doi:10.11887/j.cn.202503012 http://journal.nudt.edu.cn

深度强化学习在导弹弹道规划中的应用

张 敬[1],李 彤[1*],李建锋[2],谭立国[3],张士峰[4]

(1. 军事科学院 国防科技创新研究院,北京 100071;2. 哈尔滨工业大学 控制理论与制导技术研究中心,黑龙江 哈尔滨 150000;3. 哈尔滨工业大学 空间环境与物质科学研究院,黑龙江 哈尔滨 150000;4. 国防科技大学 空天科学学院,湖南 长沙 410073)

摘 要:针对导弹弹道规划问题,搭建了适用性的Gym训练环境,基于双延迟深度确定性策略梯度框架设计了智能体网络结构,根据终端约束和过程约束设计奖励函数,形成了智能弹道规划方法。通过部署于嵌入式GPU计算加速平台,进行了拉偏仿真和对比测试,结果表明:该方法在不同射程任务要求下能够满足导弹能力和过程约束,有效克服环境干扰,具有针对不同对象模型的适应性。同时,该方法计算速度极快,远超流行的GPOPS-Ⅱ工具箱,单步弹道指令计算用时在ms以下,能够支持实时在线弹道生成,为工程应用提供了有效实现途径和技术支撑。

关键词:弹道规划;深度强化学习;导弹;嵌入式GPU平台

中图分类号:TP18;TP27;V24 **文献标志码**:A

文章编号:1001-2486(2025)03-109-10

论文拓展

Application of deep reinforcement learning to missile trajectory planning

ZHANG Jing[1], *LI Tong*[1*], *LI Jianfeng*[2], *TAN Liguo*[3], *ZHANG Shifeng*[4]

(1. National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing 100071, China;
2. Center for Control Theory and Guidance Technology, Harbin Institute of Technology, Harbin 150000, China;
3. Laboratory for Space Environment and Physical Sciences, Harbin Institute of Technology, Harbin 150000, China;
4. College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: Aiming for missile trajectory planning, an applicable Gym training evironment was established. An intelligent agent network structure and its reward functions were designed based on twin delayed deep deterministic policy gradient framework and according to terminal and process constraints, forming an intelligent trajectory planning method. Through deploying the algorithm on an embedded GPU computing acceleration platform, bias simulation and comparison tests were conducted. The results show that the method can reach the requirements of missile capability and process constraints under different range tasks and effectively overcome environmental disturbances with adaptability to distinct object models. Meanwhile, the method has an extremely fast calculation speed, far surpassing the popular GPOPS-Ⅱ toolbox. The computation time for single step trajectory command is less than a millisecond so that it can support real-time online trajectory generation, which provides an effective implementation path and technical support for engineering applications.

Keywords: trajectory planning; deep reinforcement learning; missile; embedded GPU platform

导弹弹道规划是一个多约束、强非线性的优化问题,贯穿于导弹总体、气动、制导、控制、动力以及结构等各个环节的设计过程。随着最优控制问题研究的深入和技术指标要求的提升,导弹弹道规划方法已由传统基于线性系统的解析方法,逐步发展为以直接法和间接法为主的数值解法,很多学者都对规划方法做了总结归纳和详细综述[1-7]。由于间接法需要求解Hamilton边值问题,计算较为复杂和烦琐,因此,直接法成为目前主流的弹道规划方法。直接法包含多种不同类型

收稿日期:2023-05-08

基金项目:国家部委基金资助项目(2022A000300)

第一作者:张敬(1982—),男,四川泸州人,助理研究员,博士,E-mail:zhang505jing@163.com

***通信作者**:李彤(1989—),男,河北保定人,助理研究员,博士,E-mail:li_tong122@126.com

引用格式:张敬,李彤,李建锋,等. 深度强化学习在导弹弹道规划中的应用[J]. 国防科技大学学报, 2025, 47(3): 109-118.

Citation: ZHANG J, LI T, LI J F, et al. Application of deep reinforcement learning to missile trajectory planning[J]. Journal of National University of Defense Technology, 2025, 47(3): 109-118.

方法，其中应用较为广泛的是伪谱法，尤其GPOPS工具箱软件[8]的出现使得Gauss伪谱法成为最为流行的弹道规划方法。伪谱法基于多项式有限基函数进行全局插值和配点，将微分方程约束转换为代数约束，并通过多项式求导近似微分方程动力学，因而具有较高的计算效率，同时对于复杂问题也能够分段处理[9]。但是，随着弹道规划任务的复杂化，例如突发故障、空中变轨以及紧急发射等新需求，伪谱法作为数值解法对计算资源相对依赖，并且解算耗时对于初始节点和配点较为敏感，无法满足求解时间要求，另外，伪谱法无法完全保证求解收敛性，因此，伪谱法难以实现实时在线计算，不满足应用可靠性要求。

随着智能时代的来临，深度强化学习作为智能优化方法得到广泛研究。深度强化学习有效结合了深度学习和强化学习的优势，很好地解决了环境感知与探索决策问题，能够通过环境交互试错自主形成适应实际任务的动作选择策略，具有良好的长期效应和动态鲁棒性[10]。在路径规划[11]、轨迹规划[12-13]和运动规划[14]等相关领域，一些学者已经基于深度强化学习理论开展了一定研究，但该类研究目前还处于探索阶段，很多研究仍需结合传统技术方案，更多聚焦于局部细节改善，无法实现整体策略的替代，同时这些分析也主要针对应用前景和面临挑战，无法为实际应用提供支撑。随着确定性策略梯度类算法被提出，如确定性策略梯度[15]（deterministic policy gradient，DPG）、深度确定性策略梯度[16]（deep deterministic policy gradient，DDPG）以及双延迟深度确定性策略梯度[17]（twin delayed deep deterministic policy gradient，TD3）等框架，其相比随机性策略梯度不需要在状态和动作空间积分，大幅度减小了训练所需的样本数据，提高了计算效率，因此深度强化学习应用领域愈加广泛，非常适用于解决导弹弹道规划问题，并已在轨迹规划领域取得一定研究成果[18-20]。同时，GPU计算硬件平台快速迭代发展，已经在一定程度上实现了高性能、低功耗和小型化要求，如NVIDIA公司的Jetson Xavier NX和华为的Atlas 200等，也为深度强化学习在导弹弹道规划问题领域的工程应用提供了硬件支持和实现途径，给予了替代传统弹道规划技术方案的可能。

由此，本文针对导弹弹道规划问题搭建深度强化学习环境，应用TD3框架对弹道规划方法进行设计，并部署于嵌入式GPU平台，通过开展拉偏仿真和对比测试，分析说明本文所设计方法相对传统技术方案的优势性。

1 深度强化学习环境搭建

1.1 导弹动力学与运动学建模

所研究的弹道规划问题主要针对导弹射程范围内的给定射程，在考虑弹道约束条件下，快速生成满足要求的规划弹道。应建立相应数学模型对其运动进行描述，考虑到导弹所执行的任务不同，导弹各项物理参数并不一致，机动方式也不相同，因此，建立明确数学模型是导弹弹道规划方法设计的基础。由于弹道规划不涉及导弹绕质心运动且该动态过程能够用二阶系统近似，因此，基于一般导弹结构和运动特点建立适用于导弹弹道规划的三自由度动力学与运动学模型。导弹动力学模型在速度坐标系中表示如式(1)所示[21]。

$$\begin{cases} m\dfrac{\mathrm{d}V}{\mathrm{d}t} = P\cos\alpha\cos\beta - X - mg\sin\theta \\ mV\dfrac{\mathrm{d}\theta}{\mathrm{d}t} = P(\sin\alpha\cos\gamma_{\mathrm{v}} + \cos\alpha\sin\beta\sin\gamma_{\mathrm{v}}) + \\ \qquad Y\cos\gamma_{\mathrm{v}} - Z\sin\gamma_{\mathrm{v}} - mg\cos\theta \\ -mV\cos\theta\dfrac{\mathrm{d}\psi_{\mathrm{v}}}{\mathrm{d}t} = P(\sin\alpha\sin\gamma_{\mathrm{v}} - \\ \qquad \cos\alpha\sin\beta\cos\gamma_{\mathrm{v}}) + Y\sin\gamma_{\mathrm{v}} + Z\cos\gamma_{\mathrm{v}} \end{cases} \tag{1}$$

其中：m和V为导弹质量和导弹速度；θ、ψ_{v}和γ_{v}分别为弹道倾角、弹道偏角和速度倾斜角；α和β分别为攻角和侧滑角；P为推力，由地面试车推力和发动机喷管出口压力差构成；g为重力加速度大小，考虑到导弹对象飞行空域和飞行包线，将地球视为匀质圆球处理，不再将重力加速度地视为常数，而将其描述为关于海拔的函数；X、Y和Z分别为阻力、升力和侧向力，由相应气动系数、导弹动压和特征面积乘积得到。由于仅考虑弹体三自由度运动，不涉及导弹舵面操纵，因此气动系数主要与攻角α、侧滑角β以及马赫数Ma相关，采用上述三个因素为变量的三维插值表，通过插值获得相应气动系数，另外空气密度采用国家标准大气模型（见GB 1920—80）。

导弹运动学模型在惯性坐标系下表示为

$$\begin{bmatrix} \dfrac{\mathrm{d}x}{\mathrm{d}t} \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} \\ \dfrac{\mathrm{d}z}{\mathrm{d}t} \end{bmatrix} = \boldsymbol{C}(\theta,\psi_{\mathrm{v}}) \begin{bmatrix} V \\ 0 \\ 0 \end{bmatrix} \tag{2}$$

式中，$\boldsymbol{C}(\theta,\psi_{\mathrm{v}})$为弹道坐标系至惯性坐标系的方向余弦矩阵。

采用攻角α、侧滑角β和滚转角γ作为弹道

指令,可应用于侧滑转弯(skid-to-turn,STT)($\gamma=0°$)和倾斜转弯(bank-to-turn,BTT)($\beta=0°$)机动类型导弹弹道的设计。因此,欧拉角几何关系方程的建立需基于 α、β、γ、θ 和 ψ_v 求解 γ_v 以及姿态角中的俯仰角 φ 和偏航角 ψ,具体方程形式根据坐标系变换转序不同而有所区别。

1.2　弹目相对运动描述

为便于设计深度强化学习奖励函数,需建立导弹相对于指定目标点的弹目相对运动学模型。导弹和目标点分别用 M 和 T 表示,则有

$$\begin{cases} R=\sqrt{x_r^2+y_r^2+z_r^2} \\ \dot{R}=\dfrac{x_r\dot{x}_r+y_r\dot{y}_r+z_r\dot{z}_r}{\sqrt{x_r^2+y_r^2+z_r^2}} \end{cases} \tag{3}$$

其中,R 和 $\dot{R}$ 分别为弹目相对距离及其导数,$x_r=x_T-x_M$,$y_r=y_T-y_M$,$z_r=z_T-z_M$。

纵向和横向视线角 q_ε 和 q_β 及其角速率 $\dot{q}_\varepsilon$ 和 $\dot{q}_\beta$ 分别表示为

$$\begin{cases} q_\varepsilon=\arctan\left(\dfrac{y_r}{\sqrt{x_r^2+z_r^2}}\right) \\ q_\beta=\arctan\left(\dfrac{-z_r}{x_r}\right) \\ \dot{q}_\varepsilon=\dfrac{(x_r^2+z_r^2)\dot{y}_r-y_r(x_r\dot{x}_r+z_r\dot{z}_r)}{(x_r^2+y_r^2+z_r^2)\sqrt{x_r^2+z_r^2}} \\ \dot{q}_\beta=\dfrac{z_r\dot{x}_r-x_r\dot{z}_r}{x_r^2+z_r^2} \end{cases} \tag{4}$$

1.3　参数偏差与环境干扰

参数偏差主要考虑在导弹气动、质量、动力等参数上施加拉偏扰动,通常气动系数拉偏范围为 ±30%,质量拉偏范围为 ±1%,推力拉偏范围为 ±5%,发射角偏差范围为 ±15°。

环境干扰主要考虑风场扰动,风场模型包括纬圈风和经圈风,本文中忽略风速随纬度、季节等长周期变化项,假设风速仅为高度的函数,其幅值取为随纬度、季节变化峰值内的随机值。风速随高度变化之间的关系如表 1 所示[22]。

风场扰动会改变导弹相对于空气相对运动速度大小和方向,计算导弹所受气动力需要考虑由此所产生的附加速度、附加攻角和附加侧滑角。

1.4　Gym 训练环境搭建

Gym 是 OpenAI 推出的强化学习训练环境,它覆盖的场景非常多,包括 Car-Pole、Mountain-Car 以及 Atari Go 等经典实验和游戏,提供了比较全面的智能体与环境交互功能,极大方便了用户对强化学习算法的实现和验证。

表 1　风场扰动模型

Tab. 1　Model of gust disturbance

高度/km	子午线风/(m/s)	纬度风/(m/s)
30	0.325 7	5.022
35	0.222 6	4.175
40	1.316	5.092
55	6.023	7.051
60	1.712	6.423
70	2.034	3.557
80	8.281	1.613
90	−8.739	2.641
100	7.399	6.542

本文弹道规划采用 Gym 框架搭建强化学习的训练环境。Gym 环境中主要包含上文建立的导弹动力学与运动学模型、弹目相对运动模型和参数偏差与环境干扰,以及动力学模型所需的导弹质量插值表、发动机推力曲线、三轴气动力三维(攻角、侧滑角和马赫数)插值表、大气压力密度计算、重力加速度计算、坐标转换矩阵解算、欧拉角解算和二阶过渡环节模型。针对 STT 机动类型导弹弹道规划问题进行研究,因此环境观测状态(state)和智能体动作(action)分别选定为

$$\begin{cases} \boldsymbol{s}(t)=[R,\ q_\varepsilon,\ q_\beta,\ \dot{R},\ \dot{q}_\varepsilon,\ \dot{q}_\beta]^{\mathrm{T}} \\ \boldsymbol{a}(t)=[\alpha,\ \beta]^{\mathrm{T}} \end{cases} \tag{5}$$

考虑导弹模型特性、变量数量级以及输入约束,对环境状态和智能体动作做归一化处理。

$$\begin{cases} \boldsymbol{s}(t)=\left[R\times10^{-3},\ q_\varepsilon\cdot\dfrac{180}{\pi},\ q_\beta\cdot\dfrac{180}{\pi},\right. \\ \qquad\left.\dfrac{\dot{R}}{5},\ \dot{q}_\varepsilon\cdot\dfrac{180}{\pi}\times10^2,\ \dot{q}_\beta\cdot\dfrac{180}{\pi}\times10^2\right]^{\mathrm{T}} \\ \boldsymbol{a}(t)=\left[\dfrac{\alpha}{\alpha_{\max}},\dfrac{\beta}{\beta_{\max}}\right]^{\mathrm{T}} \end{cases} \tag{6}$$

由此,通过上述处理,环境状态变量数量级在 [−100, 100] 之间,智能体动作变量数量级在 [−1, 1] 之间。

环境模型的单步推进(step)通过设定固定步长采用欧拉积分求解,完成结束判断(is_done)为导弹落地或超出过程约束限制。

2　基于 TD3 框架的弹道规划方法设计

2.1　TD3 深度强化学习算法框架

由于弹道连续性与整体性,探索的样本量受到一定限制,且导弹动力学方程计算时间相对较长,因此,更适合确定性策略梯度的深度强化学习

算法。采用 TD3 框架对弹道规划方法进行设计。TD3 是一种 Actor-Critic 框架的深度强化学习算法，在 DDPG 的基础上拓展而来，能够有效解决 DDPG 框架的 Q 网络过拟合问题，防止 Q 值过估计，是目前性能最优的主流确定性策略算法，其框架如图 1 所示[17]。

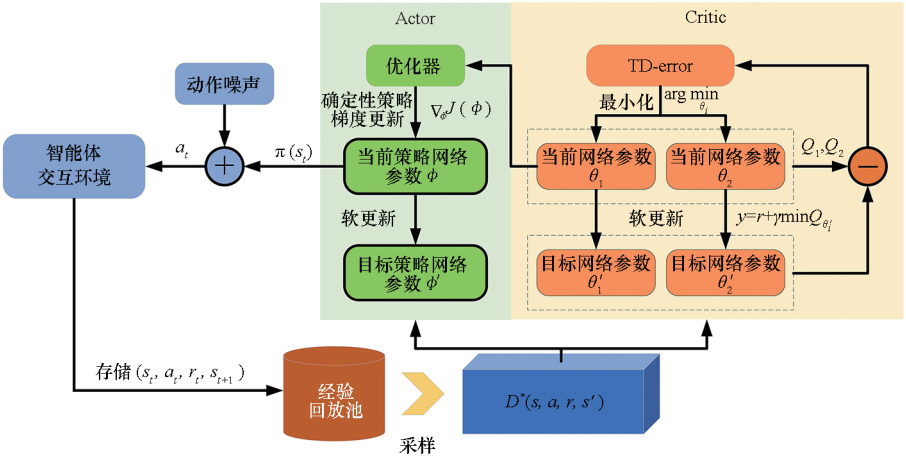


图 1 TD3 算法框架

Fig. 1 Algorithm framework of TD3

TD3 在 DDPG 基础上主要采用 3 个关键技术[23]提高算法的稳定性和性能，适用于解决弹道规划问题。

1）剪裁双 Q 学习。由于弹道规划问题中，导弹智能体的探索在经验池中分布不均匀，在采样训练过程中通常会造成 Critic 网络对 Q 值的过估计。TD3 算法目标估值计算采用双重网络中的最小值，能够有效防止 Critic 网络非均匀过估计问题。

$$y = r(s, a) + \gamma \min Q_{\theta_i'}(s', \tilde{a}') \quad i = 1, 2 \tag{7}$$

式中，y 为目标估值，$r(s, a)$、s 和 a 分别为当前时刻奖励值、环境状态和智能体动作，γ 为折扣率，$Q_{\theta_i'}(\cdot)(i=1, 2)$ 为双重 Critic 目标网络 Q 值估计，$\theta_i'(i=1, 2)$ 为双重 Critic 目标网络参数，s' 为下一时刻环境状态，$\tilde{a}'$ 为 Actor 目标网络下一时刻智能体策略动作的正则化。

2）目标策略平滑正则化。当导弹智能体落点位于目标点附近时，目标估计方差较大，容易导致过拟合问题。TD3 引入正则化方法来减少目标值方差，通过添加限幅噪声方式，在目标动作附近一个小邻域内随机生成动作，从而有利于平滑目标估值，防止自举带来的过估计，同时保持目标接近原始动作，提高目标估值准确性，保证网络训练过程的鲁棒性，同时能够改进具有故障情况的随机域，变相增加探索能力。

$$\begin{cases} \tilde{a}' = \pi_{\phi'}(s') + \varepsilon \\ \varepsilon \sim clip[N(0, \sigma), -c, c] \end{cases} \tag{8}$$

其中：$\pi_{\phi'}(\cdot)$ 为 Actor 目标网络策略；ϕ' 为 Actor 目标网络参数；ε 表示均值为 0、标准差为 σ、幅值为 c 的截断正态分布随机噪声。

3）Actor 策略延迟更新。在导弹智能体训练过程中，Critic 网络收敛是非常缓慢的，虽然随着更新步数上升，Critic 网络能够逐渐减小评估值与目标值之间的误差，但 Actor 网络保持相同更新频率会使策略动作出现离散行为，在 DDPG 框架下的导弹智能体训练几乎无法收敛。TD3 算法通过使 Actor 网络更新频率低于 Critic 网络，在保证目标估值稳定后再更新策略，能够解决 Actor 网络在训练过程中的发散问题。

2.2 智能体网络结构设计

基于 TD3 框架的弹道规划智能体网络结构需要设计 6 个神经网络，包括 1 个 Actor 网络、1 个 Actor 目标网络、2 个 Critic 网络、2 个 Critic 目标网络。

Actor 网络用于实现输入状态 s 到输出动作 a 之间的映射，即 $\pi_\phi(s)$，ϕ 为网络参数。根据环境状态维数输入神经元个数为 6，根据智能体动作维数输出神经元个数为 2，考虑到计算精度与模型复杂度，采取 4 层全连接层反向传播（back propagation，BP）神经网络结构 6 − 256 − 256 − 2。同时，为了克服梯度消失问题，提高训练速度，采用修正线性单元（rectified linear unit，ReLU）函数作为激活函数。Actor 目标网络与 Actor 网络结构一致。

Critic 网络用于实现从输入状态 s 和动作 a 到输出 Q 值估计函数 $Q(s, a)$ 之间的映射。根据环境状态维数和智能体动作维数，输入神经元个数为 8（6 + 2），输出神经元个数为 1，同样采用 4 层全连接层 BP 神经网络结构 8 − 256 − 256 − 1，并采用 ReLU 函数作为激活函数。Critic 目标网

络与 Critic 网络结构一致。

2.3　奖励函数设计

导弹弹道规划通常需要考虑终端约束、过程约束及控制量约束。终端约束是指在弹道末端需满足的条件,即最优控制问题中的边界条件;过程约束,即飞行过程中弹道参数必须满足的约束条件,包括导弹能够承受的动压、过载、气动热及机动能力等因素;控制量约束包括攻角、过载或推力等设计输入限制。由于控制量约束已在智能体设计动作范围内给予考虑,弹道约束主要考虑终端约束和过程约束。其中,终端约束包括落角约束和落点约束,过程约束包括法向过载约束和动压约束,具体表示如下。

终端约束:

$$\begin{cases} q_\varepsilon(t_\mathrm{f}) = q_\mathrm{d} \\ R(t_\mathrm{f}) = 0 \end{cases} \tag{9}$$

过程约束:

$$\begin{cases} n_y = (Y\cos\alpha + X\sin\alpha)/(mg) \leqslant n_y^{\max} \\ q = \dfrac{1}{2}\rho V^2 \leqslant q_{\max} \end{cases} \tag{10}$$

其中,t_f 为终端时间,q_d 为期望视线角,n_y 为导弹法向过载,q 为导弹动压,ρ 为空气密度。

为引导智能体到达给定目标点,同时满足上述弹道约束,奖励函数由三部分构成。

1)距离奖励:

$$r_1(t) = w_1 \cdot \frac{R(t-1) - R(t)}{\|\boldsymbol{V}(t)\|} \tag{11}$$

式中:$R(t-1)$ 和 $R(t)$ 分别为上一时刻和当前时刻弹目相对距离;$\|\boldsymbol{V}(t)\|$ 为速度矢量大小;w_1 为奖励权值系数,默认取值为 1。

2)终端约束奖励:

$$r_2(t) = w_2 \cdot \left[\mathrm{e}^{-\left(\frac{q_\varepsilon - q_\mathrm{d}}{\zeta}\right)^2} + r_2^{t_\mathrm{f}}\right] \tag{12}$$

式中,

$$r_2^{t_\mathrm{f}} = \begin{cases} r_{\mathrm{act}} & \|\boldsymbol{R}\| < d_{\min} \\ 0 & \text{其他} \end{cases} \tag{13}$$

其中:$r_2^{t_\mathrm{f}}$ 为落点约束奖励,用于激励智能体加速收敛,仅当 is_done 判断为 True 且导弹正常落地时生效;$d_{\min}$ 为弹目可接受距离;r_{act} 为设定激励值;ζ 为常数(取值为 20);指数部分用于约束落角,同时能够使规划弹道更加平滑;w_2 为奖励权值系数,默认取值为 1。

3)过程约束奖励:

$$r_3(t) = w_3 \cdot r_3^{t_\mathrm{f}} \tag{14}$$

式中,

$$r_3^{t_\mathrm{f}} = \begin{cases} -r_{\mathrm{act}} & n_y > n_y^{\max} \text{ or } q > q_{\max} \\ 0 & \text{其他} \end{cases} \tag{15}$$

其中:$r_3^{t_\mathrm{f}}$ 为过程约束奖励;该奖励用于惩罚智能体超出过程约束范围,仅当 is_done 判断为 True 且导弹超出过程约束时生效;w_3 为奖励权值系数,默认取值为 1。

最终奖励函数为上述奖励值之和,即

$$r_{\mathrm{total}}(t) = r_1(t) + r_2(t) + r_3(t) \tag{16}$$

2.4　智能体训练过程

智能体基于 TD3 算法更新过程进行训练。训练过程通过在经验池(replay buffer)中采样批次(batch size)数据,对各网络进行更新。除采用上文提到的 3 个关键技术外,Actor 网络通过最大化累积期望奖励进行更新(确定性策略梯度),任选双重 Critic 网络之一计算 Q 值,动作不施加噪声,本文选定为双重 Critic 网络中的第 1 个 Critic 网络;双重 Critic 网络均通过最小化目标估值的时间差分误差(temporal difference error,TD-error)逼近贝尔曼方程进行更新;Actor 和 Critic 目标网络均通过软更新方式进行更新。智能体基于 TD3 算法整体训练流程如算法 1[17] 所示。

算法 1　TD3 算法智能体训练伪代码

Alg. 1　Pseudo code of TD3 algorithm for agent training

输入:用参数 θ_1、θ_2 初始化 Critic 网络 Q_{θ_1}、Q_{θ_2} 和 Critic 目标网络,用 ϕ 参数初始化 Actor 网络和 Actor 目标网络,初始化经验池 D

输出:Actor 网络参数 ϕ

for $t = 1$ **to** T **do**
　选择带有噪声的动作 $a \sim \pi_\phi(s) + \varepsilon$, $\varepsilon \sim N(0, \sigma)$,观察奖励 r 和新的状态 s'
　将数据$(s, a, r, s', done)$存入经验池 D
　if is_done = True **then**
　　环境重置 *reset*
　end if
　if $t > T_{\mathrm{update}}$ **do**
　　从经验池 D 中取出 batch size 个$(s, a, r, s', done)$样本组成一组 mini-batch
　　$\tilde{a}' \leftarrow \pi_{\phi'}(s') + \varepsilon$, $\varepsilon \sim clip[N(0, \sigma), -c, c]$,计算目标动作
　　$y \leftarrow r + \gamma \min\limits_{i=1,2} Q_{\theta_i'}(s', \tilde{a}')$,计算目标 Q 值
　　$\theta_i \leftarrow \arg\min\limits_{\theta_i} |D|^{-1} \sum [y - Q_{\theta_i}(s,a)]^2$,更新双重 Critic 网络参数 θ_i
　　if t % T_{policy} **do**
　　通过确定性策略梯度更新 Actor 网络参数 ϕ:
　　$\nabla_\phi J(\phi) = |D|^{-1} \sum \nabla_a Q_{\theta_1}(s,a)\big|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$
　　目标网络参数软更新:
　　$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'(i = 1, 2)$
　　$\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
　　end if
　end if
end for

3 仿真与测试

3.1 仿真与测试环境

基于 TD3 弹道规划智能体网络训练选择搭载 NVIDIA GeForce RTX 1660Ti 独立显卡 PC 作为硬件平台，仿真与测试环境选择 Realtimes RTSO-6002 载板搭载 NVIDIA Jetson Xavier NX 核心作为算法运行的硬件部署环境，其尺寸仅为信用卡大小。软件采用 Python 语言编写，采用 PyCharm 作为编译器，由 Anaconda 进行集成开发，基于 PyTorch 强化学习架构搭建和 OpenAI Gym 环境训练，软件环境配置如表 2 所示。

表 2 软件环境配置

Tab. 2 Software environment settings

名称	版本
操作系统	Ubuntu 18.04
CUDA(cuDNN)	10.2(8.0)
Python	3.8
PyCharm	2020.1.1 arm64
Anaconda	2021.05-Linux-aarch arm64
Gym	0.23.1
PyTorch	1.10

选择导弹模型为水平垂直对称结构，导弹纵向和横向动态特性一致，因此为便于测试，弹道规划智能体网络训练仅针对二维纵向弹道，并采用最大设计射程作为固定射程进行训练，网络训练参数设置如表 3 所示。

表 3 网络训练参数

Tab. 3 Parameters of network training

参数	数值
网络学习速率	3×10^{-4}
折扣率	0.99
软更新速率	3.8
探索噪声	0.03
目标策略噪声	0.2
目标策略噪声	0.5
batch size	4×256
replay buffer	4×10^{6}
训练开始步数	5×10^{3}
延迟更新步数	2
训练总步数	1×10^{6}
训练步长	1 s
测试步长	10 ms

3.2 算法有效性测试

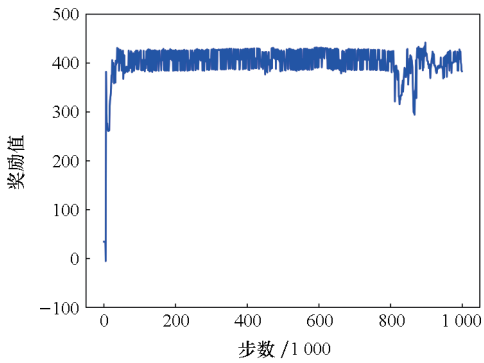
为验证所设计的弹道规划方法的有效性，选择两个特性不同的导弹模型进行训练，分别为典型岸舰导弹和反坦克导弹，其基本参数如表 4 所示。

表 4 导弹模型参数

Tab. 4 Parameters of missile model

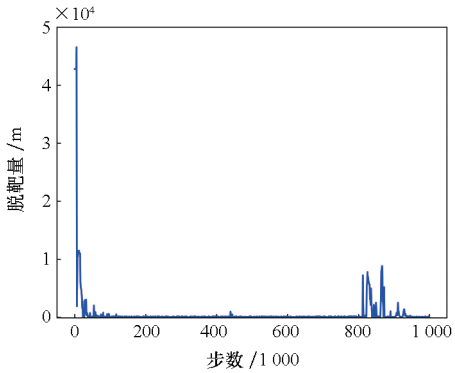
参数	岸舰导弹	反坦克导弹
发射质量/kg	1 330	61.5
特征长度/m	7.89	2
特征面积/m^2	0.096 2	0.015 7
射程范围/km	10 ~ 50	1 ~ 7
发射角/(°)	45	45
动作范围/(°)	[-10, 10]	[-20, 20]

两个导弹模型的智能体网络分别采用最大设计射程 50 km 和 7 km 进行训练，每隔 1 000 步对训练中的智能体网络进行测试，得到总奖励值和脱靶量，如图 2 和图 3 所示。



(a) 总奖励值曲线

(a) Curve of total reward

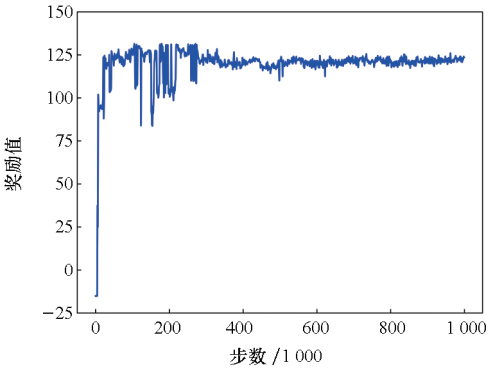


(b) 脱靶量曲线

(b) Curve of miss distance

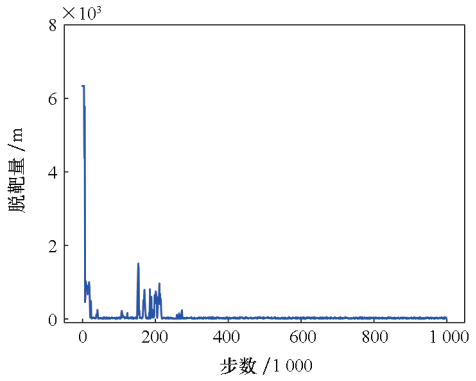
图 2 岸舰导弹模型训练过程

Fig. 2 Training process of anti-ship missile model



(a) 总奖励值曲线

(a) Curve of total reward



(b) 脱靶量曲线

(b) Curve of miss distance

图 3　反坦克导弹模型训练过程

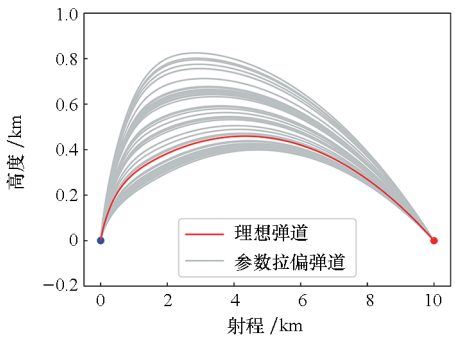
Fig. 3　Training process of anti-tank missile model

由图 2 ~ 3 可知,随着探索步数增加,两导弹模型智能体网络总奖励值均达到收敛。由脱靶量曲线可以看出,智能体到达目标点成功率逐渐增加,最终落点误差分别为 1.25 m 和 0.57 m,精度达到 m 级,同时说明弹道满足相应约束。两导弹模型智能体与环境交互结果验证了所设计的深度强化学习弹道规划方法的可行性和有效性,并且说明该方法能够应用于不同导弹模型,具有针对不同对象模型的适应性。

3.3　拉偏仿真与性能测试

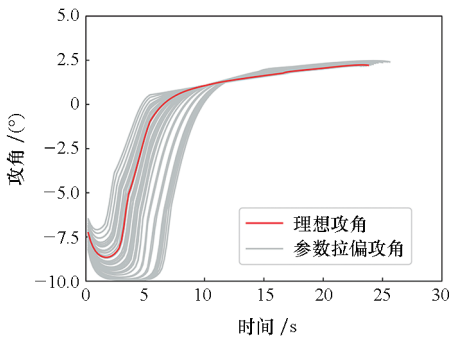
为进一步测试所设计的弹道规划方法的性能,考虑 1.3 节所述参数偏差和环境干扰,针对已训练的岸舰导弹模型智能体网络进行拉偏仿真,在设计射程范围内每隔 10 km 设定目标射程进行 100 次 Monte-Carlo 打靶性能测试,仿真测试步长为 10 ms,测试结果如图 4 ~ 8 所示。

由图 4 ~ 8 可知,岸舰导弹模型智能体网络能够克服气动、质量、推力和发射角的参数偏差以及风场环境扰动,达到给定的不同目标射程,满足终端约束和过程约束要求,生成的弹道指令攻角均



(a) 弹道曲线

(a) Curve of trajectory

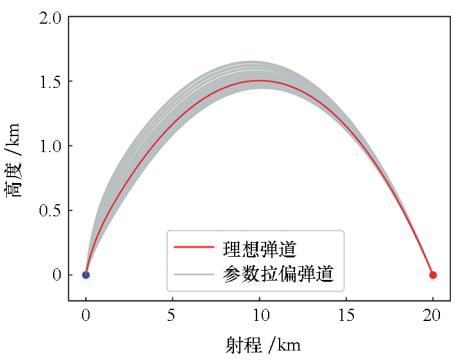


(b) 攻角曲线

(b) Curve of angle of attack

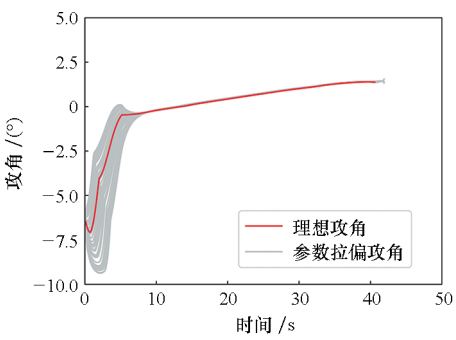
图 4　10 km 目标射程 Monte-Carlo 拉偏仿真

Fig. 4　Monte-Carlo bias simulation of 10 km target range



(a) 弹道曲线

(a) Curve of trajectory

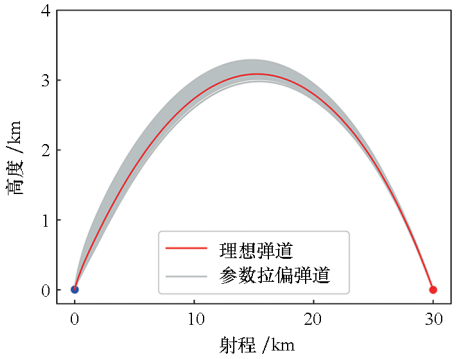


(b) 攻角曲线

(b) Curve of angle of attack

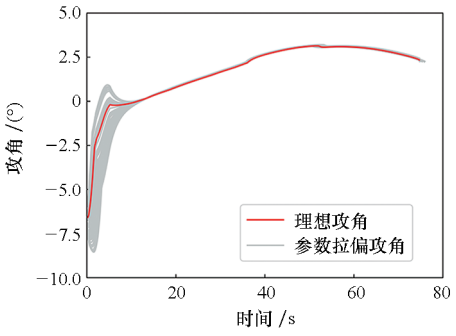
图 5　20 km 目标射程 Monte-Carlo 拉偏仿真

Fig. 5　Monte-Carlo bias simulation of 20 km target range



（a）弹道曲线

（a）Curve of trajectory

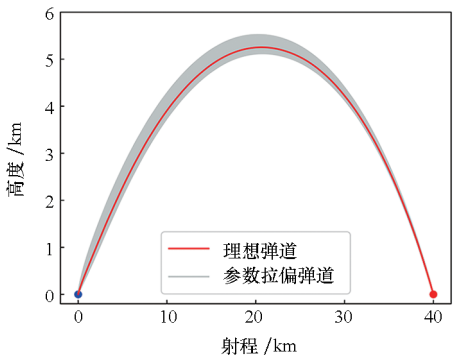


（b）攻角曲线

（b）Curve of angle of attack

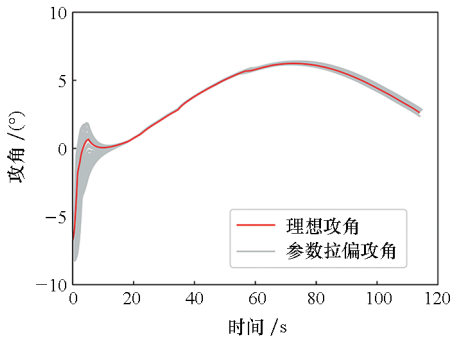
图 6　30 km 目标射程 Monte-Carlo 拉偏仿真

Fig. 6　Monte-Carlo bias simulation of 30 km target range



（a）弹道曲线

（a）Curve of trajectory

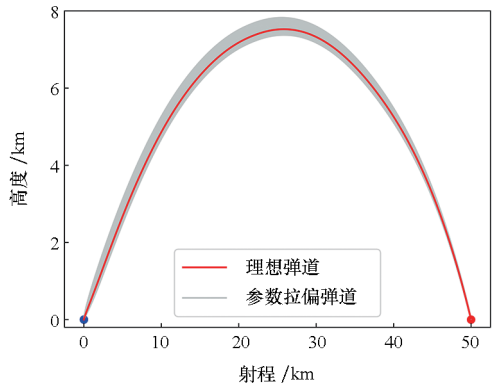


（b）攻角曲线

（b）Curve of angle of attack

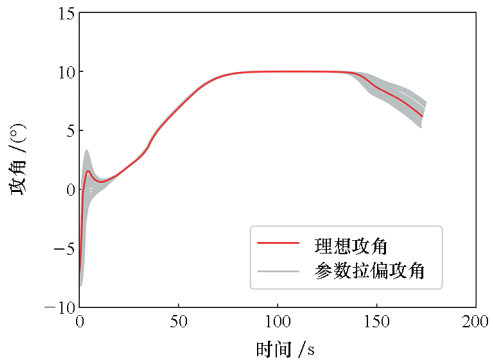
图 7　40 km 目标射程 Monte-Carlo 拉偏仿真

Fig. 7　Monte-Carlo bias simulation of 40 km target range



（a）弹道曲线

（a）Curve of trajectory



（b）攻角曲线

（b）Curve of angle of attack

图 8　50 km 目标射程 Monte-Carlo 拉偏仿真

Fig. 8　Monte-Carlo bias simulation of 50 km target range

在动作范围内，并且弹道曲线整体较为平滑，随着弹道射程的增加，攻角曲线逐渐由负向正变化，过渡平缓，符合弹道设计要求。每个目标射程 100 次仿真平均落点误差分别为 2.15 m、1.72 m、0.31 m、1.63 m 和 1.52 m，弹道落点精度较高。由此，所设计的深度强化学习弹道规划方法能够满足导弹不同射程任务下的鲁棒性要求，具备较好的工程应用前景。

3.4　运算速度对比测试

为验证所设计的弹道规划方法能够满足导弹制导控制系统计算周期和计算速度要求，本小节以 Gauss 伪谱法作为对比方法，采用搭载 Intel i9-9880H 的 PC 作为计算平台，对比 Xavier NX 平台的深度强化学习弹道规划方法，测试运行时间及弹道规划结果。Gauss 伪谱法计算效率高，应用普遍，在间接法中具有较强代表性，能够作为参考验证性能。

本文采用工具箱 GPOPS toolbox，通过 Gauss 伪谱法，针对岸舰导弹模型及其约束，在设计射程范围内每隔 10 km 设定目标射程进行弹道规划，

规划结果如图 9 和图 10 所示,弹道规划用时与已训练岸舰导弹模型智能体网络对比结果如表 5 所示。

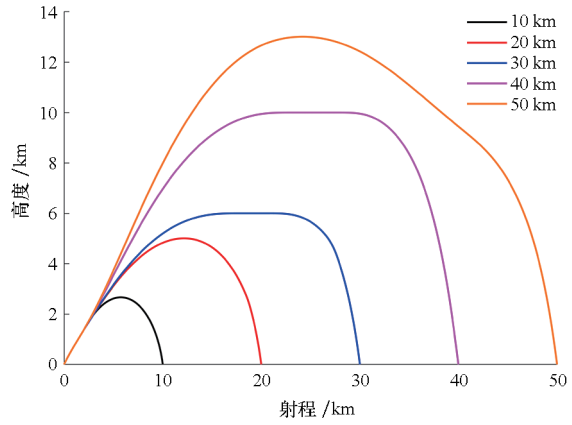


图 9　Gauss 伪谱法不同射程弹道规划结果

Fig. 9　Results of Gauss pseudo-spectral method for different ranges

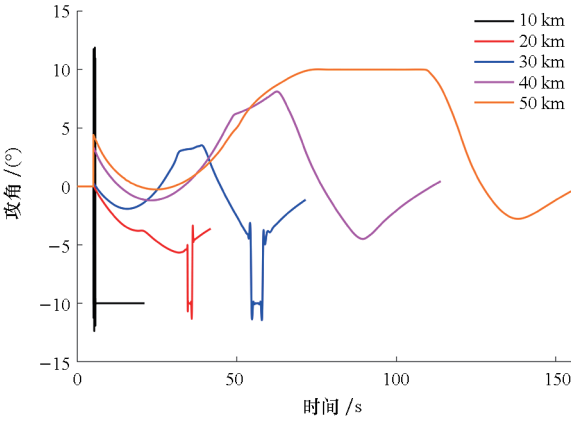


图 10　Gauss 伪谱法不同射程攻角曲线

Fig. 10　Curves of angle of attack of Gauss pseudo-spectral method for different ranges

表 5　弹道规划用时对比

Tab. 5　Comparison of trajectory planning time

目标射程/km	智能体网络规划用时/s	智能体弹道飞行时间/s	GPOPS 规划用时/s
10	0.578 6	26.54	96.867 8
20	1.264 7	42.92	105.175 1
30	1.928 8	93.16	9.711 1
40	2.397 7	113.58	135.231 6
50	3.848 6	154.57	8.207 7

由图 9 和图 10 可知,Gauss 伪谱法相比已训练的岸舰导弹模型智能体网络,所规划的弹道平滑度较差,弹道指令过渡特性不够平缓,部分指令超出了攻角约束范围,并且弹道高度有较大差异,弹道部分阶段达到导弹过载和落角约束极限,弹道能量损耗相对较大。整体而言,Gauss 伪谱法弹道规划水平低于智能体网络。

由表 5 可知,智能体网络弹道规划用时小于 GPOPS 工具箱。由于智能体网络按 10 ms 步长生成弹道指令,规划用时随着目标射程而增加,平均每个步长弹道指令生成用时 0.232 6 ms,远小于实际导弹要求的 1 ~ 10 ms 制导控制周期,满足实用性和工程性要求。相较而言,Gauss 伪谱法受初始配点影响较大,规划用时与目标射程不相关,且规划用时较长,难以支持实时在线规划。由此,智能体网络不仅能够实现整体弹道的离线快速规划,也能够支持弹道指令的在线快速生成,相比传统的弹道规划方法和工具箱,更具有运算速度优势,且嵌入式 GPU 计算加速平台依托极小尺寸的 Jetson Xavier NX,工程实用性较强。

4　结论

本文采用 Gym 搭建了导弹弹道规划问题的深度强化学习环境,基于 TD3 深度强化学习框架设计了一种导弹弹道规划方法,并部署于 Jetson Xavier NX 微小型计算平台,通过拉偏仿真和对比测试对其性能进行了分析和验证。研究表明:

1)所设计弹道规划方法具有较好的鲁棒性和适应性,其与环境的交互策略对于不同射程任务要求均能够满足导弹能力和过程约束,具备应对突发情况和环境干扰的优势;

2)所设计弹道规划方法具有快速性优势,计算速度远超过现有流行软件工具,且单步弹道指令计算用时在 ms 级以下,能够支持实时在线弹道生成;

3)所设计弹道规划方法能够部署于微小型嵌入式 GPU 计算加速硬件平台,为工程应用提供有效实现途径和技术支撑。

参考文献(References)

[1]　BETTS J T. Survey of numerical methods for trajectory optimization [J]. Journal of Guidance, Control, and Dynamics, 1998, 21(2): 193 – 207.

[2]　雍恩米, 陈磊, 唐国金. 飞行器轨迹优化数值方法综述[J]. 宇航学报, 2008, 29(2): 397 – 406.
YONG E M, CHEN L, TANG G J. A survey of numerical methods for trajectory optimization of spacecraft [J]. Journal of Astronautics, 2008, 29(2): 397 – 406. (in Chinese)

[3]　陈聪, 关成启, 史宏亮. 飞行器轨迹优化的直接数值解法综述[J]. 战术导弹控制技术, 2009, 31(2): 33 – 40.
CHEN C, GUAN C Q, SHI H L. Survey of numerical methods for direct aircraft trajectory optimization [J]. Control Technology of Tactical Missile, 2009, 31(2): 33 – 40. (in

Chinese)

[4] 陈功, 傅瑜, 郭继峰. 飞行器轨迹优化方法综述[J]. 飞行力学, 2011, 29(4): 1 – 5.
CHEN G, FU Y, GUO J F. Survey of aircraft trajectory optimization methods[J]. Flight Dynamics, 2011, 29(4): 1 – 5. (in Chinese)

[5] 黄国强, 陆宇平, 南英. 飞行器轨迹优化数值算法综述[J]. 中国科学: 技术科学, 2012, 42(9): 1016 – 1036.
HUANG G Q, LU Y P, NAN Y. A survey of numerical algorithms for trajectory optimization of flight vehicles[J]. Scientia Sinica (Technologica), 2012, 42(9): 1016 – 1036. (in Chinese)

[6] 黄长强, 国海峰, 丁达理. 高超声速滑翔飞行器轨迹优化与制导综述[J]. 宇航学报, 2014, 35(4): 369 – 379.
HUANG C Q, GUO H F, DING D L. A survey of trajectory optimization and guidance for hypersonic gliding vehicle[J]. Journal of Astronautics, 2014, 35(4): 369 – 379. (in Chinese)

[7] 崔乃刚, 郭冬子, 李坤原, 等. 飞行器轨迹优化数值解法综述[J]. 战术导弹技术, 2020(5): 37 – 51, 75.
CUI N G, GUO D Z, LI K Y, et al. A survey of numerical methods for aircraft trajectory optimization[J]. Tactical Missile Technology, 2020(5): 37 – 51, 75. (in Chinese)

[8] PATTERSON M A, RAO A V. GPOPS – Ⅱ: a MATLAB software for solving multiple-phase optimal control problems using hp-adaptive Gaussian quadrature collocation methods and sparse nonlinear programming[J]. ACM Transactions on Mathematical Software, 2014, 41(1): 1 – 37.

[9] BENSON D A, HUNTINGTON G T, THORVALDSEN T P, et al. Direct trajectory optimization and costate estimation via an orthogonal collocation method[J]. Journal of Guidance, Control, and Dynamics, 2006, 29(6): 1435 – 1440.

[10] SUTTON R S, BARTO A G. Reinforcement learning: anintroduction[M]. 2nd ed. Cambridge, Massachusetts: The MIT Press, 2018.

[11] 段建民, 陈强龙. 利用先验知识的 *Q*-Learning 路径规划算法研究[J]. 电光与控制, 2019, 26(9): 29 – 33.
DUAN J M, CHEN Q L. Prior knowledge based *Q*-Learning path planning algorithm[J]. Electronics Optics & Control, 2019, 26(9): 29 – 33. (in Chinese)

[12] LI S, WU F, LUO S Y, et al. Dynamic online trajectory planning for a UAV-enabled data collection system[J]. IEEE Transactions on Vehicular Technology, 2022, 71(12): 13332 – 13343.

[13] 李跃, 邵振洲, 赵振东, 等. 面向轨迹规划的深度强化学习奖励函数设计[J]. 计算机工程与应用, 2020, 56(2): 226 – 232.
LI Y, SHAO Z Z, ZHAO Z D, et al. Design of reward function in deep reinforcement learning for trajectory planning[J]. Computer Engineering and Applications, 2020, 56(2): 226 – 232. (in Chinese)

[14] 孙辉辉, 胡春鹤, 张军国. 移动机器人运动规划中的深度强化学习方法[J]. 控制与决策, 2021, 36(6): 1281 – 1292.
SUN H H, HU C H, ZHANG J G. Deep reinforcement learning for motion planning of mobile robots[J]. Control and Decision, 2021, 36(6): 1281 – 1292. (in Chinese)

[15] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[J]. Proceedings of Machine Learning Research, 2014, 32(1): 387 – 395.

[16] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2019 – 07 – 05)[2023 – 04 – 01]. https://arxiv.org/abs/1509.02971v6.

[17] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL]. (2018 – 10 – 22)[2023 – 04 – 01]. https://arxiv.org/abs/1802.09477v3.

[18] BAO C Y, WANG P, HE R Z, et al. Autonomous trajectory planning method for hypersonic vehicles in glide phase based on DDPG algorithm[J]. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2022, 237(8): 1855 – 1867.

[19] XU X B, CHEN Y S, BAI C C. Deep reinforcement learning-based accurate control of planetary soft landing[J]. Sensors, 2021, 21(23): 8161.

[20] WU T C, WANG H L, LIU Y H, et al. Learning-based interfered fluid avoidance guidance for hypersonic reentry vehicles with multiple constraints[J]. ISA Transactions, 2023, 139: 291 – 307.

[21] 钱杏芳, 林瑞雄, 赵亚男. 导弹飞行力学[M]. 北京: 北京理工大学出版社, 2022.
QIAN X F, LIN R X, ZHAO Y N. Missile flight mechanics[M]. Beijing: Beijing Institute of Technology Press, 2022. (in Chinese)

[22] DROB D P, EMMERT J T, CROWLEY G, et al. An empirical model of the Earth′s horizontal wind fields: HWM07[J]. Journal of Geophysical Research: Space Physics, 2008, 113: A12304.

[23] DONG H, DING Z H, ZHANG S H. Deep reinforcement learning fundamentals, research and applications: fundamentals, research and applications[M]. Berlin: Springer, 2020.