http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0375

针对高机动目标的深度强化学习智能拦截制导

张豪[1]，朱建文[1,2,*]，李小平[1]，包为民[1,3]

(1. 西安电子科技大学 空间科学与技术学院，西安 710126；2. 火箭军工程大学 导弹工程学院，西安 710025；
3. 中国航天科技集团有限公司，北京 100048)

摘 要：针对固定系数比例导引律在拦截高机动目标时脱靶量与能量损耗过大的问题，提出一种基于深度强化学习智能调参的自适应比例导引律。建立基于实时飞行状态的状态空间和包含横向、纵向导航比的动作空间，以及综合不同状态的奖励函数模型；在奖励函数的模型设计中,引入预测-校正方法，以提高动作评估的准确性；利用柔性演员-评论 (SAC) 算法，根据拦截器和目标的相对运动状态，训练获得综合考虑脱靶量和能量损耗的网络参数与制导参数决策系统。仿真结果表明：相较于传统比例导引律，所提制导策略可在保证低脱靶量的同时显著降低能量损耗，并且对未训练场景具有良好的适应性。

关 键 词：智能制导；机动目标；深度强化学习；比例导引；马尔可夫决策过程
中图分类号：V448; TJ765.3
文献标志码：A **文章编号**：1001-5965(2025)06-2060-10

高速机动目标的拦截制导问题一直是研究的热点和难点。比例导引律 (proportional navigation guidance, PNG) 以形式简单和易于实现的优点被广泛应用,其主要思想是通过零化视线角速率迫使拦截器和目标保持在碰撞三角形上。然而,随着飞行任务复杂化,制导性能指标已不仅是零脱靶量,还包括其他约束,如能量最优[1]。对于无机动目标,PNG 展现出优异的拦截性能,并在理论上已经证明导航比为 3 时能量最优[2]。而面对高速机动目标时,为了追求低脱靶量,常采用大导航比,导致拦截器对目标机动过于敏感,造成不必要的能量损耗。基于现代控制理论的制导律,如滑模制导[3]、最优制导[4]、微分对策制导[5] 和动态规划[6] 等,同样已用于解决该问题,但这些制导律往往需要剩余飞行时间或目标加速度等额外信息,尤其当考虑能量损耗时,制导指令会更加复杂。因此,研究具有能量最优的 PNG 是非常必要的。

作为比较成熟的制导律,已有许多文献对 PNG 及其变体进行了研究。文献 [7] 针对大气层外的机动目标,分析了现实真比例导引律对任意机动目标的捕获区域,并给出了拦截器不超过自身饱和过载下的比例系数范围。文献 [8] 提出了一种广义相对偏置比例制导律,以解决常系数比例法捕获域较小及碰撞角约束问题。然而,文献 [7-8] 均研究平面拦截问题,在实际应用时,需额外设计制导方案以保证两者共面,这将导致原制导律性能下降。因此,对三维 PNG 进行研究是非常重要的。文献 [9] 通过在旋转视线坐标系中建立相对运动方程,利用非线性输出调节方法研究了纯比例导引 (pure proportional navigation, PPNG) 在三维空间中的捕获能力。文献 [10] 讨论了三维 PPNG 中加速度方向对脱靶量的影响,并将其修正到最小化零控脱靶量的加速度方向,进一步增强 PPNG 的拦截性能。由于目标机动使拦截器和目标间相对运动变得异常复杂,加入其他约束将很难进行理论分析和求解。因此,对三维 PNG 的研究大多集中在最小化脱靶

收稿日期：2023-06-15；录用日期：2023-09-11；网络出版时间：2023-09-27 18:44
网络出版地址：link.cnki.net/urlid/11.2625.V.20230927.1317.001
*通信作者. E-mail：zhujianwen1117@163.com

引用格式：张豪，朱建文，李小平，等. 针对高机动目标的深度强化学习智能拦截制导 [J]. 北京航空航天大学学报，2025，51（6）：2060-2069. ZHANG H，ZHU J W，LI X P，et al. Deep reinforcement learning intelligent guidance for intercepting high maneuvering targets [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（6）：2060-2069（in Chinese）.

量上, 很少涉及能量损耗。

得益于人工智能技术的高速发展, 强化学习 (reinforcement learning, RL) 技术作为一种体现智能决策的算法, 得到了众多学者的认可。文献 [11-12] 将状态空间和动作空间离散化, 为实现零脱靶量提出了一种结合 Q-learning 和 PNG 的制导律, 但该制导律仅适用于低维空间, 且制导性能与离散程度相关。深度强化学习 (deep reinforcement learning, DRL) 结合了深度学习(deep learning, DL)的感知能力和 RL 的决策能力, 为解决连续高维空间的制导难题提供了新的可能。针对机动目标的拦截问题, 文献 [13] 基于信赖域策略优化 (trust region policy optimization, TRPO) 算法将飞行状态直接映射为制导指令, 并在训练时随机初始化拦截弹和目标状态, 以提高制导策略的泛化能力, 仿真表明, 该制导策略相比于 PNG 和增强比例导引律, 在训练场景和未学习场景下均具有更低的脱靶量。文献 [14] 将 TRPO 替换为近端策略优化 (proximal policy optimization, PPO) 算法, 并在目标函数中引入熵项, 进一步提高了拦截概率。文献 [15] 将加速度指令构造为零控脱靶量和待飞时间的函数, 利用深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 算法确定制导增益, 与飞行状态直接映射为加速度指令相比, 可以实现更快的收敛。然而, 文献 [13-15] 均研究二维拦截问题, 将其扩展到三维空间时, 势必因为飞行器纵向和侧向之间的耦合关系及动作维度的增加而增大训练难度。文献 [16] 研究了三维空间中仅有角度信息的拦截问题, 通过在 PPO 中引入循环网络实现了比增强零控脱靶量制导律更低的脱靶量, 但其奖励函数具有一定的稀疏性, 导致训练效率降低。可见, 尽管 DRL 已经取得了较好的成果, 但将其应用于三维空间拦截问题时, 由于动作维度增加及稀疏的奖励函数, 仍存在训练效率低的问题。

针对上述问题, 本文提出了一种基于柔性演员-评论(soft actor-critic, SAC)的三维自适应 PNG。相较于 DDPG, SAC 具有超参数少、训练稳定的优势, 同时具有比 TRPO 和 PPO 更高的数据利用率, 因此更加适合复杂制导问题。为此, 将相对运动模型构建为马尔可夫决策过程(Markov decision processes, MDP), 通过 DRL 将飞行状态直接映射为导航比, 并引入预测-校正方法以提高训练效率, 同时为了提高制导策略的适应性, 目标机动方式在每回合随机初始化。仿真结果表明, 提出的制导策略可以实现高精度的拦截, 并且相比于传统 PNG 具有更低的能量损耗。

1　问题描述与相对运动建模

1.1　拦截相对运动建模

三维空间中, 拦截弹和目标的相对运动关系可由图 1 描述, 其中, v_{M}和v_{T}分别为拦截弹和目标的速度, R为两者距离, θ和σ分别为速度倾角和速度方位角, λ_{D}和λ_{T}分别为视线高低角和视线方位角。

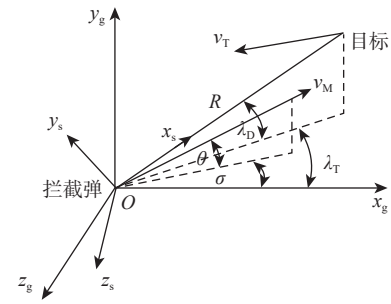


图 1　拦截弹-目标相对运动关系

Fig. 1　Interceptor-target relative motion relationship

由于末制导阶段拦截弹和目标相对距离近且飞行速度极快, 忽略两者速度损失, 假定v_{M}和v_{T}为常值。根据图 1 几何关系可得

$$\begin{cases} R=\sqrt{(x-x_{\mathrm{T}})^2+(y-y_{\mathrm{T}})^2+(z-z_{\mathrm{T}})^2} \\ \lambda_{\mathrm{D}}=\arcsin((y-y_{\mathrm{T}})/R) \\ \lambda_{\mathrm{T}}=\arctan((z-z_{\mathrm{T}})/(x-x_{\mathrm{T}})) \end{cases} \tag{1}$$

式中: (x,y,z)和$(x_{\mathrm{T}},y_{\mathrm{T}},z_{\mathrm{T}})$分别为拦截器和目标在地面系中的位置。

将式(1)两边对时间求导, 得相对运动方程为

$$\begin{cases} \dot{\lambda}_{\mathrm{D}}=\dfrac{\Delta\dot{y}R-\Delta y\dot{R}}{R\sqrt{R^2-\Delta y^2}} \\ \dot{\lambda}_{\mathrm{T}}=\dfrac{\Delta\dot{z}\Delta x-\Delta z\Delta\dot{x}}{\Delta x^2+\Delta z^2} \\ \dot{R}=\dfrac{\Delta x\Delta\dot{x}+\Delta y\Delta\dot{y}+\Delta z\Delta\dot{z}}{R} \end{cases} \tag{2}$$

式中: $\Delta x=x-x_{\mathrm{T}}$; $\Delta\dot{x}=\dot{x}-\dot{x}_{\mathrm{T}}$; Δy、$\Delta\dot{y}$、Δz和$\Delta\dot{z}$采用相同定义。

1.2　目标机动规律

目标在纵向和侧向分别采用正弦机动和余弦机动以增加突防能力:

$$\begin{cases} a_{\mathrm{D}}=-ng\sin(\omega t) & 0\leqslant t\leqslant t_{\mathrm{p}} \\ a_{\mathrm{T}}=-ng\cos(\omega t) & 0\leqslant t\leqslant t_{\mathrm{p}} \end{cases} \tag{3}$$

式中: n和ω分别为机动过载和机动频率; g 为重力加速度; $t_{\mathrm{p}}=2\pi/\omega$为机动周期。当飞行时间$t>t_{\mathrm{p}}$时, 目标不再进行机动, 即$a_{\mathrm{D}}=a_{\mathrm{T}}=0$。

1.3　比例导引律

三维制导律的经典设计方法是将三维运动解耦到两个互相垂直的平面, 即俯冲平面和转弯平

面，分别在2个平面内进行制导律求解，并在三维空间中进行综合。对PNG的改进，主要体现在比例系数和偏置项上，采用后者需要拦截器时刻计算偏置值，将非常考验弹载计算机的计算性能。因此，本文设计了如下自适应PNG：

$$\begin{cases}\dot{\gamma}_{\mathrm{D}}=K_{\mathrm{D}}\dot{\lambda}_{\mathrm{D}}\\ \dot{\gamma}_{\mathrm{T}}=K_{\mathrm{T}}\dot{\lambda}_{\mathrm{T}}\end{cases} \tag{4}$$

式中：$\dot{\gamma}_{\mathrm{D}}$和$\dot{\gamma}_{\mathrm{T}}$分别为速度方位角在俯冲平面和转弯平面内的角加速度；K_{D}和K_{T}分别为纵向和侧向导航比。采用DRL方法对其进行离线学习并在线修正。

2 基于SAC的制导系数智能调参

2.1 SAC智能调参建模

RL问题通常被构建为MDP，其通过状态、动作和奖励描述了智能体和环境的交互，通常被定义为一个五元组$<S,A,\boldsymbol{P},G,\gamma>$，其中，$S$为状态的集合，$A$为智能体动作的集合，$\boldsymbol{P}$为状态转移概率矩阵，$G$为奖励函数，$\gamma$为折扣率。在MDP中，智能体在时间$t$时刻观测到环境状态$\boldsymbol{s}_t\in S$，并根据策略$\Phi(\boldsymbol{a}_t|\boldsymbol{s}_t)$采取动作$\boldsymbol{a}_t\in A$作用于环境，环境将依据状态转移概率$P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)\in\boldsymbol{P}$反馈给智能体下一状态$\boldsymbol{s}_{t+1}$和奖励值$r(\boldsymbol{s}_t,\boldsymbol{a}_t)\in G$。$\gamma\in[0,1]$决定了未来奖励在当前时刻的折扣程度，$\gamma\to1$表示智能体更加注重未来奖励，$\gamma\to0$则优先考虑当前奖励。

2.1.1 状态空间设计

拦截弹根据状态空间中的信息感知当前状态并作出决策。繁杂的状态空间会导致智能体无法提取有用的信息，同时引起计算效率的下降，而状态信息的不足则会缩小各状态之间的差异，使智能体无法辨别，作出错误决策。因此，根据相对运动方程，设计了如式(5)所示的归一化状态空间，包括拦截弹和目标间的相对信息、拦截弹自身状态及目标的位置信息。

$$S=\left\{\bar{R},\bar{\lambda}_{\mathrm{D}},\bar{\lambda}_{\mathrm{T}},\bar{\theta},\bar{\sigma}\right\} \tag{5}$$

其中：

$$\begin{cases}\bar{R}=\dfrac{R}{R_0}\\ \bar{\lambda}_{\mathrm{D}}=\dfrac{2\lambda_{\mathrm{D}}}{\pi},\bar{\lambda}_{\mathrm{T}}=\dfrac{\lambda_{\mathrm{T}}}{2\pi}\\ \bar{\theta}=\dfrac{2\theta}{\pi},\bar{\sigma}=\dfrac{\sigma}{2\pi}\end{cases} \tag{6}$$

式中：R_0和R分别为初始时刻和当前时刻拦截弹和目标间的距离。

2.1.2 动作空间设计

DRL的目标是获得不同状态下的最优动作。根据制导需求，显然应将动作选为导航比K_{D}和K_{T}，即

$$A=\{K_{\mathrm{D}},K_{\mathrm{T}}\} \tag{7}$$

式中：$K_{\mathrm{D}}\in[0,K_{\mathrm{DMAX}}]$和$K_{\mathrm{T}}\in[0,K_{\mathrm{TMAX}}]$均为连续变量，$K_{\mathrm{DMAX}}$和$K_{\mathrm{TMAX}}$分别为纵向和侧向的最大导航比。

2.1.3 奖励函数设计

合适的奖励函数是成功应用DRL算法的关键，其不仅决定了算法优化的目标，同时也影响着算法收敛程度和速度。在本问题中，制导指标为脱靶量和能量损耗，其中，能量损耗以拦截器的总过载和表征。对于机动目标，解析给出其与飞行状态之间的关系是不现实的，若只在终端设置奖励，则会产生奖励稀疏现象，导致训练效率下降。为此，引入预测-校正方法为飞行过程中的决策提供过程奖励。在t时刻，利用策略网络计算状态$\boldsymbol{s}_t$下的导航比K_{D}和K_{T}，并以该导航比进行积分预测，根据预测的终端脱靶量和总过载和计算当前决策的过程奖励，最终将该奖励反馈给智能体，对导航比进行校正。过程奖励和终端奖励可用统一的奖励函数计算，如式(8)所示，由脱靶量项奖励R_{d}、能量损耗项奖励R_{n}及偏置项奖励R_{b}组成。若当前状态$\boldsymbol{s}_{\mathrm{c}}$为终端状态$\boldsymbol{s}_{\mathrm{e}}$，即$\boldsymbol{s}_{\mathrm{c}}=\boldsymbol{s}_{\mathrm{e}}$，根据式(8)计算得终端奖励，否则为过程奖励。

$$G=R_{\mathrm{d}}+R_{\mathrm{n}}+R_{\mathrm{b}} \tag{8}$$

脱靶量项奖励R_{d}为

$$R_{\mathrm{d}}=\begin{cases}k_{\mathrm{dp}}D_{\mathrm{p}} & \boldsymbol{s}_{\mathrm{c}}\neq\boldsymbol{s}_{\mathrm{e}}\\ k_{\mathrm{de}}D_{\mathrm{e}} & \boldsymbol{s}_{\mathrm{c}}=\boldsymbol{s}_{\mathrm{e}}\end{cases} \tag{9}$$

能量损耗项奖励R_{n}为

$$R_{\mathrm{n}}=\begin{cases}k_{\mathrm{np}}N_{\mathrm{p}} & \boldsymbol{s}_{\mathrm{c}}\neq\boldsymbol{s}_{\mathrm{e}}\\ k_{\mathrm{ne}}N_{\mathrm{e}} & \boldsymbol{s}_{\mathrm{c}}=\boldsymbol{s}_{\mathrm{e}}\end{cases} \tag{10}$$

偏置项奖励R_{b}为

$$R_{\mathrm{b}}=\begin{cases}k_{\mathrm{b}} & D_{\mathrm{e}}\leqslant D_{\mathrm{lim}}\\ 0 & D_{\mathrm{e}}>D_{\mathrm{lim}}\end{cases} \tag{11}$$

式中：D和N为脱靶量和总过载和，下标p和e分别表示预测和真实终端值；k_{dp}、k_{de}、k_{np}、k_{ne}和k_{b}为权重常数；D_{lim}为判定拦截成功的最大脱靶量。

由于过程奖励基于预测值计算，无法反映真实的终端信息，R_{d}和R_{n}中权重$k_{\mathrm{dp}}<k_{\mathrm{de}}$，$k_{\mathrm{np}}<k_{\mathrm{ne}}$。另外，设置偏置项奖励$R_{\mathrm{b}}$，当$D_{\mathrm{e}}\leqslant D_{\mathrm{lim}}$，即拦截成功时，给予正奖励$k_{\mathrm{b}}$，促使策略朝着低脱靶量方向优化。通过多次训练测试，各参数的具体值如表1所示。

2.2 SAC算法学习流程

SAC是集合了异策略、最大熵和演员-评论家的RL算法。作为异策略算法中的一员，SAC创建

表 1　奖励函数中的权重

Table 1　Weights in the reward function

k_{dp}	k_{de}	k_{np}	k_{ne}	k_{b}
−0.1	−0.5	−1	−4	100

一个大小为$|\mathcal{D}|$的经验缓存池来提高数据利用率，并在网络更新时，从中随机抽取M条经验进行反向传播，打破了数据之间的关联性。作为最大熵算法，SAC 的目的是寻找最优策略Φ^*，使 RL 中的累计奖励和熵的期望同时最大化，即

$$\Phi^* = \arg\max_{\Phi} \sum_t E_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\rho_\Phi}\left[\gamma^t\left(r(\boldsymbol{s}_t,\boldsymbol{a}_t)+\tau H(\Phi(\cdot|\boldsymbol{s}_t))\right)\right] \tag{12}$$

式中：τ为温度系数，决定了熵在策略优化时的重要性；$\rho_\Phi(\boldsymbol{s}_t)$和$\rho_\Phi(\boldsymbol{s}_t,\boldsymbol{a}_t)$分别为状态和状态-动作对在策略$\Phi$下的分布；熵度量了动作的不确定性，可根据式(13)计算：

$$H(\Phi(\cdot|\boldsymbol{s}_t)) = E_{\boldsymbol{a}\sim\Phi(\cdot|\boldsymbol{s}_t)}\left[-\log\Phi(\boldsymbol{a}|\boldsymbol{s}_t)\right] \tag{13}$$

熵项的引入大大提高了智能体对环境扰动的鲁棒性及对环境的探索能力，可有效防止过早陷入局部极小值点。SAC 采用了演员-评论家架构，即同时存在评价网络和策略网络，网络更新频率为N，具体的更新过程如图 2 所示，分为策略评估和策略提升 2 个步骤。

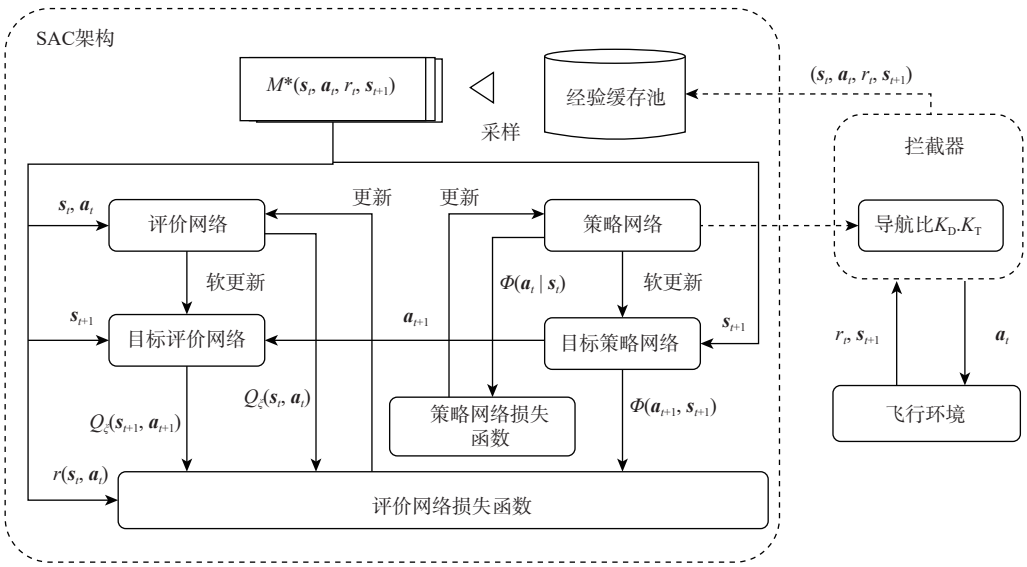


图 2　基于 SAC 的策略训练架构

Fig. 2　Strategy training architecture based on SAC

2.2.1　策略评估

策略评估是策略提升的基础，目的是获得真实的软动作价值函数。在 SAC 中，其被定义为奖励和熵和的期望，如下：

$$Q_\Phi(\boldsymbol{s},\boldsymbol{a}) = E_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\rho_\Phi}\left[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{s}_t,\boldsymbol{a}_t)+\tau\sum_{t=1}^{\infty}\gamma^t\mathcal{H}(\Phi(\cdot|\boldsymbol{s}_t))\right] \tag{14}$$

评价网络通过时序差分算法最小化损失函数J_Q来提高对式 (14) 的拟合程度：

$$J_Q(\boldsymbol{\xi}) = E_{(\boldsymbol{s}_t,\boldsymbol{a}_t,\boldsymbol{s}_{t+1})\sim\mathcal{D}\boldsymbol{a}_{t+1}\sim\pi_{\boldsymbol{\psi}}}\left[\frac{1}{2}\left(Q_{\boldsymbol{\xi}}(\boldsymbol{s}_t,\boldsymbol{a}_t)-Q'_{\boldsymbol{\xi}}(\boldsymbol{s}_t,\boldsymbol{a}_t)\right)^2\right] \tag{15}$$

式中：$Q'_{\boldsymbol{\xi}}(\boldsymbol{s}_t,\boldsymbol{a}_t)=r(\boldsymbol{s}_t,\boldsymbol{a}_t)+\gamma(Q_{\boldsymbol{\xi}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\tau\ln(\Phi_{\boldsymbol{\psi}}(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1})))$，$\boldsymbol{\psi}$和$\boldsymbol{\xi}$分别为策略网络和评价网络参数。

评价网络可通过随机梯度下降算法对其进行更新：

$$\boldsymbol{\xi}' = \boldsymbol{\xi}-\alpha\nabla J_Q(\boldsymbol{\xi}) \tag{16}$$

式中：α为学习率。

2.2.2　策略提升

策略提升将对现有策略Φ进行优化。由软动作价值定义和式 (12) 可知，最优策略中动作的概率分布应与真实的软动作价值分布相同。因此，将软动作价值化为玻尔兹曼分布形式，通过最小化两者的相对熵进行策略提升：

$$\Phi_{\mathrm{new}} = \arg\min_{\Phi'\in\Pi} D_{\mathrm{KL}}\left(\Phi'(\cdot|\boldsymbol{s}_t)\left\|\frac{\exp\left(\tau^{-1}Q^{\Phi_{\mathrm{old}}}(\boldsymbol{s}_t,\cdot)\right)}{Z^{\Phi_{\mathrm{old}}}(\boldsymbol{s}_t)}\right.\right) \tag{17}$$

式中：Π为可选策略的集合；$Z^{\Phi_{\mathrm{old}}}(\boldsymbol{s}_t)$为偏分函数。

为了降低方差，同时方便梯度回传，网络更新采用了重参数化技术，即将策略写为$\boldsymbol{a}_t=f_{\boldsymbol{\psi}}(\boldsymbol{\varepsilon}_t;\boldsymbol{s}_t)$，$\boldsymbol{\varepsilon}_t$为输入噪声[17]。化简式 (17) 可得用于策略网络的损失函数为

$$J_{\Phi}(\boldsymbol{\psi})=E_{s_t\sim\mathcal{D},a_t\sim\Phi_{\psi}}[\ln\Phi_{\psi}\left(f_{\psi}(\boldsymbol{\varepsilon}_t;\boldsymbol{s}_t)\middle|\boldsymbol{s}_t\right)-\tau^{-1}Q_{\xi}\left(\boldsymbol{s}_t,f_{\psi}(\boldsymbol{\varepsilon}_t;\boldsymbol{s}_t)\right)+\ln Z(\boldsymbol{s}_t)] \tag{18}$$

温度系数τ控制了熵的重要程度，由于策略的熵在训练过程中不断变化，固定的值将很难使算法收敛。为此，本文采用了文献[18]中改进的SAC，其将τ构造成带约束的优化问题：

$$\begin{cases}\max\limits_{\Phi_0,\cdots,\Phi_{T_f}} E\left[\sum\limits_{t=0}^{T_f}r(\boldsymbol{s}_t,\boldsymbol{a}_t)\right]\\ \text{s.t. }\forall t,\ H(\Phi_t)\geqslant H_{\text{m}}\end{cases} \tag{19}$$

式中：H_{m}为期望的最小策略熵。

通过求解式(19)，得到温度系数的损失函数$J(\tau)$，同样采用梯度下降方法对参数τ进行更新。

$$J(\tau)=E_{\boldsymbol{a}_t\sim\Phi_t}[-\tau\ln\Phi_t(\boldsymbol{a}_t|\boldsymbol{s}_t)-\tau H_{\text{m}}] \tag{20}$$

为降低网络参数剧烈波动引起的训练不稳定性，引入了目标网络和软更新技术。评价网络和策略网络对应的目标网络分别由$\boldsymbol{\xi}_t$和$\boldsymbol{\psi}_t$参数化，其网络参数通过软更新技术更新：

$$\begin{cases}\boldsymbol{\xi}_t'=(1-\epsilon)\boldsymbol{\xi}_t+\epsilon\boldsymbol{\xi}\\ \boldsymbol{\psi}_t'=(1-\epsilon)\boldsymbol{\psi}_t+\epsilon\boldsymbol{\psi}\end{cases} \tag{21}$$

式中：ϵ为平滑常数。ϵ越小，目标网络参数变化越缓慢，训练过程越稳定。

3 仿真分析与验证

为评估基于SAC制导策略的性能，对其进行训练测试。在初始时刻，拦截弹和目标在地面坐标系中的位置分别为(0, 25, 0) km和(20, 30, 20) km。拦截弹速度为1 500 m/s，初始时刻速度方向指向目标，最大过载30 g，最大导航比K_{DMAX}和K_{TMAX}均为10。目标在地面坐标系中的初始速度为(−2 000, 0, −1 500) m/s。积分步长为0.1 s，当两者距离小于350 m时，改为0.000 1 s。判定拦截成功的最大脱靶量D_{lim}为1 m。

策略网络和评价网络均采用具有3个隐藏层的全连接网络，每层的节点数如表2所示。输入层和隐藏层均采用ReLU作为激活函数，因为其不含指数项，可以有效提高计算速度，输出层无激活函数。

表2 策略网络和评价网络结构

Table 2 Architecture of the policy and critic networks

网络层	节关数	
	策略网络	评价网络
输入层	7	9
隐藏层1	256	256
隐藏层2	256	256
隐藏层3	256	256
输出层	4	1

SAC采用了随机策略，即策略网络输出每个动作的高斯分布的参数：均值和方差。根据前述动作空间设定，策略网络需要输出纵向和侧向导航比，故其输出层大小为4。同时，为了将输出值限定在预设范围内，先利用双曲正切函数压缩到[−1, 1]，再对其进行扩展。

3.1 最优导航比策略训练

针对本文中的拦截问题，利用SAC算法对制导策略进行训练。为了提高策略适应性，每回合目标机动的幅值、周期均随机改变。由于机动幅值越大，拦截成功率越低，对该场景设置更多回合数。机动幅值从[6, 7, 8, 9]序列中随机选取，每数被选择的概率是前一数的1.1倍。机动周期则在[6, 6.5, 7, 7.5, 8]中等概率选取。采样点数为10，即每回合策略网络给出10次导航比。最大训练回合数设为1 000。超参数的具体数值总结在表3中。

表3 超参数设置

Table 3 Hyperparameters setting

超参数	数值
经验缓存池大小$\lvert\mathcal{D}\rvert$	5 000
采样经验大小M	128
折扣率γ	0.98
学习率α	0.000 5
期望熵H_0	−5
温度系数τ	1
平滑常数ϵ	0.005
网络更新频率N	60

总奖励及其每50步的平滑曲线如图3所示。图4、图5和图6展示了脱靶量、总过载和及温度系数的变化。在训练前200次，由于温度系数较大，智能体更加注重探索，且此时网络参数随机初始化产生，因此，总奖励会大幅度振荡。之后，随着温度系数逐渐降低，总奖励开始收敛。500次后，总奖励维持在较高水平，脱靶量在0.05 m和0.35 m之

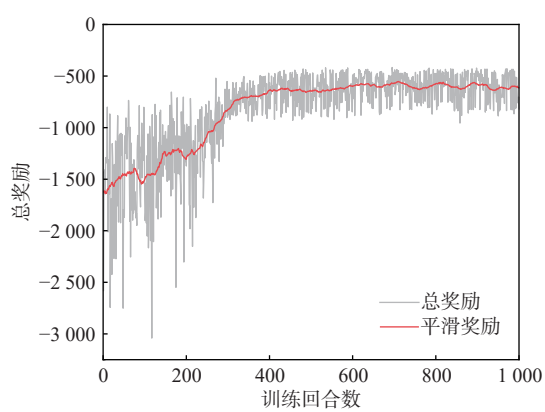


图3 总奖励曲线

Fig. 3 Curves of total reward

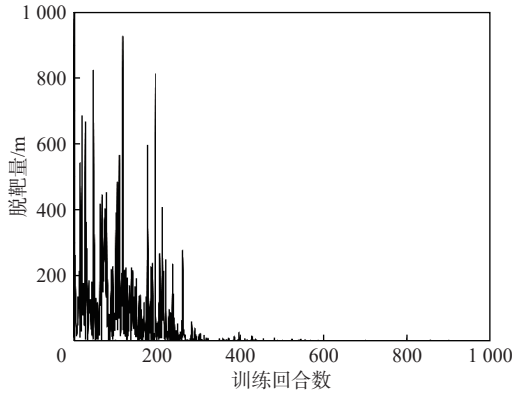


图 4　脱靶量曲线

Fig. 4　Curve of miss distance

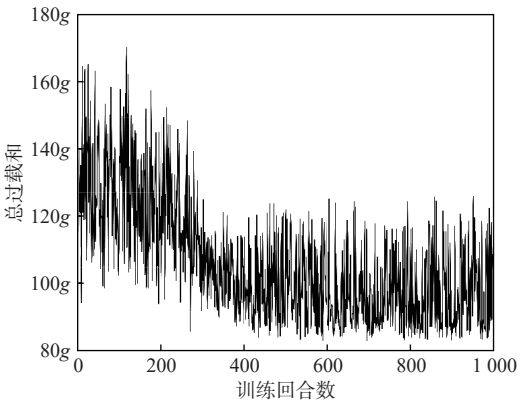


图 5　总过载和曲线

Fig. 5　Curve of total overload

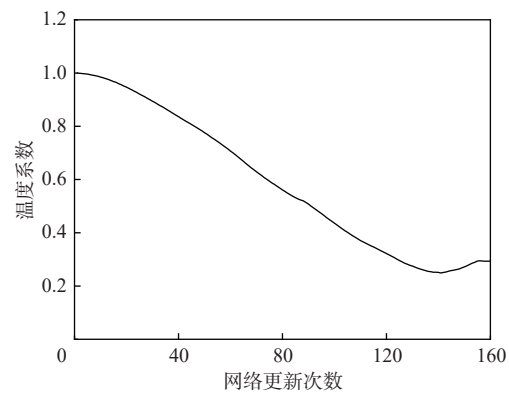


图 6　温度系数曲线

Fig. 6　Curve of temperature coefficient

间, 总过载和则因为目标机动形式改变而发生明显变化。

奖励稀疏性是影响训练效率的关键因素。图 7 对比了有无过程奖励下 4 000 次训练的终端奖励平滑曲线。可见, 当存在过程奖励时, 终端奖励从 500 次开始一直到训练后期均维持在最大值附近, 并且中间没有大幅度下降; 而无过程奖励时在 3 000 次左右才开始收敛, 并且两者收敛到相同值, 说明引入预测-校正方法不仅可以提高训练效率, 而且不会破坏策略的最优性。

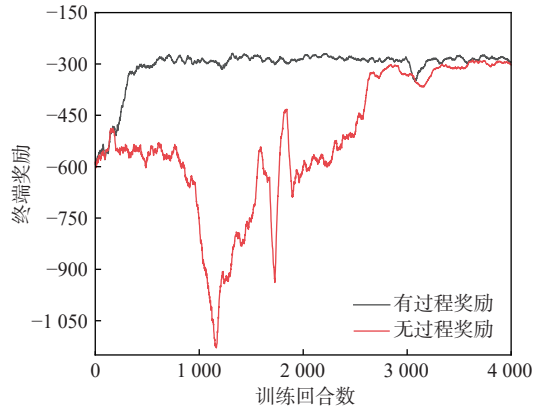


图 7　有无过程奖励下终端奖励曲线

Fig. 7　Terminal reward curves with or without process reward

3.2　制导性能对比分析

为了测试较提出制导策略 (G1) 的性能, 分别与导航比为 7, 8, 9, 10 的 PNG 和同样利用 SAC 算法训练的制导律 (G2) 进行对比分析, 其中 G2 仅考虑脱靶量, 即式 (8) 奖励函数中无能量损耗项R_n。

3.2.1　训练场景下的测试

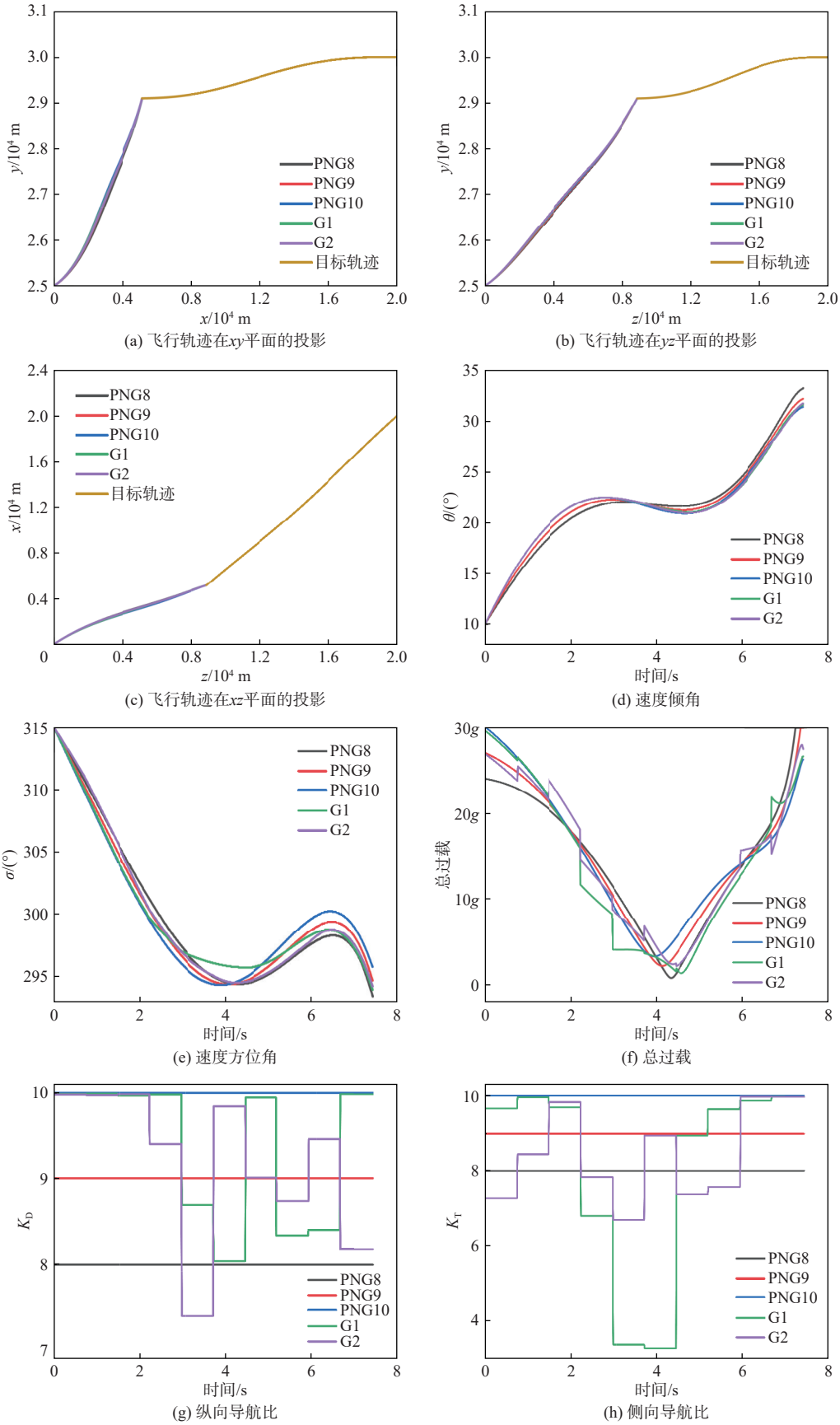
在训练过的场景下进行测试, 目标机动过载 $n=9$, 周期 $T_0=8$ s。拦截情况如表 4 所示, 其中, PNG7 表示拦截器纵向和侧向均采用导航比为 7 的 PNG, 其余类似。可见, 除导航比为 7 的 PNG 外, 其余制导律均可成功拦截目标, 并且导航比越大, 脱靶量和拦截时间越小, 但也会消耗更多能量。G1 中加入了对过载的惩罚, 实现低脱靶量的同时具有最低的能量损耗。

表 4　不同制导律下的终端参数

Table 4　Terminal parameters with different guidance laws

制导律	飞行时间/s	脱靶量/m	总过载和
PNG7	7.439	2.974	107.217 *g*
PNG8	7.435	0.780	108.323 *g*
PNG9	7.432	0.263	110.543 *g*
PNG10	7.430	0.113	112.474 *g*
G1	7.430	0.220	103.326 *g*
G2	7.432	0.248	109.267 *g*

图 8 和图 9 分别给出了拦截成功时各拦截器的飞行状态变化曲线和目标速度的变化曲线。由于目标在纵向做正弦机动, 其y_g向速度的绝对值在 $t=4$即半周期时达到最大, 之后降低, 反映在高度上即变化率先增大后减小, 而高度的变化也导致拦截器和目标间的纵向视线角速率变小, 根据式 (4), 拦截器的速度倾角θ将产生同样变化。速度方位角受目标纵向和侧向机动的双重影响, 在 4 s 后, 目标x_g向速度的绝对值先减小后增大, 而z_g向速度先增大后减小, 由 PNG 的物理意义可知, 拦截器的速度



(a) 飞行轨迹在xy平面的投影　(b) 飞行轨迹在yz平面的投影

(c) 飞行轨迹在xz平面的投影　(d) 速度倾角

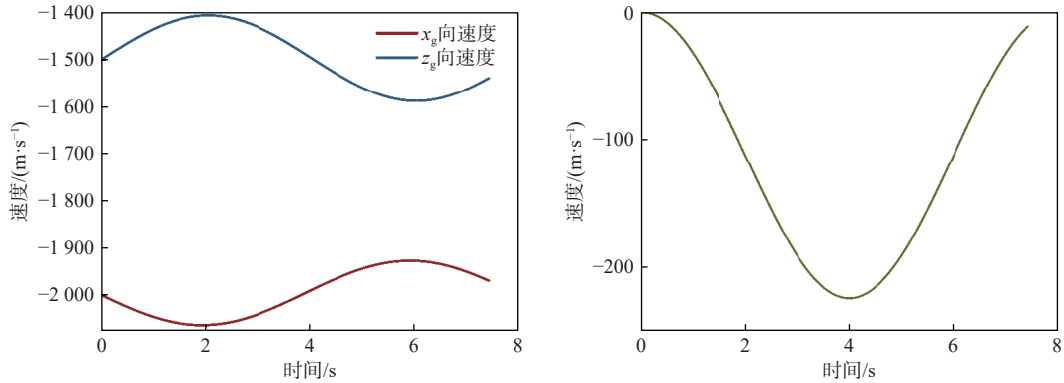
(e) 速度方位角　(f) 总过载

(g) 纵向导航比　(h) 侧向导航比

图 8　不同制导律下拦截器参数变化曲线

Fig. 8　Curves of interceptor parameters with different guidance laws

方位角曲线将先上升后下降。观察图 8(f) 可知，在目标加速度反向时，采用大导航比 PNG 的拦截器更加敏感，即具有最短的反应时间和最大的过载，可有效降低末端的拦截压力，防止过载饱和。结合



(a) 目标x_g和z_g向速度曲线　　(b) 目标y_g向速度曲线

图 9　目标速度曲线

Fig. 9　Target speed profile

图 8(g) 和图 8(h)，G1 会根据双方距离大小采用小导航比K_T以降低机动幅度，节省能量，而在末端为降低脱靶量则快速提高增大K_T。

3.2.2　未训练场景下的测试

本节将在 3 种未训练场景下测试制导策略的性能，以验证该策略的适应性。场景 1 对训练场景进行扩展，其中目标机动幅值和周期分别在 6～9 和 6～8 s 之间均匀选取。表 5 给出了上述各制导律在该场景下的 2 000 次蒙特卡罗仿真测试结果。可见，导航比为 7 的 PNG 脱靶量均值大于 1 m，意味着大多数情况下拦截失败。而本文提出的制导策略 G1 具有极低的脱靶量均值和方差，且和其余制导律相比具有最低的能量损耗。

表 5　未训练场景下蒙特卡罗结果统计

Table 5　Monte Carlo results statistics in untrained scenarios

制导律	脱靶量均值/m	脱靶量方差/m^2	总过载和均值
PNG7	1.793	2.504	105.983 g
PNG8	0.380	0.116	104.811 g
PNG9	0.192	0.012	105.242 g
PNG10	0.187	0.012	106.297 g
G1	0.189	0.012	98.532 g
G2	0.188	0.018	104.906 g

场景 2 在训练场景的基础上将正弦机动改为方波机动，机动方程如式 (22) 所示，当$t > T_0$时，$a_D = a_T = 0$。场景 3 则将拦截器最大过载降至 20g。

$$\begin{cases} a_D = -ng\mathrm{sgn}(\sin(\omega t)) & 0 \leqslant t \leqslant T_0 \\ a_T = -ng\mathrm{sgn}(\cos(\omega t)) & 0 \leqslant t \leqslant T_0 \end{cases} \tag{22}$$

图 10 和图 11 分别给出了目标方波机动和拦截器低过载时的终端参数。可见，在这两种场景下，G1 的最大脱靶量均在 PNG9 和 PNG10 之间，而能量损耗却是最小的。对场景 3 来说，降低拦截器最大过载导致末端没有充足的能力对目标机动作出反应，此时各制导律均有拦截失败的情况。

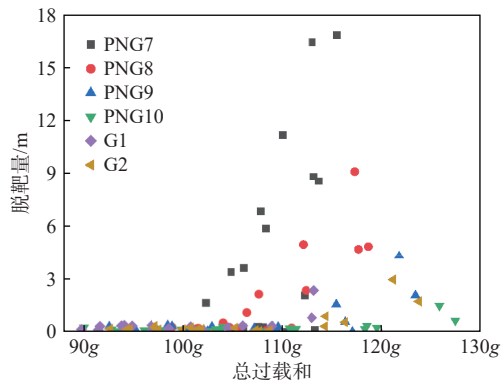


图 10　方波机动下终端参数

Fig. 10　Terminal parameters under square wave maneuvering

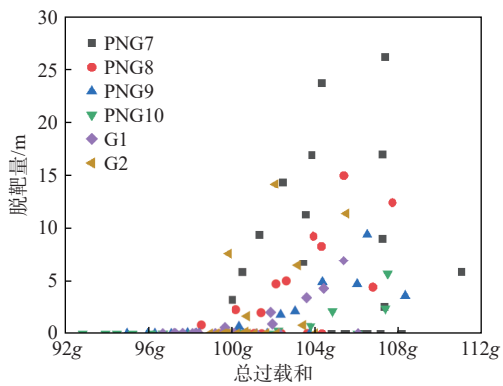


图 11　低过载下终端参数

Fig. 11　Terminal parameters under low overload

从训练场景和未训练场景中的仿真分析可见，采用本文基于 SAC 的 PNG，针对不同的机动方式、不同机动形式的目标及采用低过载拦截器均有较好的拦截性能，证明了所提制导策略的有效性和泛化性。

4　结　论

1）针对高机动目标拦截问题，同时考虑了脱靶量和能量损耗，提出了一种基于 DRL 的自适应

PNG。相较于传统PNG,在保证低脱靶量的同时,可降低能量损耗。

2)相较于平面拦截,三维空间中的拦截问题由于动作维度的增加和决策空间的复杂化导致严重的奖励稀疏现象,在构建奖励函数时引入预测-校正思想以提高训练效率。

3)仿真表明,在不同的目标机动形式、不同的机动周期和低可用过载等场景下,本文方法同样具有良好的性能,证明了该方法的适应性。

在后续的研究中,将考虑更加真实的拦截环境,同时进一步提高PNG性能。

参考文献(References)

[1] 纪毅, 王伟, 张宏岩, 等. 面向高机动目标拦截任务的空空导弹制导方法综述[J]. 航空兵器, 2022, 29(6): 15-25.
JI Y, WANG W, ZHANG H Y, et al. A survey on guidance method of air-to-air missiles facing high maneuvering targets[J]. Aero Weaponry, 2022, 29(6): 15-25(in Chinese).

[2] JEON I S, KARPENKO M, LEE J I. Connections between proportional navigation and terminal velocity maximization guidance[J]. Journal of Guidance, Control, and Dynamics, 2020, 43(2): 383-388.

[3] ZHANG B L, ZHOU D. Optimal predictive sliding-mode guidance law for intercepting near-space hypersonic maneuvering target[J]. Chinese Journal of Aeronautics, 2022, 35(4): 320-331.

[4] LI C D, WANG J, HE S M, et al. Collision-geometry-based generalized optimal impact angle guidance for various missile and target motions[J]. Aerospace Science and Technology, 2020, 106: 106204.

[5] YE D, SHI M M, SUN Z W. Satellite proximate interception vector guidance based on differential games[J]. Chinese Journal of Aeronautics, 2018, 31(6): 1352-1361.

[6] 郭建国, 胡冠杰, 郭宗易, 等. 天线罩误差下基于ADP的机动目标拦截制导策略[J]. 宇航学报, 2022, 43(7): 911-920.
GUO J G, HU G J, GUO Z Y, et al. ADP-based guidance strategy for maneuvering target interception under radome errors[J]. Journal of Astronautics, 2022, 43(7): 911-920(in Chinese).

[7] 白志会, 黎克波, 苏文山, 等. 现实真比例导引拦截任意机动目标捕获区域[J]. 航空学报, 2020, 41(8): 323947.
BAI Z H, LI K B, SU W S, et al. Capture region of RTPN guidance law against arbitrarily maneuvering targets[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(8): 323947(in Chinese).

[8] 王荣刚, 唐硕. 拦截高速运动目标广义相对偏置比例制导律[J]. 西北工业大学学报, 2019, 37(4): 682-690.
WANG R G, TANG S. Intercepting higher-speed targets using generalized relative biased proportional navigation[J]. Journal of Northwestern Polytechnical University, 2019, 37(4): 682-690(in Chinese).

[9] LI K B, SHIN H S, TSOURDOS A, et al. Capturability of 3D PPN against lower-speed maneuvering target for homing phase[J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(1): 711-722.

[10] SHIN H S, LI K B. An improvement in three-dimensional pure proportional navigation guidance[J]. IEEE Transactions on Aerospace and Electronic Systems, 2021, 57(5): 3004-3014.

[11] 张秦浩, 敖百强, 张秦雪. Q-learning强化学习制导律[J]. 系统工程与电子技术, 2020, 42(2): 414-419.
ZHANG Q H, AO B Q, ZHANG Q X. Reinforcement learning guidance law of Q-learning[J]. Systems Engineering and Electronics, 2020, 42(2): 414-419(in Chinese).

[12] 李庆波, 李芳, 董瑞星, 等. 利用强化学习开展比例导引律的导航比设计[J]. 兵工学报, 2022, 43(12): 3040-3047.
LI Q B, LI F, DONG R X, et al. Navigation ratio design of proportional navigation law using reinforcement learning[J]. Acta Armamentarii, 2022, 43(12): 3040-3047(in Chinese).

[13] 邱潇颀, 高长生, 荆武兴. 拦截大气层内机动目标的深度强化学习制导律[J]. 宇航学报, 2022, 43(5): 685-695.
QIU X Q, GAO C S, JING W X. Deep reinforcement learning guidance law for intercepting endo-atmospheric maneuvering targets[J]. Journal of Astronautics, 2022, 43(5): 685-695(in Chinese).

[14] CHEN W X, GAO C S, JING W X. Proximal policy optimization guidance algorithm for intercepting near-space maneuvering targets[J]. Aerospace Science and Technology, 2023, 132: 108031.

[15] HE S M, SHIN H S, TSOURDOS A. Computational missile guidance: a deep reinforcement learning approach[J]. Journal of Aerospace Information Systems, 2021, 18(8): 571-582.

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

[17] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018, 80: 1861-1870.

[18] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications[EB/OL]. (2019-01-29)[2023-06-01]. https://arxiv.org/abs/1812.05905v2.

Deep reinforcement learning intelligent guidance for intercepting high maneuvering targets

ZHANG Hao[1]，ZHU Jianwen[1, 2, *]，LI Xiaoping[1]，BAO Weimin[1, 3]

(1. School of Aerospace Science and Technology，Xidian University，Xi'an 710126，China；
2. College of Missile Engineering，Rocket Force University of Engineering，Xi'an 710025，China；
3. China Aerospace Science and Technology Corporation，Beijing 100048，China)

Abstract：An adaptive proportional navigation law with intelligent parameter adjustment by deep reinforcement learning is presented to address the issue of excessive miss distances and energy loss in the interception of moving targets using fixed coefficient proportional navigation law. First, a state space based on real-time flight states, an action space containing lateral and vertical gains, and a reward function model integrating different states is established. Meanwhile, a prediction-correction method is introduced to improve the accuracy of action evaluation in the model design of the reward function. Secondly, the soft actor-critic (SAC) algorithm is employed to train a network parameter and guidance parameter decision system that takes into account the miss distances and energy loss according to the relative motion states of the interceptor and the target. In comparison to traditional proportional navigation guidance, the simulation results demonstrate that the guidance technique has strong adaptability and can greatly minimize energy loss while retaining low miss distances.

Keywords：intelligent guidance；maneuvering targets；deep reinforcement learning；proportional navigation；Markov decision process

Received：2023-06-15；**Accepted**：2023-09-11；**Published Online**：2023-09-27 18：44
URL：link.cnki.net/urlid/11.2625.V.20230927.1317.001
* **Corresponding author**. E-mail：zhujianwen1117@163.com