航空科学技术
Aeronautical Science & Technology
Apr. 25 2024 Vol. 35 No.04 25-30

多无人机系统在线强化学习最优安全跟踪控制

弓镇宇，杨飞生
西北工业大学，陕西 西安 710072

摘 要：在无人机(UAV)编队跟踪任务中，虚假数据注入(FDI)攻击者可向控制指令注入误导性数据，导致无人机无法形成指定的编队构型，故需设计安全编队跟踪控制器。为此，本文利用零和图博弈对攻防过程进行建模，其中FDI攻击者和安全控制器是博弈的参与者，攻击者的目标是最大化设定的成本函数，而安全控制器的目标与之相反，求解博弈并获得最优安全控制策略依赖于求取Hamilton-Jacobi-Isaacs(HJI)方程的解。而HJI方程是耦合偏微分方程，难以直接求解，因此结合经验回放机制引入了有限时间收敛的在线强化学习算法，设计了单评价神经网络近似值函数并获得了最优安全控制策略。最终利用仿真验证了算法的有效性。

关键词：FDI攻击；多无人机；在线强化学习；优化控制；零和图博弈

中图分类号：V249.1 **文献标识码：**A DOI：10.19452/j.issn1007-5453.2024.04.004

无人机作为典型的无人系统，已经广泛应用于农林作业、电力巡检、灾后搜救、目标侦察、协同作战等领域[1]。相比于单体无人机的机载设备有限、感知范围小、任务执行容错能力差，整合了信息融合、目标分配、协同控制等技术的多无人机系统可在复杂周边环境下完成多样化任务。然而，无人机之间的信息交互依赖于通信网络，因此多无人机系统面临网络攻击的威胁。攻击者可向无人机传输信道中注入欺骗型数据以降低系统性能，甚至导致任务失败，因此设计安全控制方案以抵御虚假数据注入(FDI)攻击至关重要。

目前，主要有两种安全控制方案以应对FDI攻击，其区别在于是否引入了攻击检测机制[2-3]。在无人机系统中，Lin Hong等[4]设计了攻击检测器和线性二次高斯控制器来抵抗FDI攻击。Xiao Jiaping等[5]基于滑动新息序列提出了一种新型的攻击检测器。为了节约网络通信资源，Yin Tingting等[6]研究了事件触发机制下的无人机安全编队控制。

随着人工智能技术的发展，强化学习算法因其良好的决策优化和实时的策略选择能力而备受关注，越来越多研究者将其应用于控制问题求解[7-10]。针对安全控制问题，Wu Chengwei等[11]利用Q学习算法研究了控制信号遭受FDI攻击时的安全控制器设计。Zhou Yuanqiang等[12]结合威胁检测水平函数设计了检测机制并利用离策略算法求解了最优安全控制器。多智能体系统中，Moghadam等[13]结合离策略算法求解了非齐次博弈Riccati方程并设计了弹性控制器。考虑事件触发机制，Xu Yuanyuan等[14]结合预选器和观测器研究了传感器遭受FDI攻击时的最优控制律设计。

已有文献很少关注强化学习的快速收敛问题，Kokolakis等[15]设计了有限时间收敛的强化学习算法，但是其未考虑多无人机情形。本文在领导—跟随多智能体框架下研究了多无人机遭受网络攻击时的安全编队跟踪控制问题。考虑到FDI攻击者和安全控制器之间相互对抗的关系，引入了零和图博弈理论对攻防过程进行建模，最优攻击和最优安全控制器位于纳什均衡点。为求解博弈并进一步获取最优安全控制律，引入了有限时间收敛强化学习方法。通过单评价神经网络架构对值函数进行逼近，采用了经验回放机制以维持持续激励条件，分析了算法的收敛性以及

收稿日期：2023-06-27；退修日期：2023-11-16；录用日期：2023-12-25
基金项目：国家自然科学基金(62073269)；航空科学基金(2020Z034053002)；陕西省重点研发计划项目(2022GY-244)；重庆市自然科学基金(CSTB2022NSCQ-MSX0963)；广东省基础与应用基础研究基金(2023A1515011220)
引用格式：Gong Zhenyu, Yang Feisheng. Optimal secure tracking control in multi-UAVs based on online reinforcement learning [J]. Aeronautical Science & Technology, 2024, 35(04): 25-30. 弓镇宇，杨飞生. 多无人机系统在线强化学习最优安全跟踪控制[J]. 航空科学技术, 2024, 35(04): 25-30.

停息时间上界。

1 代数图论

在通过网络进行信息交互的多无人机系统中，每架无人机都可看作通信拓扑的节点。通信拓扑图可表示为$\mathcal{G}=(\mathcal{V},\mathcal{E},\boldsymbol{A})$，其中$\mathcal{V}=\{v_1,v_2,\cdots,v_N\}$是图的节点集，$\mathcal{E}=\{(v_i,v_j)|v_i,v_j\in\mathcal{V}\}$表示图的边集。$\boldsymbol{A}=\{a_{ij}\}$为邻接矩阵，如果$\{v_i,v_j\}\in\mathcal{E}$，则$a_{ij}>0$，否则$a_{ij}=0$。无人机$i$的邻居集为$\mathcal{N}_i=\{j|(v_i,v_j)\in\mathcal{E}\}$。定义图的入度矩阵为$\boldsymbol{D}=\mathrm{diag}\{d_i\}$，其中$d_i=\sum_{j\in\mathcal{N}_i}a_{ij}$。图的拉普拉斯矩阵可表示为$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$。考虑存在领导者的图$\overline{\mathcal{G}}$，则牵引矩阵为$\boldsymbol{G}=\mathrm{diag}\{g_i\}$。当第$i$架无人机可直接接收领导者信息时，$g_i=1$，否则$g_i=0$。若图存在一个根节点，可通过有向路径到达图中其他任意节点，则该图存在一个有向生成树。

2 问题描述

考虑包含N架无人机的多无人机系统。第i架无人机的动力学方程如下所示

$$\begin{aligned}\dot{\boldsymbol{p}}_i(t)&=\boldsymbol{v}_i(t)\\ \dot{\boldsymbol{v}}_i(t)&=-\boldsymbol{c}g_{\mathrm{e}}+\bar{\boldsymbol{u}}_i(t)\end{aligned}\tag{1}$$

式中，$\boldsymbol{p}_i(t)\in\mathbb{R}^3$、$\boldsymbol{v}_i(t)\in\mathbb{R}^3$、$\bar{\boldsymbol{u}}_i(t)\in\mathbb{R}^3$分别代表第$i$架无人机的位置矢量、速度矢量、控制输入矢量，$\boldsymbol{c}=[0\ 0\ 1]^{\mathrm{T}}$，$g_{\mathrm{e}}$是重力加速度。令系统的状态矢量为$\boldsymbol{x}_i(t)=[\boldsymbol{p}_i^{\mathrm{T}}(t)\ \ \boldsymbol{v}_i^{\mathrm{T}}(t)]^{\mathrm{T}}$，则有

$$\dot{\boldsymbol{x}}_i(t)=\boldsymbol{A}\boldsymbol{x}_i(t)+\boldsymbol{B}\bar{\boldsymbol{u}}_i(t)+\boldsymbol{C}g_{\mathrm{e}}\tag{2}$$

式中

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{0}_{3\times3}&\boldsymbol{I}_3\\ \boldsymbol{0}_{3\times3}&\boldsymbol{0}_{3\times3}\end{bmatrix},\ \boldsymbol{B}=\begin{bmatrix}\boldsymbol{0}_{3\times3}\\ \boldsymbol{I}_3\end{bmatrix},\ \boldsymbol{C}=[\boldsymbol{0}_{1\times5}\quad -1\ ]^{\mathrm{T}}$$

考虑攻击者会向无人机的控制信号中注入虚假数据，因此有

$$\bar{\boldsymbol{u}}_i(t)=\check{\boldsymbol{u}}_i(t)+\boldsymbol{w}_i(t)\tag{3}$$

式中，矢量$\check{\boldsymbol{u}}_i(t)$是安全编队跟踪控制器，矢量$\boldsymbol{w}_i(t)$是攻击信号。给定控制器为

$$\check{\boldsymbol{u}}_i(t)=\boldsymbol{u}_i(t)+\boldsymbol{c}g_{\mathrm{e}}\tag{4}$$

式中，矢量$\boldsymbol{u}_i(t)$是待设计控制器。将式(3)和式(4)代入式(2)可得

$$\dot{\boldsymbol{x}}_i(t)=\boldsymbol{A}\boldsymbol{x}_i(t)+\boldsymbol{B}\boldsymbol{u}_i(t)+\boldsymbol{B}\boldsymbol{w}_i(t)\tag{5}$$

考虑虚拟领导者可提供期望位置矢量$\boldsymbol{p}_0(t)$与期望速度矢量$\boldsymbol{v}_0(t)$，则虚拟领导者的状态矢量为$\boldsymbol{x}_0(t)=[\boldsymbol{p}_0^{\mathrm{T}}(t)\ \ \boldsymbol{v}_0^{\mathrm{T}}(t)]^{\mathrm{T}}$。令第$i$架无人机的位置跟踪误差矢量和速度跟踪误差矢量为

$$\begin{aligned}\boldsymbol{\delta}_{\mathrm{p}i}(t)&=\boldsymbol{p}_i(t)-\boldsymbol{p}_0(t)-\boldsymbol{\chi}_i\\ \boldsymbol{\delta}_{\mathrm{v}i}(t)&=\boldsymbol{v}_i(t)-\boldsymbol{v}_0(t)\end{aligned}\tag{6}$$

式中，$\boldsymbol{\chi}_i\in\mathbb{R}^3$是第$i$架无人机与领导者之间的相对位置矢量。进而引入以下假设。

假设1：无人机系统的通信拓扑图存在有向生成树，且树的根节点为领导者。

每架无人机的目标是跟踪上期望轨迹与速度，即满足$\lim\limits_{t\to\infty}\|\boldsymbol{\delta}_{\mathrm{p}i}(t)\|_2=0$，$\lim\limits_{t\to\infty}\|\boldsymbol{\delta}_{\mathrm{v}i}(t)\|_2=0$，$i=1,2,\cdots,N$。

定义系统的编队误差矢量为

$$\begin{aligned}\boldsymbol{e}_{\mathrm{p}i}(t)&=\sum_{j\in\mathcal{N}_i}a_{ij}\left(\boldsymbol{\delta}_{\mathrm{p}i}(t)-\boldsymbol{\delta}_{\mathrm{p}j}(t)\right)+g_i\boldsymbol{\delta}_{\mathrm{p}i}(t)\\ \boldsymbol{e}_{\mathrm{v}i}(t)&=\sum_{j\in\mathcal{N}_i}a_{ij}\left(\boldsymbol{\delta}_{\mathrm{v}i}(t)-\boldsymbol{\delta}_{\mathrm{v}j}(t)\right)+g_i\boldsymbol{\delta}_{\mathrm{v}i}(t)\end{aligned}\tag{7}$$

令

$$\begin{aligned}\boldsymbol{e}_i(t)&=[\boldsymbol{e}_{\mathrm{p}i}^{\mathrm{T}}(t)\ \ \boldsymbol{e}_{\mathrm{v}i}^{\mathrm{T}}(t)]^{\mathrm{T}}\\ \boldsymbol{\delta}_i(t)&=[\boldsymbol{\delta}_{\mathrm{p}i}^{\mathrm{T}}(t)\ \ \boldsymbol{\delta}_{\mathrm{v}i}^{\mathrm{T}}(t)]^{\mathrm{T}}\end{aligned}$$

则有

$$\boldsymbol{e}_i(t)=\sum_{j\in\mathcal{N}_i}a_{ij}\left(\boldsymbol{\delta}_i(t)-\boldsymbol{\delta}_j(t)\right)+g_i\boldsymbol{\delta}_i(t)\tag{8}$$

对$\boldsymbol{e}_i(t)$求导可得

$$\begin{aligned}\dot{\boldsymbol{e}}_i(t)=\sum_{j\in\mathcal{N}_i}a_{ij}\,&(\boldsymbol{A}\boldsymbol{x}_i(t)+\boldsymbol{B}\boldsymbol{u}_i(t)+\boldsymbol{B}\boldsymbol{w}_i(t)-\boldsymbol{A}\boldsymbol{x}_j(t)-\\ &\boldsymbol{B}\boldsymbol{u}_j(t)-\boldsymbol{B}\boldsymbol{w}_j(t))+g_i\left(\boldsymbol{A}\boldsymbol{x}_i(t)+\boldsymbol{B}\boldsymbol{u}_i(t)+\right.\\ &\left.\boldsymbol{B}\boldsymbol{w}_i(t)-\dot{\boldsymbol{x}}_0(t)\right)\end{aligned}$$

接下来引入博弈思想分析攻防过程。

3 安全零和图博弈

针对第i架无人机设计如下成本函数。

$$\begin{aligned}&J_i\left(\boldsymbol{e}_i(0),\boldsymbol{u}_i(t),\boldsymbol{u}_{-i}(t),\boldsymbol{w}_i(t),\boldsymbol{w}_{-i}(t)\right)=\\ &\int_0^\infty U_i\left(\boldsymbol{e}_i(t),\boldsymbol{u}_i(t),\boldsymbol{u}_{-i}(t),\boldsymbol{w}_i(t),\boldsymbol{w}_{-i}(t)\right)\mathrm{d}t=\\ &\int_0^\infty(\boldsymbol{e}_i^{\mathrm{T}}(t)\boldsymbol{Q}_{ii}\boldsymbol{e}_i(t)+\boldsymbol{u}_i^{\mathrm{T}}(t)\boldsymbol{R}_{ii}\boldsymbol{u}_i(t)+\\ &\sum_{j\in\mathcal{N}_i}\boldsymbol{u}_j^{\mathrm{T}}(t)\boldsymbol{R}_{ij}\boldsymbol{u}_j(t)-\gamma^2\boldsymbol{w}_i^{\mathrm{T}}(t)\boldsymbol{T}_{ii}\boldsymbol{w}_i(t)-\\ &\gamma^2\sum_{j\in\mathcal{N}_i}\boldsymbol{w}_j^{\mathrm{T}}(t)\boldsymbol{T}_{ij}\boldsymbol{w}_j(t))\mathrm{d}t\end{aligned}\tag{9}$$

式中，$\boldsymbol{Q}_{ii}>0$，$\boldsymbol{R}_{ii}>0$，$\boldsymbol{R}_{ij}\geqslant0$，$\boldsymbol{T}_{ii}>0$，$\boldsymbol{T}_{ij}\geqslant0$是对称权重矩阵，$\gamma>0$为攻击抑制水平，下标$-i$代表第$i$架无人机的邻居为简单起见，在后文中省去时间符号“$t$”。

从攻击者的角度而言，其目标是使得系统性能下降并影响跟踪效果，而安全控制器的目标是抵御这种负面作用。可以看出，FDI攻击者与安全控制器形成一种相互对抗的关系，可引入零和图博弈来描述该攻防过程。

攻击者和安全控制器可看作博弈参与者，攻击者的目的是最大化式(9)，而控制器的目的则与之相反。此过程可表示为

$$V_i(\boldsymbol{e}_i(0)) = \min_{\boldsymbol{u}_i} \max_{\boldsymbol{w}_i} J_i(\boldsymbol{e}_i,\boldsymbol{u}_i,\boldsymbol{u}_{-i},\boldsymbol{w}_i,\boldsymbol{w}_{-i})$$

式中，$V_i(\boldsymbol{e}_i(0))$为博弈的值。如果存在鞍点$(\boldsymbol{u}_i^*,\boldsymbol{w}_i^*)$，则有

$$V_i(\boldsymbol{e}_i(0)) = \min_{\boldsymbol{u}_i} \max_{\boldsymbol{w}_i} J_i(\boldsymbol{e}_i,\boldsymbol{u}_i,\boldsymbol{u}_{-i}^*,\boldsymbol{w}_i,\boldsymbol{w}_{-i}^*) = \max_{\boldsymbol{w}_i} \min_{\boldsymbol{u}_i} J_i(\boldsymbol{e}_i,\boldsymbol{u}_i,\boldsymbol{u}_{-i}^*,\boldsymbol{w}_i,\boldsymbol{w}_{-i}^*)$$

其等价于纳什均衡条件

$$J_i(\boldsymbol{u}_i^*,\boldsymbol{u}_{-i}^*,\boldsymbol{w}_i,\boldsymbol{w}_{-i}^*) \leqslant J_i(\boldsymbol{u}_i^*,\boldsymbol{u}_{-i}^*,\boldsymbol{w}_i^*,\boldsymbol{w}_{-i}^*) \leqslant J_i(\boldsymbol{u}_i,\boldsymbol{u}_{-i}^*,\boldsymbol{w}_i^*,\boldsymbol{w}_{-i}^*) \tag{10}$$

为求解博弈，对于值函数$V_i(\boldsymbol{e}_i)$，定义以下哈密顿函数

$$H_i(\boldsymbol{e}_i,\boldsymbol{u}_i,\boldsymbol{u}_{-i},\boldsymbol{w}_i,\boldsymbol{w}_{-i},\nabla V_i(\boldsymbol{e}_i)) = J_i(\boldsymbol{e}_i,\boldsymbol{u}_i,\boldsymbol{u}_{-i},\boldsymbol{w}_i,\boldsymbol{w}_{-i}) + \nabla V_i^{\mathrm{T}}(\boldsymbol{e}_i)\dot{\boldsymbol{e}}_i \tag{11}$$

式中，∇为微分算子。根据一阶平稳性条件可得

$$\boldsymbol{u}_i^* = -\frac{1}{2}(d_i + g_i)\boldsymbol{R}_{ii}^{-1}\boldsymbol{B}^{\mathrm{T}}\nabla V_i(\boldsymbol{e}_i) \tag{12}$$

$$\boldsymbol{w}_i^* = \frac{1}{2\gamma^2}(d_i + g_i)\boldsymbol{T}_{ii}^{-1}\boldsymbol{B}^{\mathrm{T}}\nabla V_i(\boldsymbol{e}_i) \tag{13}$$

将式(12)和式(13)代入式(11)中，可得如下Hamilton-Jacobi-Isaacs(HJI)方程。

$$H_i(\boldsymbol{e}_i,\boldsymbol{u}_i^*,\boldsymbol{u}_{-i}^*,\boldsymbol{w}_i^*,\boldsymbol{w}_{-i}^*,\nabla V_i(\boldsymbol{e}_i)) = 0 \tag{14}$$

由于HJI方程是耦合的，因此引入在线强化学习方法对其进行求解。

4 有限时间收敛在线强化学习

根据Weierstrass高阶逼近定理，给定以下评价神经网络对值函数进行逼近

$$V_i(\boldsymbol{e}_i) = \boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\phi}_i(\boldsymbol{e}_i) + \varepsilon_i(\boldsymbol{e}_i),\ \boldsymbol{e}_i \in \mathbb{E} \tag{15}$$

式中，$\boldsymbol{W}_i$，$\boldsymbol{\phi}_i(\cdot)$，$\varepsilon_i(\cdot)$和$\mathbb{E}$分别为神经网络的权重矩阵、激活函数、逼近误差和编队误差集。由于$\boldsymbol{W}_i$是理想权重，难以直接获得，所以神经网络的输出可近似为

$$\hat{V}_i(\boldsymbol{e}_i) = \hat{\boldsymbol{W}}_i^{\mathrm{T}}\boldsymbol{\phi}_i(\boldsymbol{e}_i),\ \boldsymbol{e}_i \in \mathbb{E} \tag{16}$$

式中，$\hat{\boldsymbol{W}}_i$为估计权重矩阵。因此，最优策略可近似为

$$\hat{\boldsymbol{u}}_i = -\frac{1}{2}(d_i + g_i)\boldsymbol{R}_{ii}^{-1}\boldsymbol{B}^{\mathrm{T}}\nabla\boldsymbol{\phi}_i^{\mathrm{T}}(\boldsymbol{e}_i)\hat{\boldsymbol{W}}_i \tag{17}$$

$$\hat{\boldsymbol{w}}_i = \frac{1}{2\gamma^2}(d_i + g_i)\boldsymbol{T}_{ii}^{-1}\boldsymbol{B}^{\mathrm{T}}\nabla\boldsymbol{\phi}_i^{\mathrm{T}}(\boldsymbol{e}_i)\hat{\boldsymbol{W}}_i \tag{18}$$

定义如下残差函数

$$\xi_i(t) = U_i\big(\boldsymbol{e}_i(t),\hat{\boldsymbol{u}}_i(t),\hat{\boldsymbol{u}}_{-i}(t),\hat{\boldsymbol{w}}_i(t),\hat{\boldsymbol{w}}_{-i}(t)\big) + \hat{\boldsymbol{W}}_i^{\mathrm{T}}(t)\nabla\boldsymbol{\phi}_i(\boldsymbol{e}_i(t))\dot{\boldsymbol{e}}_i(t) \tag{19}$$

在训练神经网络权重的过程中，为了维持持续激励条件，通常需引入探测噪声，但这种做法会影响系统的稳定性。而经验回放机制弥补了该缺陷，利用了历史数据，因此定义损失函数为

$$E_i(\hat{\boldsymbol{W}}_i) = \frac{1}{q+1}\left|\frac{\xi_i(t)}{1+\boldsymbol{\psi}_i^{\mathrm{T}}(t)\boldsymbol{\psi}_i(t)}\right|^{q+1} + \frac{1}{q+1}\sum_{s=t_0}^{t_{\mathrm{f}}}\left|\frac{\xi_i^{[t]}(s)}{1+\boldsymbol{\psi}_i^{\mathrm{T}}(s)\boldsymbol{\psi}_i(s)}\right|^{q+1} \tag{20}$$

式中，$0<q<1$，$\boldsymbol{\psi}_i(t) = \nabla\boldsymbol{\phi}_i(\boldsymbol{e}_i(t))\dot{\boldsymbol{e}}_i(t)$，$|\cdot|$是绝对值函数，且时间序列$0 \leqslant t_0,\cdots,t_{\mathrm{f}} < t$上的历史残差数据为

$$\xi_i^{[t]}(s) = U_i\big(\boldsymbol{e}_i(s),\hat{\boldsymbol{u}}_i(s),\hat{\boldsymbol{u}}_{-i}(s),\hat{\boldsymbol{w}}_i(s),\hat{\boldsymbol{w}}_{-i}(s)\big) + \hat{\boldsymbol{W}}_i^{\mathrm{T}}(t)\boldsymbol{\phi}_i(\boldsymbol{e}_i(s))\dot{\boldsymbol{e}}_i(s)$$

结合梯度下降原理，可得出神经网络权重的更新律为

$$\dot{\hat{\boldsymbol{W}}}_i = -\alpha_i\bar{\boldsymbol{\psi}}_i(t)\left[\!\!\left[\frac{\xi_i(t)}{1+\boldsymbol{\psi}_i^{\mathrm{T}}(t)\boldsymbol{\psi}_i(t)}\right]\!\!\right]^q - \alpha_i\sum_{s=t_0}^{t_{\mathrm{f}}}\bar{\boldsymbol{\psi}}_i(s)\left[\!\!\left[\frac{\xi_i^{[t]}(s)}{1+\boldsymbol{\psi}_i^{\mathrm{T}}(s)\boldsymbol{\psi}_i(s)}\right]\!\!\right]^q \tag{21}$$

式中，$\bar{\boldsymbol{\psi}}_i(t) = \boldsymbol{\psi}_i(t)/(1+\boldsymbol{\psi}_i^{\mathrm{T}}(t)\boldsymbol{\psi}_i(t))$，$\alpha_i > 0$是神经网络的学习率，$[\![\cdot]\!] = \mathrm{sgn}(\cdot)[\cdot]$，$\mathrm{sgn}(\cdot)$是符号函数。进一步有以下假设。

假设2：矩阵$\boldsymbol{\Psi}_i = [\boldsymbol{\psi}_i(t_0),\cdots,\boldsymbol{\psi}_i(t_{\mathrm{f}})]$由$k$组历史数据组成且是行满秩的。

定义权重误差为$\tilde{\boldsymbol{W}}_i = \hat{\boldsymbol{W}}_i - \boldsymbol{W}_i$，则权重误差的动力学方程为

$$\dot{\tilde{\boldsymbol{W}}}_i = -\alpha_i\bar{\boldsymbol{\psi}}_i(t)\left[\!\!\left[\frac{\boldsymbol{\psi}_i^{\mathrm{T}}(t)\tilde{\boldsymbol{W}}_i(t)+\epsilon_i(t)}{1+\boldsymbol{\psi}_i^{\mathrm{T}}(t)\boldsymbol{\psi}_i(t)}\right]\!\!\right]^q - \alpha_i\sum_{s=t_0}^{t_{\mathrm{f}}}\bar{\boldsymbol{\psi}}_i(s)\left[\!\!\left[\frac{\boldsymbol{\psi}_i^{\mathrm{T}}(s)\tilde{\boldsymbol{W}}_i(t)+\epsilon_i(s)}{1+\boldsymbol{\psi}_i^{\mathrm{T}}(s)\boldsymbol{\psi}_i(s)}\right]\!\!\right]^q \tag{22}$$

其中

$$\epsilon_i(t) = U_i\big(\boldsymbol{e}_i(t),\hat{\boldsymbol{u}}_i(t),\hat{\boldsymbol{u}}_{-i}(t),\hat{\boldsymbol{w}}_i(t),\hat{\boldsymbol{w}}_{-i}(t)\big) + \boldsymbol{W}_i^{\mathrm{T}}(t)\nabla\boldsymbol{\phi}_i(\boldsymbol{e}_i(t))\dot{\boldsymbol{e}}_i(t)$$

$\epsilon_i(t)$的有界性与神经网络隐层的神经元个数相关，进而与强化学习算法收敛性相关，同时有以下引理。

引理1[15]：对于任意的$\hat{\epsilon}$，存在$F(\hat{\epsilon})>0$和$M_0(\hat{\epsilon})>0$，当$M \to \infty$时满足$\sup|\epsilon_i| < F(\hat{\epsilon})$，$M_0(\hat{\epsilon}) \leqslant M$，$\epsilon_i \equiv 0$。

接下来分析算法的有限时间收敛性，我们有如下定理。

定理1：在满足假设2的前提下，令$\bar{\boldsymbol{\Psi}}_i = [\bar{\boldsymbol{\psi}}_i(t_0),\cdots,\bar{\boldsymbol{\psi}}_i(t_{\mathrm{f}})]$，给定$0<\theta_i<1$，$|\bar{\varepsilon}_i| < \bar{\varepsilon}_{\mathrm{m}}$，$\|\bar{\boldsymbol{\psi}}_i\|_2 \leqslant \bar{\boldsymbol{\psi}}_{\mathrm{m}}$。

(1)当$\epsilon_i \equiv 0$时，式(22)的零解$\tilde{\boldsymbol{W}}_i = 0$是全局有限时间稳定的，且停息时间为

$$T \leqslant \frac{\|\tilde{\boldsymbol{W}}_i(0)\|_2^{1-q}}{\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\alpha_i(1-q)}$$

(2)当$\epsilon_i\neq0$时,式(22)是最终一致有界的,且收敛界为

$$\mu_i=\left(\frac{(k+1)\bar{\epsilon}_{\mathrm{m}}^q\bar{\psi}_{\mathrm{m}}}{\theta_i\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)}\right)^{\frac{1}{q}}$$

停息时间满足

$$T\leqslant\frac{\left\|\tilde{\boldsymbol{W}}_i(0)\right\|_2^{1-q}-\mu_i^{1-q}}{\alpha_i\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)(1-\theta_i)(1-q)}$$

证明:给定Lyapunov函数为

$$\tilde{V}_i=\frac{1}{2\alpha_i}\tilde{\boldsymbol{W}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i \tag{23}$$

对其进行求导可得

$$\dot{\tilde{V}}_i=-\tilde{\boldsymbol{W}}_i^{\mathrm{T}}(t)\bar{\boldsymbol{\psi}}_i(t)\left[\!\left[\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(t)\tilde{\boldsymbol{W}}_i(t)+\bar{\epsilon}_i(t)\right]\!\right]^q-\sum_{s=t_0}^{t_f}\tilde{\boldsymbol{W}}_i^{\mathrm{T}}(t)\bar{\boldsymbol{\psi}}_i(s)\left[\!\left[\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(s)\tilde{\boldsymbol{W}}_i(t)+\bar{\epsilon}_i(s)\right]\!\right]^q \tag{24}$$

式中,$\bar{\epsilon}_i(t)=\epsilon_i(t)/(1+\boldsymbol{\psi}_i^{\mathrm{T}}(t)\boldsymbol{\psi}_i(t))$。

(1)当$\epsilon_i\equiv0$时,可得

$$\dot{\tilde{V}}_i=-\tilde{\boldsymbol{W}}_i^{\mathrm{T}}(t)\bar{\boldsymbol{\psi}}_i(t)\left[\!\left[\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(t)\tilde{\boldsymbol{W}}_i(t)\right]\!\right]^q-\sum_{s=t_0}^{t_f}\tilde{\boldsymbol{W}}_i^{\mathrm{T}}(t)\bar{\boldsymbol{\psi}}_i(s)\left[\!\left[\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(s)\tilde{\boldsymbol{W}}_i(t)\right]\!\right]^q\leqslant-\sum_{s=t_0}^{t_f}\left|\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(s)\tilde{\boldsymbol{W}}_i(t)\right|^{q+1} \tag{25}$$

由于

$$-\sum_{s=t_0}^{t_f}\left|\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(s)\tilde{\boldsymbol{W}}_i(t)\right|^{q+1}=-\left\|\bar{\boldsymbol{\Psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i\right\|_{q+1}^{q+1}$$

根据范数的单调性可得

$$\left\|\bar{\boldsymbol{\Psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i\right\|_2\leqslant\left\|\bar{\boldsymbol{\Psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i\right\|_{q+1}$$

所以有

$$\dot{\tilde{V}}_i\leqslant-\left\|\bar{\boldsymbol{\Psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i\right\|_2^{q+1}\leqslant-\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\left\|\tilde{\boldsymbol{W}}_i\right\|_2^{q+1} \tag{26}$$

式中,$\sigma_{\min}(\cdot)$代表最小奇异值。进而有

$$\dot{\tilde{V}}_i\leqslant-\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)(2\alpha_i)^{\frac{q+1}{2}}\tilde{V}_i^{\frac{q+1}{2}} \tag{27}$$

相应地,停息时间为

$$T\leqslant\frac{\left\|\tilde{\boldsymbol{W}}_i(0)\right\|_2^{1-q}}{\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\alpha_i(1-q)} \tag{28}$$

(2)当$\epsilon_i\neq0$时,根据引理1可得

$$\mathrm{sgn}(\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i+\bar{\epsilon}_i)=\mathrm{sgn}(\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i) \tag{29}$$

继而可知

$$|\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i|^q-|\bar{\epsilon}_i|^q\leqslant|\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}\tilde{\boldsymbol{W}}_i+\bar{\epsilon}_i|^q \tag{30}$$

根据式(29)和式(30)可得

$$\dot{\tilde{V}}_i\leqslant-\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\left\|\tilde{\boldsymbol{W}}_i\right\|_2^{q+1}+\sum_{s=t_0}^{t_f}\left|\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(s)\tilde{\boldsymbol{W}}_i(t)\right|\left|\bar{\epsilon}_i(s)\right|^q+\left|\bar{\boldsymbol{\psi}}_i^{\mathrm{T}}(t)\tilde{\boldsymbol{W}}_i(t)\right|\left|\bar{\epsilon}_i(t)\right|^q \tag{31}$$

进而有

$$\dot{\tilde{V}}_i\leqslant-(1-\theta_i)\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\left\|\tilde{\boldsymbol{W}}_i\right\|_2^{q+1}-\theta_i\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\left\|\tilde{\boldsymbol{W}}_i\right\|_2^{q+1}+(k+1)\bar{\epsilon}_{\mathrm{m}}^q\bar{\psi}_{\mathrm{m}}\left\|\tilde{\boldsymbol{W}}_i\right\|_2$$

因此有

$$\dot{\tilde{V}}_i\leqslant-(1-\theta_i)\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)\left\|\tilde{\boldsymbol{W}}_i\right\|_2^{q+1},\ \left\|\tilde{\boldsymbol{W}}_i\right\|_2\geqslant\mu_i \tag{32}$$

根据比较引理可得出停息时间为

$$T\leqslant\frac{\left\|\tilde{\boldsymbol{W}}_i(0)\right\|_2^{1-q}-\mu_i^{1-q}}{\alpha_i\sigma_{\min}^{q+1}(\bar{\boldsymbol{\Psi}}_i)(1-\theta_i)(1-q)} \tag{33}$$

此定理证毕。

5 仿真结果

假设系统的通信拓扑图如图1所示。图1中,1、2、3、4为无人机,0为虚拟领导者,对应的拉普拉斯矩阵和牵引矩阵分别为

$$\boldsymbol{L}=\begin{bmatrix}1&-1&0&0\\-1&1&0&0\\0&0&1&-1\\0&0&-1&1\end{bmatrix}\quad\boldsymbol{G}=\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}$$

给定攻击抑制水平为8,参数q=0.1,所有神经网络的学习率为0.1,激活函数为多项式函数。虚拟领导者的期望轨迹和期望速度(矢量)为$\boldsymbol{p}_0=[15\sin(0.1t),10\cos(0.1t),10t]^{\mathrm{T}}$,$\boldsymbol{v}_0=[1.5\cos(0.1t),-\sin(0.1t),10]^{\mathrm{T}}$。仿真结果如图2~图7所示。

由图2~图4可得,所有无人机均能到达预定的编队位

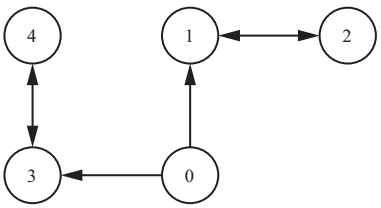

图1 通信拓扑图

Fig.1 Communication topology

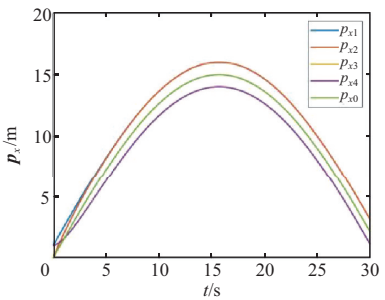

图2 无人机位置在x方向的分量

Fig.2 Components of UAVs' positions in x direction

置并跟踪上期望轨迹。由图5~图7可得，所有无人机都能达到期望速度。因此，文中给出的最优安全控制方案可抵御FDI攻击并实现编队跟踪。

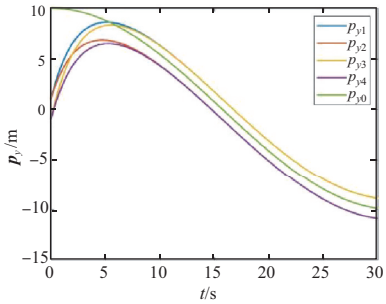

图3 无人机位置在y方向的分量

Fig.3 Components of UAVs' positions in y direction

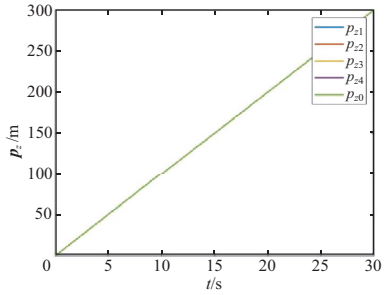

图4 无人机位置在z方向的分量

Fig.4 Components of UAVs' positions in z direction

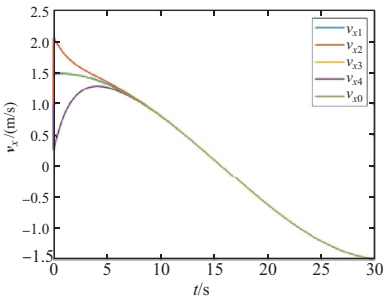

图5 无人机速度在x方向的分量

Fig.5 Components of UAVs' velocities in x direction

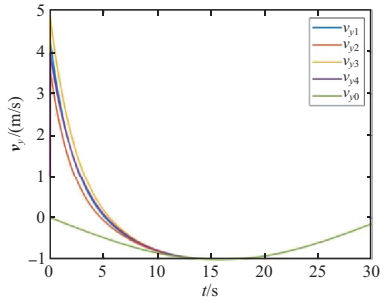

图6 无人机速度在y方向的分量

Fig.6 Components of UAVs' velocities in y direction

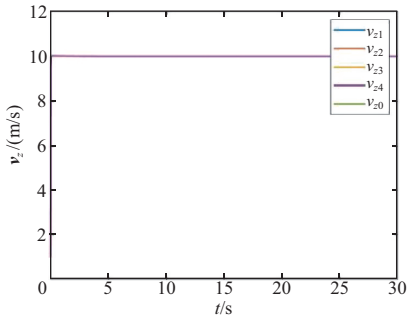

图7 无人机速度在z方向的分量

Fig.7 Components of UAVs' velocities in z direction

6 结束语

本文研究了无人机控制信号遭受FDI攻击时的最优安全跟踪控制问题，利用零和图博弈理论对攻击者与安全控制器之间的攻防进行了建模，引入了有限时间收敛的强化学习方法和单评价神经网络架构，在线求解了最优安全控制器。在后续工作中，可基于无模型在线强化学习方法实现最优安全编队跟踪控制器设计。 AST

参考文献

[1] Wang Haijun, Zhao Haitao, Zhang Jiao, et al. Survey on unmanned aerial vehicle networks: A cyber physical system perspective [J]. IEEE Communications Surveys & Tutorial, 2020, 22(2): 1027-1070.

[2] Li Xiaomeng, Zhou Qi, Li Panshuo, et al. Event-triggered consensus control for multi-agent systems against false data-injection attacks [J]. IEEE Transactions on Cybernetics, 2020, 50(5): 1856-1866.

[3] Tan Yushun, Liu Qingyi, Liu Jinliang, et al. Observer-based security control for interconnected semi-Markovian jump systems with unknown transition probabilities [J]. IEEE Transactions on Cybernetics, 2022, 52(9): 9013-9025.

[4] Lin Hong, Sun Pei, Cai Chenxiao, et al. Secure LQG control for a quadrotor under false data injection attacks [J]. IET Control Theory & Applications, 2022, 16(9): 925-934.

[5] Xiao Jiaping, Feroskhan M. Cyber attack detection and isolation for a quadrotor UAV with modified sliding innovation sequences [J]. IEEE Transactions on Vehicular Technology, 2022, 71(7): 7202-7214.

[6] Yin Tingting, Gu Zhou, Park J H, et al. Event-based intermittent formation control of multi-UAV systems under

deception attacks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 12: 1-12.

[7] Lewis F L, Vrabie J, Vamvoudakis K G, et al. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers[J]. IEEE Control Systems Magazine, 2012, 32(6): 76-105.

[8] Peng Zhinan, Luo Rui, Hu Jiangping, et al. Optimal tracking control of nonlinear multiagent systems using internal reinforce Q-learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(8): 4043-4055.

[9] Xie Kedi, Yu Xiao, Lan Weiyao. Optimal output regulation for unknown continuous-time linear systems by internal model and adaptive dynamic programming [J]. Automatica, 2022, 146: 1-7.

[10] Wei Qinglai, Zhu Liao, Li Tao, et al. A new approach to finite-horizon optimal control for discrete-time affine nonlinear systems via a pseudolinear method [J]. IEEE Transactions on Automatic Control, 2022, 67(5): 2610-2617.

[11] Wu Chengwei, Li Xiaolei, Pan Wei, et al. Zero-sum game based optimal secure control under actuator attacks[J]. IEEE Transactions on Automatic Control, 2021, 66(8): 3773-3780.

[12] Zhou Yuanqiang, Vamvoudakis K G, Haddad W M, et al. A secure control learning framework for cyber-physical systems under sensor and actuator attacks [J]. IEEE Transactions on Cybernetics, 2021, 51(9): 4648-4660.

[13] Moghadam R, Modares H. Resilient autonomous control of distributed multiagent systems in contested environments [J]. IEEE Transactions on Cybernetics, 2019, 49(11): 3957-3967.

[14] Xu Yuanyuan, Li Tieshan, Yang Yue, et al. Simplified ADP for event-triggered control of multiagent systems against FDI attacks[J]. IEEE Transactions on Systems, Man, and Cybernetics: System, 2023, 53(8): 4672-4683.

[15] Kokolakis N–M T, Vamvoudakis K G. Safety-aware pursuit-evasion games in unknown environments using Gaussian processes and finite-time convergent reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024,35(3): 3130-3143.

Optimal Secure Tracking Control in Multi-UAVs Based on Online Reinforcement Learning

Gong Zhenyu, Yang Feisheng

Northwestern Polytechnical University, Xi'an 710072, China

Abstract: In Unmanned Aerial Vehicle (UAV) formation tracking missions, False Data Injection (FDI) attackers can inject misleading data into the control commands, resulting in the fact that UAVs can not form the specified formation configuration, so there is a need to design a secure formation tracking controller. The attack-defense process was modeled as a zero-sum graphical game, in which the FDI attacker and the secure controller were viewed as game players. The attacker aims to maximize the cost function yet the secure controller serves a contrary purpose. Solving the game and acquiring the optimal secure control policy rely on solving the Hamilton-Jacobi-Isaacs (HJI) equation. The HJI equation is a coupled partial differential equation, which is difficult to solve directly. Therefore, the finite-time convergent online reinforcement learning algorithm that combines the experience replay mechanism was introduced and the critic-only neural network was utilized to approximate the value function for obtaining the optimal secure control policy. A numerical simulation was given to show the effectiveness of the raised scheme.

Key Words: FDI attack; multi-UAVs; online reinforcement learning; optimal control; zero-sum graphical game

Received: 2023-06-27; **Revised:** 2023-11-16; **Accepted:** 2023-12-25

Foundation item: National Natural Science Foundation of China (62073269); Aeronautical Science Foundation of China (2020Z034053002); Key Research and Development Program of Shaanxi (2022GY-244); Natural Science Foundation of Chongqing (CSTB2022NSCQ-MSX0963); Guangdong Basic and Applied Basic Research Foundation(2023A1515011220)