



基于强化学习方法的多智能体追逃博弈任务分配

李尚恒¹, 刘昊^{1,2*}, 任梓铭³, 李亚璠⁴, 刘大卫⁴

1. 北京航空航天大学人工智能学院, 北京 100191

2. 中关村国家实验室, 北京 100194

3. 北京航空航天大学宇航学院, 北京 100191

4. 中国兵器科学研究院, 北京 100089

* E-mail: liuhao13@buaa.edu.cn

收稿日期: 2024-06-20; 接受日期: 2024-11-18; 网络版发表日期: 2025-04-10

摘要 本文研究了在模型非线性动力学和外部干扰影响下的多智能体追逃博弈中的任务分配问题. 本文设计了一种最优任务分配价值函数, 将博弈中的任务分配问题转化为多个追逐者的轨迹跟踪问题. 本文提出了一种离策略积分强化学习方法, 使用多智能体系统的输入和输出数据, 迭代求解非线性哈密顿-雅可比-贝尔曼方程, 并引入神经网络来拟合价值函数和最优策略. 在不了解智能体模型的情况下, 求解最优控制策略和任务执行的总成本. 仿真结果验证了所提任务分配方法的有效性.

关键词 多智能体, 任务分配, 最优控制, 博弈, 强化学习

1 引言

近年来, 多无人机系统作为一种典型的多智能体系统, 因其可靠性和协同性而受到广泛关注. 多智能体追逃博弈在导弹制导、物资运输、协同拦截等领域得到广泛应用^[1]. 在执行复杂且规模庞大的任务时, 多无人机系统面临着巨大的挑战. 为了实现资源的有效调度和高效利用, 设计一种多智能体任务分配算法显得尤为重要. 在多智能体追逃博弈过程中, 无人机受到模型参数不确定及外界扰动的影响^[2], 这使得复杂环境下多智能体追逃博弈中的多任务分配问题成为一个极具挑战性的难题.

在过去几年中, 研究者们针对多无人机系统提出

了多种多任务分配算法, 以应对不同场景下的挑战. 文献[3]通过设计基于分布式反协调博弈的分配算法, 解决了动态拓扑和节点高流动性带来的问题. 文献[4]则提出了一种元启发式优化方法, 应对系统参数不确定性的影响. 文献[5]则引入延迟决策策略, 提高了系统的任务完成率和预算利用率. 文献[6-9]尝试通过设计如鸽群式优化算法、边际效益启发算法、遗传算法等多种启发式算法来解决多智能体系统的复杂多任务分配问题. 然而, 现有的方法往往基于精确且完整的智能体动力学模型, 将博弈中的任务分配问题转化为求解密尔顿-雅可比-贝尔曼(HJB)方程. 由于无人机的复杂的质量分布以及非线性、强耦合的动态特征, 使得无人机模型参数不确定, 导致非线性的HJB方程难以

引用格式: 李尚恒, 刘昊, 任梓铭, 等. 基于强化学习方法的多智能体追逃博弈任务分配. 中国科学: 技术科学, 2025, 55: 906-913

Li S H, Liu H, Ren Z M, et al. Task assignment in multi-agent games via reinforcement learning (in Chinese). Sci Sin Tech, 2025, 55: 906-913, doi: 10.1360/SST-2024-0154

求解。因此, 针对多智能体追逃博弈中多任务分配问题, 仍需设计一种全新的算法, 以综合考虑智能体动态的高度非线性特性和模型参数不确定的影响, 实现更高效的任务分配。

强化学习作为一种不依赖模型信息的智能方法被应用到多智能体追逃博弈中, 以解决多任务分配问题。文献[10]提出了一种梯度递归强化学习方法, 以适应具有输入饱和约束的系统。文献[11,12]中引入了一种深度确定性策略梯度算法来解决多任务分配问题。文献[13,14]为系统设计了两种分布式强化学习框架, 分别用于缩小决策空间和实现全局优化。文献[15]中的强化学习算法旨在解决多无人机系统间的协同决策问题。上述基于强化学习的任务分配是基于特定的任务场景下的方法, 没有考虑外界干扰对算法收敛性带来的影响。

本文提出了一种离策略强化学习算法来解决多无人机系统在非线性动力学特性和外界扰动条件下的任务分配问题。构造多智能体追逃博弈的任务分配准则来最小化多个任务的执行成本。在此基础上, 建立了追逐者的最优轨迹跟踪控制器, 并建立了多智能体追逃博弈的局部位置误差模型。

本文的主要贡献如下: 首先, 针对具有非线性和强耦合特性的四旋翼无人机系统, 提出了一种最优控制方法, 能够在无人机模型参数不确定和外界扰动的条件下, 求解最优控制策略并评估任务执行代价。其次, 提出了一种以最小执行代价为目标的任务分配方法, 在已知单个任务执行代价的基础上进行分配, 不仅能够最大化追逐者与逃跑者的匹配数量, 还能有效地最小化总任务执行代价。

2 预备知识与问题描述

2.1 图博弈理论

在博弈环境中, 智能体之间的通信拓扑可用四个加权图来描述^[16], 且每个节点只能从邻居智能体获取信息。定义集合 $\mathcal{G}_p = \{\mathcal{V}_p, \mathcal{E}_p, \mathcal{A}_p\}$ 存在 N 个节点, 其中 $\mathcal{V}_p = \{v_1, v_2, \dots, v_N\}$ 表示追逐者集合, $\mathcal{E}_p \subset \mathcal{V}_p \times \mathcal{V}_p$ 表示边的集合, $\mathcal{A}_p = [a_{ik}] \in \mathbb{R}^{N \times N}$ 表示加权邻接矩阵。如图1所示, 追逐者 i 在 $\mathcal{G}_p$ 的通信拓扑结构可以用 a_{ik} 表示, 如果 $(v_i, v_j) \in \mathcal{E}_p$ 则 $a_{ik} = 1$, 否则 $a_{ik} = 0$ 。 $d_i^p = \sum_{k=1}^N a_{ik}$ 表

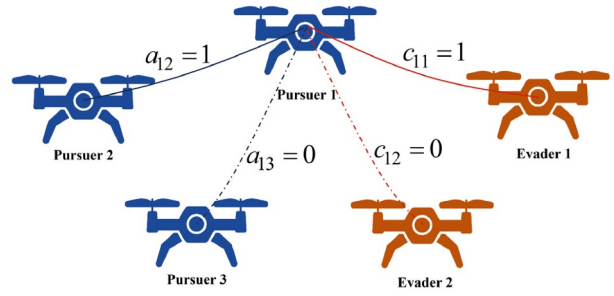


图1 (网络版彩图)图博弈理论示意图

Figure 1 (Color online) Diagram of graph game theory.

示追逐者在 $\mathcal{G}_p$ 中的邻居数量。

同样, 在博弈中 $\mathcal{G}_{pe} = \{\mathcal{V}_{pe}, \mathcal{E}_{pe}, \mathcal{A}_{pe}\}$ 由 $N+M$ 个节点组成, 表示 N 个追逐者和 M 个逃跑者的信息。 $\mathcal{V}_{pe} = \{\mathcal{V}_p, \mathcal{V}_e\}$ 表示所有追逐者和逃跑者的集合, $\mathcal{E}_{pe} \subset \mathcal{V}_p \times \mathcal{V}_e$ 表示边的集合, $\mathcal{A}_{pe} = [c_{ij}] \in \mathbb{R}^{N \times M}$ 表示加权邻接矩阵。如果追逐者 i 和逃跑者 j 之间存在信息拓扑, 则 $c_{ij} = 1$; 否则 $c_{ij} = 0$ 。 $d_i^{pe} = \sum_{j=1}^M c_{ij}$ 表示追逐者 i 在 $\mathcal{G}_{pe}$ 中的邻居数量^[17]。

2.2 四旋翼无人机模型

本文使用的智能体模型为四旋翼无人机模型。 $\mathbf{L}_I = \{l_I^x, l_I^y, l_I^z\}$ 表示以地面惯性系, $\mathbf{L}_B = \{l_B^x, l_B^y, l_B^z\}$ 表示四旋翼无人机的机体系, $\mathbf{R}_{ib}$ 和 $\mathbf{C}_{\theta i}$ 分别表示从 $\mathbf{L}_B$ 到 $\mathbf{L}_I$ 的旋转矩阵和科里奥利项^[18]。根据文献[19], 四旋翼飞行器的非线性动力学模型表示为

$$\begin{aligned} m_i \ddot{\mathbf{q}}_i &= \mathbf{R}_{ib} \mathbf{F}_i + \mathbf{d}_{q_i} \\ \mathbf{J}_{\theta i} \ddot{\boldsymbol{\Theta}}_i &= -\mathbf{C}_{\theta i}(\boldsymbol{\Theta}_i, \dot{\boldsymbol{\Theta}}_i) \boldsymbol{\Theta}_i + \boldsymbol{\tau}_i + \mathbf{d}_{\theta i} \end{aligned} \quad (1)$$

其中, $\mathbf{q}_i = [q_{i,x}, q_{i,y}, q_{i,z}]^T \in \mathbb{R}^3 (i = 1, 2, \dots, N)$ 为位置向量, $\boldsymbol{\Theta}_i = [\phi_i, \theta_i, \psi_i]^T \in \mathbb{R}^3$ 为 $\mathbf{L}_I$ 下的欧拉角向量, 满足滚转角 $\phi_i \in (-\pi/2, \pi/2)$ 、俯仰角 $\theta_i \in (-\pi/2, \pi/2)$ 和偏航角 $\psi_i \in [0, 2\pi)$ 。 m_i 为质量, $\mathbf{J}_{ei}$ 为四旋翼飞行器的惯性矩阵。

2.3 问题描述

多无人机系统的对抗过程可模拟为多智能体追逃博弈, 每个追逐者都会被分配一个攻击者作为目标。因

此, 每个追逐者的控制策略等同于对目标的轨迹跟踪策略. 根据文献[18], (1)可改写为

$$\begin{aligned}\ddot{\mathbf{q}}_i &= \mathbf{b}_{qi} \mathbf{u}_{qi} - g \mathbf{c}_{3,3} + \Delta_{qi} \\ \ddot{\boldsymbol{\theta}}_i &= \mathbf{b}_{\theta i} \mathbf{u}_{\theta i} - (\mathbf{C}(\boldsymbol{\theta}_i, \dot{\boldsymbol{\theta}}_i) \boldsymbol{\theta}_i) / \mathbf{J}_{\theta i} + \Delta_{\theta i}\end{aligned}\quad (2)$$

其中 $\mathbf{b}_{qi} = m_i^{-1} k_{wi} \mathbf{I}_3$, $\mathbf{b}_{\theta i} = \mathbf{J}_{\theta i}^{-1} \text{diag}(l_{ti} k_{wi}, l_{ti} k_{wi}, k_{ti})$. 位置环外部干扰 Δ_{qi}

$$\Delta_{qi} = \mathbf{b}_{qi} (\mathbf{u}_{zi} \mathbf{R}_{ib} \mathbf{c}_{3,3} - \mathbf{u}_{qi}) + \mathbf{d}_{qi} \quad (3)$$

姿态环外部干扰满足

$$\Delta_{\theta i} = \mathbf{d}_{\theta i} \quad (4)$$

定义追逐者的状态为 $\mathbf{x}_p$, 攻击者的状态为 $\mathbf{x}_j$, 则追逐者追逐目标的任务执行代价表示为 $\sigma(\mathbf{e}(\mathbf{x}_p, \mathbf{x}_j), \mathbf{u}(\mathbf{x}_p, \mathbf{x}_j))$, 其中 $\mathbf{e}(\mathbf{x}_p, \mathbf{x}_j)$ 为过程状态误差, $\mathbf{u}(\mathbf{x}_p, \mathbf{x}_j)$ 为过程控制输入. 多智能体追逃博弈的准则可以表示为

$$V^* = \min \sum_{i=1}^N \sigma(\mathbf{e}(\mathbf{x}_p, \mathbf{x}_j), \mathbf{u}^*(\mathbf{x}_p, \mathbf{x}_j)). \quad (5)$$

目标状态确定后, 反馈控制策略 $\mathbf{u}^*$ 也可同时确定. 然后, 将多智能体追逃博弈中的多任务分配问题转换为求解模型参数不确定性和外界干扰下的最优策略和最优代价.

3 任务分配方法设计

3.1 位置控制器设计

在多智能体追逃博弈中, 假设追逐者和逃跑者均为四旋翼无人机模型, $\mathbf{x}_p, \mathbf{x}_e \in \mathbb{R}^{6 \times 1}$ 代表追逐者和逃跑者的状态, $\mathbf{u}_p, \mathbf{u}_e \in \mathbb{R}^{3 \times 1}$ 代表追逐者和逃跑者的控制输入. 追逐者的动态模型可表示为

$$\begin{aligned}\dot{\mathbf{x}}_p &= \mathbf{A}_p \mathbf{x}_p + \mathbf{B}_p \mathbf{u}_p - g \mathbf{c}_{6,6} + \mathbf{D}_p \Delta_p \\ \mathbf{y}_p &= \mathbf{C}_p \mathbf{x}_p\end{aligned}\quad (6)$$

其中 $\mathbf{A}_i = [\mathbf{0}_{6 \times 3}, \mathbf{c}_{6,1}, \mathbf{c}_{6,2}, \mathbf{c}_{6,3}] \in \mathbb{R}^{6 \times 6}$, $\mathbf{B}_i = [\mathbf{0}_3, \mathbf{b}_i]^T$, $\mathbf{C}_i = [\mathbf{c}_{3,1}, \mathbf{c}_{3,2}, \mathbf{c}_{3,3}, \mathbf{0}_{3,3}]$, $\mathbf{D}_i = [\mathbf{0}_{3 \times 1}, \mathbf{I}_3]^T$. 追逃博弈中的攻击者的动态可表示为

$$\begin{aligned}\dot{\mathbf{x}}_e &= \mathbf{A}_e \mathbf{x}_e + \mathbf{B}_e \mathbf{u}_e \\ \mathbf{y}_e &= \mathbf{C}_e \mathbf{x}_e\end{aligned}\quad (7)$$

结合(6)和(7), 可以建立多智能体追逃博弈问题的增广系统:

$$\begin{aligned}\dot{\mathbf{X}} &= \bar{\mathbf{A}} \mathbf{X} + \bar{\mathbf{B}} \mathbf{u} + \bar{\mathbf{C}} + \bar{\mathbf{D}} \Delta_m \\ \mathbf{e} &= \mathbf{H} \mathbf{X},\end{aligned}\quad (8)$$

其中 $\mathbf{X} = [\mathbf{x}_p, \mathbf{x}_e]^T$, $\bar{\mathbf{A}} = \text{diag}(\mathbf{A}_p, \mathbf{A}_e)$, $\bar{\mathbf{B}} = [\mathbf{B}_p^T, \mathbf{0}_{3 \times 6}]^T$,

$$\bar{\mathbf{C}} = [-g \mathbf{c}_{6,6}, \mathbf{0}_{6 \times 1}]^T, \mathbf{H} = [\mathbf{c}_{3,1}, \mathbf{c}_{3,2}, \mathbf{c}_{3,3}, -\mathbf{c}_{3,1}, -\mathbf{c}_{3,2}, -\mathbf{c}_{3,3}, \mathbf{0}_{3 \times 3}].$$

根据真实环境中动态模型扰动的有界性, 本章中的扰动项 Δ_m 仍然满足有界扰动假设:

$$\int_t^\infty e^{-\beta(\tau-t)} (\mathbf{e}^T \mathbf{Q} \mathbf{e} + \mathbf{u}^T \mathbf{R} \mathbf{u}) d\tau \leq \zeta^2 \int_t^\infty e^{-\beta(\tau-t)} \Delta_m^T \Delta_m d\tau, \quad (9)$$

其中 $\mathbf{Q} > 0$, $\mathbf{R} > 0$, $\beta > 0$, $\zeta > 0$, e 为自然对数的底数. 在此假设下, 外部扰动 Δ_m 对系统的影响可以被抑制到上确界 ζ . 追逃博弈在对抗过程中的代价函数为

$$V = \int_t^\infty e^{-\beta(\tau-t)} r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m) d\tau, \quad (10)$$

其中 $r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m)$ 考虑了过程状态和控制输入, 并将干扰作为惩罚项进行惩罚, 满足:

$$r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m) = \mathbf{e}^T \mathbf{Q} \mathbf{e} + \mathbf{u}^T \mathbf{R} \mathbf{u} + \Phi(\mathbf{X}) - \zeta^2 \Delta_m^T \Delta_m. \quad (11)$$

对于追逐者的价值函数, 追逐者不仅希望迅速收敛状态误差 $\mathbf{e}$, 还希望在拦截时不要离初始位置太远. 因此, 追逐者与初始状态之间的距离被作为惩罚项. 以 $\mathbf{X}_0$ 作为 $\mathbf{X}$ 的初始状态, 可以定义 $\Phi(\mathbf{X})$ 为二次型:

$$\Phi(\mathbf{X}) = (\mathbf{X} - \mathbf{X}_0)^T \mathbf{Q}_X (\mathbf{X} - \mathbf{X}_0), \quad (12)$$

其中 $\mathbf{Q}_X \geq 0$. 在定义了多智能体追逃博弈的代价函数之后, 多智能体追逃博弈问题的关键在于最小化 V 来求解 $\mathbf{u}^*$:

$$\mathbf{u}^* = \underset{\Delta_m}{\text{argmin}} \max V. \quad (13)$$

根据最优控制理论, (10)对应的哈密顿函数为

$$H(\mathbf{V}, \mathbf{u}, \Delta_m) = r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m) - \beta V + (\Delta V)^T \dot{\mathbf{X}} = 0, \quad (14)$$

其中 $\Delta V = \partial V / \partial \mathbf{X}$. V^* 为待求解的最优代价, 满足 $V^* =$

$\min_{\mathbf{u}} \int_t^\infty r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m) d\tau$. 本文采用Min-Max策略, 控制 $\mathbf{u}$ 最小化任务执行代价, 扰动 Δ_m 最大化任务执行代价. 根据最优控制理论的静态条件, 将哈密顿函数(14)分别对控制 $\mathbf{u}$ 和扰动 Δ_m 求偏导:

$$\begin{aligned}\partial H(V^*, \mathbf{u}, \Delta_m) / \partial \mathbf{u} &= 0, \\ \partial H(V^*, \mathbf{u}, \Delta_m) / \partial \Delta_m &= 0.\end{aligned}\quad (15)$$

相应的最优反馈控制策略以及外界扰动:

$$\begin{aligned}\mathbf{u}^* &= -\mathbf{R}^{-1}\mathbf{B}^T\Delta\mathbf{V}^*/2, \\ \Delta_m^* &= \mathbf{D}^T\Delta\mathbf{V}^*/(2\zeta^2).\end{aligned}\quad (16)$$

将最优反馈控制策略(16)代入哈密顿函数(10)可得

$$\begin{aligned}H(\mathbf{V}, \mathbf{u}, \Delta_m) \\ \triangleq \mathbf{e}^T\mathbf{Q}\mathbf{e} + \frac{1}{4\zeta^2}(\Delta\mathbf{V}^*)^T\mathbf{D}\mathbf{D}^T\Delta\mathbf{V}^* \\ -\Phi(\mathbf{X}) - \frac{1}{4}(\Delta\mathbf{V}^*)^T\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\Delta\mathbf{V}^* - \beta\mathbf{V}^* \\ +(\Delta\mathbf{V}^*)^T(\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{u} + \mathbf{C} + \mathbf{D}\Delta_m) = 0.\end{aligned}\quad (17)$$

然而, 由于四旋翼无人飞行器动态模型的非线性、强耦合性、模型参数的不确定性以及外部干扰因素, 求解公式(17)非常困难. 为了实现在各种干扰因素下不依赖模型精确参数的求解, 引入积分强化学习算法对(17)迭代求解.

3.2 最优任务执行代价

首先, 为增广系统(8)引入一个迭代项:

$$\begin{aligned}\dot{\mathbf{X}} &= \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{u} + \mathbf{C} + \mathbf{D}\Delta_m \\ &+ \mathbf{B}(\mathbf{u} - \mathbf{u}^n) + \mathbf{D}(\Delta_m - \Delta_m^n).\end{aligned}\quad (18)$$

对于(18), 当估计值在迭代过程中逼近真实值时, $\mathbf{B}(\mathbf{u} - \mathbf{u}^n) + \mathbf{D}(\Delta_m - \Delta_m^n)$ 将趋于0. 结合(14)和(16), 可得

$$\begin{aligned}\dot{V}^n &= (\Delta V^n)^T \dot{\mathbf{X}} \\ &= (\Delta V^n)^T (\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{u} + \mathbf{C} + \mathbf{D}\Delta_m \\ &+ \mathbf{B}(\mathbf{u} - \mathbf{u}^n) + \mathbf{D}(\Delta_m - \Delta_m^n)) \\ &= \beta V + (2\zeta^2\Delta_m^{n+1})^T (\Delta_m - \Delta_m^n) \\ &- r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m) - 2(\mathbf{u}^{n+1})^T \mathbf{R}(\mathbf{u} - \mathbf{u}^n).\end{aligned}\quad (19)$$

将(19)两边乘以 $e^{-\beta t}$, 然后对两边进行时间 $[t: t+T]$ 上的积分, 可以得到积分贝尔曼方程:

$$\begin{aligned}e^{-\beta T}V(\mathbf{X}(t+T)) - V(\mathbf{X}(t)) \\ = -\int_t^{t+T} e^{-\beta(\tau-t)} r(\mathbf{X}, \mathbf{e}, \mathbf{u}, \Delta_m) d\tau \\ + \int_t^{t+T} e^{-\beta(\tau-t)} 2\zeta^2 (\Delta_m^{n+1})^T (\Delta_m - \Delta_m^n) d\tau \\ - \int_t^{t+T} e^{-\beta(\tau-t)} 2(\mathbf{u}^{n+1})^T \mathbf{R}(\mathbf{u} - \mathbf{u}^n) d\tau.\end{aligned}\quad (20)$$

将增广系统的状态表示为 $\mathbf{X} = [x_1, x_2, \dots, x_n]^T$, 强化学习的中间变量定义为 $\Psi_{xx} = [e^{-\beta T}\mathbf{x}(t+T) - \mathbf{x}(t)]$, 其中 $\mathbf{x} = [x_1^2, x_1x_2, \dots, x_1x_n, \dots, x_{n-1}^2, \dots, x_{n-1}x_n, x_n^2]^T$. 为拟

合不确定性项, 引入了三组神经网络, 设置如下:

$$\begin{aligned}\hat{V}^n(\mathbf{X}) &= \mathbf{P}^T\mathbf{x}, \\ \hat{\mathbf{u}}^{n+1}(\mathbf{X}) &= \mathbf{W}_u^{n+1}\mathbf{X}, \\ \hat{\Delta}_m^{n+1}(\mathbf{X}) &= \mathbf{W}_\Delta^{n+1}\mathbf{X},\end{aligned}\quad (21)$$

其中 $\hat{V}^n(\mathbf{X})$, $\hat{\mathbf{u}}^{n+1}(\mathbf{X})$ 和 $\hat{\Delta}_m^{n+1}(\mathbf{X})$ 是神经网络的近似估计值. $\mathbf{P} \in \mathbb{R}^{n(n+1)/2 \times 1}$, $\mathbf{W}_u^{n+1} \in \mathbb{R}^{3 \times n}$ 和 $\mathbf{W}_\Delta^{n+1} \in \mathbb{R}^{3 \times n}$ 为权重矩阵. 基于神经网络(21), 可以将式(20)的不确定项进行近似替换, 可以重构为

$$\begin{aligned}e^{-\beta T}V^n(\mathbf{X}(t+T)) - V^n(\mathbf{X}(t)) &= \mathbf{P}^T\Psi_{xx} \\ &= -\int_t^{t+T} e^{-\beta(\tau-t)} \left\{ \mathbf{X}^T\mathbf{H}^T\mathbf{Q}\mathbf{H}\mathbf{X} + \mathbf{X}^T(\mathbf{W}_u^n)^T\mathbf{R}\mathbf{W}_u^n\mathbf{X} \right\} d\tau \\ &+ \int_t^{t+T} e^{-\beta(\tau-t)} \left\{ -\zeta^2\mathbf{X}^T(\mathbf{W}_\Delta^n)^T\mathbf{W}_\Delta^n\mathbf{X}_{pi} \right\} d\tau \\ &+ \int_t^{t+T} e^{-\beta(\tau-t)} \left\{ \mathbf{X}^T\mathbf{Q}_x\mathbf{X} - 2\mathbf{X}_0^T\mathbf{Q}_x\mathbf{X} + \mathbf{X}_0^T\mathbf{Q}_x\mathbf{X}_0 \right\} d\tau \\ &+ \int_t^{t+T} e^{-\beta(\tau-t)} 2\zeta^2\mathbf{X}^T(\mathbf{W}_\Delta^{n+1})^T (\Delta_m - \mathbf{W}_\Delta^n\mathbf{X}) d\tau \\ &- \int_t^{t+T} e^{-\beta(\tau-t)} 2\mathbf{X}^T(\mathbf{W}_u^{n+1})^T \mathbf{R}(\mathbf{u} - \mathbf{K}_u^n\mathbf{X}) d\tau.\end{aligned}\quad (22)$$

在式(22)中, $(V^n, \mathbf{u}^{n+1}, \Delta_m^{n+1})$ 为未知的神经网络近似项, 所有其他项均可从系统的输入和输出信息中获得. 根据文献[20], 在此数据采集过程中存在连续激励条件, 神经网络参数值可使用最小二乘法求解. 从式(20)、(21)和(22)中可以看出, 在无人机动力学模型的精确参数未知的条件下, 利用神经网络近似得到执行代价的值 V^* , 以及相应的最优反馈控制律 $\mathbf{u}^*$.

定理 3.1: 贝尔曼方程(20)的解是 $(V^n, \mathbf{u}^{n+1}, \Delta_m^{n+1})$ 当且仅当(14)和(15)的解也是 $(V^n, \mathbf{u}^{n+1}, \Delta_m^{n+1})$, 且 $(V^n, \mathbf{u}^{n+1}, \Delta_m^{n+1})$ 是贝尔曼方程(20)的唯一解.

证明: 从贝尔曼方程(20)可以得到

$$\begin{aligned}d(e^{-\beta t}V^n)/dt \\ = -e^{-\beta t}r(\mathbf{e}, \mathbf{u}^n, \Delta_m^n) - 2e^{-\beta t}(\mathbf{u}^{n+1})^T \mathbf{R}(\mathbf{u} - \mathbf{u}^n) \\ + e^{-\beta t}(2\zeta^2\Delta_m^{n+1})^T (\Delta_m - \Delta_m^n).\end{aligned}\quad (23)$$

假设 $(V', \mathbf{u}', \Delta_m')$ 是(20)的另一个解, 也满足(23).

$$\begin{aligned}d(e^{-\beta t}V')/dt \\ = -e^{-\beta t}r(\mathbf{e}, \mathbf{u}^n, \Delta_m^n) - 2e^{-\beta t}(\mathbf{u}')^T \mathbf{R}(\mathbf{u} - \mathbf{u}^n) \\ + e^{-\beta t}(2\zeta^2\Delta_m')^T (\Delta_m - \Delta_m^n).\end{aligned}\quad (24)$$

那么将(23)和(24)做差可得

$$\begin{aligned} & d(e^{-\beta t} V^n - e^{-\beta t} V') / dt \\ &= -2e^{-\beta t} (\mathbf{u}^{n+1} - \mathbf{u}')^T \mathbf{R}(\mathbf{u} - \mathbf{u}') \\ &+ 2\zeta^2 e^{-\beta t} (\Delta_m^{n+1} - \Delta_m')^T (\Delta_m - \Delta_m^n). \end{aligned} \quad (25)$$

通过让 $\mathbf{u} = \mathbf{u}^n$ 和 $\Delta_m = \Delta_m^n$, 可以得出

$$d(e^{-\beta t} V^n - e^{-\beta t} V') / dt = 0. \quad (26)$$

式(26)可以得出 $e^{-\beta t} V^n - e^{-\beta t} V' = c$, c 是一个常数, 满足 $c = V^n(0) - V'(0) = 0$, $V^n = V'$.

$$\begin{aligned} & -2e^{-\beta t} (\mathbf{u}^{n+1} - \mathbf{u}')^T \mathbf{R}(\mathbf{u} - \mathbf{u}') \\ &+ 2\zeta^2 e^{-\beta t} (\Delta_m^{n+1} - \Delta_m')^T (\Delta_m - \Delta_m^n) = 0. \end{aligned} \quad (27)$$

由于(30)对任何给定的 $\mathbf{u}$ 和 Δ_m 成立, 因此可以得出 $\mathbf{u}^{n+1} = \mathbf{u}'$ 和 $\Delta_m^{n+1} = \Delta_m'$. 那么 $(\mathbf{V}^n, \mathbf{u}^{n+1}, \Delta_m^{n+1})$ 是贝尔曼方程(20)的唯一解, 从而完成了证明.

3.3 任务分配

第3.2节通过强化学习理论拟合四旋翼无人飞行器的任务执行代价和控制输入. 在求解所有追逐者和攻击者匹配对的最优任务执行代价后, 四旋翼无人机群的任务分配问题转化为基于最优轨迹跟踪控制的离散最优分配问题, 并利用最优控制策略 $\mathbf{u}^*$ 来追逐目标, 实现总的任务执行代价 V_{total}^* 最小.

$$\begin{aligned} & \max \sum_{i=1}^N \sum_{j=1}^M \frac{1}{V_{ij}} \mathbf{T}_{ij}, \\ & s.t. \sum_{i=1}^M \mathbf{T}_{ij} \leq 1, i \in [1, \dots, N], \\ & \sum_{i=1}^N \mathbf{T}_{ij} \leq 1, j \in [1, \dots, M], \end{aligned} \quad (28)$$

其中, V_{ij} 表示3.2节中追逐者 i 和逃跑者 j 的最小任务执

行代价矩阵. $\mathbf{T}_{ij}$ 为匹配矩阵, $\mathbf{T}_{ij} = 1$ 表示追逐者 i 和逃跑者 j 进行追逐匹配, $\mathbf{T}_{ij} = 0$ 表示未匹配. 通过强化学习方法求解拟合任务代价的神经网络参数, 并构建所有匹配对的任务执行代价矩阵 V_{ij} , 然后通过(28)进行任务分配. 在(28)中, $1/V_{ij}$ 保证了最小的任务执行代价, $\mathbf{T}_{ij}$ 保证了最多的匹配对, 使更多的逃跑者被抓捕. 总的执行流程如图2所示.

4 仿真

为了验证任务分配算法的有效性, 本章设计了多智能体追逃仿真, 并使用式(1)对异构无人飞行器进行建模. 仿真由三个追逐者和三个逃跑者组成. 逃跑者的初始位置设置为 $\mathbf{q}_{e1}(0) = [-14, -4, 4]^T \text{m}$, $\mathbf{q}_{e2}(0) = [-9, +7, 5]^T \text{m}$, $\mathbf{q}_{e3}(0) = [-10, -1, 6]^T \text{m}$. 同样, 将追逐者的初始位置设置为 $\mathbf{q}_{p1}(0) = [-18, -6, 3]^T \text{m}$, $\mathbf{q}_{p2}(0) = [-12, +7, 8]^T \text{m}$, $\mathbf{q}_{p3}(0) = [-14, 0, 4]^T \text{m}$.

为了模拟外部环境扰动对模型的影响, 将等效干扰定义为 $\Delta_q = (-1)^i [0.2\sin(t), 0.2\cos(t), 0.2\sin(t)]^T$, 价值函数中的加权矩阵定义为 $\mathbf{Q} = 9\mathbf{I}_6$, $\mathbf{R} = \mathbf{I}_3$ 和 $\mathbf{Q}_x = \text{diag}(\mathbf{0}_3, \mathbf{I}_3)$. 外部扰动的上界相关的参数为 $\zeta = 2$. 仿真环境步长设置为 $T = 0.1$.

在收集用于强化学习算法的四旋翼无人机模型的输入输出信息的阶段, 初始控制策略 $\mathbf{u}_0 = \mathbf{W}_u^0 \mathbf{X}$ 使用标准PD控制器生成, 其中探测噪声被定义为满足的一组正弦信号的叠加 $\mathbf{u}_e = 10 \sum_{m=1}^{100} \sin(\omega t)$, ω 是属于 $(-2\pi, 2\pi]$ 的随机数.

$$\mathbf{W}_u^0 = \begin{bmatrix} 1.4 & 0 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & -0.6 & 0 & 0 \\ 0 & 2.2 & 0 & 0 & 0.8 & 0 & 0 & -2.1 & 0 & 0 & -0.7 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1.1 & 0 & 0 & -1.6 & 0 & 0 & -0.8 \end{bmatrix}. \quad (29)$$

强化学习算法中的神经网络是在模拟环境中训练的. 首先, 给定初始稳定的控制率 $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}_e$, 其中, $\mathbf{u}_0$ 为初始稳定的控制策略, $\mathbf{u}_e$ 为输入探测噪声, 在系统的跟踪误差收敛时收集输入和状态数据. 将收集到的数据通过式(22)迭代求解. $\mathbf{P}$, $\mathbf{W}_u$ 和 $\mathbf{W}_\Delta$ 的收敛结果如图3

所示, 可以发现在前几次迭代中收敛. 使用学习得到的最优控制策略进行博弈, 如图4所示, 所有追逐者都会选择距离最近的躲避者, 且博弈过程中状态误差如图5所示. 可以观察到, 所有的追逐者都被分配到离自己最近的逃跑者, 并在追逐过程中逐渐减小状态差.

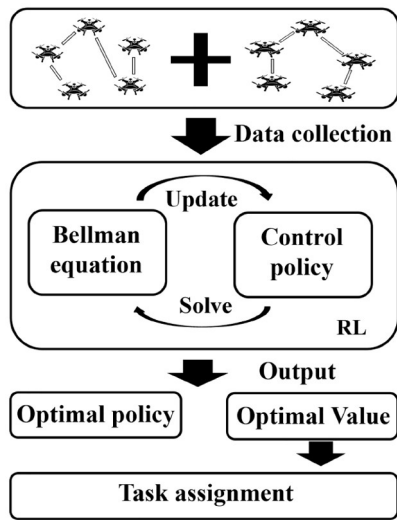


图 2 任务分配方法流程

Figure 2 Task assignment method flowchart.

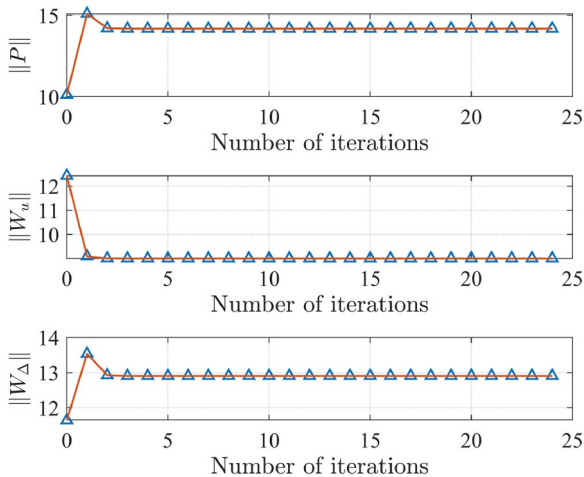


图 3 (网络版彩图)神经网络参数收敛图

Figure 3 (Color online) Neural networks parameters convergence diagram.

5 总结

本文研究了受模型参数不确定性和外部扰动影响的多智能体追逃博弈中的任务分配问题。本文提出了一种新型的任务执行代价,并将任务分配问题转化为多个追逐者的轨迹跟踪问题。为了应对非线性哈密顿-

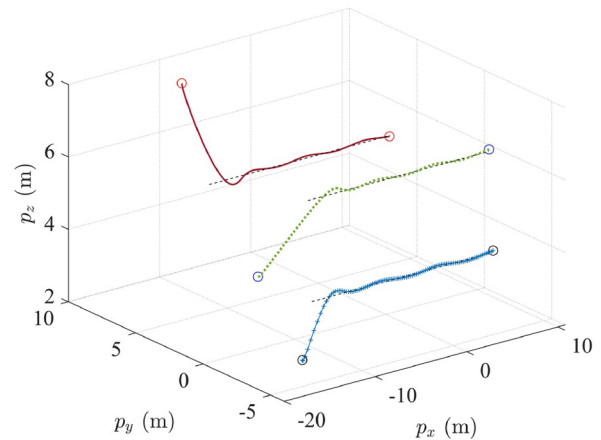


图 4 (网络版彩图)多无人机追逃博弈轨迹图

Figure 4 (Color online) 3-D trajectories of games.

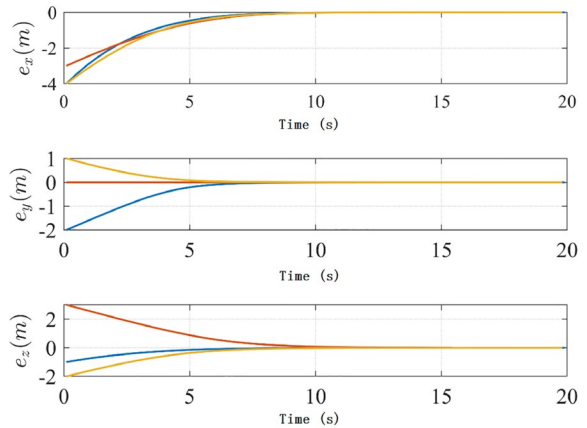


图 5 (网络版彩图)博弈过程的状态误差

Figure 5 (Color online) State errors in the game process.

雅可比-贝尔曼方程求解的难题,本文引入了一种离策略积分强化学习方法,通过使用多智能体系统的输入输出数据,结合神经网络拟合价值函数和最优策略,实现了在未知智能体模型条件下的最优控制策略求解。仿真实验验证了所提出方法的有效性,表明该方法能够在存在参数不确定性和外界扰动的情况下,有效实现多智能体的任务分配,且能够显著降低任务执行的总成本。本文作为离策略强化学习方法,控制策略并不能在线更新,后续将会对在线方法进一步研究。

参考文献

- 1 Zhang H, Zhang J, Yang G H, et al. Leader-based optimal coordination control for the consensus problem of multiagent differential games via

- fuzzy adaptive dynamic programming. *IEEE Trans Fuzzy Syst*, 2015, 23: 152–163
- 2 Guo L, Yu X, Zhang X, et al. Safety control system technologies for UAVs: Review and prospect (in Chinese). *Sci Sin Inform*, 2020, 50: 184–194
[郭雷, 余翔, 张霄, 等. 无人机安全控制系统技术: 进展与展望. 中国科学: 信息科学, 2020, 50: 184–194]
- 3 Tang F, Fadlullah Z M, Kato N, et al. AC-POCA: Anticoordination game based partially overlapping channels assignment in combined UAV and D2D-based networks. *IEEE Trans Veh Technol*, 2018, 67: 1672–1683
- 4 Chen Y, Yang D, Yu J. Multi-UAV task assignment with parameter and time-sensitive uncertainties using modified two-part wolf pack search algorithm. *IEEE Trans Aerosp Electron Syst*, 2018, 54: 2853–2872
- 5 Ding Y, Zhang L, Guo L. Dynamic delayed-decision task assignment under spatial-temporal constraints in mobile crowdsensing. *IEEE Trans Netw Sci Eng*, 2022, 9: 2418–2431
- 6 Duan H, Zhao J, Deng Y, et al. Dynamic discrete pigeon-inspired optimization for multi-UAV cooperative search-attack mission planning. *IEEE Trans Aerosp Electron Syst*, 2021, 57: 706–720
- 7 Xin B, Wang Y, Chen J. An efficient marginal-return-based constructive heuristic to solve the sensor–weapon–target assignment problem. *IEEE Trans Syst Man Cybern Syst*, 2019, 49: 2536–2547
- 8 Lo V M. Heuristic algorithms for task assignment in distributed systems. *IEEE Trans Comput*, 1988, 37: 1384–1397
- 9 Bai W, Zhou Q, Li T, et al. Adaptive reinforcement learning neural network control for uncertain nonlinear system with input saturation. *IEEE Trans Cybern*, 2020, 50: 3433–3443
- 10 Gao A, Wang Q, Liang W, et al. Game combined multi-agent reinforcement learning approach for UAV assisted offloading. *IEEE Trans Veh Technol*, 2021, 70: 12888–12901
- 11 Zhao W, Meng Q, Chung P W H. A heuristic distributed task allocation method for multivehicle multitask problems and its application to search and rescue scenario. *IEEE Trans Cybern*, 2016, 46: 902–915
- 12 Lou J, Tang Z, Jia W. Energy-efficient joint task assignment and migration in data centers: A deep reinforcement learning approach. *IEEE Trans Netw Serv Manage*, 2023, 20: 961–973
- 13 Sahni J, Vidyarthi P. A cost-effective deadline-constrained dynamic scheduling algorithm for scientific workflows in a cloud environment. *IEEE Trans Cloud Comput*, 2018, 6: 2–18
- 14 Kachar K G, Gorodetsky A A. Dynamic multiagent assignment via discrete optimal transport. *IEEE Trans Control Netw Syst*, 2022, 9: 151–162
- 15 Vidal R, Shakernia O, Kim H J, et al. Probabilistic pursuit-evasion games: Theory, implementation, and experimental evaluation. *IEEE Trans Robot Automat*, 2002, 18: 662–669
- 16 Zhao W, Liu H, Lewis F L. Robust formation control for cooperative underactuated quadrotors via reinforcement learning. *IEEE Trans Neural Netw Learn Syst*, 2021, 32: 4577–4587
- 17 Lopez V G, Lewis F L, Wan Y, et al. Solutions for multiagent pursuit-evasion games on communication graphs: Finite-time capture and asymptotic behaviors. *IEEE Trans Automat Contr*, 2020, 65: 1911–1923
- 18 Liu H, Ma T, Lewis F L, et al. Robust formation control for multiple quadrotors with nonlinearities and disturbances. *IEEE Trans Cybern*, 2020, 50: 1362–1371
- 19 Liu H, Tian Y, Lewis F L, et al. Robust formation tracking control for multiple quadrotors under aggressive maneuvers. *Automatica*, 2019, 105: 179–185
- 20 Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 2012, 48: 2699–2704

Task assignment in multi-agent games via reinforcement learning

LI ShangHeng¹, LIU Hao^{1,2}, REN ZiMing³, LI YaFan⁴ & LIU DaWei⁴

¹ School of Artificial Intelligence, Beihang University, Beijing 100191, China

² Zhongguancun Laboratory, Beijing 100194, China

³ School of Astronautics, Beihang University, Beijing 100191, China

⁴ China Research and Development Academy of Machinery Equipment, Beijing 100089, China

This paper investigates the task assignment problem in multi-agent pursuit–evasion games under the influence of model nonlinear dynamics and external disturbances. An optimal task assignment value function is proposed that transforms the task allocation problem in games into a multiple pursuer trajectory tracking problem. An off-policy integral reinforcement learning method is proposed using input and output data from multi-agent systems. A neural network is introduced to fit the value function and derive the optimal strategy, and the nonlinear Hamilton–Jacobi–Bellman equation is solved iteratively. The optimal control policy and total cost of task execution are solved without the knowledge of the agent model. Simulation results verify the effectiveness of the proposed task assignment method.

multi-agent system, task assignment, optimal control, game, reinforcement learning

doi: [10.1360/SST-2024-0154](https://doi.org/10.1360/SST-2024-0154)