文章编号：2097-1974(2025)01-0065-08 DOI：10.7654/j.issn.2097-1974.20250109

基于强化学习的飞行器博弈制导方法

倪炜霖[1]，刘佳琪[2]，邵 节[2]，刘 鹏[2]，梁海朝[1]
（1. 中山大学航空航天学院，深圳，518000；2. 北京航天长征飞行器研究所，北京，100076）

摘要：针对飞行器与伴飞防御飞行器协同躲避拦截器攻击的主动反拦截博弈对抗问题，基于深度强化学习算法提出一种飞行器主动防御智能制导方法，该方法具有在目标飞行器机动能力不足情况下博弈成功率较高的特点。针对强化学习训练过程中的稀疏奖励问题，提出了一种奖励函数塑造方法，提高了强化学习算法收敛效率和训练稳定度。最后，通过数值仿真对所提出方法的有效性进行验证，仿真结果表明，所提出的方法能够实现飞行器博弈对抗成功，且相比于传统博弈制导方法具有更高的博弈成功率。

关键词：博弈对抗；深度强化学习；奖励函数塑造；稀疏奖励；主动反拦截
中图分类号：V11 **文献标识码**：A

Game Guidance Method for Flight Vehicle based on Reinforcement Learning

NI Weilin[1], LIU Jiaqi[2], SHAO Jie[2], LIU Peng[2], LIANG Haizhao[1]
(1. School of Aeronauties, Sun Yat-sen University, Shenzhen, 518000;
2. Beijing Institute of Space Long March Vehicle, Beijing, 100076)

Abstract: Aiming at the active anti-interception game confrontation between hypersonic aircraft and accompanying defense aircraft to avoid interceptor attacks, an active defense intelligent guidance method for hypersonic aircraft is proposed based on deep reinforcement learning algorithm. In the case of insufficient maneuverability of the target aircraft, this method can achieve a higher success rate. Aiming at the sparse reward problem in the reinforcement learning training process, a reward function shaping method is proposed, which improves the convergence efficiency and training stability of the reinforcement learning algorithm. Finally, the effectiveness of the proposed method is verified by numerical simulation. The simulation results show that the proposed method can successfully achieve flight vehicle game confrontation, and has a higher game success rate than traditional game guidance methods.

Keywords: game theory; reinforcement learning; reward shaping; sparse reward; active anti-interception

0 引 言

高超声速飞行器是一种能够以大于5倍声速的速度持续飞行，并能够完成指定任务的飞行器[1]。面对飞行器防御系统的防御威胁，飞行器如何利用飞行器攻防博弈理论指导其在末制导阶段成功逃逸，具有重大研究价值。

末制导博弈对抗是一种典型的双边最优问题，很多学者开展了研究。古典制导律[2-3]仅适用于简单博弈对抗场景；最优制导方法具有较高的制导精度和良好的收敛性能，但飞行器需要实时获取博弈成员的运动状态，且对状态估计误差较为敏感[4]。此外，上述方法本质上都仅为单边博弈制导方法，无法同时应用于逃逸方与拦截方。微分对策方法的提出解决了传统最优制导方法的问题，一方面，微分对策方法是一种双边最优控制理论，可同时用于拦截方与逃逸方；另一方面，制导律解算主要基于参与方各自的最大机动能力，对加速度估计精度要求不高，因此是现阶段飞行器博弈问题的主要研究方向之一[5]。但是，采用微分对策方法解决飞行器博弈对抗问题仍存在以下不足：一是动力学模型需要以常微分方程形式描述，对于实际复杂应用场景而言建模难度大；二是计算量大，解算时间长，实际飞行器计算能力无法满足其对计算资源的需求；三是需要状态反馈量大，实际飞行器传感器可能无法完全得到其所需状态量[6]。

随着智能技术的发展，利用强化学习（Reinforcement Learning，RL）方法解决飞行器博弈对抗问

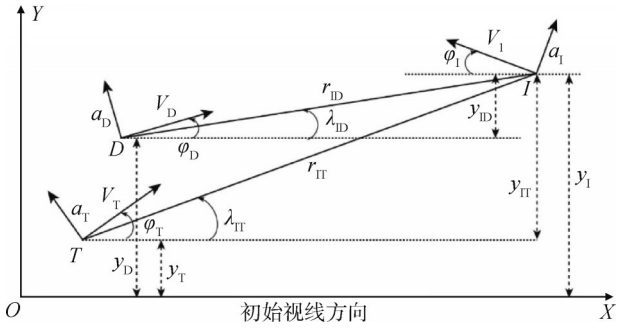
收稿日期：2024-07-26；修回日期：2024-12-12
基金项目：国家自然科学基金(No.62003375)

题被视为一种全新且可行的途径。相比传统博弈方法，强化学习方法具有训练后计算量小、运算速度快、鲁棒性强和所需状态反馈量与先验信息较少等特点[6-7]。张阳康等[8]提出了基于引导策略搜索算法的有模型强化学习制导方法，提高了行星着陆器着陆制导方法的鲁棒性。文献［9］采用元强化学习算法，在不同的飞行器着陆场景中训练智能体，得到适用性较强的神经网络策略模型。上述关于强化学习制导的研究，大多是针对非对抗类型的环境，而当智能体面对飞行器博弈对抗这类博弈环境复杂、博弈成员对抗程度高、博弈场景随机性强等特点的环境时，智能体将会难以获得有益的奖励信息，进而导致智能体产生稀疏奖励问题[10]或出现高原现象[11]。如何解决飞行器智能博弈对抗中的稀疏奖励问题，并得到有效且稳定的飞行器智能攻防博弈制导方法是本文的研究重点。

针对上述问题，本文开展了飞行器智能攻防博弈制导方法研究，基于双竞争深度Q学习网络（Dueling Double DQN，D3QN）深度强化学习算法，提出了一种主动防御飞行器的智能攻防博弈制导方法，并利用奖励函数塑造方法，将飞行器博弈过程中的稀疏奖励转化为连续奖励以解决训练过程中的稀疏奖励问题，以提高强化学习算法收敛效率和训练稳定度。数值仿真验证结果证明，飞行器采用本文所提出的智能博弈制导方法能显著提高其在机动能力不足情况下的博弈成功率。

1 问题描述

考虑一个典型主动防御博弈对抗场景[12]，场景中包含3个成员：目标飞行器（Target，T）、防御飞行器（Defender，D）、拦截飞行器（Interceptor，I）。其中，目标飞行器与防御飞行器采取主动反拦截策略对抗迎面来袭的拦截飞行器，而拦截飞行器在场景中需要避免被防御飞行器碰撞同时靠近目标。由于在攻防对抗最后阶段，两方飞行器相对速度大，制导时间短，因此可将三维运动分解为俯仰平面和横侧向平面的运动；又由于俯仰平面是飞行器的重要运动面，仅在俯仰平面内进行机动博弈降低了飞行器的控制难度，减少了飞行器的射程损失，减轻了飞行器的结构载荷[13]，且飞行器在俯仰平面内的博弈机动策略也可映射用于其在横侧向平面内的博弈对抗；综上，仅考虑飞行器在俯仰平面进行机动是飞行器博弈对抗研究中常用并合理的简化。俯仰平面内主动防御博弈对抗场景如图1所示。



*OXY*坐标系—惯性参考坐标系；*O*—位于目标飞行器初始位置的地平面的原点；*OX*轴—飞行器初始视线方向，在俯仰平面内垂直于*OY*轴；r_{IT}，r_{ID}—拦截飞行器与目标飞行器、防御飞行器间的相对距离；λ_{IT}，λ_{ID}—拦截飞行器与目标飞行器、防御飞行器间的视线角；V_I，V_T，V_D—各飞行器的速度；a_I，a_T，a_D—各飞行器对应的侧向加速度；φ_I、φ_T，φ_D—飞行器航向角；y_I，y_D，y_T—飞行器纵向距离。

图1 主动防御博弈对抗场景

Fig.1 Active defense game confrontation scenario

飞行器博弈对抗制导律通常基于如下假设[14]：

a）飞行器的速度与最大过载为常值；

b）飞行器动力学环节可近似为一阶传递函数的形式；

c）忽略目标飞行器与防御飞行器间通信的时延。

在所建立的参考坐标系中，目标飞行器考虑仅采用气动力进行机动，动力学方程为

$$\begin{bmatrix} a_{xT} \\ a_{yT} \end{bmatrix} = \begin{bmatrix} -\sin\phi_T g \\ -\cos\phi_T g \end{bmatrix} + \frac{q_T A_T}{m_T} \begin{bmatrix} \cos\alpha_T C_{xb} - \sin\alpha_T C_{yb} + C_x \\ \sin\alpha_T C_{xb} + \cos\alpha_T C_{yb} \end{bmatrix} \tag{1}$$

式中 a_{xT}，a_{yT}分别为目标飞行器的轴向和侧向加速度；C_{xb}，C_{yb}和C_x分别为目标飞行器的轴向力气动系数、侧向力气动系数和波阻系数；α_T为目标飞行器攻角；q_T为目标飞行器动压；A_T为目标飞行器参考面积；m_T为目标飞行器质量。

防御飞行器与拦截飞行器采用直接力与气动力复合的方式进行机动，由于直接力远大于气动力的作用，因此忽略气动力，其动力学方程在所建立参考坐标系中可表示为

$$\begin{bmatrix} a_{xi} \\ a_{yi} \end{bmatrix} = \begin{bmatrix} -\sin\phi_i g \\ -\cos\phi_i g \end{bmatrix} + \begin{bmatrix} 0 \\ u_i \end{bmatrix}, i = \{D, I\} \tag{2}$$

$$\dot{a}_{yi} = \frac{1}{\varepsilon_i}\left(u_i - a_{yi}\right), i = \{D, I\} \tag{3}$$

式中 ε_i为飞行器的控制响应时间。

因此，研究问题可以描述为：设计一种基于深度强化学习算法的主动反拦截协同博弈制导律，旨在实现目标飞行器与防御飞行器之间的协同机动，使得目标飞行器在防御飞行器展开反拦截的同时，能够巧妙地躲避拦截飞行器的拦截。

2　智能攻防策略设计

2.1　基于D3QN的智能博弈制导方法

深度强化学习（Deep Reinforcement Learning，DRL）是机器学习中的重要分支，结合了强化学习与机器学习，令智能体与环境交互，通过试错的方式学习做出更好的决策。基于Q-Learning的深度强化学习方法又被称为基于价值的深度强化学习算法，利用最优动作-价值函数$Q_\pi^*(s_t,\ a_t)$，通过观察智能体状态S_t，执行动作空间A中价值最大的动作。在DRL中，动作-价值函数$Q_\pi^*(s_t,\ a_t)$可由神经网络$Q(s,\ a;\ \theta)$表示，其中θ为模型参数，被称为深度Q网络（Deep Q Network，DQN）。

双竞争深度Q学习网络（Dueling Double DQN，D3QN）是一种基于值函数的强化学习算法[15]，在DQN的基础上结合了Dueling DQN和Double DQN两种算法的优点：一方面，参考Dueling DQN算法，有效提高动作价值估计的准确性[16]；另一方面，D3QN算法参考Double DQN算法，建立当前收益函数Q和目标收益函数Q'两个动作价值函数，有效避免了Q值过高估计的问题[17]，是一种性能优越的离散动作空间DRL算法。鉴于上述优点，本文基于D3QN算法，提出一种针对飞行器主动防御博弈对抗场景的智能攻防博弈制导方法。本文将目标飞行器与防御飞行器视为智能体，并对飞行器主动防御博弈场景进行MDP参数表述，即在单次博弈回合中，智能体通过观测自身在t时刻的状态S_t，由收益函数$Q(\cdot)$得到动作空间A中各动作的价值，其收益函数可表示为

$$Q(S,A,w,a,b)=V(S,w,a)+A(S,A,w,b) \tag{4}$$

式中　$V(\cdot)$为智能体动作价值函数；$A(\cdot)$为智能体动作优势函数；w为公共部分的网络参数；a，b分别为价值函数和优势函数独有部分的网络参数。

再通过ε-greedy算法选择动作空间A中的动作a_t执行：

$$\begin{aligned}&A^*\leftarrow\arg\max_a Q(s,a)\\&\text{For all } a\in A(s):\\&\pi(a|s)\leftarrow\begin{cases}1-\varepsilon+\varepsilon/|A(s)| & a=A^*\\ \varepsilon/|A(s)| & a\neq A^*\end{cases}\end{aligned} \tag{5}$$

式中　A^*为当前网络的最优动作；$A(s)$为在状态s情况下智能体可执行的动作集合；$|A(s)|$为该动作集合的大小；$\pi(\cdot)$为该智能体所采取的策略。上述算法令智能体有ε的概率选择非当前网络的最优动作，以给予智能体探索空间。

随后，将智能体动作a_t输入由各飞行器动力学模型与主动防御博弈场景运动学模型组成的训练环境，得到下一时刻的飞行器状态S_{t+1}，与对应的奖励r_t；最后，算法会将上述t时刻产生的信息$[S_t\ \ a_t\ \ r_t\ \ S_{t+1}]^{\mathrm{T}}$放入回放记忆单元；之后，重复上述步骤，直至回合结束。

在训练过程中，智能体在多个博弈回合中积累的记忆，将会通过网络反向传播，定期更新目标值网络与当前值网络参数，且智能体将会不断重复进行博弈，更新记忆，直至智能体网络参数收敛且保持稳定。

智能博弈流程如图2所示，其中针对飞行器三方攻防博弈场景所设计的状态空间$\boldsymbol{S}_t$为

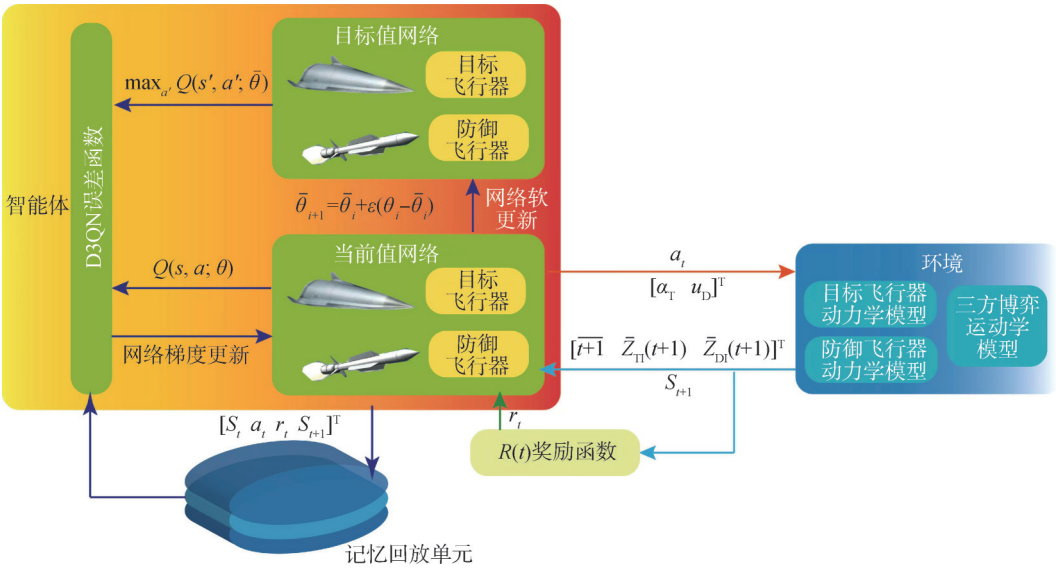


图2　智能博弈流程

Fig.2　Intelligent game process

$$\boldsymbol{O}=\boldsymbol{S}_t=\frac{\boldsymbol{S}_{t0}-\boldsymbol{S}_{\min}}{\boldsymbol{S}_{\max}-\boldsymbol{S}_{\min}}=\left[\bar{t}\quad \bar{Z}_{\mathrm{TI}}(t)\quad \bar{Z}_{\mathrm{DI}}(t)\right]^{\mathrm{T}} \tag{6}$$

$$\begin{cases}\boldsymbol{S}_{\max}=\left[t_{\mathrm{f}}\quad |y_{\mathrm{T0}}-y_{\mathrm{I0}}|\quad |y_{\mathrm{D0}}-y_{\mathrm{I0}}|\right]^{\mathrm{T}}\\ \boldsymbol{S}_{\min}=\left[0\quad -|y_{\mathrm{T0}}-y_{\mathrm{I0}}|\quad -|y_{\mathrm{D0}}-y_{\mathrm{I0}}|\right]^{\mathrm{T}}\\ \boldsymbol{S}_{t0}=\left[t\quad Z_{\mathrm{IT}}(t)\quad Z_{\mathrm{ID}}(t)\right]^{\mathrm{T}}\end{cases} \tag{7}$$

其中，为了将不同参数去量纲化并缩小数值差别，使网络快速收敛，需要对状态空间进行归一化操作[18]。$\boldsymbol{S}_{\max}$，$\boldsymbol{S}_{\min}$分别为状态空间预估最大值和最小值；$\boldsymbol{S}_{t0}$为t时刻归一化前状态量；$\boldsymbol{S}_t$为实际归一化后t时刻的状态量；t为时间；y_{T0}，y_{D0}，y_{I0}分别为初始时刻目标飞行器、防御飞行器和拦截飞行器在y方向的坐标；$Z_{\mathrm{IT}}(t)$，$Z_{\mathrm{ID}}(t)$分别为拦截飞行器相对目标飞行器与防御飞行器在t时刻的零控脱靶量，表示从当前时刻到制导结束，博弈双方均不再输出制导指令，在制导结束时脱靶量的大小[19]，其具体计算方法如下：

$$Z_{\mathrm{IT}}(t)=y_{\mathrm{IT}}(t)+\dot{y}_{\mathrm{IT}}(t)t_{\mathrm{go2}}+a_{\mathrm{I}}\varepsilon_{\mathrm{I}}^2\varphi\left(\frac{t_{\mathrm{go2}}}{\varepsilon_{\mathrm{I}}}\right)-a_{\mathrm{T}}\varepsilon_{\mathrm{T}}^2\varphi\left(\frac{t_{\mathrm{go2}}}{\varepsilon_{\mathrm{T}}}\right) \tag{8}$$

$$Z_{\mathrm{ID}}(t)=y_{\mathrm{ID}}(t)+\dot{y}_{\mathrm{ID}}(t)t_{\mathrm{go1}}+a_{\mathrm{I}}\varepsilon_{\mathrm{I}}^2\varphi\left(\frac{t_{\mathrm{go1}}}{\varepsilon_{\mathrm{I}}}\right)-a_{\mathrm{D}}\varepsilon_{\mathrm{D}}^2\varphi\left(\frac{t_{\mathrm{go1}}}{\varepsilon_{\mathrm{D}}}\right) \tag{9}$$

式中 t_{go1}，t_{go2}分别为对应博弈场景剩余时间；ε为飞行器过载响应时间。

其中，

$$\varphi(x)=\mathrm{e}^{-x}+x-1 \tag{10}$$

此外，针对飞行器三方攻防博弈场景所设计的动作空间$\boldsymbol{a}_t$如下：

$$\boldsymbol{a}_t=\left[\alpha_{\mathrm{T}}\quad u_{\mathrm{D}}\right]^{\mathrm{T}} \tag{11}$$

式中 α_{T}为目标飞行器的攻角控制信号；u_{D}为防御飞行器的过载控制信号。

2.2 基于塑造技术的非稀疏奖励函数设计

本文受零控脱靶量概念的启发，基于主动防御博弈对抗场景建模所得到的拦截飞行器与目标飞行器之间的零控脱靶量$Z_{\mathrm{IT}}(t)$和拦截飞行器与防御飞行器之间的零控脱靶量$Z_{\mathrm{ID}}(t)$，塑造随智能体状态$s(t)$连续变化的奖励函数。具体塑造过程如下：

首先，定义函数$R(Z,\ k)$：

$$R(Z,k)=\begin{cases}\mathrm{sign}(\Delta)\cdot\Delta^{0.1}\\ \Delta=|Z|-k\end{cases} \tag{12}$$

式中 Z为零控脱靶量；k为杀伤半径。R函数的特点为，当$|Z|<k$时，R值为负；当$|Z|>k$时，R值为正，且当Δ趋近于0时，函数梯度绝对值趋近于无穷，有利于智能体在期望策略处收敛；当$|Z|=k$时，R值为0。

设定离散奖励函数$R_S(t)$如下：

$$R_S(t)=\begin{cases}\alpha\cdot\exp\left[-d_{\mathrm{DI}}^2(t)\right] & t=t_{\mathrm{fID}}\\ \beta\cdot R\left[d_{\mathrm{TI}}(t),k_{\mathrm{TI}}\right] & t=t_{\mathrm{fIT}}\end{cases} \tag{13}$$

式中 $d(t)$为t时刻各飞行器之间距离；k_{TI}为目标飞行器与拦截飞行器之间的杀伤半径；α，β为人工设定的超参数；t_{f}对应博弈对场景终端时间。由式（13）可知，当防御飞行器与拦截飞行器之间的博弈场景结束时，即$t=t_{\mathrm{fID}}$时，奖励值随着d_{DI}的减小而增大，当$d_{\mathrm{DI}}=0$时，奖励值为α；当目标飞行器与拦截飞行器之间的博弈场景结束时，即$t=t_{\mathrm{fIT}}$时，若此时$d_{\mathrm{TI}}>k_{\mathrm{TI}}$，代表目标飞行器成功逃逸，奖励值为正，反之，奖励值为负，视为对智能体的“惩罚”。

而对于博弈过程中的奖励值，本文塑造连续奖励函数$R_C(t)$如下：

$$R_C(t)=\begin{cases}-R\left[Z_{\mathrm{ID}}(t),k_{\mathrm{ID}}\right] & t<t_{\mathrm{fID}}\\ R\left[Z_{\mathrm{IT}}(t),k_{\mathrm{IT}}\right] & t_{\mathrm{fID}}<t<t_{\mathrm{fIT}}\end{cases} \tag{14}$$

式中 $Z_{\mathrm{IT}}(t)$与$Z_{\mathrm{ID}}(t)$可根据式（8）与（9）计算得到；k_{ID}，k_{IT}分别为拦截飞行器与防御飞行器和目标飞行器之间的杀伤半径。由式（12）、式（14）与式（16）可知，当$t<t_{\mathrm{fID}}$时，若$|Z_{\mathrm{ID}}(t)|<k_{\mathrm{ID}}$，则代表防御飞行器成功对拦截飞行器造成威胁，奖励值为正数，视为“奖励”，其他情况下奖励值为负数，视为“惩罚”，且奖励值与$|Z_{\mathrm{ID}}(t)|$呈负相关；当$t_{\mathrm{fID}}\leqslant t<t_{\mathrm{fIT}}$时，若$|Z_{\mathrm{IT}}(t)|<k_{\mathrm{IT}}$，则代表拦截飞行器成功对目标飞行器造成威胁，奖励值为负数，视为“惩罚”，其他情况下奖励值为正数，视为“奖励”，且奖励值与$|Z_{\mathrm{IT}}(t)|$呈正相关。

综上，本文所采用的塑造后的奖励函数$r(t)$为

$$r(t)=R_S(t)+R_C(t)=\begin{cases}-R\left[Z_{\mathrm{ID}}(t),\ k_{\mathrm{ID}}\right] & t<t_{\mathrm{fID}}\\ \alpha\cdot\exp\left[-d_{\mathrm{DI}}^2(t)\right] & t=t_{\mathrm{fID}}\\ R\left[Z_{\mathrm{IT}}(t),\ k_{\mathrm{IT}}\right] & t_{\mathrm{fID}}<t<t_{\mathrm{fIT}}\\ \beta\cdot R\left[d_{\mathrm{TI}}(t),\ k_{\mathrm{TI}}\right] & t=t_{\mathrm{fIT}}\end{cases} \tag{15}$$

3 数值仿真验证

本文在CPU为Intel Core Xeon Platinum 8270、主频2.70 GHz，GPU GTX2080的硬件环境下进行数值仿真验证。

考虑一枚携带有防御飞行器的目标飞行器，飞行器在55 km的高度受到拦截飞行器的拦截威胁，拦截飞行器导引弹头探测距离为100 km，因此博弈场景纵深距离为100 km。在场景中，目标飞行器提前发现拦截飞行器的拦截威胁，释放防御飞行器进行反拦截，并保持防御飞行器相对目标飞行器纵向位置约500 m、初始横向位置约5 km的编队飞行。博弈场景示意如图3所示。

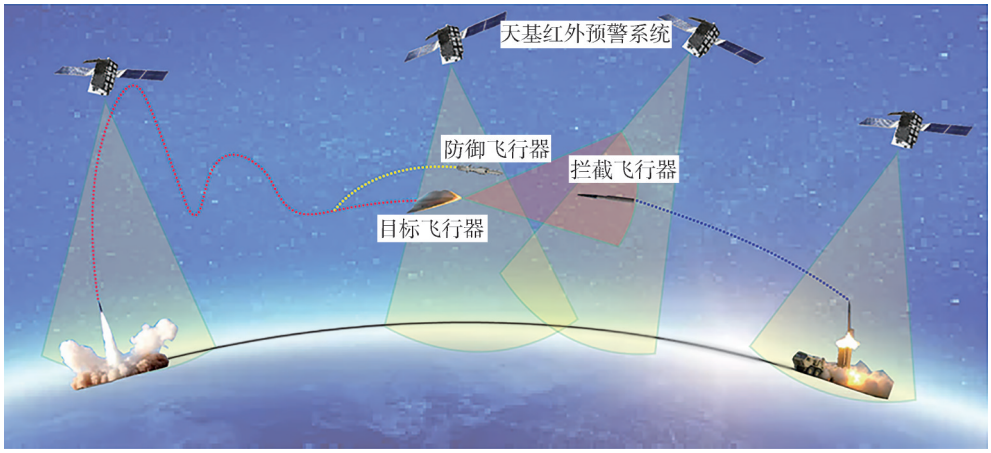


图3　博弈场景示意

Fig.3　Game background

3.1　拦截飞行制导律

考虑拦截飞行器采用微分对策制导方法进行躲避与拦截，基于上述机动策略与所建立的主动防御博弈对抗场景模型，结合微分对策理论，在主动防御博弈对抗场景中，拦截飞行器需要在避免被防御飞行器碰撞同时靠近目标，其具体实现为：拦截飞行器会实时判断其与目标飞行器和防御飞行器之间的零控脱靶量，若某时刻拦截飞行器与防御飞行器之间的零控脱靶量小于杀伤半径，则拦截飞行器判断其受到了防御飞行器的拦截威胁，执行躲避策略；其余情况下，拦截飞行器执行拦截策略。拦截飞行器微分对策制导律计算公式如下：

$$u_{\mathrm{I}}=\begin{cases}\operatorname{sgn}\left[Z_{\mathrm{ID}}(t)\right] & Z_{\mathrm{ID}}(t)<k\\ \operatorname{sgn}\left[Z_{\mathrm{IT}}(t)\right] & Z_{\mathrm{ID}}(t)\geqslant k\end{cases}\tag{16}$$

3.2　有效性验证

场景中，敌方拦截飞行器具有6g机动过载能力，且拦截飞行器的控制响应优于目标飞行器，并采用如式（16）所示微分对策方法进行制导；防御飞行器具有3g机动过载能力，控制响应时间与拦截飞行器相当。主动防御博弈对抗数值仿真初始条件与约束具体设计如表1与表2所示。

表1　目标飞行器初始条件与约束

Tab.1　Initial conditions and constraints of the target

参数	参数值	参数	参数值
质量/kg	400	初始x轴坐标/km	0
攻角变化范围/(°)	−5~15	初始y轴坐标/km	55
最大攻角角速度/((°)·s^{-1})	10	初始x方向速度/(km·s^{-1})	3
气动参考面积/m^2	2	初始y方向速度/(km·s^{-1})	0

表2　防御飞行器与拦截飞行器初始条件

Tab.2　Defender and interceptor initial conditions

参数	防御飞行器	拦截飞行器
最大侧向加速度/g	3	6
系统响应时间/s	0.02	0.02
初始x轴坐标/km	5	100
初始y轴坐标/km	54.5	54.5~55.5
初始x轴速度/(km·s^{-1})	3	−2
初始y轴速度/(km·s^{-1})	0	0
杀伤半径/m	0.4	0.75

设定在场景中双方飞行器具备相对完善的信息获取与探测手段，先验知识准确且均能够获知对方飞行器的状态参数，在此条件下采用D3QN算法对智能体进行博弈对抗训练，设定算法超参数如表3所示。

表3　D3QN算法参数

Tab.3　D3QN algorithm parameters

超参数	数值
衰减系数	0.99
学习率	3×10^{-4}
记忆容量	2^{18}
批次尺寸	2^{10}
软更新系数	5×10^{-3}
贪婪度	0.1
记忆复用次数	2

值函数网络及其对应目标网络的网络结构相同，均由三层全连接层构成，节点数为1 024，并采用ReLU作为激活函数。设置奖励函数中的超参数如下：

$$\begin{cases}\alpha=4\\ \beta=2\end{cases}\tag{17}$$

整个训练过程训练总回合数5 000次，回合累计奖励值随训练回合数变化曲线如图4所示。

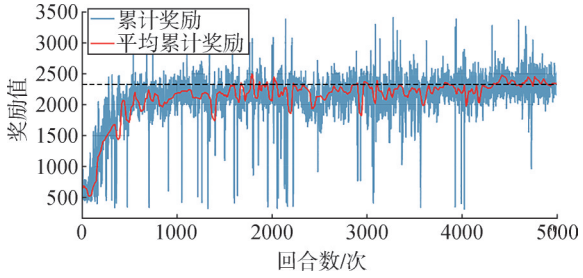


图4 强化学习模型训练过程

Fig.4 Reinforcement learning model training process

由图4可知，随着训练进行，回合累计奖励值呈现稳定但缓慢上升的趋势，并在约4 200回合之后奖励值稳定于2 400上下，智能体在该模型下训练的收敛性得到验证。经过5 000回合的训练平均回合累计奖励值曲线达到峰值2 476.73。

取最终训练模型进行200次蒙特卡罗打靶，目标飞行器逃逸成功率达到89.0%，脱靶量为1 141.67 m，初步证明所提出智能攻防博弈制导方法的有效性，采用本文所设计的智能博弈制导方法，目标飞行器在拦截飞行器机动能力与控制响应都占优的情况下可实现稳定逃逸。

图5、图6为博弈仿真结果，由图5与图6可知，在俯仰平面内，智能博弈对抗策略可描述为：在博弈回合开始后约17.5 s内，目标飞行器维持1g左右的弱机动以配合防御飞行器对拦截飞行器展开拦截；当拦截飞行器意识到自身受到拦截威胁并执行躲避策略时，目标飞行器开始采取更大的机动能力，朝着拦截飞行器来袭方向的反方向进行逃逸机动；最终，尽管拦截飞行器成功躲避了防御飞行器的拦截，但却无法及时拦截目标飞行器，使得目标飞行器成功逃逸。

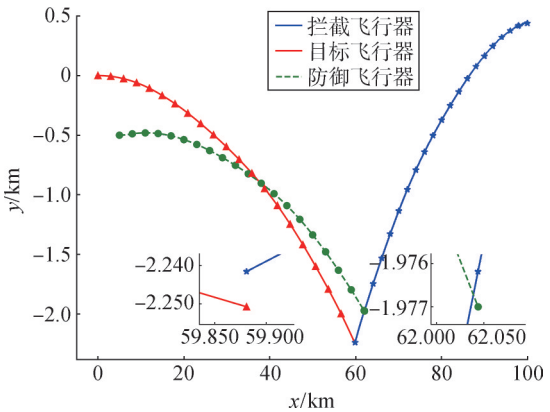


图5 飞行器运动轨迹

Fig.5 Aircraft motion trajectory

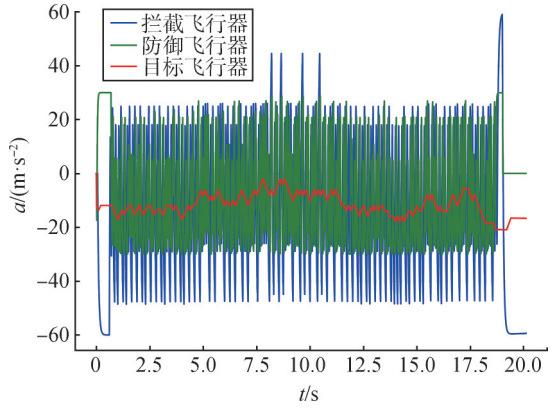


图6 飞行器侧向加速度

Fig.6 Lateral acceleration of vehicle

3.3 智能博弈制导方法效能分析

为证明所提出智能攻防博弈制导方法相比传统攻防博弈制导方法，在目标飞行器和防御飞行器机动能力不足的情况下具有更强的反拦截能力，本节设计目标飞行器采用4种程序式机动方法[20]：正弦机动、方波机动、阶跃机动和随机机动，具体机动策略如表4所示。

表4 目标飞行器程序机动策略

Tab.4 Target aircraft program maneuvering strategy

机动方法	制导律计算
正弦机动	$\alpha = \arg_{u2\alpha}\left[2\sin\left(2\pi/T\cdot t\right)g + \cos\theta_v g\right]$
方波机动	$\alpha = \arg_{u2\alpha}\left\{2\mathrm{sign}\left[\sin\left(2\pi/T\cdot t\right)\right]g + \cos\theta_v g\right\}$
阶跃机动	$\alpha = \alpha_{\min}$
随机机动	$\alpha = \arg_{u2\alpha}\left[2\mathrm{rand}\left(-1,1\right)g + \cos\theta_v g\right]$

注：a_y为目标飞行器侧向加速度；θ_v为目标飞行器速度倾角；$\cos\theta_v g$为重力平衡项；$\mathrm{rand}(-1,1)$为-1~1内的随机数；$\arg_{u2\alpha}(\cdot)$为过载控制量与攻角控制量之间的转换程序，基于目标飞行器气动特性搭建。

此外，考虑防御飞行器采用比例导引[7]方法进行制导，制导律计算方法如下：

$$u_{\mathrm{D}} = NV_C\dot{q} \tag{18}$$

式中 N为比例系数；V_C为防御飞行器与拦截飞行器之间的相对速度；$\dot{q}$为防御飞行器的视线角角速度。

博弈场景假设与第3.1节一致，为了更全面地探究所提出智能博弈方法相比传统博弈方法随拦截飞行器机动能力变化时的效能变化，考虑拦截飞行器机动能力2~8g，基于蒙特卡罗打靶法，对比每种工况下目标飞行器和防御飞行器采用传统方法与智能方法的目标飞行器逃逸成功率。具体打靶仿真结果如图7所示。

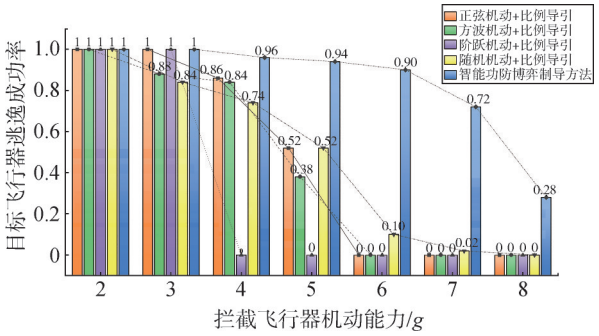


图7　采用传统博弈方法与采用智能博弈方法的博弈成功率

Fig.7　The success rate of traditional game method and intelligent game method

由图7可知，当拦截飞行器的机动能力为2~3g时，即与目标飞行器机动能力接近时，采用4种传统博弈方法与智能博弈方法都能实现较高的目标飞行器逃逸成功率。其中，“正弦机动+比例导引”和“阶跃机动+比例导引”被证明是较为有效的博弈策略。然而，当拦截飞行器机动能力逐步提升至6g，即目标飞行器机动能力的2~3倍时，传统博弈方法中仅有“随机机动+比例导引”实现了10%的逃逸成功率，而采用智能博弈方法目标飞行器仍有90%的逃逸成功率。进一步提升拦截飞行器的机动能力至7g时，采用传统博弈方法的逃逸成功率仅剩2%，然而采用智能博弈方法仍可保持72%的逃逸成功率，目标飞行器仍能稳定逃逸。当拦截飞行器的制导能力达到目标飞行器的约4倍时，传统博弈制导方法已无法成功逃逸，但采用智能博弈制导方法仍有28%的逃逸成功率。综上所述，在机动能力不足的情况下，目标飞行器与防御飞行器采用所提出的智能博弈方法相比传统博弈方法可以达到更高的博弈成功率。

需要指出的是，鉴于第1节中提出的合理假设与第3节中完成的仿真验证，在智能博弈对抗策略中，目标飞行器始终采用机动过载不超过2.5g的弱机动能力。这种要求在飞行器的横侧向运动中同样可以得到满足。因此，本文在纵向平面内对智能博弈方法的评估结果具有代表性，并且相关的验证结论可以推广应用于目标飞行器在横侧向平面的博弈对抗问题。飞行器在俯仰和横侧向平面内的耦合机动博弈策略也是未来值得深入研究的方向。

4　结束语

本文针对飞行器与拦截器末制导段的攻防博弈问题展开了研究，提出了一种飞行器主动防御博弈智能制导方法。为解决传统基于解析制导方法在弱机动情况下博弈成功率较低的问题，基于双竞争深度Q学习网络深度强化学习算法提出了一种飞行器主动防御博弈对抗智能制导方法，并利用奖励函数塑造方法，基于飞行器间零控脱靶量，设计了一种整型分数指数非稀疏奖励函数，提高强化学习算法收敛效率和训练稳定度。数值仿真结果表明，所提出的方法能够实现飞行器在机动能力不足情况下的博弈对抗成功，且相比于传统攻防博弈制导方法具有更高的博弈成功率。

参　考　文　献

[1] 谭毅伦, 闫杰. 针对高超声速飞行器的非线性动态逆最优控制[J]. 导弹与航天运载技术, 2011(1): 36-39.
TAN Yilun, YAN Jie. Nonlinear dynamics inversion optimal control for hypersonic vehicle[J]. Missiles and Space Vehicles, 2011(1): 36-39.

[2] SIOURIS G. Comparison between proportional and augmented proportional navigation[J]. Nachrichtentechnische Zeitschrift, 1974, 27(7): 278-280.

[3] SIOURIS G M. Missile guidance and control systems[M]. Berlin: Springer Science & Business Media, 2004.

[4] ANDERSON G M. Comparison of optimal control and differential game intercept missile guidance laws[J]. Journal of Guidance Control, 1981, 4(2): 109-115.

[5] 赵亮博, 朱广生, 张耀, 等. 智能飞行器追逃博弈中的关键技术及发展趋势[J]. 飞航导弹, 2021(12): 134-139.
ZHAO Liangbo, ZHU Guangsheng, ZHANG Yao, et al. Key technologies and development trends of intelligent aircraft pursuit and escape game[J]. Aerodynamic Missile Journal, 2021(12): 134-139.

[6] GAUDET B, LINARES R, FURFARO R. Deep reinforcement learning for six degree-of-freedom planetary landing[J]. Advances in Space Research, 2020, 65(7): 1723-1741.

[7] GAUDET B, FURFARO R. Missile homing-phase guidance law design using reinforcement learning[C]. Minneapolis: Proceedings of the AIAA Guidance, Navigation, and Control Conference, 2006.

[8] 张阳康, 孙晨, 泮斌峰. 行星软着陆GPS有模型强化学习制导方法[J]. 飞控与探测, 2021(5): 34-43.
ZHANG Yangkang, SUN Chen, PAN Binfeng. Guidance method of planetary soft landing with GPS model-based reinforcement learning [J]. Flight Control & Detection, 2021(5): 34-43.

[9] GAUDET B, FURFARO R, LINARES R. Reinforcement meta-learning for angle-only intercept guidance of maneuvering targets[J]. Aerospace Science and Technology, 2020, 99(4): 105746.

[10] RIEDMILLER M, HAFNER R, LAMPE T, et al. Learning by playing: solving sparse reward tasks from scratch[J]. Machine Learning Research, 2018, 80: 4344-4353.

[11] AINSWORTH M, SHIN Y. Plateau phenomenon in gradient descent training of RELU networks: explanation, quantification, and avoidance[J]. SIAM Journal on Scientific Computing, 2021, 43(5): 3438-3468.

[12] LIANG H, JIANYING W, YONGHAI W, et al. Optimal guidance against active defense ballistic missiles via differential game strategies[J]. Chinese Journal of Aeronautics, 2020, 33(3): 978-989.

[13] 王建华, 刘鲁华, 王鹏, 等. 高超声速飞行器纵向平面滑翔飞行制导控制方法[J]. 国防科技大学学报, 2017, 39(1): 58-66.
WANG Jianhua, LIU Luhua, WANG Peng, et al. Longitudinal integrated guidance and control scheme for hypersonic vehicle in glide phase[J]. Journal of National University of Defense Technology, 2017, 39(1): 58-66.

[14] 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): 761-773.

[15] HAN B A, YANG J J. Research on adaptive job shop scheduling problems based on dueling double DQN[J]. IEEE Access, 2020(8): 186474-186495.

[16] SEWAK M. Deep q network (DQN), double DQN, and dueling DQN: a step towards general artificial intelligence[M]. Berlin: Springer, 2019.

[17] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. Computerence, 2015. DOI: 10.48550/arXiv.1509.06461.

[18] 柳小桐. BP神经网络输入层数据归一化研究[J]. 机械工程与自动化, 2010(3): 122-126.
LIU Xiaotong. Study on data normalization in BP neural network[J]. Mechanical Engineering & Automation, 2010(3): 122-126.

[19] 张士熊, 刘新学, 李斌, 等. 基于微分对策的拦截末段突防导弹机动突防制导律研究[J]. 导弹与航天运载技术, 2015(2): 81-84.
ZHANG Shixiong, LIU Xinxue, LI Bin, et al. Study on maneuver penetration guidance law of ballistic missile based on differential games in the terminal of interception[J]. Missiles and Space Vehicles, 2015(2): 81-84.

[20] 鲜勇, 李少朋, 雷刚, 等. 弹道导弹中段机动突防技术研究综述[J]. 飞航导弹, 2015(9): 43-46.
XIAN Yong, LI Shaopeng, LEI Gang, et al. Review on midcourse maneuvering penetration technology of ballistic missiles[J]. Aerodynamic Missile Journal, 2015(9): 43-46.

作 者 简 介

倪炜霖（2000—），男，博士研究生，主要研究方向为飞行器智能博弈对抗技术。

刘佳琪（1963—），男，研究员，主要研究方向为雷达电子战和飞行器设计。

邵 节（1991—），男，高级工程师，主要研究方向为飞行器设计。

刘 鹏（1986—），男，高级工程师，主要研究方向为飞行器设计。

梁海朝（1986—），男，博士，教授，主要研究方向为先进飞行器智能与仿生博弈对抗、智能制导控制。