Journal of System Simulation

Volume 35 | Issue 11 Article 10

11-30-2023

Imitative Generation of Optimal Guidance Law Based on Reinforcement Learning

Zhengxuan Jia
Beijing Simulation Center, Beijing 100854, China, danny2006_2007@126.com

Tingyu Lin
Beijing Simulation Center, Beijing 100854, China

Yingying Xiao
Beijing Simulation Center, Beijing 100854, China

Guoqiang Shi
Beijing Simulation Center, Beijing 100854, China

See next page for additional authors

Imitative Generation of Optimal Guidance Law Based on Reinforcement Learning

Abstract

Abstract: Under the background of high-speed maneuvering target interception, an optimal guidance law generation method for head-on interception independent of target acceleration estimation is proposed based on deep reinforcement learning. In addition, its effectiveness is verified through simulation experiments. As the simulation results suggest, the proposed method successfully achieves head-on interception of high-speed maneuvering targets in 3D space and largely reduces the requirement for target estimation with strong uncertainty, and it is more applicable than the optimal control method.

Keywords

reinforcement learning, optimal guidance, imitation learning, head-on interception, guidance and control

Authors

Zhengxuan Jia, Tingyu Lin, Yingying Xiao, Guoqiang Shi, Hao Wang, Bi Zeng, Yiming Ou, and Pengpeng Zhao

Recommended Citation

Jia Zhengxuan, Lin Tingyu, Xiao Yingying, et al. Imitative Generation of Optimal Guidance Law Based on Reinforcement Learning[J]. Journal of System Simulation, 2023, 35(11): 2410-2418.

第 35 卷第 11 期 系统仿真学报© Vol. 35 No. 11
2023 年 11 月 Journal of System Simulation Nov. 2023

基于强化学习的最优控制指令模仿生成方法

贾政轩[1]，林廷宇[1]，肖莹莹[1]，施国强[1]，王豪[2]，曾贲[2]，欧一鸣[1]，赵芃芃[1]
(1. 北京仿真中心，北京 100854；2. 北京电子工程总体研究所，北京 100854)

摘要：以高速机动目标拦截为问题背景，基于深度强化学习提出了一种不依赖目标加速度估计的逆轨拦截最优控制指令生成方法，并通过仿真实验进行了有效性验证。从仿真实验结果看，提出的方法实现了三维空间高速机动目标逆轨拦截并大幅削减了对带有强不确定性目标估计的要求，相比最优控制方法具有更强的适用性。

关键词：强化学习；最优制导；模仿学习；逆轨拦截；制导控制
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2023)11-2410-09
DOI: 10.16182/j.issn1004731x.joss.22-0632
引用格式: 贾政轩, 林廷宇, 肖莹莹, 等. 基于强化学习的最优控制指令模仿生成方法[J]. 系统仿真学报, 2023, 35(11): 2410-2418.
Reference format: Jia Zhengxuan, Lin Tingyu, Xiao Yingying, et al. Imitative Generation of Optimal Guidance Law Based on Reinforcement Learning[J]. Journal of System Simulation, 2023, 35(11): 2410-2418.

Imitative Generation of Optimal Guidance Law Based on Reinforcement Learning

Jia Zhengxuan[1], *Lin Tingyu*[1], *Xiao Yingying*[1], *Shi Guoqiang*[1], *Wang Hao*[2],
Zeng Bi[2], *Ou Yiming*[1], *Zhao Pengpeng*[1]
(1. Beijing Simulation Center, Beijing 100854, China; 2. Beijing Institute of Electronic System Engineering, Beijing 100854, China)

Abstract: Under the background of high-speed maneuvering target interception,*an optimal guidance law generation method for head-on interception independent of target acceleration estimation is proposed based on deep reinforcement learning.* In addition, its effectiveness is verified through simulation experiments. As the simulation results suggest, the proposed method successfully achieves head-on interception of high-speed maneuvering targets in 3D space and largely reduces the requirement for target estimation with strong uncertainty, and it is more applicable than the optimal control method.

Keywords: reinforcement learning; optimal guidance; imitation learning; head-on interception; guidance and control

0 引言

随着科学技术的不断进步，现代化战争正逐步演化为体系化攻防对抗，作为其中的重要火力打击手段之一，战术弹道导弹[1](tactical ballistic missile, TBM)再入大气层时，通常远高于防御武器的速度，为成功拦截防御带来了一定的困难。

针对以TBM目标为代表的高速机动目标再入速度高、机动能力强的特点，为有效对目标进行打击，需要采用带有较强末端角度约束的逆轨拦截方式。逆轨拦截是指防御武器接近目标时以反目标速度的方向迎击目标。这种打击方式一方面便于导引头截获和稳定跟踪目标，对于提高引战配合效率和战斗部杀伤概率极为有利；另一方面能够通过将导弹引入良好的弹目相对位置，形成

收稿日期：2022-06-09 修回日期：2022-08-03
第一作者：贾政轩(1990-)，男，工程师，硕士，研究方向为深度强化学习、建模仿真、智能决策等。E-mail：danny2006_2007@126.com

良好的拦截态势，削减高速目标机动拉开的防御武器能量需求缺口，降低低速导弹打击高速目标难度，达到更好的作战效果。

目前，已有针对高速目标打击的制导律生成方法问题的大部分研究是围绕比例制导[2-6]、滑模制导[5-6]和最优制导[7-10]等方法进行制导律设计。在这些研究中，研究人员均对弹目交会时刻的期望角度约束进行了考虑，并通过虚拟目标导引[2]、指令反馈[4]、寻优约束条件[7-9, 11-12]等手段对制导过程终点的交会角度限制进行了约束，进而实现在弹目交会时对目标的逆轨拦截，达到较好的打击效果。

上述研究大多从理论分析出发，均显式或隐式地引入了前提假设，从而在其方法的应用上带来了不同程度的限制，特别是对目标与导弹运动能力的限制。例如，部分文献假设打击的目标为固定目标[2, 3, 5]或匀速运动目标[7]，或者假设导弹运动速度不变[2-5]。这为所设计的制导律带来了较大的限制约束。文献[5-6]均考虑了对高速机动目标打击问题，但均围绕二维平面内的制导律生成问题开展研究。文献[8-9]在三维空间对高速机动目标的逆轨拦截问题进行了研究，分别采用最优制导与高斯伪谱法进行制导律设计，得到了较好的逆轨拦截效果。然而在制导律设计中，文献[8-9]均引入了目标运动加速度作为设计输入。面向目标机动样式和探测信息均存在不确定性的场景，上述方法对探测与估计部分提出了更高的要求。

近年来，由深度学习[13-14]崛起助推的新一代人工智能技术掀起了新一轮智能化浪潮，相继攻克视频游戏[15]、围棋[16]、星际争霸II[17]等，揭示了其在复杂问题求解方面的巨大潜力，也为解决上述问题的求解带来新的契机。当前，基于深度强化学习的智能制导律生成方法方面已有部分研究[6,18-20]，但是，一方面现有研究中较多所采用的拦截导弹模型较为简化，一些研究甚至未考虑气动部分的模型[20-21]；另一方面，现有研究也鲜有聚焦于高速目标逆轨拦截问题。

针对此，本文提出一种基于深度强化学习的最优控制律模仿生成方法，以弹目相对运动及导弹自身动力学状态为输入，以模仿最优导弹导引控制(即最优制导)过程为奖励生成函数，借助深度神经网络强非线性拟合能力，构建面向三维空间高速机动目标逆轨拦截的导弹导引控制指令(即制导指令)生成。该方法具有如下特点：①可实现在三维空间中对目标的逆轨拦截；②可实现对高速机动目标的逆轨拦截；③制导指令生成过程不再需要以目标加速度作为决策输入，从而大幅削减了对带有强不确定性目标估计的要求，具有更好的适用性。

1 最优制导策略模仿生成方法

面向三维空间高速机动目标拦截的导弹最优制导策略模仿生成方法，主要包括以下几个部分：最优制导指令模仿生成框架、状态空间模型构建、决策空间模型构建、奖励函数设计、训练算法设计。

1.1 最优制导指令模仿生成框架

本文的核心思想在于借助对最优制导律的模仿，基于有限的弹目状态数据，建立有效打击三维空间高速机动目标的制导指令与状态观测之间的映射，实现不依赖目标加速度估计前提下对目标进行有效打击。

针对此目标，在现有方法中可采用模仿学习方法实现。模仿学习[22]方法主要从专家示例数据中进行学习，借助专家示例数据教会智能体在复杂场景，特别是包含一些难以进行引导信号量化描述的场景下进行决策。当前，模仿学习方法主要包括行为克隆类方法[23]和对抗式模仿学习类[24]方法，形成两类模仿学习框架，前者通过最小化智能体策略与专家策略动作分布差异实现对专家策略的模仿，后者则基于专家数据构建奖励函数[25]，进而通过最大化该奖励函数实现对专家策略的模仿。在这两类框架中，专家策略数据通常

DOI: 10.16182/j.issn1004731x.joss.22-0632

采用离线收集的方式构建，然而在本文场景下，即便采用不同的场景，由于专家策略(即最优制导律)是确定的，能够获取的用于智能体训练的数据仍然非常有限。尽管在行为克隆类方法中有研究人员提出了DAgger方法[26]，但其仍然难以规避其“策略分布匹配”带来的复合误差。

基于此，本文以对抗式模仿学习类算法思想为基础，融入DAgger方法思想，提出一种基于深度强化学习的最优制导律模仿生成方法，其不同于行为克隆方法和逆强化学习方法，本文方法以深度学习框架为基础，通过奖励函数引导智能体对最优制导律进行模仿。总体框架如图1所示。

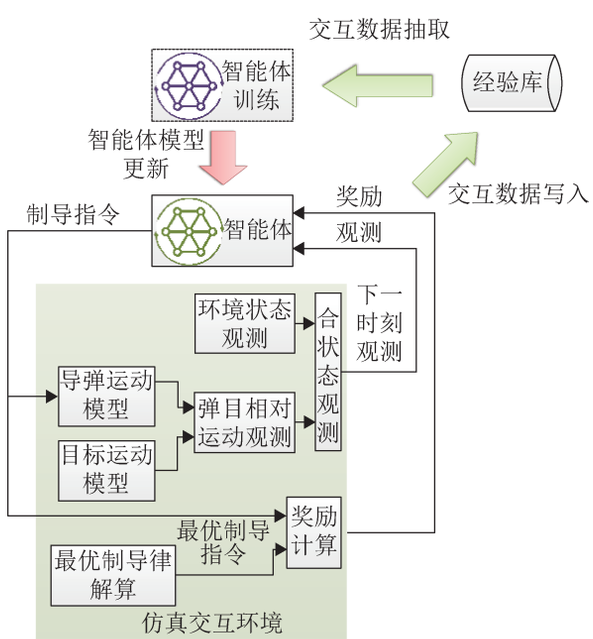


图1 基于深度强化学习的最优制导律模仿生成方法总体框架

Fig. 1 Framework of imitative generation method of optimal guidance law based on deep reinforcement learning

具体而言，如图1所示，在仿真交互过程中，在每个仿真时刻，智能体基于其当前网络参数与合状态观测输入生成制导指令分布，并通过采样给出制导指令，该指令传入导弹动力学模型调整导弹运动状态，进而改变弹目状态，得到下一时刻合状态观测。同时，仿真交互环境基于最优制导律进行解算，通过比较智能体给出的制导指令和最优制导律给出的制导指令解算得到奖励函数值，从而得到当前时刻合状态观测、智能体制导指令、最优制导律制导指令、奖励函数值、当前轮仿真是否结束以及下一时刻合状态观测构成的训练数据，存入经验库。

从上述过程可以看出，由于采用深度强化学习为基础框架对最优制导过程进行模仿，与行为克隆方法不同，本文提出方法在对每个时刻制导指令生成的学习中，考虑的是从当前时刻到打击结束整个过程中模仿的效能，而不是在局部某个时刻上的模仿效果。这种学习模式能够形成对整条弹道性能的综合考虑，且弹道制导指令具有较好的连续性和平滑性。

智能体训练过程则与普通深度强化学习过程相同，基于从经验库采样抽取的训练数据批，采用梯度下降类方法，以最大化累积奖励为目标，对智能体网络参数进行更新。

1.2 状态空间模型构建

在前一节，本文给出了三维空间高速机动目标打击的最优制导律模仿生成方法总体框架；智能体基于从仿真交互环境获得的状态观测，生成当前时刻导弹的制导指令。区别于实际状态，状态观测仅为基于实际观测手段获得的部分真实状态或其融合值。事实上，在实际任务中，导弹和目标的所有状态信息并非都能够毫无偏差地完全掌握，制导控制系统往往需要基于导引头提供的弹目相对位置关系及惯测系统提供的自身位姿信息，根据所设计的导引规律，进行制导指令解算。例如，比例导引规律即基于视线角变化率进行制导指令解算。因此，本节将对深度强化学习框架下的状态信息观测模型进行构建，智能体所生成的制导指令将基于该观测模型的观测结果给出。

基于导引头及惯测系统能力，本文选择如表1所示测量变量作为观测变量提供给智能体用于制导指令生成。

表1 测量变异汇总

Table 1 Summary of measured variables

变量	含义	变量	含义
$(r_x \quad r_y \quad r_z)$	弹目坐标差在地心坐标系三轴上投影	ψ_{m}	弹道偏角
$(v_{\mathrm{rx}} \quad v_{\mathrm{ry}} \quad v_{\mathrm{rz}})$	弹目速度差在地心坐标系三轴上投影	$\theta_{\mathrm{m}xy}$	弹道倾角在纵向平面内投影
q_1	纵向平面视线角	$(v_{\mathrm{mx}} \quad v_{\mathrm{my}} \quad v_{\mathrm{mz}})$	导弹速度在地心坐标系三轴上投影
q_2	横向平面视线角	ρ	局部大气密度
θ_{m}	弹道倾角	v_{s}	局部声速

从上述观测状态信息可知，本文方法中，智能体用于决策的观测输入信息均来自于可以直接测量的数据，而不包含类似目标加速度等需要进行复杂估计的量，从而在针对机动目标打击时具有更好的适用性。

1.3 决策空间模型构建

考虑到最优制导律给出的制导指令是导弹在垂直速度矢量平面上两轴方向上的加速度值，是 $\mathbb{R}^2$ 空间上的实数点，为实现对最优制导律的良好模仿，本文采用多元高斯分布对决策空间进行建模，并使用神经网络模型对分布参数进行拟合。

具体而言，以神经网络对式(1)所示的多元高斯分布概率密度函数中的均值向量 $\boldsymbol{\mu}_{\theta_1}$ 和协方差矩阵 $\boldsymbol{\Sigma}_{\theta_2}$ 进行拟合，构建决策分布。

$$\boldsymbol{\pi}_{\theta=[\theta_1,\theta_2]}(x)=\frac{1}{(\sqrt{2\pi})^n\det\left(\boldsymbol{\Sigma}_{\theta_2}\right)^{1/2}}\cdot \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu}_{\theta_1})^{\mathrm{T}}\boldsymbol{\Sigma}_{\theta_2}{}^{-1}(x-\boldsymbol{\mu}_{\theta_1})\right) \tag{1}$$

为简化网络模型训练复杂度，本文假设导弹在法向两轴上的加速度值彼此独立，则式(1)中协方差矩阵 Σ_{θ_2} 为对角矩阵。在此假设下，决策分布的概率密度函数为

$$\boldsymbol{\pi}_{\theta=[\theta_1,\theta_2]}(x)=\frac{1}{(\sqrt{2\pi})^n\left(\prod_{i=1}^{n}\left(\boldsymbol{\Sigma}_{\theta_2}\right)_{ii}\right)^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(x_i-(\boldsymbol{\mu}_{\theta_2})_i\right)^2}{\left(\boldsymbol{\Sigma}_{\theta_2}\right)_{ii}}\right)= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\left(\boldsymbol{\Sigma}_{\theta_2}\right)_{ii}}}\exp\left(-\frac{1}{2}\frac{(x_i-(\boldsymbol{\mu}_{\theta_1})_i)^2}{\left(\boldsymbol{\Sigma}_{\theta_2}\right)_{ii}}\right) \tag{2}$$

1.4 奖励函数设计

在本文所提出的框架下，智能体基于不同弹目状态对最优制导律给出指令的拟合与生成是通过奖励函数引导实现的。因此，为实现对最优制导律给出指令更好地拟合与学习，需对奖励函数模型进行详细设计。

在本文问题场景下，智能体需对最优制导律给出的两个轴的加速度指令进行同步学习，因此，需在奖励函数中对两个轴的加速度指令同步予以响应。基于此，本文中面向最优制导律模仿的奖励函数设计为

$$r=1-\frac{\left|a_y-a_{1y}\right|+\left|a_z-a_{1z}\right|}{4a_{\max}} \tag{3}$$

式中：a_y 为智能体给出的 y 轴方向加速度；a_z 为智能体给出的 z 轴方向加速度；a_{1y} 为最优制导律给

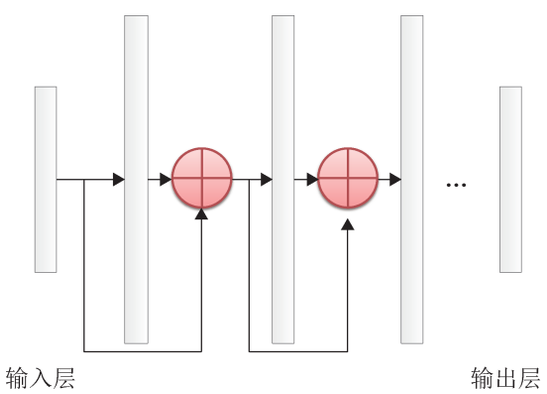
DOI: 10.16182/j.issn1004731x.joss.22-0632

出的y轴方向加速度；a_{1z}为最优制导律给出的z轴方向加速度；$a_{\max}$为y方向和z方向导弹的加速度绝对值的最大值。

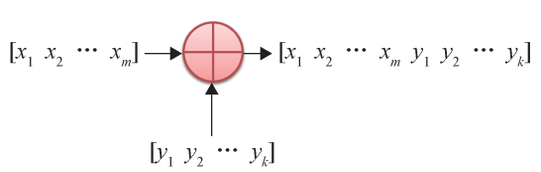
在此奖励函数下，通过深度强化学习可实现对两轴加速度的同步学习，而不会出现在一个轴上学习较好而在另一个轴上学习较差的问题。

1.5 训练算法设计

训练算法设计主要包括智能体网络结构设计与智能体训练算法设计。在智能体网络结构设计方面，为使特征信息得到充分利用与挖掘，本文采用类似于残差结构的短接合并方式搭建网络结构，如图2所示。



(a) 智能体神经网络层设计



(b) 智能体神经网络单层输入输出融合设计

图2 智能体结构设计

Fig. 2 Proximal Policy optimization with covariance matrix adaptation

在训练算法设计方面，本文采用PPOCMA[27] (proximal policy optimization with covariance matrix adaptation)算法进行实现。具体而言，相比PPO (proximal policy optimization)算法，策略网络更新过程调整为

$$\mathcal{L}_\theta=-\mathbb{E}_{s\sim\rho_s,\ a\sim\pi_\theta}A^\pi(s,a)\ \ln\left(\pi_\theta(s|a)\right) \tag{4}$$

式中：$\mathbb{E}$为取期望；ρ_s为状态s的分布；π_θ为动作a的分布；$A^\pi(s,a)$为策略π下在状态s选择动作a时的优势函数值。同时，基于假设优势函数在当前均值策略附近呈现线性特征，将负值优势函数连同其状态–动作对以当前均值策略为中心进行对称，得到正值优势函数及其对应的状态–动作对，实现其中有效信息的利用。具体而言，对应于(s_i,a_i,A_i)的状态–动作–优势函数组，以$\mu(s_i)$为对称中心进行对称映射，从而得到新的$(s_i,2\mu(s_i)-a_i,-A_i\varphi(s_i,a_i))$状态–动作–优势函数组。最终以该状态–动作–优势函数组为替换数据训练策略网络参数。

2 仿真实验分析

在前述章节中，本文提出了基于深度强化学习的最优制导律模仿生成方法。本节将通过仿真实验，从交会精度、交会角度、空间轨迹以及制导指令等方面本文方法与最优制导律的对比，对本文提出方法的有效性进行验证。

本文仿真实验中目标与导弹的数学模型、各仿真参数以及目标运动场景采用文献[8]中的仿真场景及参数，此处不再进行赘述，可具体参考文献[8]中的详细论述。

在上述仿真实验配置下，采用本文方法进行智能体训练，所得回合累积奖励曲线如图3所示。

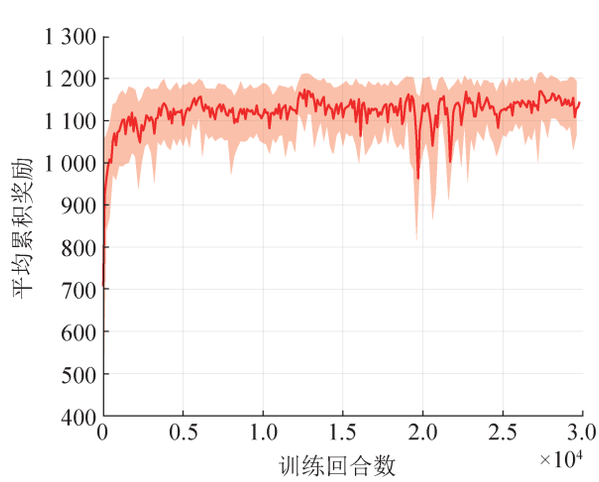


图3 本文方法智能体训练累积奖励曲线

Fig. 3 Accumulated episode reward curve of agent

从图 3 累积奖励曲线随训练演化过程可知，本文方法构建的智能体能够对最优制导律进行有效学习。同时，该指标为单轮仿真累积奖励，其表征了在一个仿真轮次(回合)中的不同状态上，本文方法训练生成的智能体均能对最优制导律方法进行较好地模仿，实现对整条弹道性能的综合平衡考虑。

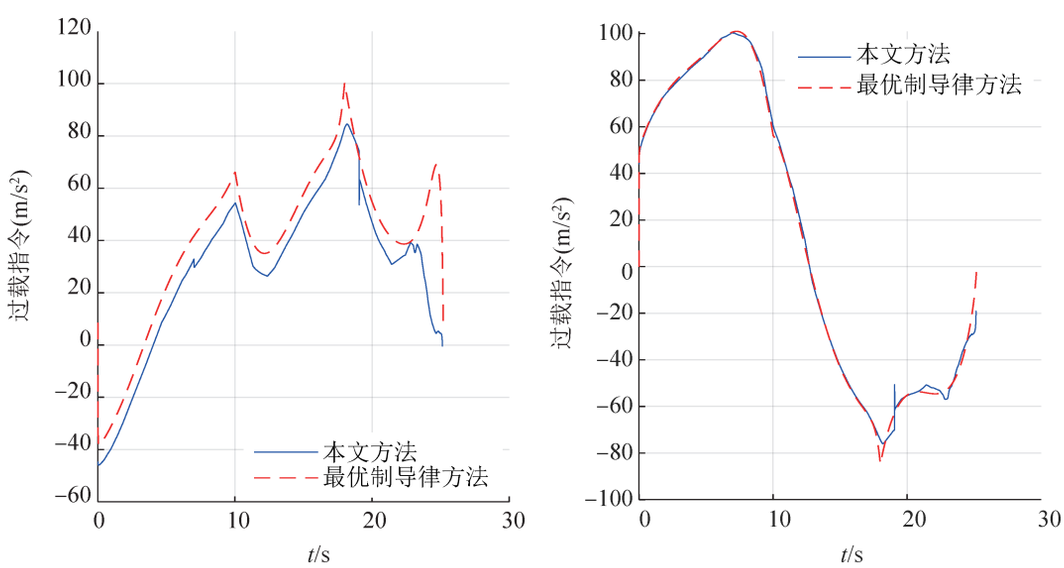
表 2 所示结果为本文方法与最优制导律方法所得的制导律性能对比。从表中结果可知，采用本文方法对最优制导律方法进行模仿，可以达到与最优制导律方法接近的制导律性能水平。具体而言，在末时刻弹目距离与脱靶量方面，最优制导律方法在脱靶量上比本文方法略好，但在交会时刻弹目距离上，本文方法具有更好的性能。与此类似，在交会角偏差方面，交会角偏差 1 为(θ_t+θ_m，交会角偏差 2 为 $\psi_t-(\psi_m-180°)$。本文方法在弹道倾角偏差上较最优制导律方法更小，而在弹道偏角偏差上则更大。从上述结果可以看出，在机动目标打击中，本文方法在交会时刻弹目距离、脱靶量、交会角偏差上均能够达到与最优制导律方法持平的性能水平，验证了本文方法的有效性。

表 2　制导律性能对比

Table 2　Performance comparison of different guidance laws

方法	末时刻距离/m	脱靶量/m	交会角偏差 1(°)	交会角偏差 2(°)
最优制导律方法[8]	11.210 5	0.582 8	−1.438 4	0.128 8
本文方法	3.424 4	3.237 8	0.894 7	−1.417 7

图 4~6 分别给出了本文方法与最优制导律方法在制导指令和空间轨迹方面的对比。制导指令对比结果方面，本文方法训练的智能体在 y 和 z 两轴方向上均可形成对最优制导律方法的较好模仿。尽管在 y 方向上存在一定的差异，但在最终行为和性能表现上相差不多，且总体上变化幅度较小。从轨迹曲线结果可知，本文方法训练的智能体所生成的制导指令制导下导弹的飞行轨迹与采用最优制导律生成的制导指令制导下导弹的飞行轨迹基本重合，验证了本文方法可对最优制导律方法形成较好的模仿，实现对目标的有效打击。



(a) y向制导指令　　(b) z向制导指令

图 4　本文方法与最优制导律方法制导指令对比

Fig. 4　Comparison of guidance signal generated by proposed method and optimal guidance law

DOI: 10.16182/j.issn1004731x.joss.22-0632

第 35 卷第 11 期 系统仿真学报 Vol. 35 No. 11
2023 年 11 月 Journal of System Simulation Nov. 2023

图7给出了本文方法与最优制导律方法制导下导弹弹道倾角与弹道偏角曲线对比。从图中所示结果可知，在对机动目标进行拦截过程中，本文方法制导下导弹的弹道偏角与最优制导律方法完全重合，而弹道倾角仅在末时刻出现差异(且差异值在5°以内)，验证了本文方法对导弹的制导能力可以达到与最优制导律方法相近，甚至相同的水平。

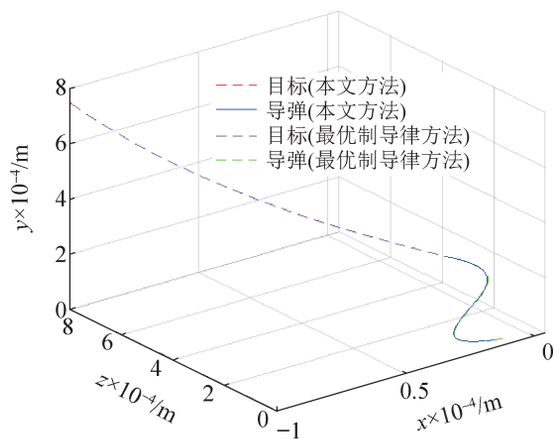
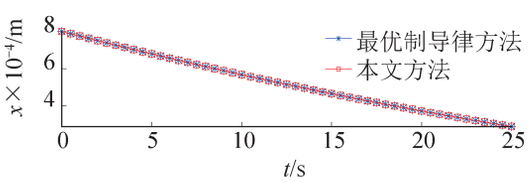
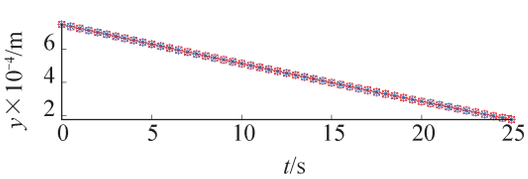


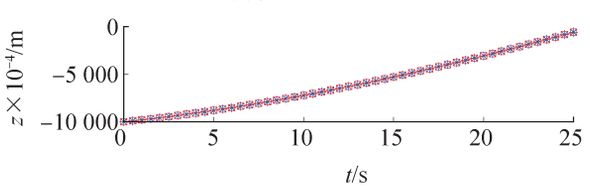
图5 本文方法与最优制导律方法空间轨迹对比(1)
Fig. 5 Comparison of trajectories generated by proposed method and optimal guidance law (1)



(a) x轴方向弹目相对位移

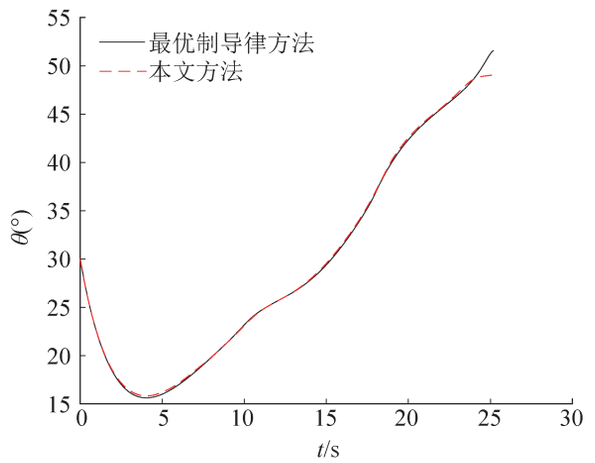


(b) y轴方向弹目相对位移

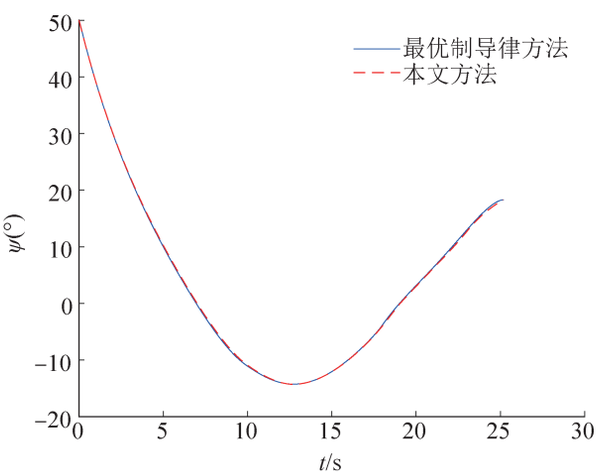


(c) z轴方向相对位移

图6 本文方法与最优制导律方法空间轨迹对比(2)
Fig. 6 Comparison of trajectories generated by proposed method and optimal guidance law (2)



(a) 导弹弹道倾角



(b) 导弹弹道偏角

图7 本文方法与最优制导律方法弹道倾角与弹道偏角对比
Fig. 7 Comparison of missile inclination angle and deflection angle by proposed method and optimal guidance law

图8给出了本文方法与最优制导律方法制导下导弹的速度曲线对比。从该结果中可以看到，本文方法与最优制导律方法制导下，导弹的行为基本一致。此外，在末时刻，本文方法制导下，导弹具有更高的速度，即更高的动能，可带来更好的气动操纵性。因此，尽管脱靶量相比最优制导律方法略差，但本文方法达到了更小的末时刻弹目距离。结合图4所示过载曲线进行分析可知，本文方法能够达到上述效果，主要是因为在z向过载指令与最优制导律方法基本重合的基础上，本文方法相比最优制导律方法给出了趋势相似但绝对值更低的y向过载指令，从而减少了导弹动能的损失。

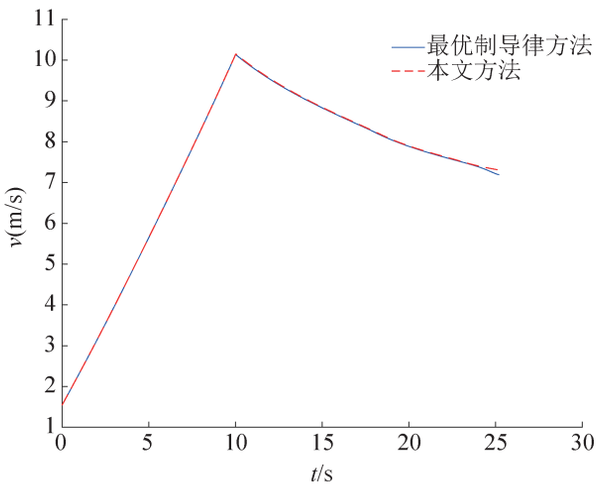


图8 本文方法与最优制导律方法速度曲线对比

Fig. 8 Comparison of velocity curves by proposed method and optimal guidance law

综合上述分析结果来看，本文方法可以在无需对目标机动加速度进行估计的前提下，达到与最优制导律方法接近的制导律性能水平。同时，从图8分析结果可以推断，以本文方法模拟最优制导律所得智能体模型为基础，进一步进行强化提升，可以得到相比最有制导律方法性能更优的制导指令生成智能体。

3 结论

本文以高速机动目标打击为问题背景，针对目标逆轨拦截中制导指令生成方法开展研究，提出一种基于深度强化学习的最优制导律模仿生成方法，构建并训练了仅以弹目相对运动及导弹自身动力学状态为输入的制导指令生成智能体并实现了：①在三维空间中对高速机动目标的逆轨拦截；②规避以目标加速度作为决策输入，大幅削减了对带有强不确定性目标估计的要求，增强制导指令生成方法适用性。同时，从对仿真实验结果的分析可知，本文方法所得智能体仍具备进一步强化提升，进而能够给出相比最优制导律方法性能更优的制导指令。

本文后续工作将主要聚焦于以下两个方面：一是在状态观测中引入噪声扰动并进一步扩充对抗场景，在更复杂的场景、更苛刻的要求下对最优制导律进行模仿；二是在模仿学习的基础上，通过基于模仿结果的强化，进一步提升制导律性能，达到更复杂场景下更优性能制导指令的动态生成。

参考文献：

[1] 吴帅, 周晓华, 汪莉莉, 等. 基于实际采样的导弹弹道建模与仿真[J]. 系统仿真学报, 2019, 31(4): 811-817.
Wu Shuai, Zhou Xiaohua, Wang Lili, et al. Modeling and Simulation of Missile Trajectory Based on Practical Sampling[J]. Journal of System Simulation, 2019, 31(4): 811-817.

[2] 顾文锦, 雷军委, 潘长鹏. 带落角限制的虚拟目标比例导引律设计[J]. 飞行力学, 2006, 24(2): 43-46.
Gu Wenjin, Lei Junwei, Pan Changpeng. Design of the Climbing Trajectory Using Virtual Target's Proportional Navigation Method with the Control of Terminal Azimuth of a Missile[J]. Flight Dynamics, 2006, 24(2): 43-46.

[3] Lee C H, Kim T H, Tahk M J. Interception Angle Control Guidance Using Proportional Navigation with Error Feedback[J]. Journal of Guidance, Control, and Dynamics, 2013, 36(5): 1556-1561.

[4] 闫梁, 赵继广, 李辕. 带约束碰撞角的顺/逆轨制导律设计[J]. 北京航空航天大学学报, 2015, 41(5): 857-863.
Yan Liang, Zhao Jiguang, Li Yuan. Guidance Law with Angular Constraints for Head-pursuit or Head-on Engagement[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(5): 857-863.

[5] Li Yuan, Yan Liang, Zhao Jiguang, et al. Combined Proportional Navigation Law for Interception of High-speed Targets[J]. Defence Technology, 2014, 10(3): 298-303.

[6] 司玉洁, 熊华, 李喆. 拦截机动目标的三维自适应神经网络制导律[J]. 系统仿真学报, 2021, 33(2): 453-460.
Si Yujie, Xiong Hua, Li Zhe. Three-dimensional Adaptive Neural Network Guidance Law Against Maneuvering Targets[J]. Journal of System Simulation, 2021, 33(2): 453-460.

[7] 熊少锋, 魏明英, 赵明元, 等. 考虑导弹速度时变的角度约束最优中制导律[J]. 控制理论与应用, 2018, 35(2): 248-257.
Xiong Shaofeng, Wei Mingying, Zhao Mingyuan, et al. Impact Angle Constrained Optimal Midcourse Guidance Law for Missiles of Time-varying Speed[J]. Control Theory & Applications, 2018, 35(2): 248-257.

[8] 熊少锋, 魏明英, 赵明元, 等. 逆轨拦截机动目标的三维最优制导律[J]. 宇航学报, 2020, 41(1): 80-90.

DOI: 10.16182/j.issn1004731x.joss.22-0632

Xiong Shaofeng, Wei Mingying, Zhao Mingyuan, et al. Three Dimensional Optimal Guidance Law Against Maneuvering Targets for Head-on Engagement[J]. Journal of Astronautics, 2020, 41(1): 80-90.

[9] 孟克子, 周荻. 多约束条件下的最优中制导律设计[J]. 系统工程与电子技术, 2016, 38(1): 116-122.
Meng Kezi, Zhou Di. Design of Optimal Midcourse Guidance Law with Multiple Constraints[J]. Systems Engineering and Electronics, 2016, 38(1): 116-122.

[10] Taub I, Shima T. Intercept Angle Missile Guidance Under Time Varying Acceleration Bounds[J]. Journal of Guidance, Control, and Dynamics, 2013, 36(3): 686-699.

[11] Bai Guoyu, Shen Huairong, Chen Jingpeng, et al. Novel Guidance Law for Interception for Maneuvering Target with High-speed[C]//Proceedings of 2016 3rd International Conference on Engineering Technology and Application. Lancaster, PA, USA: DEStech Publications, 2016: 735-742.

[12] 周慧波, 宋申民, 刘海坤. 具有攻击角约束的非奇异终端滑模导引律设计[J]. 中国惯性技术学报, 2014, 22(5): 606-611, 618.
Zhou Huibo, Song Shenmin, Liu Haikun. Nonsingular Terminal Sliding Mode Guidance Law with Impact Angle Constraint[J]. Journal of Chinese Inertial Technology, 2014, 22(5): 606-611, 618.

[13] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553): 436-444.

[14] 郭圣明, 贺筱媛, 吴琳, 等. 基于强制稀疏自编码神经网络的作战态势评估方法研究[J]. 系统仿真学报, 2018, 30(3): 772-784, 800.
Guo Shengming, He Xiaoyuan, Wu Lin, et al. Situation Assessment Approach for Air Defense Operation System Based on Force-sparsed Stacked-auto Encoding Neural Networks[J]. Journal of System Simulation, 2018, 30(3): 772-784, 800.

[15] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level Control Through Deep Reinforcement Learning[J]. Nature, 2015, 518(7540): 529-533.

[16] Silver D, Huang A, Maddison C J, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search [J]. Nature, 2016, 529(7587): 484-489.

[17] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster Level in StarCraft II Using Multi-agent Reinforcement Learning[J]. Nature, 2019, 575(7782): 350-354.

[18] Furfaro R, Linares R. Waypoint-based Generalized ZEM/ZEV Feedback Guidance for Planetary Landing Via a Reinforcement Learning Approach[C]//3rd International Academy of Astronautics Conference on Dynamics and Control of Space Systems. Escondido, CA, USA: Univelt Inc., 2017: 401-416.

[19] Liang Chen, Wang Weihong, Liu Zhenghua, et al. Learning to Guide: Guidance Law Based on Deep Meta-learning and Model Predictive Path Integral Control[J]. IEEE Access, 2019, 7: 47353-47365.

[20] Gaudet B, Furfaro R. Missile Homing-phase Guidance Law Design Using Reinforcement Learning[C]//AIAA Guidance, Navigation, and Control Conference. Reston, VA, USA: AIAA, 2012: AIAA 2012-4470.

[21] Chen Yadong, Wang Jianan, Wang Chunyan, et al. Three-dimensional Cooperative Homing Guidance Law with Field-of-view Constraint[J]. Journal of Guidance, Control, and Dynamics, 2020, 43(2): 389-397.

[22] Hussein A, Gaber M M, Elyan E, et al. Imitation Learning: A Survey of Learning Methods[J]. ACM Computing Surveys, 2018, 50(2): 21.

[23] Micheal B, Claude S. A Framework for Behavioural Cloning[M]. [S.l.]: [s.n.], 1995: 103-129.

[24] Abbeel P, Ng A Y. Apprenticeship Learning Via Inverse Reinforcement Learning[C]//Proceedings of the Twenty-First International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2004: 1.

[25] Ng A Y, Russell S J. Algorithms for Inverse Reinforcement Learning[C]//Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000: 663-670.

[26] Stéphane Ross, Gordon G J, Bagnell J A. A Reduction of Imitation Learning and Structured Prediction to No-regret Online Learning[C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Sardinia, Italy: PMLR, 2011: 627-635.

[27] Perttu Hämäläinen, Babadi A, Ma Xiaoxiao, et al. PPO-CMA: Proximal Policy Optimization with Covariance Matrix Adaptation[C]//2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). Piscataway, NJ, USA: IEEE, 2020: 1-6.