文章编号：2097-1974(2025)02-0060-09 DOI：10.7654/j.issn.2097-1974.20250208

基于强化学习的飞行器轨迹跟踪制导与编队保持问题研究

滕庆骅[1]，惠俊鹏[2]，李天任[1]，杨 奔[1]
（1. 中国运载火箭技术研究院研究发展中心，北京，100076；2. 北京航天长征飞行器研究所，北京，100076）

摘要：飞行器的智能化升级对制导能力提出了新的需求，传统算法在有偏差条件下跟踪空间三维轨迹的表现不佳。基于TD3强化学习算法设计了飞行器轨迹跟踪制导方式。通过偏差形式的动作空间、奖励函数中的惩罚项、距离变化率的导引，解决了算法训练难收敛、控制量波动过大、中末交班点偏差累积大等问题。相比传统LQR算法，强化学习制导算法的制导精度、偏差适应性均有较大提升，且具备良好的泛用性，能够应用于小规模编队保持问题。

关键词：TD3算法；标准轨迹制导；强化学习制导；编队保持；蒙特卡罗仿真
中图分类号：V448 **文献标识码**：A

Research on Aircraft Standard Trajectory Tracking Guidance and Formation Keeping based on Reinforcement Learning

TENG Qinghua[1], HUI Junpeng[2], LI Tianren[1], YANG Ben[1]
(1. Research & Development Center, China Academy of Launch Vehicle Technology, Beijing, 100076;
2. Beijing Institute of Space Long March Vehicle, Beijing, 100076)

Abstract: The intelligent upgrade of the aircraft has put forward new requirements for guidance capabilities, and traditional algorithms perform poorly in tracking spatial three-dimensional trajectories under biased conditions. An aircraft trajectory tracking guidance method is designed based on the TD3 reinforcement learning algorithm. Through the action space in the form of deviation, the penalty term in the reward function and the guidance of the rate of change of distance, problems such as difficult convergence of algorithm training, large fluctuations in control quantity, and large cumulative deviation at the middle and final shift points are solved. Compared with the traditional LQR algorithm, the reinforcement learning guidance algorithm has significantly improved guidance accuracy and deviation adaptability, and has good versatility, which can be applied to small-scale formation maintenance issues.

Keywords: TD3 Algorithm; standard trajectory guidance; reinforcement learning; formation keeping; Monte Calo simulation

0 引 言

随着飞行器智能化程度的提升，飞行器逐步具备更强的态势感知与决策能力，可以针对绕飞、避障等任务需求，在线规划飞行轨迹，而这对飞行器制导能力提出了新的要求。

飞行器制导方式主要包括标准轨迹制导与预测校正制导两大类[1]，标准轨迹制导是指预先计算出符合要求的标准轨迹，制导系统根据实际测量的飞行状态和标准轨迹的关系，计算出所需要的控制参数，控制飞行器按照标准轨迹飞行。预测校正制导不需要标准轨迹，基于飞行器当前状态，在线计算预测终端状态，并基于终端状态与目标点的偏差来校正制导指令[2-8]。

标准轨迹制导跟踪的标准轨迹可以是三维空间中完整的飞行状态量序列，也可以是表征飞行器运动特性的广义飞行剖面，例如D-V剖面[9]，H-V剖面[10]等。目前的在线轨迹规划方法已具备在复杂约束条件下生成三维空间中标准轨迹的能力。

文献［11］以能量为自变量推导了动力学方程，采用倾侧角的正弦、余弦值作为控制量，使用二阶锥规划求解了优化轨迹。文献［12］以时间为自变量，

收稿日期：2024-05-21；修回日期：2024-05-31

采用序列凸优化方法进行轨迹规划，相比高斯伪谱法，计算速度可提升10~20倍。文献［13］通过B样条离散、回溯搜索等策略，提高了算法的稳定性、快速性，7秒内即可完成一条轨迹的规划，具备在线应用的能力。

传统的标准轨迹跟踪方法例如线性二次型[14]、比例-微分控制等方法会随着飞行误差的累积，导致制导精度越来越低，且需要根据每条轨迹或某段的特性设计制导律参数，泛用性不高。强化学习方法可根据飞行器状态信息直接给出控制量信息，消除了传统方法对飞行器附加的一些不必要的约束，有望更充分地发挥飞行器的制导能力，以在线规划出的轨迹飞行。

强化学习按照智能体的动作选择方式，可以分为基于价值与基于策略两大类。前者的典型算法有DQN[15]、Double-DQN[16]、Duel-DQN[17]，后者的典型算法有TRPO[18]、PPO[19]。此外，将基于价值方法和基于策略方法结合，研究者提出了Actor-Critic框架，典型算法有DDPG[20]、TD3[21]、SAC[22]。文献［23］基于OpenAI Gym设计了多个基准环境，测试了不同强化学习算法的性能表现。

强化学习已经开始应用在飞行器制导中，解决了很多问题。文献［24］基于Q-Learning算法训练横向倾侧角翻转策略，使飞行器相比原始的预测-校正制导具备更强的机动能力。文献［25］基于Q-Learning算法训练比例制导律的比例系数，获得了更好的制导精度。文献［26］基于PPO算法，根据视线角及其变化率，调整飞行器的机动推力指令进行制导。目前在飞行器制导中，强化学习常与原有制导算法结合，发挥各自优势，且通常只对终端状态制导，较少有对完整空间三维轨迹的持续跟踪。

综合文献的结论，在处理飞行器制导这种非线性连续问题时，TD3与SAC算法表现良好，结合本文的测试，最终选用TD3算法作为制导设计的工具。本文针对动作空间覆盖控制量可达完整范围时算法无法收敛的问题，基于相对标准轨迹中控制量的偏差值设计动作空间，使算法仿真500轮即可收敛。针对动作波动过大不满足飞行器实际控制要求的问题，在奖励中设计惩罚项，抑制了攻角、倾侧角的不合理波动。针对飞行过程中偏差逐渐累积，导致中末交班点偏移过大的问题，设计额外的距离变化率导引，提高了中末交班点精度。

经仿真验证，基于强化学习的飞行器轨迹跟踪制导的精度、对偏差的修正能力、不同轨迹的泛用性满足要求，且能通过菱形与一字型两种编队保持问题的测试。

1　建　模

圆球旋转地球模型假设下，飞行器满足动力学方程：

$$\begin{cases}\dot{r}=V\sin\gamma\\ \dot{\theta}=\dfrac{V\cos\gamma\sin\psi}{r\cos\phi}\\ \dot{\phi}=\dfrac{V\cos\gamma\cos\psi}{r}\\ \dot{v}=-\dfrac{D}{m}-g\sin\gamma+\omega_{\mathrm{e}}^{2}r\left(\cos^{2}\phi\sin\gamma-\sin\phi\cos\phi\cos\psi\cos\gamma\right)\\ \dot{\gamma}=\dfrac{L\cos\sigma}{mV}-\left(\dfrac{g}{V}-\dfrac{V}{r}\right)\cos\gamma+2\omega_{\mathrm{e}}\cos\phi\sin\psi+\\ \qquad\dfrac{\omega_{\mathrm{e}}{}^{2}r}{V}\left(\cos^{2}\phi\cos\gamma+\sin\phi\cos\phi\cos\psi\sin\gamma\right)\\ \dot{\psi}=\dfrac{L\sin\sigma}{mV\cos\gamma}+\dfrac{V\cos\gamma\sin\psi\tan\phi}{r}+\\ \qquad2\omega_{\mathrm{e}}\left(\sin\phi-\cos\phi\cos\psi\tan\gamma\right)+\dfrac{\omega_{\mathrm{e}}{}^{2}r\sin\phi\cos\phi\sin\psi}{V\cos\gamma}\end{cases}\tag{1}$$

式中　r，θ，ϕ，v，γ，ψ，m，L，D分别为飞行器的地心距、经度、纬度、速度大小、速度倾角、航向角、质量、升力、阻力。

升力与阻力的表达式为

$$\begin{cases}L=\dfrac{1}{2}\rho V^{2}SC_{\mathrm{L}}\\ D=\dfrac{1}{2}\rho V^{2}SC_{\mathrm{D}}\end{cases}\tag{2}$$

式中　ρ为当地大气密度；S为飞行器参考面积；C_{L}，C_{D}为飞行器的升力系数与阻力系数，由飞行器的攻角与马赫数插值求得。飞行器的m，S，C_{L}，C_{D}均取自国外某飞行器，升阻比大于4。

2　算法设计

2.1　标准轨迹设计

设计标准轨迹作为制导目标，基于凸优化方法生成标准轨迹，考虑到编队控制需要基于时间同步各飞行器的位置，以时间为自变量进行轨迹规划，得到以200个离散点表示的状态量与控制量序列。以0.1 s的步长对控制量序列进行线性插值，积分得到完整的标准轨迹如图1所示。

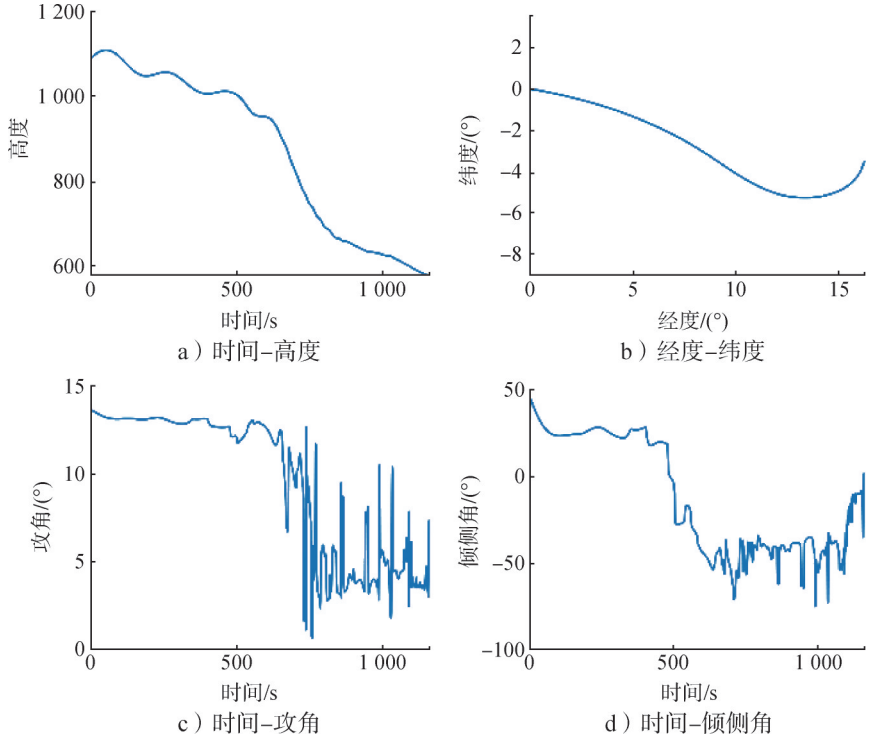


图1 标准轨迹

Fig.1 Standard trajectory

2.2 轨迹跟踪制导设计

轨迹跟踪制导的算法流程见图2。

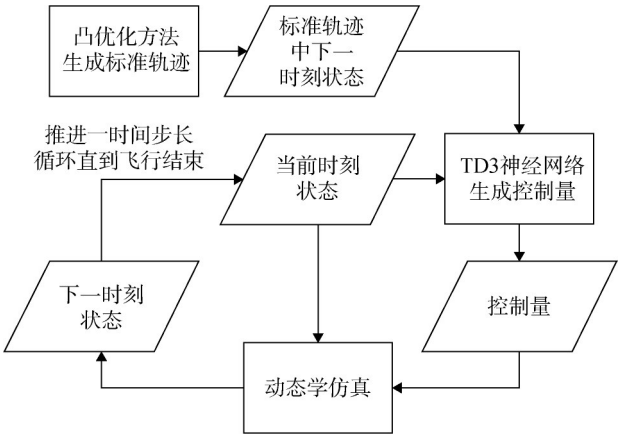


图2 轨迹跟踪制导流程

Fig.2 Trajectory tracking guidance process

基于飞行器制导场景连续动作空间、非线性的特点，选用了TD3强化学习算法进行训练和测试。

2.3 TD3算法原理

双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）算法是一种基于深度强化学习的算法，主要用于解决连续动作空间的强化学习问题。TD3算法是深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）算法的改进版本，主要针对DDPG在训练过程中可能出现的过高值估计问题进行了优化，使用双重Critic网络来评估当前策略的性能。在训练过程中，TD3算法的Actor网络间隔一定步数更新一次，Critic网络每一步更新一次，有助于提高算法的稳定性和性能。TD3算法在目标策略中添加了随机噪声，以平滑目标值函数可以更好地处理连续动作空间中的不确定性和非线性问题。

TD3算法一共有6个神经网络，它们的更新关系如图3所示。

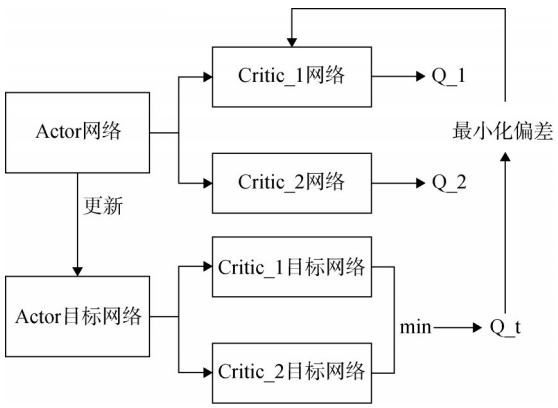


图3 TD3算法网络关系

Fig.3 TD3 algorithm network relationships

2.4　TD3制导算法设计

基于TD3算法与飞行器制导仿真环境，设计算法的状态空间、动作空间、状态转移函数、奖励函数。

2.4.1　状态空间设计

序列凸优化算法给出的标准轨迹是随时间变化的一系列离散点，每个点包括飞行器位置、速度与控制量信息，是$N\times 9$的矩阵。对于单步飞行器制导，至少需要知晓飞行器当前位置、速度信息，以及从标准轨迹中读取到的下一时刻飞行器目标位置、速度信息，才可给出对应的攻角、倾侧角控制指令。基于以上考虑，设计状态空间为以下12维向量。

$$\boldsymbol{S}_{\text{tate}}=\left[h,\theta,\phi,v,\gamma,\psi,h_{\text{next}},\theta_{\text{next}},\phi_{\text{next}},v_{\text{next}},\gamma_{\text{next}},\psi_{\text{next}}\right] \tag{3}$$

以上状态量的数值差别巨大，为防止数据差异过大导致梯度爆炸或无法学习，根据实际飞行场景中参数的变化范围，对状态空间进行归一化，使所有量的数值处在［−10，10］内。

2.4.2　动作空间设计

飞行器制导的控制量包括攻角与倾侧角，攻角影响气动力的大小，倾侧角影响气动力的方向。由于偏差的存在，标准轨迹中的攻角、倾侧角并非最优值。当动作空间设计为攻角、倾侧角时，算法在完整的取值范围内寻找最优解，难度过高，经测试无法收敛。因此将动作空间设计为相对标准轨迹的攻角、倾侧角偏移量，充分利用标准轨迹的信息，加快算法收敛。

$$\text{Action}=\left[\delta\alpha,\delta\sigma\right] \tag{4}$$

实际使用的攻角、倾侧角如下：

$$\begin{cases}\alpha=\alpha_{\text{ref}}+\delta\alpha\\ \sigma=\sigma_{\text{ref}}+\delta\sigma\end{cases} \tag{5}$$

定义动作偏移量的范围如下：

$$\begin{cases}\delta\alpha\in\left[-3,3\right]\\ \delta\sigma\in\left[-20,20\right]\end{cases} \tag{6}$$

参考状态空间的方式，对动作空间进行归一化处理，使其值处于［−10，10］范围内。

在状态转移函数中，再对α和σ的范围进行限制，以防止$\alpha_{\text{ref}}+\delta\alpha$或$\sigma_{\text{ref}}+\delta\sigma$的范围超过限制。

2.4.3　状态转移函数设计

在飞行器制导过程中，通过飞行器的运动学方程，得到智能体的状态转移函数。基于当前状态中的位置、速度与动作给出的攻角、倾侧角，可以计算出升力、阻力，由此计算出位置与速度的变化率，通过积分得到下一时间步长飞行器的状态量。

飞行器实际飞行过程中，控制量不能在完整可达范围内随意变化，而是需要响应时间，因此在状态转移函数中，加入对攻角、倾侧角范围和变化率的限制。攻角与倾侧角的取值范围如下：

$$\begin{cases}\alpha\in\left[-5,25\right]\\ \sigma\in\left[-90,90\right]\end{cases} \tag{7}$$

攻角与倾侧角变化率限制范围如下：

$$\begin{cases}\dot{\alpha}\in\left[-5,5\right]\\ \dot{\sigma}\in\left[-10,10\right]\end{cases} \tag{8}$$

2.4.4　奖励函数设计

飞行器制导的任务是在偏差的影响下让飞行器尽量贴合标准轨迹飞行到中末交班点。以代价函数方式设计奖励函数，总奖励函数为多种惩罚项和的负数。

$$\text{Reward}=-\sum\text{cost} \tag{9}$$

衡量制导效果最重要的参数是制导精度。在整个飞行过程中，每个制导周期根据射程，在标准轨迹中插值找到对应的位置，即高度、经度、纬度，与飞行器实际位置进行比较，位置的偏移量作为代价放进奖励函数的惩罚项中。

$$\text{cost}_{(x-x_{\text{ref}})}=\sum_{i=1}^{n}x_i-x_{i,\text{ref}} \tag{10}$$

在强化学习探索动作值的过程中，常出现动作值在上下限约束之间来回跳动的现象，反映在控制量上为攻角、倾侧角的跳跃抖动，不符合控制系统要求。为了抑制动作值的跳动，在奖励函数中加入对攻角、倾侧角变化率积分值的惩罚项。

$$\text{cost}_{\alpha,\sigma}=\int\delta\dot{\alpha}\text{d}t+\int\delta\dot{\sigma}\text{d}t \tag{11}$$

飞行器制导的最终目的不只是跟随标准轨迹，还需要到达指定的中末交班点。若飞行前期偏差积累过多，在飞行后期只朝标准轨迹修正不能保证最优接近中末交班点。因此，在飞行过程的后30%时间，增加对中末交班点距离变化量的奖励值，以促使飞行器向中末交班点飞行。

$$\text{cost}_{m2\text{e}}=-\left(v_{m2\text{e}}-v_{\text{last},m2\text{e}}\right) \tag{12}$$

综上所述，最终为飞行器制导强化学习算法设计的奖励函数为

$$\text{Reward}=-\left(\text{cost}_{(x-x_{\text{ref}})}+\text{cost}_{\alpha,\sigma}+\text{cost}_{m2\text{e}}\right) \tag{13}$$

3　数值仿真验证与分析

本文针对TD3算法经过试验调试后选取性能最佳的超参数。Actor与Critic的学习率设置为0.000 3，6个神经网络均为2层，第1层大小为400，第2层大

小为300。

基于标准轨迹，训练1 600轮后，奖励函数收敛曲线如图4所示。

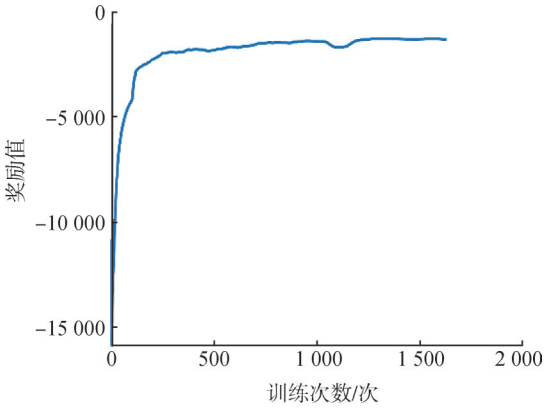


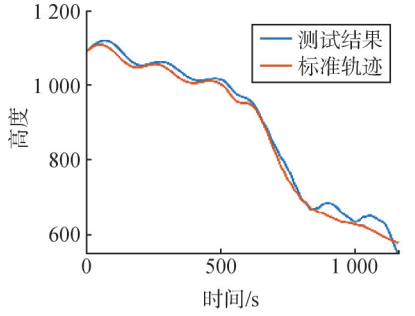
图4 基于强化学习的飞行器轨迹跟踪制导训练收敛曲线

Fig.4 Convergence curve for vehicle trajectory tracking guidance training based on reinforcement learning

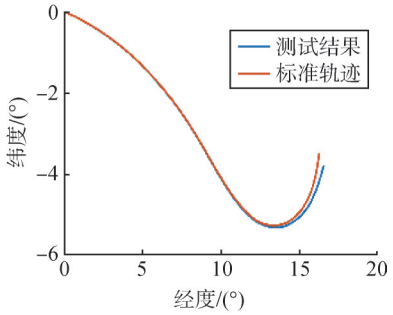
在训练300轮后，奖励函数已趋近收敛。

3.1 典型偏差条件下的标准轨迹跟踪制导仿真结果

以质量0.333%、大气密度3.33%、气动系数5%作为典型偏差条件，对仿真环境进行拉偏，使用传统LQR算法跟踪标准轨迹，效果如图5所示。



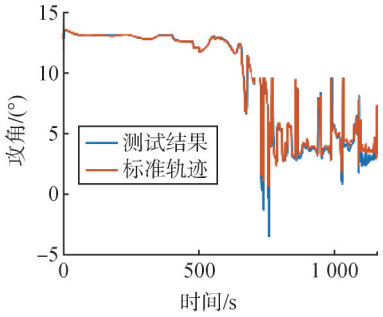
a）时间–高度



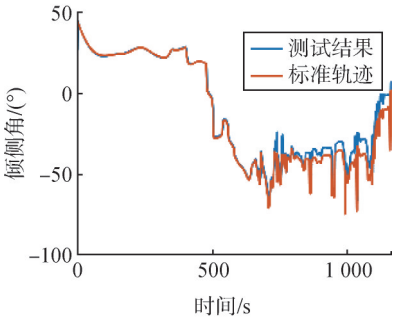
b）经度–纬度

图5 典型偏差条件下LQR算法跟踪标准轨迹结果

Fig.5 Results of LQR algorithm tracking standard trajectory under typical deviation conditions



c）时间–攻角

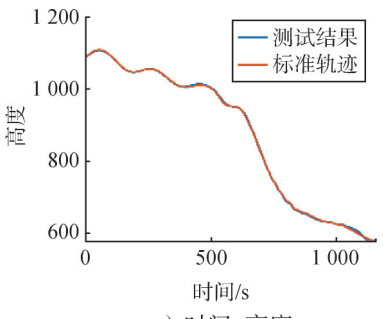


d）时间–倾侧角

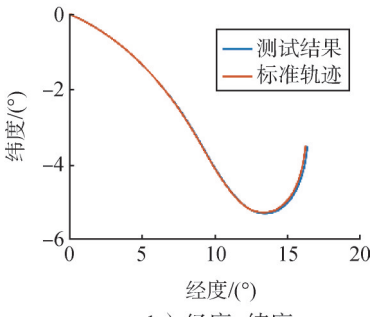
续图5

典型偏差条件下，记录LQR算法跟踪的位置累计偏差值与中末交班点的偏差值，作为基准值。

在同样的典型偏差条件下，对强化学习算法训练得到的制导模型进行测试，结果如图6所示。



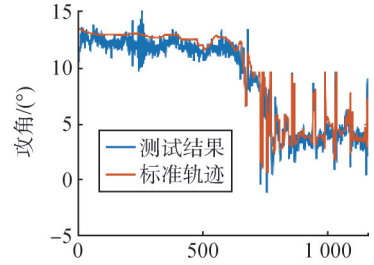
a）时间–高度



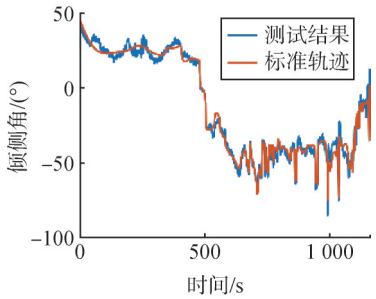
b）经度–纬度

图6 典型偏差条件下强化学习算法跟踪标准轨迹结果

Fig.6 Results of TD3 algorithm tracking standard trajectory under typical deviation conditions



c）时间–攻角



d）时间–倾侧角

续图6

在典型偏差条件下，强化学习算法跟踪的位置平均偏差为LQR算法的74.07%，中末交班点偏差为LQR算法的31.35%。

根据以上结果，在典型偏差条件下，相比传统LQR算法，强化学习算法轨迹跟踪效果更好。

3.2　正态分布偏差下的蒙特卡罗打靶仿真结果

基于工程经验，以质量最高1%、大气密度最大10%、气动系数最大15%的范围进行拉偏，每条轨迹对应4个拉偏参数。基于正态分布$N(0,\ \sigma)$计算偏差百分比，取3σ为偏差最大值，超过最大值的偏差截断为最大值。拉偏参数设置如下：

$$\begin{cases}\varepsilon_m \sim N(0,0.00333)\\ \varepsilon_\rho \sim N(0,0.0333)\\ \varepsilon_{\mathrm{CL}} \sim N(0,0.05)\\ \varepsilon_{\mathrm{CD}} \sim N(0,0.05)\end{cases} \tag{14}$$

在仿真计算中，飞行器实际计算值和理论值的关系为

$$\begin{cases}m_{\mathrm{true}} = m(1+\varepsilon_m)\\ \rho_{\mathrm{true}} = \rho(1+\varepsilon_\rho)\\ CL_{\mathrm{true}} = CL(1+\varepsilon_{\mathrm{CL}})\\ CD_{\mathrm{true}} = CD(1+\varepsilon_{\mathrm{CD}})\end{cases} \tag{15}$$

在随机偏差条件下，进行1 000次打靶测试，记录LQR算法跟踪的位置累计偏差平均值与中末交班点的偏差平均值，作为基准值。中末交班点的经纬度偏差分布如图7所示。

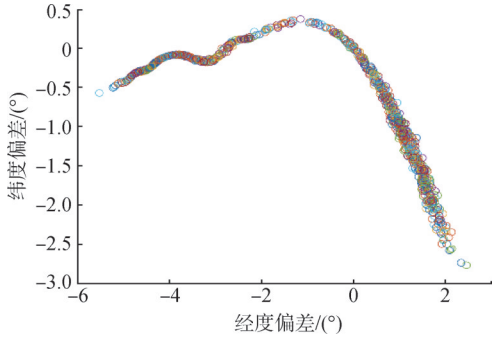


图7　LQR制导中末交班点经纬度偏差分布

Fig.7　Latitude and longitude deviation distribution at the final handover point in LQR guidance

在随机偏差条件下，进行1 000次打靶测试，强化学习算法跟踪的位置累计偏差平均值为LQR算法的27.72%，中末交班点的偏差平均值为LQR算法的17.25%。TD3制导中末交班点经纬度偏差分布如图8所示。

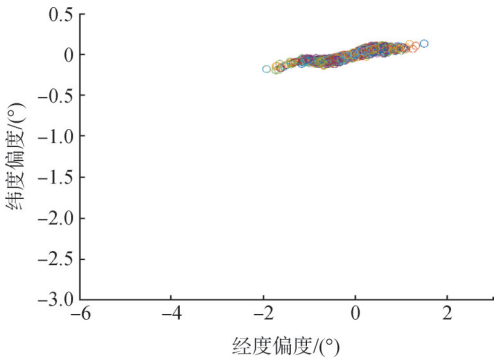
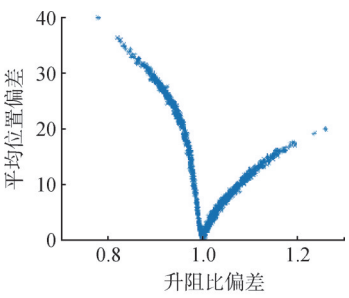


图8　TD3制导中末交班点经纬度偏差分布

Fig.8　Latitude and longitude deviation distribution at the final handover point in TD3 guidance

根据仿真结果，在随机偏差条件下，强化学习算法跟踪的精度更高，中末交班点散布更集中。

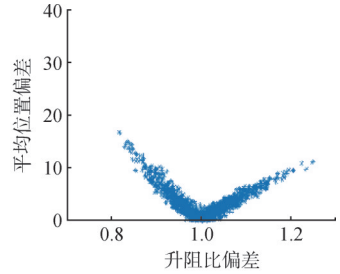
分析偏差对制导精度的影响，发现升力、阻力系数的偏差相互叠加产生的升阻比偏差对制导精度影响最大，绘制升阻比-跟踪精度的图像，如图9所示。



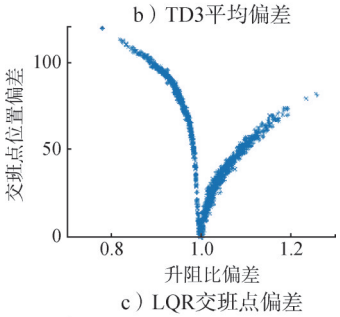
a）LQR平均偏差

图9　不同升阻比偏差下轨迹跟踪制导效果

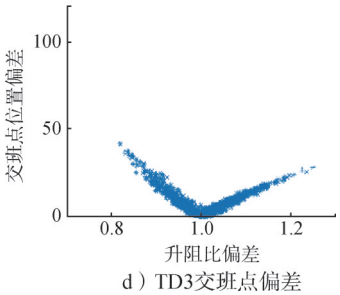
Fig.9　Results of trajectory tracking guidance under different lift-to-drag ratio deviations



b）TD3平均偏差



c）LQR交班点偏差



d）TD3交班点偏差

续图9

升阻比偏差最小时，制导精度最高。升阻比降低对制导精度的影响大于升阻比升高的影响。

3.3 典型偏差条件下的两种编队保持仿真结果

采用虚拟中心法[27]，以菱形与一字形两种编队，在典型偏差条件下测试制导算法的编队保持能力。飞行器集群的中心为虚拟中心，以标准轨迹飞行。

对于菱形编队，4个飞行器位于菱形的4个角上，在标准轨迹的基础上对经度、纬度的位置进行拉偏，菱形的边长为固定值，得到飞行器实际跟踪的轨迹。

在典型偏差条件下，使用TD3制导算法进行轨迹跟踪制导，得到实际飞行轨迹如图10所示。

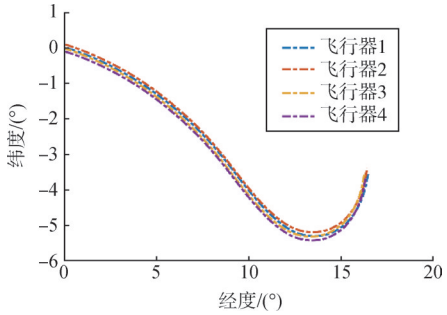


图10 菱形编队保持飞行结果

Fig.10 Results of diamond formation maintenance test

根据飞行轨迹可看出，4个飞行器全程以菱形编队飞行，以相对虚拟中心的距离偏差作为编队保持结果的评判指标，4个飞行器相对虚拟中心平均距离偏差9.44%，最大偏差25.71%，小于30%，满足菱形编队保持要求。

对于一字形编队，4个飞行器一字排开，在标准轨迹的基础上对纬度的位置进行拉偏，每相邻两个飞行器间距离相同，得到飞行器实际跟踪的轨迹。

在典型偏差条件下，使用TD3制导算法进行轨迹跟踪制导，得到实际飞行轨迹如图11所示。

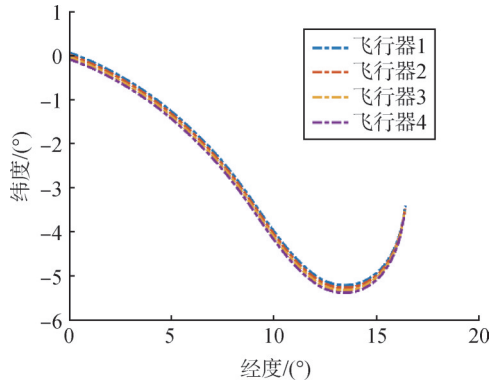


图11 一字编队保持飞行结果

Fig.11 Results of line formation maintenance test

根据飞行轨迹可以看出，4个飞行器全程以一字构型飞行，以相邻两个飞行器间的距离偏差作为编队保持结果的评判指标，每相邻两个飞行器间平均距离偏差14.61%，最大偏差27.91%，小于30%，满足一字型编队保持要求。

3.4 不同轨迹的泛用性测试蒙特卡罗打靶仿真结果

为了验证算法对于不同轨迹的泛用性，重新生成训练集与测试集，更换不同的目标点，使用序列凸优化方法生成21条不同的轨迹，如图12所示。

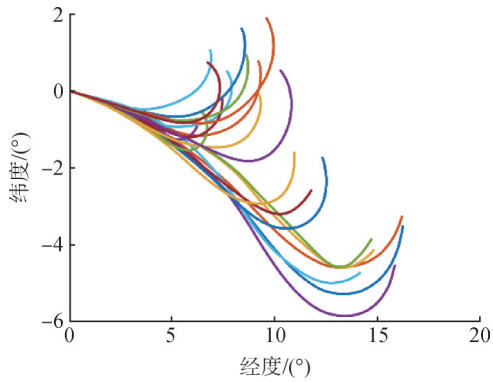


图12 不同标准轨迹经度-纬度

Fig.12 Latitude-longitude maps of different standard trajectories

取其中20条轨迹为训练集，1条轨迹为测试集，两集合互斥，因此使用训练集训练的算法在测试集上即可验证泛用性。从20条轨迹中，每轮随机抽取一条对制导算法进行训练，收敛曲线如图13所示。

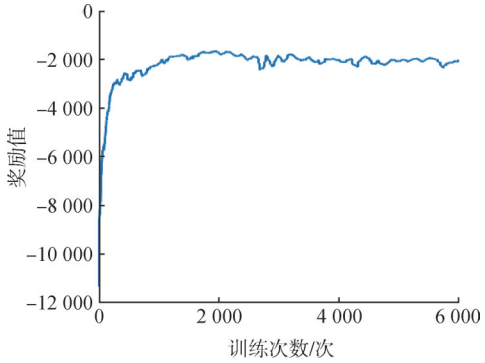


图13　多条轨迹随机抽取训练收敛曲线

Fig.13　Convergence curves for training with randomly selected multiple trajectories

相比单条标准轨迹训练时，300轮以上基本收敛的情况，多条轨迹训练时收敛速度变慢，1 000轮以后收敛。在其他条件相同的情况下，两者收敛后的奖励值相近，多条轨迹跟踪的精度略低于单条轨迹，但差距不大。使用测试集的1条轨迹进行蒙特卡罗打靶测试，结果如图14所示。

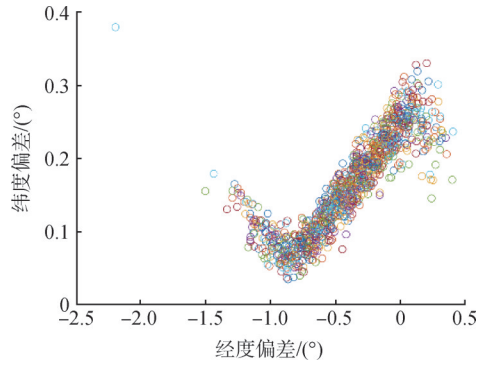


图14　泛用性打靶测试中末交班点分布

Fig.14　Distribution of final handover points in the versatile targeting test

在随机偏差条件下，进行1 000次打靶测试，算法跟踪的位置累计偏差平均值为LQR算法的30.18%，相比单轨迹测试结果增加了8.85%，中末交班点的偏差平均值为LQR算法的20.34%，相比单轨迹测试结果增加了17.95%。

相比单轨迹训练与测试结果，多轨迹的测试集结果偏差只增大了不到20%，与LQR算法相比仍有巨大优势。因此，强化学习算法具备良好的泛用性，具备在相似轨迹中直接复用的能力。

4　结　论

为了充分发挥新型飞行器制导能力，满足以在线规划轨迹飞行的要求，本文基于TD3算法设计了基于强化学习的飞行器轨迹跟踪制导方式。通过偏差量形式的动作空间，加快了算法训练的收敛。通过奖励函数的惩罚项，抑制了攻角、倾侧角的过大波动。通过距离变化率导引，提高了中末交班点的制导精度。

仿真结果表明，基于强化模型的轨迹跟踪制导算法相比传统LQR算法，中末交班点精度更高，对偏差的适应性更好，在正态分布偏差下，飞行过程平均制导偏差为LQR算法的27.72%，中末交班点偏差为LQR算法的17.25%，具备良好的泛用性，在更换不同的训练集与测试集的试验中，偏差增加不超过20%。在编队保持测试中，算法飞行全程能维持菱形与一字构型在容许误差范围内，菱形构型距离偏差不超过25.71%，一字构型距离偏差不超过27.19%。综上，在标准轨迹跟踪制导中，强化学习方法可行且效果良好，具备基于在线规划的三维轨迹进行小规模编队保持的能力，后续工作可继续探索算法在编队生成、编队切换等方向的应用潜力并进行改进。

参　考　文　献

[1] 张远龙, 谢愈. 滑翔飞行器弹道规划与制导方法综述[J]. 航空学报, 2020, 41(1): 50-62.
ZHANG Yuanlong, XIE Yu. Overview of glide vehicle trajectory planning and guidance methods[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(1): 50-62.

[2] DUKEMAN G. Profile-following entry guidance using linear quadratic regulator theory[C]. Monterey: AIAA Guidance, Navigation, and Control Conference and Exhibit, 2002.

[3] 汪轶俊, 梁艳迁, 周鼎, 等. 运载火箭自适应制导及在线轨迹重构方法研究[J]. 上海航天(中英文), 2023, 40(1): 1-10.
WANG Yijun, LIANG Yanqian, ZHOU Ding, et al. Research on self-adaptive guidance and online trajectory reconfiguration methods for launch vehicles[J]. Aerospace Shanghai (Chinese & English), 2023, 40(1): 1-10.

[4] MEASE K D, TEUFEL P, SCHONENBERGER H, et al. Reentry trajectory planning for a reusable launch vehicle[C]. Portland: AIAA Atmospheric Flight Mechanics Conference and Exhibit, 1999.

[5] 尹中杰, 王磊, 杨建东, 等. 多约束航迹规划与跟踪制导律[J]. 上海航天(中英文), 2023, 40(6): 136-143.
YIN Zhongjie, WANG Lei, YANG Jiandong, et al. Multi-constraint trajectory planning and tracking guidance law[J]. Aerospace Shanghai (Chinese & English), 2023, 40(6): 136-143.

[6] LU Ping. Predictor-corrector entry guidance for low-lifting vehicles [J]. Journal of Guidance, Control, and Dynamics, 2008, 31(4): 1067-1075.

[7] XUE S B, LU P. Constrained predictor-corrector entry guidance[J]. Journal of Guidance, Control, and Dynamics, 2010, 33(4): 1273-1281.

[8] XU M L, CHEN K J, LIU L H, et al. Quasi-equilibrium glide adap-

tive guidance for hypersonic vehicles[J]. Science China Technological Sciences, 2012, 55(3): 856-866.
[9] LU P. Entry guidance and trajectory control for reusable launch vehicle[J]. Journal of Guidance Control & Dynamics, 1997, 20(1): 143-149.
[10] LI D W, YANG B. Reentry guidance for reusable launching vehicle [J]. Journal of Solid Rocket Technology, 2010, 33(2): 119-124.
[11] LIU X, SHEN Z, LU P. Entry trajectory optimization by second-order cone programming[J]. Journal of Guidance, Control, and Dynamics, 2015, 39(2): 227-241.
[12] WANG Z, GRANT M J. Constrained trajectory optimization for planetary entry via sequential convex programing[C]. Washington, D. C.: AIAA Atmospheric Flight Mechanics Conference, 2016.
[13] 杨奔, 李天任, 马晓媛. 基于序列凸优化的多约束轨迹快速优化[J]. 航天控制, 2020, 38(3): 25-30.
YANG Ben, LI Tianren, MA Xiaoyuan. Rapid optimization of multi-constrained trajectories based on sequential convex optimization[J]. Aerospace Control, 2020, 38(3): 25-30.
[14] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge: MIT Press, 2018.
[15] VOLODYMYR M, KORAY K, DAVID S, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015(518): 529-533.
[16] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-Learning[C]. Phoenix: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[17] WANG Ziyu, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]. New York: The 33rd International Conference on Machine Learning, 2016.
[18] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[EB/OL].（2017-04-20）[2024-05-10]. http://arxiv. org/abs/1502. 05477v5.
[19] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL].（2017-08-28）[2024-05-10]. https://doi.org/10.48550/arXiv.1707.06347.
[20] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL].（2019-07-05）[2024-05-10]. http://arxiv.org /abs/1509.02971.
[21] FUJIMOTO S, VAN H H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL].（2018-10-22）[2024-05-10]. http://arxiv.org/abs/1802.09477?context=stat.
[22] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[EB/OL].（2017-07-21）[2024-05-10]. http://arxiv.org /abs/1702.08165.
[23] WANG T, BAO X, CLAVERA I, et al. Benchmarking model-based reinforcement learning[EB/OL].（2019-07-03）[2024-05-10]. http://arxiv.org/abs/1907.02057?context=stat.ML.
[24] 李天任, 杨奔, 汪韧, 等. 基于Q-Learning算法的再入飞行器制导方法[J]. 战术导弹技术, 2019(5): 44-49.
LI Tianren, YANG Ben, WANG Ren, et al. Reentry vehicle guidance method based on Q-Learning algorithm[J]. Tactical Missile Technology, 2019(5): 44-49.
[25] 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 Qinhao, AO Baiqiang, ZHANG Qinxue. Q-Learning reinforcement learning guidance law[J]. Journal of Systems Engineering and Electronics, 2020, 42(2): 414-419.
[26] GAUDET B, FURFARO R, LINARES R. Reinforcement learning for angle-only intercept guidance of maneuvering targets[J]. Aerospace Science and Technology, 2019, 99(4): 1-10.
[27] 赵恩娇. 多飞行器编队控制及协同制导方法[D]. 哈尔滨: 哈尔滨工业大学, 2018.
ZHAO Enjiao. Multi-aircraft formation control and coordinated guidance methods[D]. Harbin: Harbin Institute of Technology, 2018.

作 者 简 介

滕庆骅（1999—），男，硕士研究生，主要研究方向为飞行器制导与控制。
惠俊鹏（1981—），男，研究员，主要研究方向为飞行器总体设计。
李天任（1993—），男，工程师，主要研究方向为飞行器制导与控制。
杨 奔（1994—），男，工程师，主要研究方向为飞行器制导与控制。

（上接第36页）

[22] 王林, 张晓卫. 微型计算机算法与程序——扩展BASIC[M]. 上海: 上海科学技术文献出版社, 1983.
WANG Lin, ZHANG Xiaowei. Microcomputer algorithms and programs: extended BASIC[M]. Shanghai: Shanghai Scientific and Technical Literature Press, 1983.
[23] 郑宗成, 王振堂. 实用预测方法BASIC程序[M]. 广州: 中山大学出版社, 1985.
ZHENG Zongcheng, WANG Zhentang. Practical forecasting methods with BASIC programs[M]. Guangzhou: Sun Yat-sen University Press, 1985.
[24] 毛宗秀. BASIC语言常用数理统计方法程序汇编[M]. 杭州: 浙江科学技术出版社, 1983.
MAO Zongxiu. BASIC language program collection for common statistical methods[M]. Hangzhou: Zhejiang Science and Technology Press, 1983.
[25] 安鸿志, 顾岚. 统计模型与预报算法[M]. 北京: 气象出版社, 1986.
AN Hongzhi, GU Lan. Statistical models and prediction algorithms [M]. Beijing: China Meteorological Press, 1986.
[26] 徐士良. FORTRAN常用算法程序集[M]. 北京: 清华大学出版社, 1992.
XU Shiliang. FORTRAN algorithm program collection[M]. Beijing: Tsinghua University Press, 1992.
[27] 张启锐. 实用回归分析[M]. 北京: 地质出版社, 1988.
ZHANG Qirui. Applied regression analysis[M]. Beijing: Geological Publishing House, 1988.
[28] 黄俊钦. 静、动态数学模型的实用建模方法[M]. 北京: 机械工业出版社, 1988.
HUANG Junqin. Practical modeling methods for static and dynamic mathematical models[M]. Beijing: China Machine Press, 1988.
[29] 王行愚. 控制论基础[M]. 上海: 华北工学院出版社, 1989.
WANG Xingyu. Fundamentals of cybernetics[M]. Shanghai: North China Institute of Technology Press, 1989.

作 者 简 介

耿卫国（1971—），男，研究员，主要研究方向为液体火箭发动机试验与测试。
王晓磊（1974—），女，高级工程师，主要研究方向为液体火箭发动机试验与测试。