DOI:10.11918/202503020

强信息约束下的飞行器智能协同机动决策方法

丘沛桓,倪炜霖,吴志刚,梁海朝

(中山大学 航空航天学院,广东 深圳 518107)

摘 要:为实现高超声速飞行器在“目标-拦截者-防御者”多角色博弈场景下对拦截飞行器的逃逸,其需要与防御飞行器执行协同机动策略。然而,由于探测装置限制,高超声速飞行器面临非完美、非完备和非完整等强信息约束下的协同机动决策问题。为此,结合多智能体深度强化学习算法,提出了一种端到端协同机动决策方法,使高超声速飞行器能够在强信息约束下进行协同机动,进而成功逃逸。首先,将研究场景建模为分布式部分可观测马尔可夫决策过程,并提出一种观测信息共享堆叠机制,用于设计受强信息约束的局部观测状态空间。其次,针对多智能体强化学习稀疏奖励问题,构造一种结合博弈关系与零控脱靶量的多智能体合作决策奖励函数,提高多智能体系统在复杂博弈场景中的训练效率。最后,设计由基础智能体网络和顶层值分解网络构成的多智能体协同决策网络架构,能够从非完美、非完备和非完整信息中提取飞行器的时空轨迹特征,实现智能体系统的策略协调与飞行器的协同机动决策。结果表明,搭载所提出的智能协同机动决策方法的高超声速飞行器能够在强信息约束下的多角色博弈场景中成功逃逸,并在典型博弈场景与蒙特卡洛测试等数值仿真中展现了出色的效能和鲁棒性。

关键词:协同机动决策;高超声速飞行器;强化学习;部分可观测问题;多智能体

中图分类号:V11 **文献标志码**:A **文章编号**:0367-6234(2026)04-0011-12

Intelligent cooperative maneuver decision-making approach for vehicles under strong information constraints

QIU Peihuan, NI Weilin, WU Zhigang, LIANG Haizhao

(School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzhen 518107, Guangdong, China)

Abstract: To achieve the escape of a hypersonic vehicle from an interceptor in a multi-role game scenario of “target-interceptor-defender”, it is necessary to execute a cooperative maneuver strategy with the defender. However, due to the limitations of the detection device, hypersonic vehicles face the problem of cooperative maneuver decision-making with imperfect, incomplete, and intermittent strong information constraints. To address this, this paper proposed an end-to-end cooperative maneuver decision-making approach by integrating a multi-agent deep reinforcement learning algorithm, enabling hypersonic vehicles to make cooperative maneuver decisions under strong information constraints and achieve successful evasion. First, the research scenario was modeled as a decentralized partially observable Markov decision process, and an observation information sharing stacking mechanism was proposed for the design of local observation state spaces under the strong information constraints. Second, to address the sparse reward problem in multi-agent deep reinforcement learning, a cooperative decision-making reward function was constructed by integrating game relationships and zero-effort miss distance, enhancing training efficiency in complex game scenarios. Finally, a multi-agent cooperative decision-making network architecture was designed, comprising the agent's basic networks and the top value decomposition network. This architecture extracted spatio-temporal trajectory features from imperfect, incomplete, and intermittent information, enabling policy coordination among agents and cooperative maneuver decision-making for vehicles. Research results demonstrate that hypersonic vehicles equipped with the proposed intelligent cooperative maneuver decision-making approach can successfully evade in multi-role game scenarios under strong information constraints. The proposed approach exhibits outstanding performance and robustness in numerical simulations, including typical game scenarios and Monte Carlo tests.

Keywords: cooperative maneuver decision-making; hypersonic vehicle; reinforcement learning; partially observable problem; multi-agent

收稿日期:2025-03-05;**录用日期**:2025-05-06;**网络首发日期**:2026-01-29
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20260129.1033.002
基金项目:国家自然科学基金(62388101)
作者简介:丘沛桓(2001—),男,硕士研究生;梁海朝(1986—),教授,博士生导师
通信作者:梁海朝,lianghch5@mail.sysu.edu.cn

作为提高飞行器生存能力的关键手段,飞行器博弈技术在近年来备受关注。其中,基于飞行器主动性实现威胁规避的机动决策方法已成为研究的重点之一[1-2]。在“目标-拦截者-防御者”(target-interceptor-defender,TID)多角色博弈场景中,目标飞行器通过与防御飞行器的协同机动,可有效降低逃逸所需的机动能力要求,从而显著提高生存能力[3]。因此,开展飞行器协同机动决策方法研究具有重要的意义。

在传统的机动决策方法研究中,基于最优控制[4-5]与微分对策[6-7]的方法通过将追逃博弈问题转化为最优化问题进行求解。然而,这些方法通常依赖对手位置、速度和加速度等完备运动学信息的假设,并需要掌握对手机动过载能力和响应速度等精确的先验信息。事实上,飞行器通常无法准确获取对手的先验信息,并且受限于探测装置的能力,其仅能通过受观测噪声干扰的非完备运动学信息进行机动决策。

为解决上述非完美与非完备信息的问题,文献[8-9]提出了一种融合滤波估计器与传统最优解析算法的混合方法。该方法首先利用滤波估计器预估对对手的运动状态进行预估,然后再利用传统最优解析方法求解机动策略。该方法能够在预设范围内有效应对对手拦截制导策略,并在一定程度上减轻非完备信息的影响。然而,由于滤波估计器对初始参数高度敏感,在对手采用非预设拦截制导方法时,其决策可靠性显著下降,在复杂博弈场景难以确保目标飞行器的成功逃逸。针对传统方法的局限性,深度强化学习(deep reinforcement learning,DRL)算法开始被应用于飞行器博弈场景[10-12]。DRL算法通过智能体与环境的交互,利用奖励反馈优化策略,使其能够应对动态决策场景,进而输出可信机动决策。然而,上述研究仅将目标飞行器和防御飞行器作为单一智能体进行强化学习机动决策方法设计,未充分考虑二者在任务角色、机动能力,以及探测数据等方面的差异,限制了其在多角色博弈中的独立决策能力。

考虑到单智能体强化学习方法的局限性,一些研究人员通过多智能体强化学习(multi-agent reinforcement learning,MARL)算法进行机动决策方法的设计。针对无人机协同问题,文献[13]采用多智能体深度确定性策略梯度算法设计了一种多无人机协同追捕决策方法,实现对快速机动目标的协同围捕。针对多航天器博弈场景,文献[14]在非完备信息的约束下基于MARL提出了一种多航天器轨道博弈的决策方法,显著提高了博弈成功率并降低了燃料损失。在飞行器拦截技术的研究中,文献[15]提出了一种基于近端策略优化方法的MARL协同拦截方法,实现飞行器在不同拦截决策场景的自主学习与任务分配。尽管上述基于MARL的方法在其研究场景中取得了出色的效果,但均未考虑探测信息的非完整性问题。由于飞行器机动过程中探测装置的视线角方向随飞行器姿态变化,拦截飞行器可能超出视场范围,导致探测信息时序上的缺失。上述探测信息的非完美、非完备和非完整特性为基于MARL的协同机动决策方法设计带来更大的挑战。

因此,非完美、非完备和非完整等信息约束的处理成为多智能体强化学习应用于协同机动决策方法并实现飞行器在TID场景下成功逃逸的关键。为应对该问题,本文以高超声速飞行器为研究对象,提出了高超声速飞行器值分解式多智能体强化学习(hypersonic vehicle value decomposition MARL,HV^2D)算法,并基于此算法设计了一种端到端协同机动决策方法。首先,为建立多智能体合作任务型系统,将TID多角色博弈场景建模为分布式部分可观测马尔可夫决策过程,并通过观测信息共享堆叠机制进行强信息约束下的局部观测状态空间设计。其次,构造了一种结合博弈关系与零控脱靶量的多智能体合作决策奖励函数。然后,设计由基础智能体网络和顶层值分解网络组成的多智能体协同决策网络架构。最后,基于HV^2D的协同机动决策方法以非完美、非完备和非完整信息为输入,输出目标飞行器和防御飞行器的可信决策指令,在强信息约束下的多角色博弈场景中,实现目标飞行器的成功逃逸,规避拦截飞行器的威胁。

1 问题描述

1.1 TID多角色博弈场景

考虑飞行器拦截末段的TID多角色博弈场景,一枚高超声速目标飞行器(T)感知到高机动性拦截飞行器(I)的对向高速接近时,释放防御飞行器(D)进行协同机动,以规避拦截威胁。各飞行器在博弈场景中的空间关系如图1所示,图1中:$i=\{\mathrm{T},\mathrm{I},\mathrm{D}\}$,$OXYZ$坐标系为北-天-东惯性坐标系,$x_i$、$y_i$、$z_i$为各飞行器在各坐标轴上的具体位置,$V_i$为各飞行器的速度,$\rho_{\mathrm{IT}}$、$\rho_{\mathrm{ID}}$分别为拦截飞行器与目标飞行器、防御飞行器的相对空间距离,θ_i、φ_i分别为各飞行器的飞行路径角和飞行航向角,LOS ID、LOS IT分别为飞行器间的视线方向(line of sight,LOS),θ^j_{LOS}、φ^j_{LOS},

($j=\{\mathrm{IT},\mathrm{ID}\}$)为各 LOS 的俯仰角和方位角，FOV 为飞行器的探测装置视场角(field of view，FOV)。

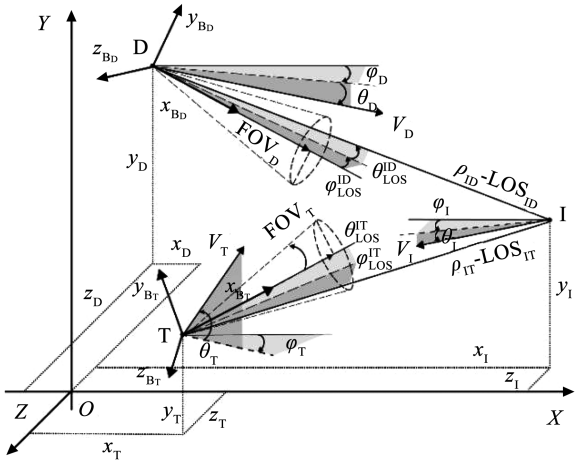


图 1　多角色博弈场景几何关系

Fig. 1　Geometric relationships in multi-role game scenario

本文在拦截末段场景中，飞行器间相对速度高，博弈时间短，可忽略地球曲率及其自转影响，飞行器运动方程如下：

$$\begin{cases}\dot{x}=V\cdot\cos\theta\cdot\cos\varphi\\ \dot{y}=V\cdot\sin\theta\\ \dot{z}=-V\cdot\cos\theta\cdot\sin\varphi\\ \dot{V}=-(D+mg\cdot\sin\theta)/m\\ \dot{\theta}=(L\cdot\cos\gamma-mg\cdot\cos\theta)/(mV)\\ \dot{\varphi}=L\cdot\sin\gamma/(mV\cdot\cos\theta)\end{cases}\tag{1}$$

式中：g 为地球的重力加速度，γ 为倾侧角，m 为飞行器的质量，L、D 分别为飞行器的升力和阻力。

1.2　场景信息约束

在 TID 博弈场景中，强信息约束问题主要包括非完美、非完备和非完整等信息约束。

首先，由于拦截飞行器的机动能力和制导策略未知，并且探测信息受噪声干扰，目标飞行器和防御飞行器难以获得完美的拦截飞行器机动特征与运动状态信息，从而形成非完美信息问题。

同时，受探测装置的能力限制，目标飞行器与防御飞行器仅能获取与拦截飞行器的相对距离 ρ、视线俯仰角和视线方位角 θ_{LOS}，φ_{LOS} 等信息，造成了非完备探测信息问题[16]，由各飞行器间的几何关系可得

$$\begin{cases}\rho_{\mathrm{I}i}=\sqrt{(x_{\mathrm{I}}^{Li}-x_i^{Li})^2+(y_{\mathrm{I}}^{Li}-y_i^{Li})^2+(z_{\mathrm{I}}^{Li}-z_i^{Li})^2}\\ \theta_{\mathrm{LOS}}^{Li}=\arctan\left[\dfrac{y_{\mathrm{I}}^{Li}-y_i^{Li}}{\sqrt{(x_{\mathrm{I}}^{Li}-x_i^{Li})^2+(z_{\mathrm{I}}^{Li}-z_i^{Li})^2}}\right]\\ \varphi_{\mathrm{LOS}}^{Li}=\arctan\left(\dfrac{z_{\mathrm{I}}^{Li}-z_i^{Li}}{x_{\mathrm{I}}^{Li}-x_i^{Li}}\right)\end{cases}\tag{2}$$

式中：$i=\{\mathrm{T},\mathrm{D}\}$，$j=\{\mathrm{T},\mathrm{I},\mathrm{D}\}$，$x_j^{Li}$，$y_j^{Li}$，$z_j^{Li}$ 为各飞行器的位置矢量在目标飞行器和防御飞行器的观测坐标系中沿各轴的分量，可通过下式计算：

$$[x_j^{Li},y_j^{Li},z_j^{Li}]^{\mathrm{T}}=\boldsymbol{\Lambda}_i^{\mathrm{GL}}\cdot[x_j,y_j,z_j]^{\mathrm{T}}\tag{3}$$

式中：$i=\{\mathrm{T},\mathrm{D}\}$，$j=\{\mathrm{T},\mathrm{I},\mathrm{D}\}$，$\boldsymbol{\Lambda}_i^{\mathrm{GL}}$ 为目标飞行器和防御飞行器对应的惯性坐标系与观测坐标系间的坐标变换矩阵。

此外，飞行器在博弈过程中不断机动，其姿态的变化使探测装置的视场方向随之改变。当拦截飞行器超出目标飞行器和防御飞行器的视场范围时，即 $\theta_{\mathrm{LOS}}>\mathrm{FOV}$、$\varphi_{\mathrm{LOS}}>\mathrm{FOV}$，目标或防御飞行器的探测装置无法获取相关信息，使探测数据在时序上出现缺失，造成非完整探测信息问题[17]。

1.3　问题表述

基于上述场景与约束条件，所研究的问题可表述为：在非完美、非完备和非完整等信息约束下的 TID 多角色博弈场景中，设计一种适应于高超声速飞行器的多智能体强化学习协同机动决策方法，引导目标飞行器与防御飞行器实现协同机动，确保目标飞行器规避拦截飞行器的威胁并实现成功逃逸。

2　多智能体合作任务型系统搭建

2.1　分布式部分可观测马尔可夫决策过程构建

本文通过将 TID 多角色博弈场景建模为分布式部分可观测马尔可夫决策过程(decentralized partially observable MDP，Dec-POMDP)[18]，以求解多智能体系统的全局最优决策，从而解决所研究的飞行器机动决策问题。

Dec-POMDP 可表示为一个 8 元组 $\{\boldsymbol{N},\boldsymbol{S},\boldsymbol{U},\boldsymbol{O},P,Z,R,\gamma\}$。其中：$\boldsymbol{N}=\{1,\cdots,n\}$ 为智能体的集合，$\boldsymbol{S}$ 为所有全局环境状态集合，其中 $\boldsymbol{s}_t\in\boldsymbol{S}$ 为当前时刻 t 的全局环境状态；智能体 $i\in\{1,\cdots,n\}$ 可根据策略 $\boldsymbol{\pi}^i$ 选择动作 $\boldsymbol{u}_t^i$，$\boldsymbol{u}_t=(\boldsymbol{u}_t^1,\cdots,\boldsymbol{u}_t^n)\in\boldsymbol{U}$ 为所有智能体的联合动作，$\boldsymbol{U}$ 为所有动作集合；$\boldsymbol{O}$ 为所有局部观测状态集合，其中 $\boldsymbol{o}_t=(\boldsymbol{o}_t^1,\cdots,\boldsymbol{o}_t^n)\in\boldsymbol{O}$ 为所有智能体的局部观测状态 $\boldsymbol{o}_t^i$ 组成联合观测状态集合；$P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{u}_t)$、$Z(\boldsymbol{o}_t|\boldsymbol{s}_{t+1},\boldsymbol{u}_t)$ 分别为全局环境状态转移函数与局部观测状态转移函数，$R(\boldsymbol{s}_t,\boldsymbol{u}_t)$ 为多智能体合作任务型系统中所有智能体共享的奖励函数，$\gamma\in[0,1)$ 为折扣因子。

在 Dec-POMDP 中，智能体只能局部观测当前真实状态 $\boldsymbol{s}_t$，因此引入联合信念状态 $\boldsymbol{b}_t(\boldsymbol{s}_t)=\prod_{i=1}^{n}\boldsymbol{b}_t^i(\boldsymbol{s}_t)$ 来表示所有智能体对全局环境状态 $\boldsymbol{s}_t$ 的联合概率分布，其中 $\boldsymbol{b}_t^i(\boldsymbol{s}_t)$ 为智能体 i 在时刻 t 对全局环境状态 $\boldsymbol{s}_t$ 的局部信念状态。此外，引入动作-

观测历史 $\boldsymbol{\tau}_t^i=(\boldsymbol{u}_0^i,\boldsymbol{o}_0^i,\cdots,\boldsymbol{u}_t^i,\boldsymbol{o}_t^i)$ 以描述智能体与环境交互的动作和局部观测历史,其与联合信念状态的关系如下:

$$\boldsymbol{b}_t^i(\boldsymbol{s}_t)=P(\boldsymbol{s}_t\mid\boldsymbol{\tau}_t^i)=P(\boldsymbol{s}_t\mid\boldsymbol{u}_0^i,\boldsymbol{o}_0^i,\cdots,\boldsymbol{u}_t^i,\boldsymbol{o}_t^i)\tag{4}$$

在 Dec-POMDP 中,最优策略函数 V^* 的贝尔曼最优方程形式如下:

$$V^*=\max_{\boldsymbol{u}_t}\Big[\sum\boldsymbol{b}(\boldsymbol{s}_t)R(\boldsymbol{s}_t,\boldsymbol{u}_t)+\gamma\sum P(\boldsymbol{o}_t\mid\boldsymbol{b}_t,\boldsymbol{u}_t)\sum P(\boldsymbol{s}_{t+1}\mid\boldsymbol{s}_t,\boldsymbol{u}_t,\boldsymbol{o}_t)V^*(\boldsymbol{b}_{t+1})\Big]\tag{5}$$

为使多智能体合作任务型系统能够准确反映博弈环境的动态变化,将研究场景的全局环境状态空间定义为:

$$\boldsymbol{S}=[t,\bar{\boldsymbol{s}}_{\mathrm{T}},\bar{\boldsymbol{s}}_{\mathrm{I}},\bar{\boldsymbol{s}}_{\mathrm{D}}]^{\mathrm{T}}\tag{6}$$

$$\bar{\boldsymbol{s}}_i=[\boldsymbol{x}_i,\boldsymbol{v}_i,\boldsymbol{a}_i]^{\mathrm{T}}\tag{7}$$

式中:$i=\{\mathrm{T},\mathrm{I},\mathrm{D}\}$,$t$ 为时间,$\bar{\boldsymbol{s}}_{\mathrm{T}}$、$\bar{\boldsymbol{s}}_{\mathrm{I}}$、$\bar{\boldsymbol{s}}_{\mathrm{D}}$ 分别为目标、防御和拦截飞行器的运动学信息,$\boldsymbol{x}$、$\boldsymbol{v}$、$\boldsymbol{a}$ 分别为位置、速度和加速度信息。

在协同机动时,目标飞行器利用攻角和倾侧角进行机动,而防御飞行器采取直接力进行机动,因此动作空间设计为

$$\boldsymbol{U}=[\bar{\alpha},\bar{\gamma},\bar{u}_{\mathrm{D}}^{y},\bar{u}_{\mathrm{D}}^{z}]^{\mathrm{T}}\tag{8}$$

式中:$\bar{\alpha}$、$\bar{\gamma}$分别为目标飞行器的攻角和倾侧角的期望信号,$u_{\mathrm{D}}^i(i=\{y,z\})$ 为防御飞行器在其体坐标系对应方向过载的期望信号。

2.2 局部观测状态空间设计

局部观测状态空间是 Dec-POMDP 系统中各智能体进行决策的直接信息来源,决定了智能体对环境感知的准确性,因此局部观测状态空间的设计需要能够反映各博弈对象的动态特征。然而,在非完美、非完备和非完整等强约束信息下,局部观测状态空间的设计面临运动学信息不完备和时序信息不完整等问题,使得智能体难以准确地感知博弈环境的动态变化。

为解决上述问题,结合飞行器状态信息在时空维度的高度耦合性,提出一种观测信息共享堆叠机制。该机制通过飞行器探测信息的共享以及智能体局部观测状态的堆叠,构建出二维局部观测状态空间,从而在强信息约束下提高智能体的环境感知能力,具体表达式如下:

$$\boldsymbol{O}^i\triangleq\{\boldsymbol{o}_t^i,\boldsymbol{o}_{t-1}^i,\cdots,\boldsymbol{o}_{t-m}^i\}\in\boldsymbol{O}\tag{9}$$

$$\boldsymbol{o}_t^i=[t,\bar{\boldsymbol{s}}_t^i,\boldsymbol{x}_t^i,\boldsymbol{\Theta}(\boldsymbol{o}_t^{\mathrm{I}i})]\tag{10}$$

式中:$\boldsymbol{o}_t^i$ 为智能体 i 在 t 时刻的局部观测信息簇,通过时空维度的 $m+1$ 层堆叠得到智能体 i 的局部观测状态空间 $\boldsymbol{O}^i$;$\bar{\boldsymbol{s}}_t^i$ 为智能体 i 的运动学信息;$\boldsymbol{x}_t^i$ 为通过共享得到的除智能体 i 外的位置信息;$\boldsymbol{\Theta}(\cdot)$ 为根据式(2)得出的变换函数;$\boldsymbol{o}_t^{\mathrm{I}i}$ 为探测装置获取的拦截飞行器信息,由所有智能体进行共享,表达式如下:

$$\boldsymbol{o}^{\mathrm{I}i}=\bar{\boldsymbol{o}}^{\mathrm{I}i}+\boldsymbol{\omega}=[\rho^{\mathrm{I}i}\ \theta_{\mathrm{LOS}}^{\mathrm{I}i}\ \varphi_{\mathrm{LOS}}^{\mathrm{I}i}]^{\mathrm{T}}+[\delta\rho\ \delta\theta\ \delta\varphi]^{\mathrm{T}}\tag{11}$$

式中:$\bar{\boldsymbol{o}}^{\mathrm{I}i}$为拦截飞行器的真实信息,$\boldsymbol{\omega}\sim\boldsymbol{U}(\boldsymbol{0}_3,\boldsymbol{\Sigma})\in\mathbf{R}^3$ 为探测装置的观测噪声,其中 $\boldsymbol{\Sigma}=[\sigma_\rho\ \sigma_{\mathrm{LOS}}\ \sigma_{\mathrm{LOS}}]^{\mathrm{T}}$ 为噪声振幅(σ_ρ/m,σ_{LOS}/mrad)。当出现非完整信息问题时,式(11)可变为

$$\boldsymbol{o}^{\mathrm{I}i}=[\rho_{t_{\mathrm{Miss}}}\ \theta_{t_{\mathrm{Miss}}}\ \varphi_{t_{\mathrm{Miss}}}]^{\mathrm{T}}+[\delta\rho\ \delta\theta\ \delta\varphi]^{\mathrm{T}}\tag{12}$$

式中,t_{Miss}为失去拦截飞行器信息的时间,此时 $\bar{\boldsymbol{o}}^{\mathrm{I}i}$保持最后一次获取拦截飞行器信息时的状态。

2.3 多智能体合作决策奖励函数

针对多智能体强化学习的稀疏奖励问题[19],构造了多智能体合作决策奖励函数 $R(t)$,以提高智能体系统在高动态和强对抗的 TID 博弈场景中的训练效率。该奖励函数由终端奖励与过程连续奖励组成,即

$$R(t)=R_{\mathrm{s}}(t)+R_{\mathrm{c}}(t)\tag{13}$$

首先,根据 TID 场景中各飞行器的博弈关系,构造终端奖励函数 $R_{\mathrm{s}}(t)$ 如下:

$$R_{\mathrm{s}}(t)=\begin{cases}r_{\mathrm{s}1}, & t=t_{\mathrm{ID}}^{\mathrm{f}}\ \text{and I is Damaged}\\ -r_{\mathrm{s}2}, & t=t_{\mathrm{IT}}^{\mathrm{f}}\ \text{and T is Damaged}\\ r_{\mathrm{s}3}, & t=t_{\mathrm{IT}}^{\mathrm{f}}\ \text{and T is Alive}\end{cases}\tag{14}$$

式中:$t_{\mathrm{ID}}^{\mathrm{f}}$为对应飞行器的相遇时间;$R_{\mathrm{s}}(t)$ 为拦截飞行器和防御飞行器相遇时,若拦截飞行器被防御飞行器成功摧毁,所有智能体将获得正奖励 $r_{\mathrm{s}1}$;当拦截飞行器和目标飞行器相遇时,若拦截飞行器成功摧毁目标飞行器,所有智能体将获得负奖励 $-r_{\mathrm{s}2}$;反之,目标飞行器成功逃逸,所有智能体将获得正奖励 $r_{\mathrm{s}3}$。

然后,利用零控脱靶量的时序连续性及其对飞行器空间位置的表征能力[6],构造博弈过程中的连续奖励函数 $R_{\mathrm{c}}(t)$ 为

$$R_{\mathrm{c}}(t)=r_{\mathrm{c}1}(t)+r_{\mathrm{c}2}(t)\tag{15}$$

式中:$R_{\mathrm{c}}(t)$将随飞行器机动决策而连续变化;$r_{\mathrm{c}1}(t)$可促使智能体控制防御飞行器缩小与拦截飞行器的零控脱靶量,引导防御飞行器对拦截飞行器进行防御,以保护目标飞行器;$r_{\mathrm{c}2}(t)$可促使智能体控制目标飞行器增大与拦截飞行器的零控脱靶量,引导目标飞行器逃逸。具体表达式如下:

$$r_{\mathrm{c}1}(t)=A_1\cdot\exp\{[B_1\cdot Z_{\mathrm{ID}}(t)/R_{\mathrm{K,ID}}]^{C_1}\}\tag{16}$$

$$r_{\mathrm{c}2}(t)=A_2\cdot\log\{B_2\cdot[Z_{\mathrm{IT}}(t)/R_{\mathrm{K,IT}}]^{C_2}\}\tag{17}$$

式中:γ 为马尔可夫决策过程中的折扣因子,$R_{\mathrm{K},j}$为飞行器间的拦截毁伤半径,A_i、B_i、C_i 为超参数($i=$

$\{1,2\}$），各超参数使得奖励函数 $r_c(t)$ 能够保证各智能体在博弈过程中获取平滑的奖励反馈，$Z_j=\sqrt{(Z_j^y)^2+(Z_j^z)^2}$，$j=\{\mathrm{IT},\mathrm{ID}\}$，其中 Z_j^y、Z_j^z 分别为 OXY 平面与 OXZ 平面下对应飞行器间的零控脱靶量，其表达式如下：

$$Z_{\mathrm{I}i}^{\mathrm{K}}=\mathrm{K}_{\mathrm{I}i}+t_{\mathrm{I}i}^{\mathrm{go}}\dot{\mathrm{K}}_{\mathrm{I}i}+a_{\mathrm{I}}^{\mathrm{K}}\tau_{\mathrm{I}}^2\Gamma\left(\frac{t_{\mathrm{I}i}^{\mathrm{go}}}{\tau_{\mathrm{I}}}\right)-a_i^{\mathrm{K}}\tau_i^2\Gamma\left(\frac{t_{\mathrm{I}i}^{\mathrm{go}}}{\tau_i}\right) \tag{18}$$

其中：

$$t_{\mathrm{I}i}^{\mathrm{go}}=\rho_{\mathrm{I}i}^0/(V_{\mathrm{I}}+V_i)-t$$

$$\Gamma(\varpi)=\mathrm{e}^{-\varpi}+\varpi-1$$

式中：$i=\{\mathrm{T},\mathrm{D}\}$，$j=\{\mathrm{T},\mathrm{I},\mathrm{D}\}$，$\mathrm{K}=\{y,z\}$，$\tau_j$ 为各飞行器的时间响应常数。

3　机动决策方法设计

3.1　多智能体协同决策网络架构

在所建立的多智能体合作任务型系统中，所有智能体共享系统奖励，为确保单智能体利益与整体系统利益的一致性，本文提出的多智能体强化学习算法 HV²D 采用值函数分解机制量化各智能体对整体系统利益的贡献，以处理目标飞行器与防御飞行器之间的合作博弈关系。HV²D 由基础智能体网络和顶层值分解网络组成。

3.1.1　基础智能体网络设计

结合卷积神经网络（convolutional neural network，CNN）与双深度决斗循环 Q 网络（dueling double deep recurrent Q network，D³RQN）架构，基础智能体网络通过飞行器时空博弈特征的提取，增强智能体在强信息约束下的决策能力。

首先，为应对非完美信息约束，CNN 网络对具有信息非完备特性的智能体 i 的局部观测空间 O^i 进行特征提取，生成蕴含飞行器机动特征的一维张量：

$$\bar{\boldsymbol{z}}_t^i=\mathrm{Conv}(\boldsymbol{O}_t^{i,\,\mathrm{stak}})=\boldsymbol{W}\otimes\boldsymbol{O}_t^{i,\mathrm{stak}}+\boldsymbol{b} \tag{19}$$

$$\boldsymbol{z}_t^i=f(\bar{\boldsymbol{z}}_t^i)=f(\mathrm{Conv}(\boldsymbol{O}_t^{i,\,\mathrm{stak}})) \tag{20}$$

式中：$f(\cdot)$ 为 ReLU 激活函数，$\mathrm{Conv}(\cdot)$ 为 CNN 卷积过程。

随后，D³RQN 结合机动特征张量 $\boldsymbol{z}_t^i$ 和具有飞行器轨迹时序特性的历史隐藏状态 $\boldsymbol{h}_{t-1}^i$，对时序上存在缺失的非完整信息进行数据推理，得到当前时刻的隐藏轨迹状态 $\boldsymbol{h}_t^i$，并对当前智能体策略进行评估，输出智能体 i 的局部动作–价值函数 Q^i。基于 ε – 贪婪策略，智能体选择相应的飞行器机动决策 $\boldsymbol{u}_t^i$，并更新环境状态：

$$\boldsymbol{h}_t^i=\mathrm{GRU}(\tanh(\sigma(\boldsymbol{z}_t^i,\boldsymbol{h}_{t-1}^i))) \tag{21}$$

$$\boldsymbol{u}^{i,\,*}=\underset{\boldsymbol{u}^i\in U}{\mathrm{argmax}}\,Q^i(\boldsymbol{h}_t^i,\boldsymbol{U};p,q) \tag{22}$$

$$Q^i(\boldsymbol{h}_t^i,\boldsymbol{u}_t^i;p^i,q^i)=\boldsymbol{V}(\boldsymbol{h}_t^i;p^i)+\left[\boldsymbol{\Lambda}(\boldsymbol{h}_t^i;q^i)-\frac{1}{\mathrm{len}(\boldsymbol{U})}\sum_{\boldsymbol{u}_t^i\in\boldsymbol{U}}\boldsymbol{\Lambda}(\boldsymbol{h}_t^i,\boldsymbol{u}_t^i;q^i)\right] \tag{23}$$

式中：$*$ 为哈达马乘积，p、q 分别为状态价值函数 $\boldsymbol{V}$ 和优势函数 $\boldsymbol{\Lambda}$ 的网络参数，$\mathrm{len}(\boldsymbol{U})$ 为多智能体动作空间的维度。

通过上述设计，基于 CNN-D³RQN 架构的基础智能体网络如图 2 所示。

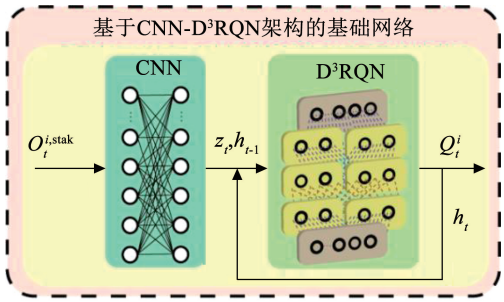


图 2　基础智能体网络结构示意

Fig. 2　Basic network structure diagram of agents

3.1.2　顶层值分解网络设计

为解决高超声速飞行器复杂博弈场景中各智能体间的协调合作问题，并确保各智能体能够共同优化整体系统目标，设计了顶层值分解网络以保持单智能体策略与整体系统策略的一致性。顶层值分解网络由线性网络和绝对激活函数构成的超网络[20]组成，将各智能体的局部动作–价值函数 Q^i 作为输入，结合全局环境状态 $\boldsymbol{S}$ 生成顶层网络的权重 $\boldsymbol{W}$ 和偏置 $\boldsymbol{B}$。最终，通过这些参数对符合个人全局最大约束条件[21]的全局动作–价值函数进行分解，实现对各智能体的信用分配：

$$\underset{\boldsymbol{u}_t}{\mathrm{argmax}}\,Q_{\mathrm{tot}}(\boldsymbol{h}_t,\boldsymbol{u}_t)=\begin{pmatrix}\underset{\boldsymbol{u}_t^1}{\mathrm{argmax}}\,Q^1(\boldsymbol{h}_t^1,\boldsymbol{u}_t^1)\\ \vdots\\ \underset{\boldsymbol{u}_t^n}{\mathrm{argmax}}\,Q^n(\boldsymbol{h}_t^n,\boldsymbol{u}_t^n)\end{pmatrix} \tag{24}$$

$$\partial Q_{\mathrm{tot}}/\partial Q^i\geqslant 0,\quad\forall\, i\in\mathrm{Agent} \tag{25}$$

式中，全局动作–价值函数用于衡量多智能体系统在执行各智能体的局部策略时的期望总回报。

综上所述，引导智能体协的同决策的智能体顶层网络结构设计如图 3 所示。

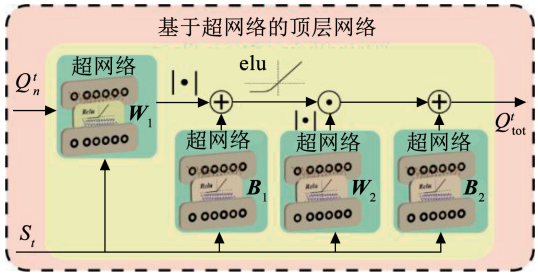


图 3　智能体顶层网络结构示意

Fig. 3　Top network structure diagram of agents

3.2 基于 HV²D 的端到端协同机动决策方法

基于 HV²D 算法的端到端协同机动决策方法采用集中式训练分布式执行的运行框架。在训练阶段，利用局部观测状态空间与全局环境状态等输入，智能体系统结合基础智能体网络和顶层值分解网络共同优化多智能体系统的策略；在执行阶段，飞行器仅需根据自身的局部观测状态空间，采用自身基础智能体网络作出决策。

3.2.1 集中训练阶段

在训练阶段，为适配所提出的多智能体协同决策网络架构，结合飞行器时空博弈特征，将 HV²D 算法的记忆回放池扩展为 $\{\boldsymbol{S}_t, \boldsymbol{O}_t^{\text{stak}}, \boldsymbol{u}_t, \boldsymbol{h}_t, R_t, \boldsymbol{S}_{t+1}, \boldsymbol{O}_{t+1}^{\text{stak}}\}$，其中：$\boldsymbol{S}_t$ 为当前时刻全局环境状态，$\boldsymbol{O}_t$ 为当前时刻的局部观测状态空间，$\boldsymbol{u}_t$ 为当前时刻联合动作张量，$\boldsymbol{h}_t$ 为当前时刻联合历史隐藏状态，R_{t+1} 为当前时刻奖励，$\boldsymbol{S}_{t+1}$ 为下一时刻全局环境状态，$\boldsymbol{O}_{t+1}$ 为下一时刻的局部观测状态空间。在训练过程中，从经验回放区中采样一批多智能体系统的记忆数据 $\{\bar{\boldsymbol{S}}_t, \bar{\boldsymbol{O}}_t^{\text{stak}}, \bar{\boldsymbol{u}}_t, \bar{\boldsymbol{h}}_t, \bar{r}_{t+1}, \bar{\boldsymbol{S}}_{t+1}, \bar{\boldsymbol{O}}_{t+1}^{\text{stak}}\}$，由各智能体 $i \in [1, \cdots, n]$ 通过基础网络计算当前与目标局部动作-价值函数 $Q_j^i(\bar{\boldsymbol{h}}^i, \bar{\boldsymbol{u}}^i; \alpha_j^i, \beta_j^i)$。其中：$\alpha_j^i$、$\beta_j^i$ 为对应网络的全连接层参数，$j = \{\text{cur}, \text{tag}\}$，$\bar{\boldsymbol{u}}^{i,*}$ 为当前局部网络的最优动作，表达式如下：

$$\bar{\boldsymbol{u}}_t^{i,*} = \underset{\bar{\boldsymbol{u}}_t}{\operatorname{argmax}}\, Q_{\text{cur}}^i(\bar{\boldsymbol{h}}_t^i, \bar{\boldsymbol{u}}_t^i; \alpha_{\text{cur}}^i, \beta_{\text{cur}}^i) \tag{26}$$

在顶层值分解网络处理所有智能体的局部动作-价值函数后，通过 Huber 损失计算全局动作-价值函数的预测与目标值间的误差分别为：

$$\boldsymbol{y} = \bar{r} + \gamma \underset{\bar{\boldsymbol{u}}^*}{\operatorname{argmax}}\, Q_{\text{tot}}(\bar{\boldsymbol{h}}_{t+1}, \bar{\boldsymbol{u}}_{t+1}^*, \bar{\boldsymbol{S}}_{t+1}; \alpha_{\text{tag}}, \beta_{\text{tag}}) \tag{27}$$

$$\text{Loss} = L_\delta[\boldsymbol{y}, Q_{\text{tot}}(\bar{\boldsymbol{h}}_t, \bar{\boldsymbol{u}}_t, \bar{\boldsymbol{S}}_t; \alpha_{\text{cur}}, \beta_{\text{cur}})] \tag{28}$$

式中，下一时刻的联合历史隐藏状态 $\bar{\boldsymbol{h}}_{t+1} = [\bar{\boldsymbol{h}}_{t+1}^1, \bar{\boldsymbol{h}}_{t+1}^2, \cdots, \bar{\boldsymbol{h}}_{t+1}^n]$ 可由下式得到：

$$\bar{\boldsymbol{h}}_{t+1}^i = \text{D}^3\text{RQN}(\text{CNN}(\bar{\boldsymbol{O}}_{t+1}^i), \bar{\boldsymbol{h}}_t^i) \tag{29}$$

基于损失 Loss，利用梯度下降法调整当前网络的所有参数 $[\boldsymbol{\theta}_t^{\text{tot}}, \boldsymbol{\theta}_t^1, \cdots, \boldsymbol{\theta}_t^n] \in \boldsymbol{\theta}_t$，即

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \cdot \text{Loss} \cdot \nabla_{\boldsymbol{\theta}_t} Q_{\text{cur}}(\bar{\boldsymbol{h}}_t, \bar{\boldsymbol{u}}_t, \bar{\boldsymbol{S}}_t; \boldsymbol{\theta}_t) \tag{30}$$

式中 η 为学习率。最后，对目标网络的所有参数 $\boldsymbol{\theta}'_{t+1}$ 进行软更新为

$$\boldsymbol{\theta}'_{t+1} = \boldsymbol{\theta}'_t + \kappa(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}'_t) \tag{31}$$

式中 κ 为软更新系数。

综上所述，在训练阶段，HV²D 算法通过经验采样与网络更新，优化多智能体系统的联合策略，其训练流程如图 4 所示，训练伪代码如算法 1 所示。

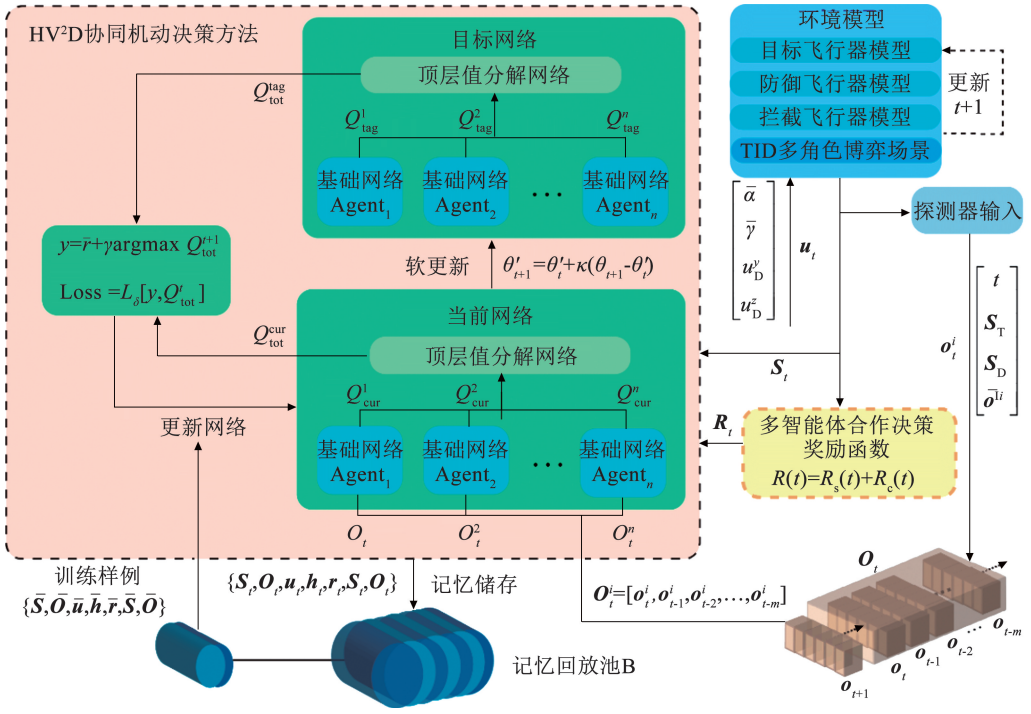


图 4 基于 HV²D 算法的协同机动决策方法的训练框图

Fig. 4 Training block diagram of cooperative maneuver decision-making approach based on HV²D algorithm

算法 1　高超声速飞行器值分解多智能体强化学习算法

1. 通过随机参数 $\boldsymbol{\theta}$ 初始化智能体 $i\in[1,\cdots,n]$ 的当前顶层网络 $Q_{\text{tot}}^{\text{cur}}$ 和当前基础网络 Q_{cur}^{i}，并复制到目标网络 $Q_{\text{tot}}^{\text{tag}}$，$Q_{\text{tag}}^{i}$。

2. 初始化记忆回放池 **B** 与 ε −贪婪策略搜索。

3. for（episode = 1；$\text{episode}_{\max}$；episode + = 1）：

4. 重置环境并获取初始全局状态空间 $\boldsymbol{s}_0$ 和每个智能体的初始局部观测空间 $\boldsymbol{o}_0^i$。

5. for（$t=1;t_{\text{IT}}^{\text{f}};t+=\text{d}t$）：

6. for（$i=1;n;i+=1$）：

7. 输入每个智能体的局部观测空间 $\boldsymbol{O}^i$，并使用 CNN 网络提取状态特征 $\boldsymbol{z}_t^i$：

8. 通过 D^3RQN 网络处理特征 $\boldsymbol{z}_t^i$ 和上一时刻输出张量 $\boldsymbol{h}_{t-1}^i$，得到当前时刻输出张量 $\boldsymbol{h}_t^i$。

9. 各智能体基础网络通过 $\boldsymbol{h}_t^i$ 获取局部 Q 并使用 $\boldsymbol{\varepsilon}$ − 贪婪策略选择动作 $\boldsymbol{u}_t^i$。

10. End for

11. 各智能体收到奖励 $\boldsymbol{r}_t$ 后，环境接收到动作 $\boldsymbol{u}_t$ 更新为 $\boldsymbol{s}_{t+1}$，并更新环境数据至记忆回放池 **B**。

12. End for

13. 从记忆回放池 **B** 中随机抽取采样一批数据 $\{\bar{\boldsymbol{S}}_t,\bar{\boldsymbol{O}}_t,\bar{\boldsymbol{u}}_t,\bar{\boldsymbol{h}}_t,\bar{r}_{t+1},\bar{\boldsymbol{S}}_{t+1},\bar{\boldsymbol{O}}_{t+1}\}$ 用于智能体训练。

14. 通过基础网络计算 Q_{cur}^{i} 和 Q_{tag}^{i}。

15. 使用顶层网络对所有智能体的局部 Q 函数 Q^i 进行评价。

16. 使用 Huber 算法计算损失值。

17. 使用梯度下降法更新所有当前网络的参数，最小化损失。

18. 通过软更新方式将所有目标网络进行更新。

19. End for

3.2.2　分布执行阶段

在完成集中训练后，基于 HV^2D 算法的端到端协同机动决策方法进入分布执行阶段。在该阶段，将基础智能体网络部署至对应的飞行器上，各飞行器通过探测装置获取非完美、非完备和非完整信息，由基础智能体网络独立完成可信协同决策。通过智能体网络对输入信息的处理，端到端输出飞行器可信机动决策 $\boldsymbol{U}_t$，从而引导目标飞行器与防御飞行器实现协同机动。分布执行阶段的协同机动决策方法运行流程如图 5 所示。

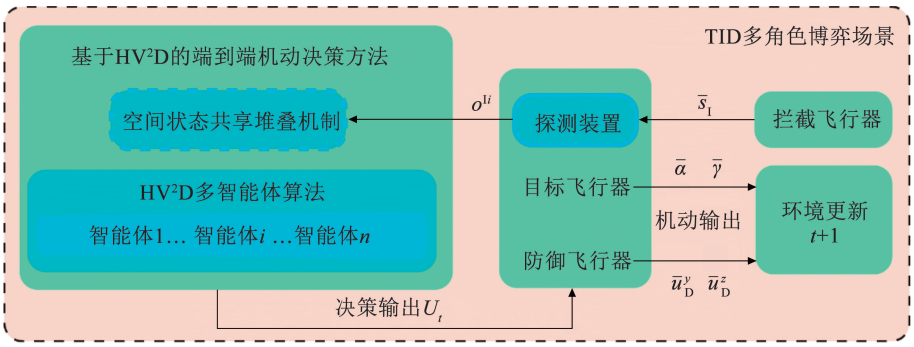


图 5　基于 HV^2D 算法的协同机动决策方法的运行框图

Fig. 5　Operational block diagram of cooperative maneuver decision-making approach based on HV^2D algorithm

4　仿真试验

本文通过数值仿真验证了所提出的智能协同机动决策方法的性能。考虑临近空间中的高超声速飞行器 TID 多角色博弈场景，首先设定目标、防御和拦截飞行器（T，D，I）等飞行器参数，其中拦截飞行器在响应速度、机动能力和探测信息获取等方面具有优势。随后，通过在典型工况下的数值仿真和蒙特卡洛仿真测试，验证了基于 HV^2D 的端到端协同机动决策方法在应对非完美、非完备和非完整等信息约束问题时的有效性、效能和鲁棒性。

4.1　场景设定

设定如下仿真验证场景：1 枚高超声速目标飞行器在约 55 km 的高度面临 1 枚拦截飞行器的威胁后，释放防御飞行器执行反拦截措施，3 枚飞行器构成 TID 多角色博弈场景。考虑到高超声速飞行器探测距离与机动能力，仿真场景的初始轴向距离约为 50 km。将式(11)中的相对距离和视线角观测噪声标准偏差分别设定为 $\sigma_\rho=10$ m 和 $\sigma_{\text{LOS}}=1\times10^{-3}$ rad，将目标飞行器和防御飞行器的标准探测装置视场角设定为 $\text{FOV}_i=3°,i=\{\text{T},\text{D}\}$。详细参数见表 1、2。

表 1　目标飞行器的详细参数

Tab. 1　Detailed parameters of target

参数	参数值
攻角/(°)	−5 ~ 15
倾侧角/(°)	−75 ~ 75
时间响应常数/s	0.1
探测装置视场角/(°)	3
初始位置/km	(0,55,0)
初始速度/(km·s^{-1})	(3,0,0)

表 2 防御飞行器和拦截飞行器的详细参数

Tab. 2 Detailed parameters of defender and interceptor

参数	防御飞行器	拦截飞行器
最大过载/g	4	6
时间响应常数/s	0.05	0.02
毁伤半径/m	0.75	0.75
探测装置视场角/(°)	3	
初始位置/km	(0,55,0)	(50,55,0)
初始速度/(km·s^{-1})	(3,0,0)	(-2,0,0)

在仿真试验中,目标飞行器和防御飞行器均采用基于 HV^{2}D 的协同机动决策方法。一方面,拦截飞行器采取多种制导策略,包括比例导引(proportional navigation,PN)制导方法[22]、基于最优控制的制导方法(opcimal guiance,OG)[23],以及基于微分对策的制导方法(differential game guidance,DG)[6];另一方面,为提高拦截飞行器在执行拦截任务时的生存能力,考虑其面对防御飞行器威胁时优先进行躲避。

4.2 训练效率测试

为验证在高动态强对抗的多角色博弈场景下,所提出的 HV^{2}D 算法和所构造的多智能体合作任务型奖励函数对多智能体训练效率的提高,设计了 3 种训练工况(表 3),进行训练效率测试。

表 3 MARL 智能算法训练工况设置

Tab. 3 MARL algorithm training condition setting

工况	多智能体强化学习算法	奖励函数
工况 1.1	HV^{2}D	多智能体合作任务型奖励函数
工况 1.2	Qmix	
工况 1.3	HV^{2}D/Qmix	稀疏奖励函数

表 3 中,工况 1.1 采用了所提出的基于 HV^{2}D 算法的协同机动决策方法以及多智能体合作任务型奖励函数进行训练。工况 1.2 在使用多智能体合作任务型奖励函数的同时,采用典型多智能体强化学习算法(Q-value mix,Qmix)[21]作为对照算法。在工况 1.3 中,两种多智能体强化学习算法均采用稀疏奖励函数。

图 6 展示了不同工况下的多智能体学习过程曲线,其中黑色实线为工况 1.3 中两种多智能体强化学习算法在使用稀疏奖励时获得的奖励值,系统始终只能获得博弈失败的负数奖励值,两者均无法成功训练;蓝色实线及点划线分别为工况 1.2 中 Qmix 算法在训练过程中获得的平均奖励值及最大逃逸成功率,Qmix 算法在训练初期通过合作任务型奖励函数获得了奖励信息并尝试优化其策略,但受限于强信息约束,系统最终陷入局部最优策略,导致最大逃逸成功率仅为 63%;而红色实线和点划线分别为 HV^{2}D 算法在训练过程中获得的平均奖励值和最大逃逸成功率,在同样应用合作任务型奖励函数后,HV^{2}D 算法在约 16 500 步后完成训练,最大逃逸成功率达到了 97%。上述结果表明,所提出的奖励函数和 HV^{2}D 算法提高了多智能体系统在高超声速飞行器复杂博弈场景中的训练效率。

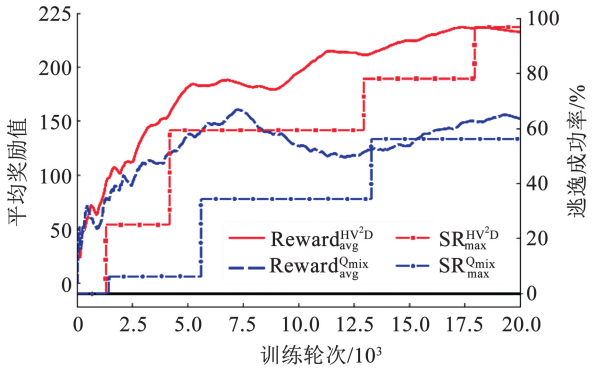


图 6 MARL 智能算法学习曲线

Fig. 6 MARL algorithm learning curves

4.3 有效性仿真测试

在多智能体协同决策网络架构所设定的场景基础上进行有效性仿真试验,选择拦截飞行器采用 DG 制导方法的仿真结果为例,分析基于 HV^{2}D 算法的协同机动决策方法所制定的机动策略,仿真结果如图 7 所示。

图 7(a)展示了各飞行器在所建立场景中的博弈轨迹及其 OXY 平面和 OXZ 内的投影;图 7(b)、7(c)分别为各飞行器在其体系 Y 轴方向和 Z 轴方向的过载变化图。根据图 7(a)~7(e)显示,目标飞行器感知到拦截飞行器的威胁后立即释放防御飞行器,并与拦截飞行器展开协同机动,具体决策过程如下。

1)在博弈场景开始至第 2.0 s 期间,目标飞行器进行小幅度机动,配合防御飞行器进行前期探测并为后续协同机动做准备;同时,防御飞行器进行震荡机动,并伴飞目标飞行器。

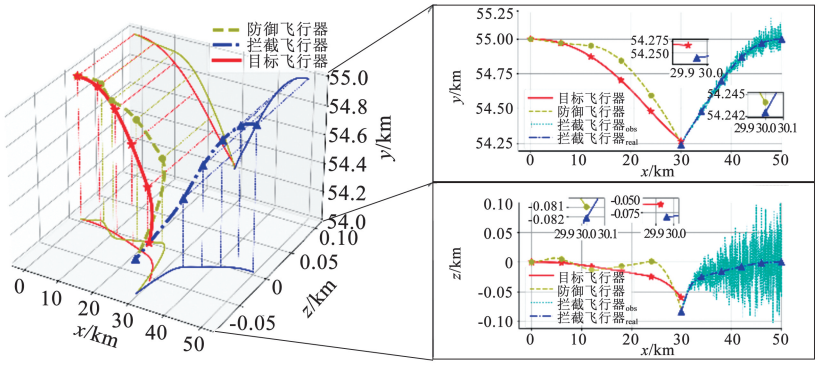
2)在博弈场景第 3.0~5.0 s 期间,目标飞行器沿体坐标系的 y 轴负方向大幅度机动,尝试避开拦截飞行器;同时,防御飞行器在保持约 1.5 s 的零机动飞行后,分别在体坐标系的 y 轴与 z 轴方向施加 -4.0 g 与 1.5 g 的过载,继续伴飞目标飞行器。

3)在第 6.0 s 后至博弈场景结束,目标飞行器与防御飞行器展开最终的协同机动。目标飞行器增大整体的机动幅度;为配合目标飞行器,防御飞行器在体坐标系的 y 轴方向上施加约 2.0 s 的 -2.5 g 以及约 1.0 的 -4.0 g 过载,并在 z 轴方向施加 -4.0 g

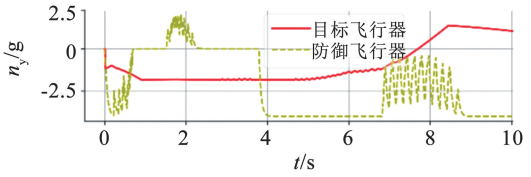
的最大过载，使拦截飞行器进入其防御范围。

另一方面，在博弈场景开始阶段，拦截飞行器分别于体坐标系的 y 轴和 z 轴方向施加约 6.0 g 的最大过载；当拦截飞行器感知到目标飞行器的机动后，施加 −2.0 g 的过载，使目标飞行器保持在其可拦截范围内；在博弈场景第 8.0 s 后，拦截飞行器感知到防御飞行器的反拦截威胁，在躲避防御飞行器的同时继续拦截目标飞行器。

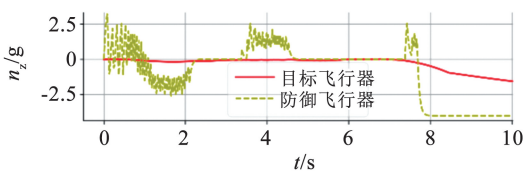
最终，防御飞行器虽未成功防御拦截飞行器，但使其无法到达目标飞行器的拦截窗口；而目标飞行器成功逃逸，逃逸距离约为 39.6 m。该结果验证了所提出的智能协同机动决策方法的有效性。



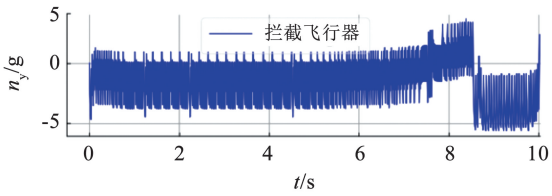
(a) 飞行器博弈轨迹



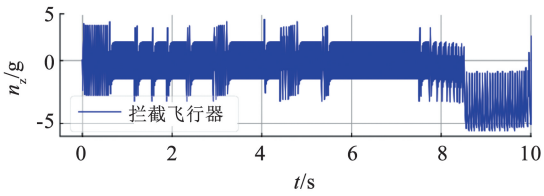
(b) 目标飞行器和防御飞行器在体系Y方向的过载



(c) 目标飞行器和防御飞行器在体系Z方向的过载



(d) 拦截飞行器在体系Y方向的过载



(e) 拦截飞行器在体系Z方向的过载

图 7　拦截飞行器采用 DG 制导方法时的有效性测试仿真结果

Fig. 7　Simulation results of effectiveness test when the interceptor uses the DG

4.4　效能仿真测试

为验证所提出的智能端到端协同机动决策方法在强信息约束下的高超声速飞行器博弈场景中的效能，本文增大场景变量的随机性，进行了 1 000 次的蒙特卡洛仿真测试，同样将 Qmix 算法作为对照组。每次测试中，拦截飞行器制导方法与初始位置、目标飞行器与防御飞行器的探测信息噪声与视场角等变量在一定范围内随机，详细参数见表 4，测试结果如图 8 所示。

表 4　蒙特卡洛仿真工况设置

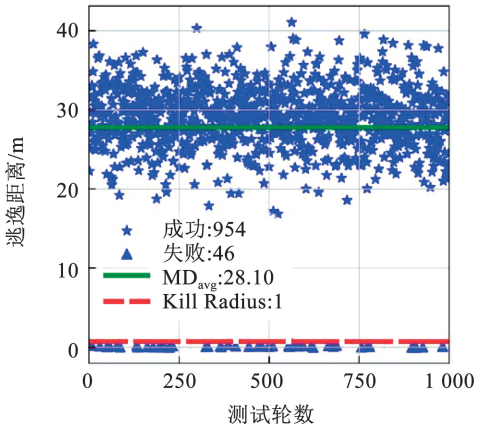
Tab. 4　Monte Carlo simulation condition setting

随机变量	参数值
拦截飞行器制导方法	PN、OG、DG
拦截飞行器初始位置/km	$[50\pm1.0, 55\pm0.2, 0\pm0.2]$
距离探测噪声/m	$\sigma_\rho = 20$
视线角探测噪声/rad	$\sigma_{\mathrm{LOS}} = 2\times10^{-3}$
探测装置视场角/(°)	2 ~ 4

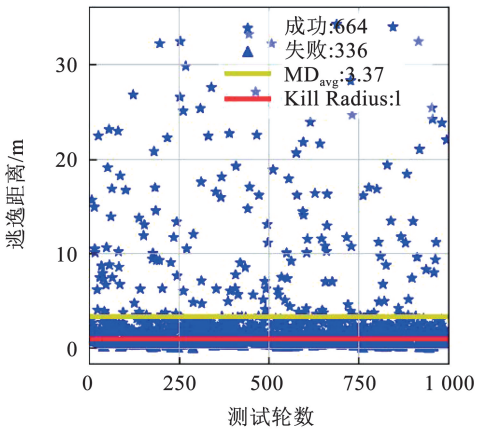
结果表明，飞行器在信息约束更强的博弈场景下，采用基于 Qmix 算法时，目标飞行器的逃逸成功率仅为 66.4%，平均逃逸距离为 3.37 m；而基于 HV^{2}D 的协调机动决策方法能够使飞行器保持出色的博弈效果，目标飞行器的逃逸成功率为 95.4%，平均逃逸距离为 28.12 m，远高于对照组的效能。因此，所提出的智能协同机动决策方法具有出色的效能。

4.5　鲁棒性仿真测试

为进一步探究提出的机动决策方法的鲁棒性，本文分别针对非完美、非完备和非完整等强信息约束，设置了不同范围的探测装置噪声、拦截飞行器初始位置和视场角大小等 3 类测试场景，并通过蒙特卡洛仿真进行了鲁棒性测试。



(a) HV²D算法



(b) Qmix算法

图 8 典型场景下的蒙特卡洛测试结果

Fig. 8 Monte Carlo test result in typical scenarios

4.5.1 非完美信息问题的鲁棒性探究

为分析探测装置噪声对机动决策效果的影响，设定目标飞行器和拦截飞行器的相对距离和视线角观测噪声标准差分别在[0,10,20,30,40] m 与[0,1,2,3,4] $\times 10^{-3}$ rad 的范围内变化，生成 25 种工况。在各工况下进行 1 000 次蒙特卡洛仿真测试。

图 9 为不同探测噪声工况下的 1 000 次蒙特卡洛仿真测试结果。

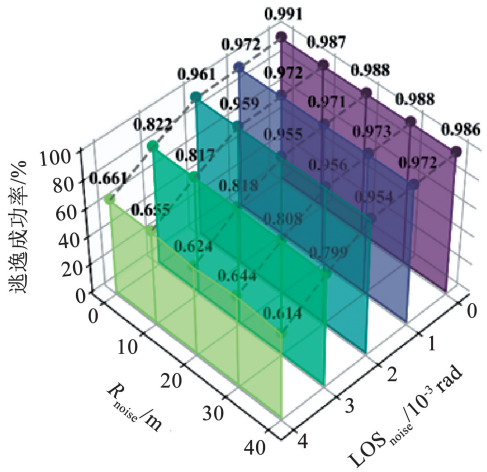


图 9 不同探测噪声下的鲁棒性测试结果

Fig. 9 Robustness test results under different detection noises

由图 9 测试结果可知，本文提出的智能协同机动决策方法能够在不同噪声水平下保持良好的机动决策能力。即使在相对距离噪声达到 40 m 且视线角噪声达到 4×10^{-3} rad 的强噪声条件下，目标飞行器的逃逸成功率维持在 61.4%。可见，所提出方法在处理存在探测装置噪声的非完美信息约束时具有较强的鲁棒性。

4.5.2 非完备信息问题的鲁棒性探究

针对拦截飞行器初始位置的不确定性对机动决策效果的影响，将拦截飞行器标准初始位置设定为[50 ± 0.5,55 ± 0.1,0 ± 0.1] km，并通过逐步扩大其标准初始位置的变化范围至 1 ~ 5 倍，具体见表 5。随后，对每种工况的不同方向变化分别进行 1 000 次蒙特卡洛仿真测试。

表 5 拦截飞行器初始位置工况设置

Tab. 5 Initial position condition setting of interceptor

工况	x 轴方向位置/km	y 轴方向位置/km	z 轴方向位置/km
工况 2.1	50 ± [0.5,1.0,1.5,2.0,2.5]	55	0
工况 2.2	50	55 ± [0.1,0.2,0.3,0.4,0.5]	0
工况 2.3	50	55	0 ± [0.1,0.2,0.3,0.4,0.5]
工况 2.4	50 ± [0.5,1.0,1.5,2.0,2.5]	55 ± [0.1,0.2,0.3,0.4,0.5]	0 ± [0.1,0.2,0.3,0.4,0.5]

图 10 为拦截飞行器初始位置范围变化的 1 000 次蒙特卡洛仿真测试结果。由图 10 测试结果可知，随着拦截飞行器初始位置随机变化范围的增加，所提出的决策方法依然展现出出色的适应能力。在 3 个方向同时扩展至 4 倍标准变化范围的苛刻条件下，目标飞行器的逃逸成功率依然高达 92.3%。因此，所提出方法在应对受拦截飞行器未知初始位置所影响的非完备探测信息约束时具有较强的鲁棒性。

4.5.3 非完整信息问题的鲁棒性探究

为探讨视场角范围变化对机动决策效果的影响，将目标飞行器和防御飞行器的视场角分别设置为[1°,2°,3°,4°,5°]，生成 25 种工况，并对各工况进行 1 000 次蒙特卡洛仿真测试，结果如图 11 所示。

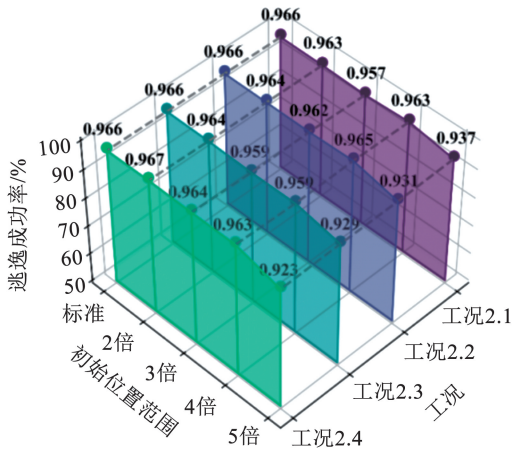


图 10　拦截飞行器初始位置范围变化的鲁棒性测试结果

Fig. 10　Robustness test results of interceptor with different initial position ranges

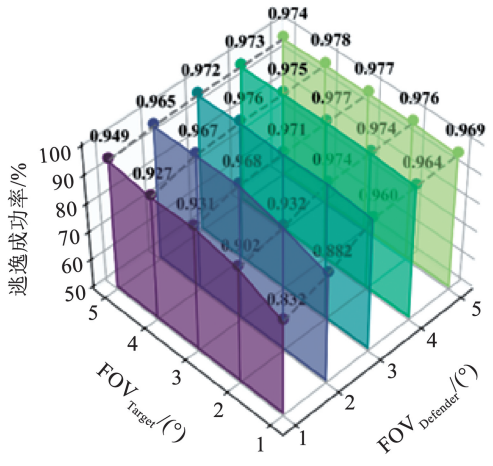


图 11　不同视场角大小的鲁棒性测试结果

Fig. 11　Robustness test results for different fields of view

在不同视场角条件下，本文提出的智能协同机动决策方法均表现出稳定的机动决策能力。即使在目标飞行器和防御飞行器视场角均缩小至 1°的狭窄视场条件下，目标飞行器仍具有 83.2% 的逃逸成功率。这验证了所提出方法在视线角范围限制导致的非完整探测信息约束下具备出色的鲁棒性。

综上所述，本文提出的智能协同机动决策方法在应对非完美、非完备、非完整等信息约束时，始终保持较强的鲁棒性。

5　结　论

1）针对 TID 多角色博弈场景中非完美、非完备和非完整等信息约束下的协同机动决策问题，设计了一种适用于高超声速飞行器的 HV^{2}D 多智能体强化学习算法，并基于此算法提出了一种端到端协同机动决策方法，能够生成有效的协同机动决策指令，使目标飞行器成功逃逸拦截飞行器的威胁。

2）为应对强信息约束下智能体难以准确感知博弈环境动态特征的问题，基于观测信息共享堆叠机制设计了多智能体系统的局部观测状态空间；并结合飞行器博弈关系与零控脱靶量构造多智能体合作决策奖励函数，有效提高了 HV^{2}D 算法在强对抗与高动态的多角色博弈场景中收敛能力。

3）设计了一种多智能体协同决策网络架构，由基于 CNN-D^{3}RQN 的基础智能体网络以及基于超网络的顶层值分解网络构成，从强信息约束中提取飞行器的时空轨迹特征，支持多智能体的协同策略生成，并引导飞行器做出可信机动决策。

4）数值仿真结果表明，提出的端到端智能协同机动决策方法在强信息约束下的高超声速飞行器复杂博弈场景中具有出色的性能，并在典型场景仿真和蒙特卡洛测试中表现出优越的效能和鲁棒性。

参考文献

[1] CHEN Jieqing, SUN Ruisheng, LU Yu. Cooperative game penetration guidance for multiple hypersonic vehicles under safety critical framework [J]. Chinese Journal of Aeronautics, 2024, 37(1): 247. DOI: 10.1016/j.cja.2023.08.023

[2] 郭建国，陆东陈，周敏. 飞行器博弈制导进程与展望[J]. 航空兵器，2024，31(2)：8
GUO Jianguo, LU Dongchen, ZHOU Min. Analysis of the progress of aircraft game guidance [J]. Aero Weaponry, 2024, 31(2): 8. DOI: 10.12132/ISSN.1673－5048.2024.0022

[3] PERELMAN A, SHIMA T, RUSNAK I. Cooperative differential games strategies for active aircraft protection from a homing missile [J]. Journal of Guidance, Control, and Dynamics, 2011, 34(3): 761. DOI: 10.2514/1.51611

[4] JIA Zhen, YE Dong, XIAO Yan, et al. Approximate analytical approach for spacecraft pursuit-evasion game with reachability analysis [J]. IEEE Transactions on Aerospace and Electronic Systems, 2025, 61(4): 9058. DOI: 10.1109/TAES.2025.3552073

[5] MISHLEY A, SHAFERMAN V. Near-optimal evasion from acceleration bounded modern pursuers [J]. Journal of Guidance, Control, and Dynamics, 2025, 48(4): 793. DOI: 10.2514/1.G008704

[6] LIANG Haizhao, WANG Jianying, WANG Yonghai, et al. Optimal guidance against active defense ballistic missiles via differential game strategies[J]. Chinese Journal of Aeronautics, 2020, 33(3): 978. DOI: 10.1016/j.cja.2019.12.009

[7] LI Jianqing, ZHAO Qiancheng, LI Chaoyong, et al. A maneuvering strategy based on motion camouflage in three-player differential game [J]. Aerospace Science and Technology, 2024, 155: 109642. DOI: 10.1016/j.ast.2024.109642

[8] LI Zhenyu, ZHU Hai, LUO Yazhong. An escape strategy in orbital pursuit-evasion games with incomplete information [J]. Science China Technological Sciences, 2021, 64(3): 559. DOI: 10.1007/s11431－020－1662－0

[9] TANG Xu, YE Dong, HUANG Lei, et al. Pursuit-evasion game switching strategies for spacecraft with incomplete-information [J].

Aerospace Science and Technology, 2021, 119: 107112. DOI: 10.1016/j.ast.2021.107112

[10] ZHOU Yaoming, YANG Fan, ZHANG Chaoyue, et al. Cooperative decision-making algorithm with efficient convergence for UCAV formation in beyond-visual-range air combat based on multi-agent reinforcement learning[J]. Chinese Journal of Aeronautics, 2024, 37(8): 311. DOI: 10.1016/j.cja.2024.04.008

[11] 倪炜霖, 王永海, 徐聪, 等. 基于强化学习的高超飞行器协同博弈制导方法[J]. 航空学报, 2023, 44(增刊2): 729400
NI Weilin, WANG Yonghai, XU Cong, et al. Cooperativegame guidance method for hypersonic vehicles based on reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(Sup 2): 729400. DOI: 10.7527/S1000-6893.2023.29400

[12] 李永丰, 史静平, 章卫国, 等. 深度强化学习的无人作战飞机空战机动决策[J]. 哈尔滨工业大学学报, 2021, 53(12): 33
LI Yongfeng, SHI Jingping, ZHANG Weiguo, et al. Maneuver decision of UCAV in air combat based on deep reinforcement learning[J]. Journal of Harbin Institute of Technology, 2021, 53(12): 33. DOI: 10.11918/202005108

[13] ZHOU Wenhong, LI Jie, LIU Zhihong, et al. Improving multi-target cooperative tracking guidance for UAV swarms using multi-agent reinforcement learning[J]. Chinese Journal of Aeronautics, 2022, 35(7): 100. DOI: 10.1016/j.cja.2021.09.008

[14] 王英杰, 袁利, 汤亮, 等. 信息非完备下多航天器轨道博弈强化学习方法[J]. 宇航学报, 2023, 44(10): 1522
WANG Yingjie, YUAN Li, TANG Liang, et al. Reinforcement learning method for multi-spacecraft orbital game with incomplete information[J]. Journal of Astronautics, 2023, 44(10): 1522. DOI: 10.3873/j.issn.1000-1328.2023.10.005

[15] 高树一, 林德福, 郑多, 等. 针对集群攻击的飞行器智能协同拦截策略[J]. 航空学报, 2023, 44(18): 271
GAO Shuyi, LIN Defu, ZHENG Duo, et al. Intelligent cooperative interception strategy of aircraft against cluster attack[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(18): 271. DOI: 10.7527/S1000-6893.2023.28301

[16] GOETZ L P, ALBRIGHT J D. Airborne pulse-Doppler radar[J]. IRE Transactions on Military Electronics, 1961, MIL-5(2): 116. DOI: 10.1109/iret-mil.1961.5008329

[17] DONG Wei, WANG Chunyan, WANG Jianan, et al. Unified method for field-of-view-limited homing guidance[J]. Journal of Guidance, Control, and Dynamics, 2022, 45(8): 1415. DOI: 10.2514/1.G006710

[18] OLIEHOEK F A, AMATO C. A concise introduction to decentralized POMDPs[M]. Cham, Switzerland: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-28929-8

[19] WANG Hongbo, ZHANG Yao. Impulsive maneuver strategy for multi-agent orbital pursuit-evasion game under sparse rewards[J]. Aerospace Science and Technology, 2024, 155: 109618. DOI: 10.1016/j.ast.2024.109618

[20] CHAUHAN V K, ZHOU Jiandong, LU Ping, et al. A brief review of hypernetworks in deep learning[J]. Artificial Intelligence Review, 2024, 57(9): 250. DOI: 10.1007/s10462-024-10862-8

[21] RASHID T, SAMVELYAN M, DE WITT C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(1): 7234. DOI:10.5555/3455716.3455894

[22] YUAN P J, CHERN J S. Ideal proportional navigation[J]. Journal of Guidance, Control, and Dynamics, 1992, 15(5): 1161. DOI: 10.2514/3.20964

[23] COTTRELL R G. Optimal intercept guidance for short-range tactical missiles[J]. AIAA Journal, 1971, 9(7): 1414. DOI: 10.2514/3.6369

(编辑 张 红)