 377-386.

[11] TAPIA R A. The Kantorovich theorem for Newton's method[J]. The American Mathematical Monthly, 1971, 78(4): 389-392.

[12] FINLAYSON B A. The method of weighted residuals and variational principles[M]. New York: Academic Press, Inc., 2014.