控制与决策
Control and Decision

基于零和博弈的多智能体网络鲁棒包容控制

于镝

引用本文:
于镝. 基于零和博弈的多智能体网络鲁棒包容控制[J]. 控制与决策, 2021, 36(8): 1841−1848.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1348

基于零和博弈的多智能体网络鲁棒包容控制

于 镝

(北京信息科技大学 自动化学院，北京 100192)

摘 要: 针对受扰非线性多智能体网络,研究分布式鲁棒包容控制方法. 采用微分博弈理论将有界$\mathcal{L}_2$增益包容控制问题描述成多玩家零和博弈问题. 对于每个跟随者,当至少有一个领航者与其存在有向路径通信时,基于局部邻居信息定义每个跟随者的性能指标,从而得出包容误差$\mathcal{L}_2$有界且零和博弈解存在的结论. 在系统动态完全未知的情况下,采用积分强化学习算法和执行-评价-干扰网络,在线得到近似最优策略. 仿真结果表明了所提出方案的有效性和正确性.

关键词: 多智能体网络；鲁棒包容控制；零和博弈；积分强化学习

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1348 **开放科学(资源服务)标识码(OSID):**

引用格式: 于镝. 基于零和博弈的多智能体网络鲁棒包容控制[J]. 控制与决策, 2021, 36(8): 1841-1848.

Robust containment control of multi-agent networks based on zero-sum game

YU Di

(College of Automation，Beijing Information Science and Technology University，Beijing 100192，China)

Abstract: The distributed robust containment control methods are investigated for disturbed nonlinear multi-agent networks. Applying the differential game theory, the bounded $\mathcal{L}_2$ gain containment control problem is described as multi-player zero-sum game one. When there exists at least one leader that has a directed path from it to each follower, its performance index is defined based on the information of local neighbors. Furthermore, it is proved that the containment errors are $\mathcal{L}_2$ bounded and there exists Nash equilibrium solution. With the completely unknown system dynamics, the integral reinforcement learning method and critic-actor-disturbance neural networks are used to solve the approximate optimal strategy online. Simulation results verify the effectiveness and correctness of the proposed scheme.

Keywords: multi-agent networks；robust containment control；zero-sum game；integral reinforcement control

0 引 言

多智能体网络协调控制的研究作为控制领域的前沿课题,深受研究人员的青睐,而且已在诸多工程领域中得到了广泛成功的应用. 例如,自组装机器人聚集、无人机火灾救援、卫星姿态调整和智能电网分配等. 作为典型的协调控制,包容控制的潜在应用前景涵盖了危险物资搬运、火灾救援等军事和民用方面. 在包容控制系统中,存在多个领航者,并且跟随者的运动限定在领航者所围成的最小几何空间内. 迄今为止,在多智能体网络包容控制研究方面已经涌现出很多优秀的研究成果[1-4].

目前,大多数研究成果均要求系统动态已知且非最优控制. 而在实际应用中,救援和搬运机器人需在尽可能短的时间内且能量损耗最小的情况下,将人员或物资转移到安全地点. 因此,它们必须适应不可预测、连续变化的环境,在安全任务中学习采取最优行动得到最优性能. 博弈理论为多智能体网络动态优化问题的求解提供了极其合适的工具. 博弈论为动态交互网络提供了表示多参与者决策控制问题的环境,从而网络中智能体之间的策略交互可建模为多玩家同时运动的博弈[5]. 针对线性离散网络,文献[6]基于博弈论思想,解决了数据驱动的多智能体网络一致问题. 而针对非线性多智能体网络,文献[7-8]在给出领航-跟随非线性微分图博弈描述的基础上,采用评价-执行框架和梯度下降法实现了最优控制策略的估计,并且设计依赖系统动态的算法以实现分布式跟踪

收稿日期: 2019-09-25；**修回日期:** 2020-04-29.
基金项目: 北京信息科技大学学科群建设项目(5121911003).
责任编委: 左志强.
†通讯作者. E-mail: yudizlg@aliyun.com.

控制.

实际应用中,网络个体经常受到外部的干扰,例如测量噪声、敌对方对网络个体的攻击以及外部环境的变化所导致动态的不确定性. 为了保证网络个体顺利完成任务或者在受到攻击后具有防御性或复原性,研究人员主要采用零和博弈框架来研究多智能体网络分布式鲁棒控制. 零和博弈是竞争类博弈,其意味着当一个玩家赢时,另一个玩家就输. 在控制系统中,零和博弈与干扰抑制的H_∞问题联系紧密. 文献[9]研究具有未知动态的受限输入非线性系统有限域内的跟踪问题,其中采用零和博弈理论以及离策略控制方法实现系统在有限时间内跟踪上目标系统. 针对非线性连续系统,文献[10]将H_∞跟踪问题转化为有界L_2增益跟踪问题. 由跟踪Hamilton-Jacobi-Issacs(HJI)方程得出纳什平衡解,并分析了系统的稳定性,同时给出了保证跟踪误差局部渐近稳定时折扣因子的上界. 在系统动态未知的情况下,通过离策略强化学习算法求解跟踪HJI方程的解.

上述成果均限于单个系统. 基于零和博弈理论和梯度下降法,文献[11-12]求解近似最优控制策略,分别解决了多个轮式机器人的同步问题和线性多智能体网络的干扰抑制问题. 文献[13]针对非线性多智能体网络,结合零和博弈理论和自适应动态规划思想,构造评价神经网络在线逼近协调代价函数,从而实现网络跟踪控制. 但上述成果的最优策略均依赖于系统动态. 在实际应用中,外部环境的复杂性很难获得精确的系统动态信息,因此,本文受文献[10,12]的启发,采用零和博弈理论和积分强化学习(integral reinforcement learning, IRL)思想,给出包容误差$\mathcal{L}_2$有界以及零和博弈Nash平衡解存在的条件,并在提出基于模型的策略迭代学习算法的基础上,设计无模型策略迭代算法在线执行近似最优控制策略,从而实现多智能体网络鲁棒包容控制. 本文从以下3个方面对现有成果进行了拓展: 1) 与文献[9-10]相比,考虑多智能体网络的鲁棒包容控制,比单个系统的跟踪控制要复杂得多; 2) 与文献[6-8]相比,考虑受扰多智能体网络的协调控制,更具实际意义; 3) 与文献[11-13]相比,考虑基于无模型策略迭代算法的多智能体网络近似最优鲁棒包容控制,降低了对系统动态的限制.

1 问题描述

考虑由n个智能体所构成的网络,用$F=\{1,2,\dots,m\}$和$L=\{m+1,\dots,n\}$分别代表跟随者和领航者索引集合,则$\mathcal{V}$包括跟随者集合$\mathcal{V}_F=\{\nu_i,i\in F\}$和领航者集合$\mathcal{V}_L=\{\nu_i,i\in L\}$.

1.1 网络动态

考虑m个跟随者,其动态描述为

$$\dot{x}_i=f(x_i)+g(x_i)u_i+k(x_i)\omega_i,\ i\in F. \tag{1}$$

其中: $x_i\in\boldsymbol{R}^p,u_i\in\boldsymbol{R}^q$和$\omega_i\in\boldsymbol{R}^l$分别代表第$i$个跟随者的状态矢量、控制输入矢量和有界干扰矢量; $f(x_i)\in\boldsymbol{R}^p,g(x_i)\in\boldsymbol{R}^{p\times q}$和$k(x_i)\in\boldsymbol{R}^{p\times l}$分别表示转移动态、输入动态和干扰动态,且均为紧集$\chi\in\boldsymbol{R}^p$上的未知局部Lipschitz函数, $f(0)=0$. 为了研究方便,令$g(x_i)$和$k(x_i)$均为有界常矢量.

领航者的动态描述为

$$\dot{x}_i=h_i(x_i),\ i\in L. \tag{2}$$

其中: $x_i\in\boldsymbol{R}^p$代表第i个领航者的状态矢量; $h(x_i)\in\boldsymbol{R}^p$表示状态$x_i$的连续未知函数,并且$0<\|h(x_i)\|<h_M,\forall x_i\in\boldsymbol{R}^p$且$h_i(0)=0$. 令跟随者和领航者状态矢量为$x_F=[x_1^{\mathrm{T}},\dots,x_m^{\mathrm{T}}]^{\mathrm{T}}$和$x_L=[x_{m+1}^{\mathrm{T}},\dots,x_n^{\mathrm{T}}]^{\mathrm{T}}$,跟随者的控制矢量为$u_F=[u_1,\dots,u_m]^{\mathrm{T}}$,且$f(x_F)=[f^{\mathrm{T}}(x_1),f^{\mathrm{T}}(x_2),\dots,f^{\mathrm{T}}(x_m)]^{\mathrm{T}},g(x_F)=[g^{\mathrm{T}}(x_1),g^{\mathrm{T}}(x_2),\dots,g^{\mathrm{T}}(x_m)]^{\mathrm{T}}$.

1.2 网络拓扑

令领航者之间无通信,且领航者与跟随者之间通信是单向的,即领航者发送信息,则跟随者之间的网络拓扑和领航者与跟随者之间的网络拓扑能够决定整个网络通信. 由此对Laplacian阵$\mathcal{L}$进行结构划分,有

$$\mathcal{L}=\begin{bmatrix}\mathcal{T} & \mathcal{T}_d\\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)}\end{bmatrix}.$$

其中: $\mathcal{T}\in\boldsymbol{R}^{m\times m},\mathcal{T}_d\in\boldsymbol{R}^{m\times(n-m)}$.

假设1 对于每个跟随者,至少存在一个领航者与其存在有向路径通信.

1.3 网络误差

定义网络误差为

$$e_i=\sum_{j=1}^{n}a_{ij}(x_i-x_j),\ i\in F, \tag{3}$$

则网络误差动态为

$$\begin{aligned}\dot{e}_i=&\sum_{j=1}^{n}a_{ij}(\dot{x}_i-\dot{x}_j)=\\ &\Phi_i+d_ig(x_i)u_i-\sum_{j\in F}a_{ij}g(x_j)u_j+\\ &d_ik(x_i)\omega_i-\sum_{j\in F}a_{ij}k(x_j)\omega_j.\end{aligned} \tag{4}$$

其中: $d_i=\sum_{j\in N_i}a_{ij},N_i=\{\nu_j,a_{ij}\neq0,j=1,2,\dots,n\},\Phi_i=\sum_{j\in F}a_{ij}(f(x_i)-f(x_j))+\sum_{k\in L}a_{ik}(f(x_i)-$

$h_k(x_k)$). 由此可知,每个跟随者的网络误差动态由其自身和其所有邻居的信息所决定.

由网络拓扑和网络误差定义可得整个网络的误差动态,可描述为$E = \mathcal{T}x_F + \mathcal{T}_d x_L$,其中$E = [e_1^{\mathrm{T}}, \dots, e_m^{\mathrm{T}}]^{\mathrm{T}}$. 由文献[14]中引理3.1可知,跟随者的期望状态矢量可表示为$x_d = -\mathcal{T}^{-1}\mathcal{T}_d x_L$,其中$x_d = [x_{d1}^{\mathrm{T}}, \dots, x_{dm}^{\mathrm{T}}]^{\mathrm{T}}$. 令$e_c = x_F - x_d$代表包容误差,其中$e_c = [e_{c1}^{\mathrm{T}}, \dots, e_{cm}^{\mathrm{T}}]^{\mathrm{T}}$,则网络误差和包容误差满足$E = \mathcal{T}e_c$. 本文的控制目的是在有干扰情况下,设计分布式控制策略使得包容误差$\mathcal{L}_2$有界,从而使得跟随者渐近收敛到领航者所围成的凸包中. 下面先给出两个定义.

定义1　设X是实矢量空间$V \subseteq R^n$. 用$\mathrm{Co}(X)$表示X的凸包,即

$$\mathrm{Co}(X) = \Big\{ \sum_{i=1}^{k} \alpha_i x_i | x_i \in X, \alpha_i \in R,\ \alpha_i \geqslant 0,\ \sum_{i=1}^{k} \alpha_i = 1,\ k = 1, 2, \dots \Big\}.$$

定义2　考虑由动态(1)和(2)所构成的多智能体网络,对于所有的跟随者有$\omega_i(t) \neq 0, i \in F$以及给定的$\gamma > 0$,寻找最优控制策略满足如下有界$\mathcal{L}_2$增益条件:

$$\int_{t_0}^{\infty} (e_i^{\mathrm{T}} Q_i e_i + u_i^{\mathrm{T}} R_i u_i)\mathrm{d}t \leqslant \gamma^2 \int_{t_0}^{\infty} \omega_i^{\mathrm{T}} P_i \omega_i \mathrm{d}t + V_i(e_i(t_0)). \tag{5}$$

其中: V_i为有界函数且$V_i(t_0) = 0$,Q_i、R_i、P_i均为对称正定矩阵.

2　主要结果

2.1　多玩家零和博弈

为每个跟随者定义性能指标

$$J_i(e_i(t_0), u_i, u_{-i}, \omega_i, \omega_{-i}) = \int_{t_0}^{\infty} (e_i^{\mathrm{T}} Q_i e_i + u_i^{\mathrm{T}} R_i u_i - \gamma^2 \omega_i^{\mathrm{T}} P_i \omega_i)\mathrm{d}t,\ i \in F. \tag{6}$$

其中: $u_{-i} = \{u_j : j \in N_i\}$,$\omega_{-i} = \{\omega_j : j \in N_i\}$. 有界$\mathcal{L}_2$增益包容控制问题与下列多玩家零和博弈问题是等价的:

$$V_i(e_i(t_0)) = \min_{u_i} \max_{\omega_i} J_i(e_i(t_0), u_i, u_{-i}, \omega_i, \omega_{-i}). \tag{7}$$

若博弈意义上的鞍点(u_i^*, ω_i^*)存在,则该博弈具有唯一解,即若

$$V_i^*(e_i(t_0)) = \min_{u_i} \max_{\omega_i} J_i(e_i(t_0), u_i, u_{-i}^*, \omega_i, \omega_{-i}^*) = \max_{\omega_i} \min_{u_i} J_i(e_i(t_0), u_i, u_{-i}^*, \omega_i, \omega_{-i}^*), \tag{8}$$

则V_i^*称为该博弈值. 由式(7)可见,在控制策略最小化性能指标的同时,干扰却要最大化,即

$$\begin{aligned} &J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i^*, \omega_{-i}^*) \leqslant \\ &J_i(e_i(t_0), u_i, u_{-i}^*, \omega_i^*, \omega_{-i}^*), \\ &J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i^*, \omega_{-i}^*) \geqslant \\ &J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i, \omega_{-i}^*). \end{aligned} \tag{9}$$

于是,与式(8)等价的Nash平衡条件为

$$\begin{aligned} &J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i, \omega_{-i}^*) \leqslant \\ &J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i^*, \omega_{-i}^*) \leqslant \\ &J_i(e_i(t_0), u_i, u_{-i}^*, \omega_i^*, \omega_{-i}^*), \end{aligned} \tag{10}$$

其中$u_i, \omega_i, i \in F$.

对于第i个跟随者,定义其值函数为

$$V_i(e_i(t_0)) = \int_{t_0}^{\infty} (e_i^{\mathrm{T}} Q_i e_i + u_i^{\mathrm{T}} R_i u_i - \gamma^2 \omega_i^{\mathrm{T}} P_i \omega_i)\mathrm{d}t. \tag{11}$$

由此可得下列Bellman方程:

$$\begin{aligned} &H_i(e_i, \nabla V_i, u_i, u_{-i}, \omega_i, \omega_{-i}) \equiv \\ &\nabla V_i \Big(\Phi_i + d_i g(x_i) u_i - \sum_{j \in F} a_{ij} g(x_j) u_j + \\ &d_i k(x_i) \omega_i - \sum_{j \in F} a_{ij} k(x_j) \omega_j \Big) + \\ &e_i^{\mathrm{T}} Q_i e_i + u_i^{\mathrm{T}} R_i u_i - \gamma^2 \omega_i^{\mathrm{T}} P_i \omega_i, \end{aligned} \tag{12}$$

其中$\nabla V_i = \partial V_i / \partial e_i$. 由静止条件可得

$$u_i^*(t) = -\frac{1}{2} d_i R_i^{-1} g^{\mathrm{T}}(x_i) \nabla V_i^*, \tag{13}$$

$$\omega_i^*(t) = \frac{1}{2\gamma^2} d_i P_i^{-1} k^{\mathrm{T}}(x_i) \nabla V_i^*, \tag{14}$$

则耦合HJI方程为

$$(\nabla V_i^*)^{\mathrm{T}} \Pi_i + e_i^{\mathrm{T}} Q_i e_i + \Xi_i = 0. \tag{15}$$

其中

$$\begin{aligned} \Pi_i = &\Phi_i - \frac{d_i^2}{2} g(x_i) R_i^{-1} g^{\mathrm{T}}(x_i) \nabla V_i^* + \\ &\frac{d_i^2}{2\gamma^2} k(x_i) P_i^{-1} k^{\mathrm{T}}(x_i) \nabla V_i^* - \\ &\frac{d_j}{2} \sum_{j \in F} a_{ij} g(x_j) R_j^{-1} g^{\mathrm{T}}(x_j) \nabla V_j^* - \\ &\frac{d_j}{2\gamma^2} \sum_{j \in F} a_{ij} k(x_j) P_j^{-1} k^{\mathrm{T}}(x_j) \nabla V_j^*, \end{aligned} \tag{16}$$

$$\begin{aligned} \Xi_i = &\frac{d_i^2}{4} (\nabla V_i^*)^{\mathrm{T}} g(x_i) R_i^{-1} g^{\mathrm{T}}(x_i) \nabla V_i^* - \\ &\frac{d_i^2}{4\gamma^4} (\nabla V_i^*)^{\mathrm{T}} k(x_i) P_i^{-1} k^{\mathrm{T}}(x_i) \nabla V_i^*. \end{aligned} \tag{17}$$

由此,该零和博弈问题需要求解m个耦合HJI方程.

2.2 $\mathcal{L}_2$有界的包容误差和零和博弈的Nash平衡解

对于给定干扰抑制水平$\gamma > 0$和所有干扰$\omega_i(t) \in \mathcal{L}_2[t_0,\infty)$,本节给出使得有界$\mathcal{L}_2$增益条件满足的控制策略,并且给出在某些条件下,HJI方程的解满足Nash条件(10),由此解得零和博弈.

定理1 令$\gamma > \gamma^*$,假设$V_i^* > 0, i \in F$是HJI方程(15)的光滑正定解. 假定邻居智能体的策略均为最优,则:

1) 当控制策略$u_i^*(t)$如式(13)所示,且当$\omega_i(t) = 0, i \in F$时,网络误差动态渐近稳定;

2) 当所有跟随者均选择各自的最优控制策略$u_i^*(t)$时,对于所有的干扰都有有界$\mathcal{L}_2$增益条件(5)成立.

证明 因$V_i^* > 0, i \in F$是HJI方程(15)的光滑正定解,所以

$$V_i^*(e_i(t+\Delta t)) - V_i^*(e_i(t)) = \int_t^{t+\triangle t}(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^* - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i)\mathrm{d}\tau.$$

当$\triangle t \to 0$时,得

$$\frac{\mathrm{d}V_i^*(e_i)}{\mathrm{d}t} = -(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^* - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i).$$

1) 当$\omega_i(t) = 0$时,因为Q_i、R_i均为正定矩阵,所以

$$\frac{\mathrm{d}V_i^*(e_i)}{\mathrm{d}t} = -(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^*) < 0.$$

因此网络误差动态渐近稳定,由文献[14]中引理3.1得知跟随者渐近趋于期望状态.

2) 当所有跟随者均选择各自的最优控制策略$u_i^*(t)$时,由式(10)得知,对于所有的干扰都有

$$V_i^*(e_i(\infty)) - V_i^*(e_i(t_0)) = -\int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^* - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i)\mathrm{d}\tau,$$

则

$$V_i^*(e_i(\infty)) + \int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^*)\mathrm{d}\tau = \int_{t_0}^{\infty}\gamma^2\omega_i^{\mathrm{T}}P_i\omega_i\mathrm{d}\tau + V_i^*(e_i(t_0)),$$

即

$$\int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^*)\mathrm{d}\tau \leqslant \gamma^2\int_{t_0}^{\infty}\omega_i^{\mathrm{T}}P_i\omega_i\mathrm{d}\tau + V_i^*(e_i(t_0)).$$

所以有界$\mathcal{L}_2$增益条件(5)成立. □

推论1 若定理1中的条件均满足,并且假设1成立,则可得包容误差$\mathcal{L}_2$有界.

注1 由定理1以及网络误差与包容误差的关系式$E = \mathcal{T}e_c$,可得包容误差$\mathcal{L}_2$有界.

定理2 令$\gamma > \gamma^*$,假设博弈(7)具有有限值并且邻居智能体的策略均为最优. 令$V_i^* > 0, i \in F$是HJI方程(15)的光滑正定解,使得网络误差动态(4)渐近稳定,则当$u_i^*(t)$和$\omega_i^*(t)$分别为式(12)和(13)所示时,整个网络满足Nash平衡条件(10),而且博弈值为$V_i^*(e_i(t_0))$.

证明 当$V_i^* > 0, i \in m$是HJI方程(15)的光滑正定解,且使得网络误差动态(4)渐近稳定时,有$e_i(\infty) = 0, V_i^*(e_i(\infty)) = 0$. 所以

$$J_i(e_i(t_0), u_i, u_{-i}, \omega_i, \omega_{-i}) = V_i^*(e_i(\infty)) + \int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + u_i^{\mathrm{T}}R_iu_i - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i)\mathrm{d}t = V_i^*(e_i(t_0)) + \int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + u_i^{\mathrm{T}}R_iu_i - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i)\mathrm{d}t - \int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^* - \gamma^2(\omega_i^*)^{\mathrm{T}}P_i\omega_i^*)\mathrm{d}t.$$

计算$H_i(e_i, \nabla V_i^*, u_i, u_{-i}, \omega_i, \omega_{-i})$和$H_i(e_i, \nabla V_i^*, u_i^*, u_{-i}^*, \omega_i^*, \omega_{-i}^*)$,可得

$$\int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + u_i^{\mathrm{T}}R_iu_i - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i)\mathrm{d}t - \int_{t_0}^{\infty}(e_i^{\mathrm{T}}Q_ie_i + (u_i^*)^{\mathrm{T}}R_iu_i^* - \gamma^2(\omega_i^*)^{\mathrm{T}}P_i\omega_i^*)\mathrm{d}t = \int_{t_0}^{\infty}\Big((u_i - u_i^*)^{\mathrm{T}}R_i(u_i - u_i^*) - \gamma^2\int_{t_0}^{\infty}\Big((\omega_i - \omega_i^*)^{\mathrm{T}}R_i(\omega_i - \omega_i^*) - \nabla V_i^{\mathrm{T}}\sum_{j\in F}a_{ij}g_i(x_i)(u_j^* - u_j)\Big)\mathrm{d}t - \nabla V_i^{\mathrm{T}}\sum_{j\in F}a_{ij}g_i(x_i)(\omega_j^* - \omega_j)\Big)\mathrm{d}t.$$

当$u_{-i} = u_{-i}^*, \omega_{-i} = \omega_{-i}^*$时,有

$$J_i(e_i(t_0), u_i, u_{-i}^*, \omega_i, \omega_{-i}^*) = V_i^*(e_i(t_0)) + \int_{t_0}^{\infty}(u_i - u_i^*)^{\mathrm{T}}R_i(u_i - u_i^*)\mathrm{d}\tau - \gamma^2\int_{t_0}^{\infty}(\omega_i - \omega_i^*)^{\mathrm{T}}R_i(\omega_i - \omega_i^*)\mathrm{d}\tau,$$

可知满足Nash平衡条件(10),即

$$J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i, \omega_{-i}^*) \leqslant J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i^*, \omega_{-i}^*) \leqslant J_i(e_i(t_0), u_i, u_{-i}^*, \omega_i^*, \omega_{-i}^*),$$

而且博弈值

$$J_i(e_i(t_0), u_i^*, u_{-i}^*, \omega_i^*, \omega_{-i}^*) = V_i^*(e_i(t_0)).$$ □

注2 定理1从稳定性的角度出发,表明当所有跟随者取得最优控制策略时,可保证包容误差有界且实现鲁棒包容控制. 定理2从博弈论角度出发,表明最优控制策略和最优干扰策略同时满足Nash平衡条件. 定理2为定理1提供了最优化情况,即当整个网络实现Nash平衡时,多智能体网络能够在克服最坏干扰情况下,实现较高精度和耗能最小的鲁棒包容控制.

2.3 求解HJI方程的策略迭代算法

由前述可知,实现网络的鲁棒包容控制,需要求解m个耦合博弈HJI方程(15). 但时变HJI方程为非线性偏微分方程,一般很难求解,因此,本节首先提出基于模型的IRL算法,然后提出无模型的IRL算法.

算法1 基于模型的策略迭代算法.

令$V_i^0 \in V_0$为初始代价函数,其数值可由文献[15]中的引理5所确定. 因此,初始控制策略为

$$u_i^{(0)} = -\frac{1}{2}d_i R_i^{-1} g_i^{\mathrm{T}}(x_i)\nabla V_i^{(0)},$$

初始扰动策略为

$$\omega_i^{(0)} = \frac{1}{2\gamma^2}d_i P_i^{-1} k_i^{\mathrm{T}}(x_i)\nabla V_i^{(0)}.$$

令$k=0$,基于模型的策略迭代算法的步骤如下.

step 1: 根据下式求解值函数$V_i^{(k+1)}$:

$$(\nabla V_i^{(k+1)})^{\mathrm{T}}\Big(\Phi_i + d_i g(x_i)u_i - \sum_{j\in F} a_{ij}g(x_j)u_j + d_i k(x_i)\omega_i - \sum_{j\in F} a_{ij}k(x_j)\omega_j\Big) + r(e_i,u_i,\omega_i) = 0, \tag{18}$$

其中$r(e_i,u_i,\omega_i) = e_i^{\mathrm{T}}Q_i e_i + u_i^{\mathrm{T}}R_i u_i - \gamma^2\omega_i^{\mathrm{T}}P_i\omega_i$.

step 2: 由下式更新控制策略和扰动策略:

$$u_i^{(k+1)} = -\frac{1}{2}d_i R_i^{-1} g^{\mathrm{T}}(x_i)\nabla V_i^{(k+1)},$$

$$\omega_i^{(k+1)} = \frac{d_i}{2\gamma^2}P_i^{-1} k^{\mathrm{T}}(x_i)\nabla V_i^{(k+1)}.$$

step 3: 令$k=k+1$,若$\|V_i^{(k)} - V_i^{(k+1)}\| \leqslant \varepsilon$,$\varepsilon$是小的正实数,则停止并获得最优代价函数$V_i^* = V_i^{(k)}$、最优控制策略$u_i^* = u_i^{(k)}$及最优扰动策略$\omega_i^* = \omega_i^{(k)}$;否则返回step 1继续迭代.

算法1的收敛性证明如下.

首先给出如下定理.

定理3 基于算法1,迭代序列$V_i^{(k+1)}$、$u_i^{(k+1)}$和$\omega_i^{(k+1)}$都收敛到其最优值,即当$k\to\infty$时,有$V_i^{(k+1)} \to V_i^*, u_i^{(k+1)} \to u_i^*$和$\omega_i^{(k+1)} \to \omega_i^*, i\in F$.

注3 依据牛顿迭代法和Gâteaux导数与Frechet导数之间的关系,可以证明算法1的收敛性,具体可参见文献[15]中的定理1.

很显然算法1依赖于系统动态信息,然而,在复杂环境下很难获得这些信息. 因此,下面提出无模型策略迭代算法.

算法2 无模型的策略迭代算法.

受强化学习中探索未知信息和利用已有信息之间寻求平衡思想的启发,网络误差动态(4)还可写为

$$\begin{aligned}\dot{e}_i = {}& \Phi_i + d_i g(x_i)u_i^{(k)} - \sum_{j\in F} a_{ij}g(x_j)u_j^{(k)} + d_i k(x_i)\omega_i^{(k)} - \\ & \sum_{j\in F} a_{ij}k(x_j)\omega_j^{(k)} + d_i g(x_i)(u_i - u_i^{(k)}) - \\ & \sum_{j\in F} a_{ij}g(x_j)(u_j - u_j^{(k)}) + d_i k(x_i)(\omega_i - \omega_i^{(k)}) - \\ & \sum_{j\in F} a_{ij}k(x_j)(\omega_j - \omega_j^{(k)}).\end{aligned} \tag{19}$$

其中: $u_i = u_i^{(k)} + n_{ui}$, $u_j = u_j^{(k)} + n_{uj}$, $\omega_i = \omega_i^{(k)} + n_{wi}$, $\omega_j = \omega_j^{(k)} + n_{wj}, i\in F$. 探索信号$n_{ui}, n_{uj}, n_{wi}, n_{wj} \in \vartheta$且$\vartheta$为有界集. 根据文献[16],探索信号要从有界集中选取,并且要保证闭环系统的输入-状态稳定性. 其中有界集的上界可从文献[16]中的定理2中得到. 沿着轨迹(19)对V_i^{k+1}求导,可得

$$\begin{aligned}\frac{\mathrm{d}V_i^{(k+1)}}{\mathrm{d}t} = {}& (\nabla V_i^{(k+1)})^{\mathrm{T}}\Big[\Phi_i + d_i g(x_i)u_i^{(k)} - \sum_{j\in F} a_{ij}g(x_j)u_j^{(k)} + \\ & d_i k(x_i)\omega_i^{(k)} - \sum_{j\in F} a_{ij}k(x_j)\omega_j^{(k)} + d_i g(x_i)n_{ui} - \\ & \sum_{j\in F} a_{ij}g(x_j)n_{uj} + d_i k(x_i)n_{wi} - \\ & \sum_{j\in F} a_{ij}k(x_j)n_{wj}\Big].\end{aligned} \tag{20}$$

应用式(18)可得

$$\begin{aligned}\frac{\mathrm{d}V_i^{(k+1)}}{\mathrm{d}t} = {}& -r(e_i,u_i,\omega_i) - (u_i^{(k+1)})^{\mathrm{T}}R_i n_{ui} + \\ & \frac{2}{d_i}(u_i^{(k+1)})^{\mathrm{T}}R_i\sum_{j\in F} a_{ij}n_{uj} + 2\gamma^2(\omega_i^{(k+1)})^{\mathrm{T}}P_i n_{wi} - \\ & \frac{2\gamma^2}{d_i}(\omega_i^{(k+1)})^{\mathrm{T}}P_i\sum_{j\in F} a_{ij}n_{wj}.\end{aligned} \tag{21}$$

然后式(21)两端在t与$t+T$之间取积分,有

$$\begin{aligned}V_i^{(k+1)}(e_i(t+T)) = {}& V_i^{(k+1)}(e_i(t)) - \int_t^{t+T} r(e_i,u_i^{(k)},\omega_i^{(k)})\mathrm{d}\tau - \\ & \int_t^{t+T}(u_i^{(k+1)})^{\mathrm{T}}R_i\Big(n_{ui} + \frac{2}{d_i}\sum_{j\in F} a_{ij}n_{uj}\Big)\mathrm{d}\tau + \\ & 2\gamma^2\int_t^{t+T}(\omega_i^{(k+1)})^{\mathrm{T}}P_i\Big(n_{\omega i} - \frac{1}{d_i}\sum_{j\in F} a_{ij}n_{\omega j}\Big)\mathrm{d}\tau,\end{aligned} \tag{22}$$

其中T为强化采样间隔. 由式(22)可得,迭代不需要已知系统的动态信息$f(x)$、$g(x)$和$k(x)$,从而得出无模型的IRL迭代算法,其步骤如图1所示,初始条件与算法1相同.

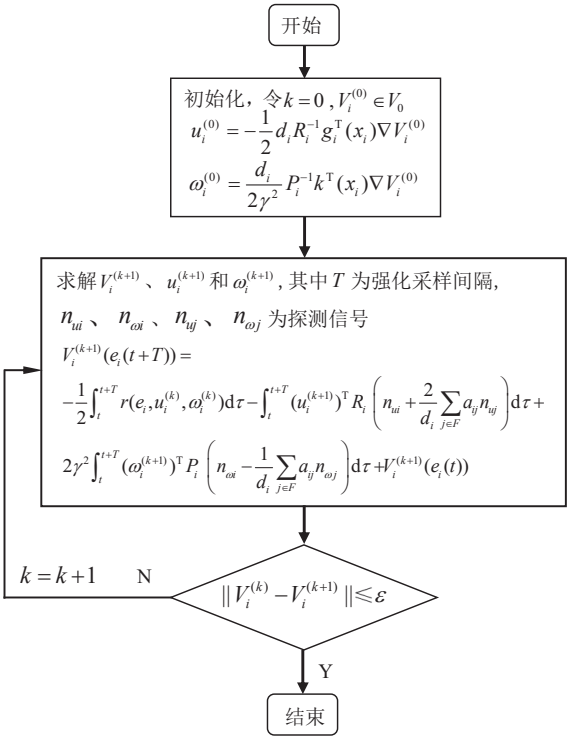

图 1 无模型IRL迭代算法流程

定理4 采用算法2,当$k\to\infty$时,有$V_i^{(k+1)}\to V_i^*,u_i^{(k+1)}\to u_i^*,\omega_i^{(k+1)}\to\omega_i^*$.

注4 依据算法2的推导过程,可以证明算法1与算法2等价. 由前面算法1的收敛性知,当$k\to\infty$时,有$V_i^{(k+1)}\to V_i^*,u_i^{(k+1)}\to u_i^*,\omega_i^{(k+1)}\to\omega_i^*$.

注5 实际上,算法2中的状态信息、控制输入信息和干扰输入信息已包含了未知动态信息,所以算法2与算法1的控制效果是等价的. 因此在系统动态未知的情况下,算法2同样可以达到鲁棒包容控制的目的,从而降低了系统动态已知的要求或者避免了辨识系统动态的过程.

2.4 算法2的在线执行

为了实现算法2,对第i个跟随者应用3个神经网络分别逼近控制策略$u_i^{(k)}$、干扰策略$\omega_i^{(k)}$和代价函数$V_i^{(k)}$. 在此种情况下,3个神经网络都具有输入-隐层-输出3层结构,并且它们的输出由下式给出:

$$\begin{aligned}&\hat{V}_i^{(k+1)}(e_i)=\hat{\theta}_i^{\mathrm{T}}\varphi(e_i),\\&\hat{u}_i^{(k+1)}(e_i)=\hat{\varpi}_i^{\mathrm{T}}\phi(e_i),\\&\hat{\omega}_i^{(k+1)}(e_i)=\hat{\vartheta}_i^{\mathrm{T}}\rho(e_i).\end{aligned}\tag{23}$$

其中:$\varphi=[\varphi_1,\dots,\varphi_{r_1}]\in R^{r_1},\phi=[\phi_1,\dots,\phi_{r_2}]\in R^{r_2}$和$\rho=[\rho_1,\dots,\rho_{r_3}]\in R^{r_3}$为合适的隐层激励函数矢量;$\hat{\theta}_i^{\mathrm{T}}\in R^{r_1}$,$\hat{\varpi}_i^{\mathrm{T}}\in R^{r_2}$和$\hat{\vartheta}_i^{\mathrm{T}}\in R^{r_3}$为常权值矢量的估值. 有

$$\begin{aligned}\delta_i(t)=&\hat{\theta}_i^{\mathrm{T}}(\varphi(e_i(t+T))-\varphi(e_i(t)))+\\&\int_t^{t+T}r(e_i,u_i^{(k)},\omega_i^{(k)})\mathrm{d}\tau+\\&\sum_{j'=1}^{q}r_{ij'}\int_t^{t+T}\hat{\varpi}_{i,j'}^{\mathrm{T}}\phi(e_i(\tau))\delta_u\mathrm{d}\tau-\\&2\gamma^2\sum_{j'=1}^{l}p_{ij'}\int_t^{t+T}\hat{\vartheta}_{i,j'}^{\mathrm{T}}\rho(e_i(\tau))\delta_\omega\mathrm{d}\tau.\end{aligned}\tag{24}$$

其中:$\delta_i(t)$是逼近误差,$\delta_u=n_{ui}+\frac{2}{d_i}\sum_{j\in F}a_{ij}n_{uj},\delta_\omega=n_{\omega i}-\frac{1}{d_i}\sum_{j\in F}a_{ij}n_{\omega j},R_i=\mathrm{diag}\{r_{i1},\dots,r_{iq}\},P_i=\mathrm{diag}\{p_{i1},\dots,p_{il}\}$. 然后重新整理式(24),可得

$$z_i(t)+\delta_i(t)=\hat{W}_i^{\mathrm{T}}y_i(t).\tag{25}$$

其中

$$z_i(t)=-\int_t^{t+T}r(e_i,u_i^{(k)},\omega_i^{(k)})\mathrm{d}\tau,$$

$$\hat{W}_i=[\theta_i^{\mathrm{T}},\varpi_{i,1}^{\mathrm{T}},\dots,\varpi_{i,q}^{\mathrm{T}},\vartheta_{i,1}^{\mathrm{T}},\dots,\vartheta_{i,l}^{\mathrm{T}}]^{\mathrm{T}},$$

$$y_i(t)=\begin{bmatrix}\varphi(e_i(t+T))-\varphi(e_i(t)\\r_{i1}\int_t^{t+T}\phi(e_i(\tau))\left(n_{ui}+\frac{2}{d_i}\sum_{j\in F}a_{ij}n_{uj}\right)\mathrm{d}\tau\\\vdots\\r_{iq}\int_t^{t+T}\phi(e_i(\tau))\left(n_{ui}+\frac{2}{d_i}\sum_{j\in F}a_{ij}n_{uj}\right)\mathrm{d}\tau\\-2\gamma^2p_{i1}\int_t^{t+T}\rho(e_i(\tau))\left(n_{\omega i}+\frac{1}{d_i}\sum_{j\in F}a_{ij}n_{\omega j}\right)\mathrm{d}\tau\\\vdots\\-2\gamma^2p_{il}\int_t^{t+T}\rho(e_i(\tau))\left(n_{\omega i}+\frac{1}{d_i}\sum_{j\in F}a_{ij}n_{\omega j}\right)\mathrm{d}\tau\end{bmatrix}.$$

为了最小化逼近误差,采用最小二乘法进行计算. 假定从时间t_1到t_K内,每隔相同的时间间隔T对系统数据进行充分的采样,共得到$K\geqslant r_1+r_2q+r_3l$组系统数据,于是可得到K组数据构成$Y_i=[y_i^{\mathrm{T}}(t_1),\dots,y_i^{\mathrm{T}}(t_K)]$和$Z_i=[z_i(t_1),\dots,z_i(t_K)]^{\mathrm{T}}$. 则最小二乘解为$\hat{W}_i=(Y_iY_i^{\mathrm{T}})^{-1}Y_iZ_i$. 因此得到$V_i^{(k+1)}$、$u_i^{(k+1)}$和$\omega_i^{(k+1)}$的近似值.

3 仿真研究

实验1 考虑由8个智能体组成的多智能体网络. 有向拓扑如图2所示.

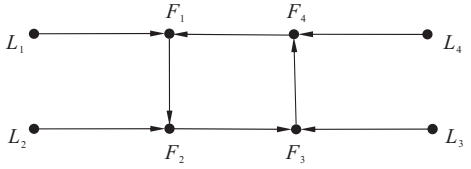

图 2 网络拓扑结构1

第i个跟随者动态由下式所描述:

$$\dot{x}_i=f(x_i)+g(x_i)u_i+k(x_i)\omega_i,\ i\in F.\tag{26}$$

其中: $x_i \triangleq [x_{i1}, x_{i2}]^{\mathrm{T}}$, $f(x_i) = [x_{i2}, -x_{i1} + 0.5(1 - x_{i1}^2)x_{i2}]^{\mathrm{T}}$, $g(x_i) = [0, -0.8]^{\mathrm{T}}$ 和 $k(x_i) = [0, 0.07]^{\mathrm{T}}$. 非线性干扰选为 $\omega_i = x_{i2}\sin(x_{i1})^3\cos(0.5x_{i2})$. 式(5)中的参数选为 $Q_i = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$, $R_i = P_i = 1, i \in F, \gamma = 0.1$. 对于第 i 个跟随者, 其评价器NN、执行器NN和干扰器NN的激励函数分别选为 $\varphi(e_i) = [e_i^2, e_i\dot{e}_i, \dot{e}_i^2, e_i^4, e_i^3\dot{e}_i, e_i^2\dot{e}_i^2, e_i\dot{e}_i^3, \dot{e}_i^4]^{\mathrm{T}}$ 和 $\phi(e_i) = \rho(e_i) = [2e_i, \dot{e}_i, 0, 3e_i^3, 3e_i^2\dot{e}_i, 2e_i\dot{e}_i^2, \dot{e}_i^3, 0]^{\mathrm{T}}$. 采样周期选为 $T = 0.01$ 且探索信号的选择与文献[16]类似. 智能体的运动轨迹曲线和包容误差变化曲线如图3和图4所示. 在图3中, 红色实心圆点代表跟随者的初始位置, 蓝色实心圆点代表动态领航者分别在不同时刻的位置. 而且, 4种不同线型的曲线代表跟随者的实际运动轨迹, 黑色方框代表领航者所围成的动态凸包. 由上述仿真结果可得, 大约在10 s左右跟随者进入领航者所围成的凸包并保持在领航者所围成的凸包中, 在其期望轨迹的小邻域内运动, 并且在25 s后跟随者的运动轨迹趋于稳定. 可见, 基于本文的控制方案和所提出的无模型IRL算法, 可以实现受扰多智能体网络的鲁棒最优包容控制, 且得到零和博弈的Nash平衡解.

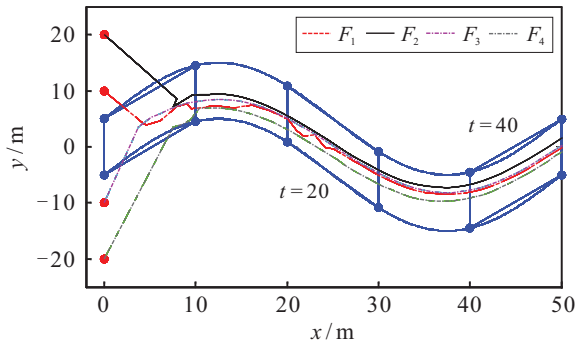

图 3　多智能体网络运动轨迹1

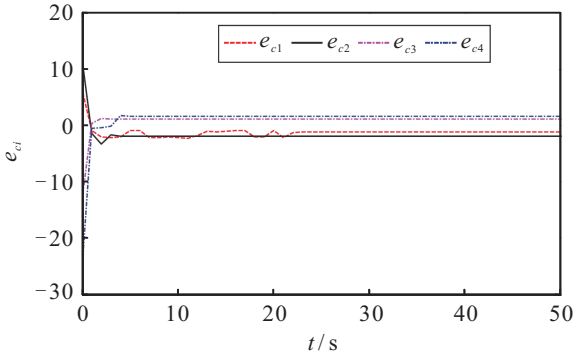

图 4　包容误差变化曲线1

实验2　考虑由4个领航者和6个跟随者所组成的多智能体网络, 有向拓扑如图5所示. 跟随者动态、领航者动态、参数选择和评价-执行-干扰框架选择同实验1. 智能体的运动轨迹曲线和包容误差变化曲线如图6和图7所示. 同样可以得出本文的控制方案有效可行.

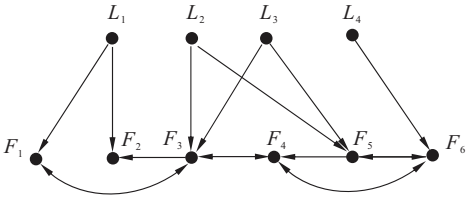

图 5　网络拓扑结构2

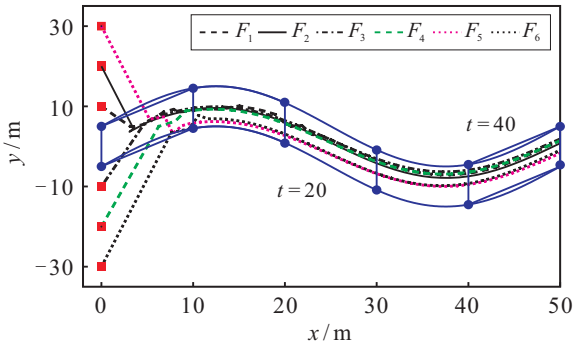

图 6　多智能体网络运动轨迹2

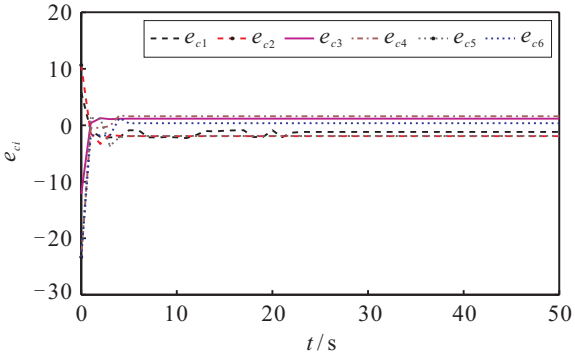

图 7　包容误差变化曲线2

4　结　论

为了使智能体学习采取最优行动而在任务中取得快速、准确和最优性能, 本文提出了受扰多智能体网络鲁棒包容控制新方法. 基于零和博弈思想和积分强化学习算法, 在证明零和博弈Nash平衡解存在且网络包容误差 $\mathcal{L}_2$ 有界的基础上, 提出了无模型策略迭代学习算法, 并且采用执行-评价-干扰网络框架, 在线实现网络的近似最优鲁棒包容控制. 下一步将针对异构非线性多智能体网络鲁棒包容控制展开研究.

参考文献(References)

[1] Li D Y, Zhang W, He W, et al. Two-layer distributed formation-containment control of multiple Euler-Lagrange systems by output feedback[J]. IEEE Transactions on Cybernetics, 2019, 49(2): 675-687.

[2] Zhu Y R, Zheng Y S, Wang L. Containment control of switched multi-agent systems[J]. International Journal of Control, 2015, 88(12): 2570-2577.

[3] Mei J, Ren W, Li B, et al. Distributed containment control

for multiple unknown second-order nonlinear systems with application to networked Lagrangian systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(9): 1885-1899.

[4] Yu D, Ji X Y. Finite-time containment control of perturbed multi-agent systems based on sliding-mode control[J]. International Journal of Systems Science, 2018, 49(2): 299-311.

[5] 谭拂晓, 刘德荣, 关新平, 等. 基于微分对策理论的非线性控制回顾与展望[J]. 自动化学报, 2014, 40(1): 1-15.
(Tan F X, Liu D R, Guan X P, et al. Review and perspective of nonlinear systems control based on differential games[J]. Acta Automatica Sinica, 2014, 40(1): 1-15.)

[6] Ren H, Zhang H G, Wen Y L, et al. Integral reinforcement learning off-policy method for solving nonlinear multi-player nonzero-sum games with saturated actuator[J]. Neurocomputing, 2019, 335: 96-104.

[7] Tatari F, Naghibi-Sistani M B, Vamvoudakis K G. Distributed learning algorithm for non-linear differential graphical games[J]. Transactions of the Institute of Measurement and Control, 2017, 39(2): 173-182.

[8] Mazouchi M, Naghibi-Sistani M B, Sani S K H. A novel distributed optimal adaptive control algorithm for nonlinear multi-agent differential graphical games[J]. IEEE/CAA Journal of Automatica Sinica, 2018, 5(1): 331-341.

[9] Zhang H G, Cui X H, Luo Y H, et al. Finite-horizon H_∞ tracking control for unknown nonlinear systems with saturating actuators[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4):1200-1212.

[10] Modares H, Lewis F L, Jiang Z P. H_∞ tracking control of completely unknown continuous-time systems via off policy reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2550-2562.

[11] Wen G X, Chen C L P, Ge S S, et al. Optimized adaptive nonlinear tracking control using actor-critic reinforcement learning strategy[J]. IEEE Transactions on Industrial Informatics, 2019, 15(9): 4969-4977.

[12] Jiao Q, Modares H, Xu S Y, et al. Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control[J]. Automatica, 2016, 69: 24-34.

[13] Sun J L, Liu C S. Distributed zero-sum differential game for multi-agent nonlinear systems via adaptive dynamic programming[C]. The 37th Chinese Control Conference. Wuhan: IEEE, 2018: 2770-2775.

[14] Yu D, Wu Q H, Song L. Finite time estimation and containment control of second order perturbed directed networks[C]. The 50th IEEE Conference on Decision and Control and European Control Conference. Orland: IEEE, 2011: 4126-4131.

[15] Wu H N, Luo B. Neural network based online simultaneous policy update algorithm for solving the HJI align in nonlinear H_∞ control[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(12): 1884-1895.

[16] Yang X, Liu D R, 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.

作者简介

于镝 (1977−), 女, 副教授, 博士, 从事多智能体协调控制、自适应动态规划等研究, E-mail: yudizlg@aliyun.com.

(责任编辑：李君玲)