http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2021.0436

基于深度强化学习的跳跃式导弹轨迹优化算法

龚开奇[1]，魏宏夔[2]，李嘉玮[2]，宋晓[1,*]，李勇[1]，李怡昕[2]，张岳[2]

(1. 北京航空航天大学 网络空间安全学院，北京 100191；2. 北京电子工程总体研究所，北京 100191)

摘 要：导弹的跳跃飞行过程可建模为时变非线性微分方程组，该方程组难以得到解析解，给导弹的轨迹优化带来很大的困难。针对该问题，提出一种基于双深度Q网络的带网络优选（NEO）策略的深度强化学习（NEO-DDQN）算法，所提算法以跳跃导弹航程最大化为优化目标，在热流密度、动压、过载及末速度等约束下，求解跳跃导弹的轨迹优化问题。设计问题的动作空间、状态空间和奖励函数；确定算法关键参数学习率的取值及合适的贪心策略，并提出NEO策略，得到所提NEO-DDQN算法；开展与最优恒定攻角（OCAOA）方案、遗传算法(GA)的对比实验。结果表明：NEO策略有效提升了算法的求解稳定性且将航程提升了2.52%；与OCAOA方案、GA相比，所提算法使跳跃导弹航程分别提高了2.61%和1.33%；所提算法还避免了直接求解复杂非线性微分方程，为轨迹优化问题提供了一种新型的基于学习的算法。

关 键 词：导弹；跳跃飞行；轨迹优化；深度强化学习；网络优选

中图分类号：V448.23; TP183

文献标志码：A **文章编号**：1001-5965(2023)06-1383-11

高超声速飞行器的相关理论技术研究是当前航空航天领域的热点研究问题[1-2]，轨迹优化是高超声速飞行器的关键技术之一[3]，其直接影响飞行器的机动能力、射程等重要性能指标，对气动布局、制导控制、动力和结构等多个分系统的优化设计也至关重要[4]。本文所研究的跳跃导弹是一种典型的高超声速飞行器，其无动力跳跃飞行阶段是大多数探测系统的盲区，因为跳跃飞行段的弹道很难预测[5]。跳跃导弹的轨迹优化是非线性最优控制问题，所对应的动力学模型是时变的非线性微分方程组，该方程组难以得到解析解[6]。目前常用的轨迹优化问题求解方法可分为3类：间接法、直接法、启发式算法。

间接法基于变分法和极小值原理，将飞行器轨迹优化问题转换为边值问题。文献[7]基于间接法生成了火箭上升段弹道优化初始猜测解，文献[8]使用间接法求解了飞行器垂直平面两点边值弹道优化问题，文献[9]利用遗传算法（genetic algorithm, GA）为间接法找到了合适的初始猜测解。间接法的优点是解的高保真度，但间接法也存在明显的缺陷。第一，最优性的必要条件，如伴随方程和横截条件等，需要利用最优控制理论的知识手工推导出来，而所涉及的实际航空航天应用中的微分方程往往非常复杂。第二，通常需要构造一个非常好的初始猜测解来实现边值问题的收敛，这也限制了间接法在工程上的应用。

为了避免间接法对初值高度敏感且在复杂约束下难以求解等问题，研究人员提出打靶法[10]、配点法[11]、伪谱法[12-13]等直接法，直接法将轨迹优化连续最优控制问题离散化，并转换为参数优化问题，再通过优化算法直接对性能指标寻优。传统优化算法在解决目标函数连续或可微、线性约束、小

收稿日期：2021-08-02；**录用日期**：2021-11-11；**网络出版时间**：2021-12-14 12:49
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20211213.1850.002.html
基金项目：国家重点研发计划 (2018YFB1702703)
*通信作者. E-mail：songxiao@buaa.edu.cn

引用格式：龚开奇，魏宏夔，李嘉玮，等. 基于深度强化学习的跳跃式导弹轨迹优化算法 [J]. 北京航空航天大学学报，2023，49（6）：1383-1393. GONG K Q，WEI H K，LI J W，et al. Trajectory optimization algorithm of skipping missile based on deep reinforcement learning [J]. Journal of Beijing University of Aeronautics and Astronautics，2023，49（6）：1383-1393（in Chinese）.

规模问题时能求得精确最优解，而求解飞行器轨迹优化这类复杂非线性、大规模多约束问题时，无需梯度信息的智能启发式算法（如GA、粒子群算法等）表现出很强的适用性[14-17]。然而，一旦问题的求解空间显著增大，此类方法的求解速度慢，同时易陷入局部最优解。

深度学习和强化学习等数据驱动方法是目前各领域研究的热点[18]，二者结合的深度强化学习方法近年来受到学者的广泛关注。目前，在飞行器的控制与优化领域，已有部分研究采用深度强化学习的方法。文献[19]设计了含稀疏奖励的奖励函数，利用深度确定性策略梯度（deep deterministic policy gradient, DDPG）算法解决了变体飞行器自主变形决策问题。为解决卫星姿态控制问题，文献[20]在DDPG算法中引入启发式搜索，提升了算法的收敛速度。文献[21]将双深度Q网络（double deep Q-network, DDQN）算法应用于高海拔长航时的太阳能无人飞行器能量最优的三维飞行策略研究，该算法与恒定高度速度相比提高了50%的飞行续航能力。文献[22]利用强化学习算法中的决斗双深度Q网络（dueling double DQN, D3QN）算法解决弹道导弹中段突防问题，该算法训练所得的深度神经网络控制器很好地规避了拦截器的攻击。深度强化学习将控制问题转换为序列决策问题进行求解，其以全新的视角解决了文献[19-22]中的连续控制问题。然而，却少有研究利用此类方法求解导弹轨迹优化这一高度非线性的最优控制问题[23]。所以，本文利用深度强化学习方法离线训练导弹的深度神经网络决策控制器，基于该控制器即可根据导弹飞行状态实时调整控制量，实现导弹的实时优化控制。

本文在过程约束如热流密度、动压、过载及终端约束如末速度约束等约束条件下，以航程最大化为目标、以攻角为控制量，研究跳跃导弹的轨迹优化问题。针对该问题，提出一种基于网络优选（network optimization, NEO）策略的DDQN算法。首先，将轨迹优化问题进行合理的分段离散处理，且由于整数级别的攻角能达到较高的控制精度，故本文采用离散的整数攻角控制，因而选用DDQN算法用于求解这一动作空间离散的问题；其次，设计问题的动作空间、状态空间和奖励函数，确定算法关键性参数学习率的取值及合适的贪心策略，提出NEO策略，得到NEO-DDQN算法；最后，对比NEO-DDQN算法与最优恒定攻角（optional constant angle of attack, OCAOA）方案、GA的优化结果。

1 问题描述

1.1 动力学描述

本文跳跃导弹外形为轴对称的回转体，三维模型如图1所示。由于导弹横纵面有相似气动特性和运动规律，且本文所研究问题为航程最大的优化问题，故问题可简化为跳跃导弹在纵平面内无动力飞行的弹道优化。

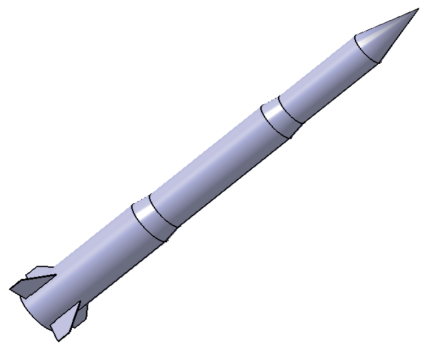


图1 跳跃导弹的三维模型

Fig. 1 Three dimensional model of skipping missile

图2为导弹的受力分析，其中O_2为地心，地面坐标系Oxy以飞行起点在地面的垂直投影O为原点，Ox为纵平面内过原点的水平方向，指向目标为正，Oy在纵平面内，垂直于Ox向上；机体坐标系Ox_1y_1以导弹质心O_1为原点，O_1x_1为弹体纵轴，指向弹头为正，O_1y_1轴在弹体纵向对称面内，与O_1x_1垂直，向上为正。根据牛顿第二定律，可得导弹在地面坐标系下的运动微分方程为

$$\begin{cases}\dfrac{\mathrm{d}x}{\mathrm{d}t}=v\cos\beta\\[2ex]\dfrac{\mathrm{d}y}{\mathrm{d}t}=v\sin\beta\end{cases}\tag{1}$$

式中：v为导弹速度。

对导弹进行受力分析，可得导弹的动力学方程为

$$\begin{cases}\dfrac{\mathrm{d}v}{\mathrm{d}t}=\dfrac{(F_x\cos\delta-F_y\sin\delta)}{m}-g\sin(\beta+\eta)\\[2ex]\dfrac{\mathrm{d}\beta}{\mathrm{d}t}=\dfrac{(F_x\sin\delta+F_y\cos\delta)}{mv}-\dfrac{g\cos(\beta+\eta)}{v}\end{cases}\tag{2}$$

式中：下标x、y为地面坐标系下的位置；δ、β、η分别为导弹攻角、导弹速度倾角、导弹质心和地心连线与地面坐标系Oy轴的夹角；m为导弹质量；g为重力加速度；F_x和F_y分别为导弹在弹体坐标系下沿纵轴和横轴的气动力，表达式分别为

$$F_x=-\frac{1}{2}SC_x\rho v^2\tag{3}$$

$$F_y=\frac{1}{2}SC_y\rho v^2\tag{4}$$

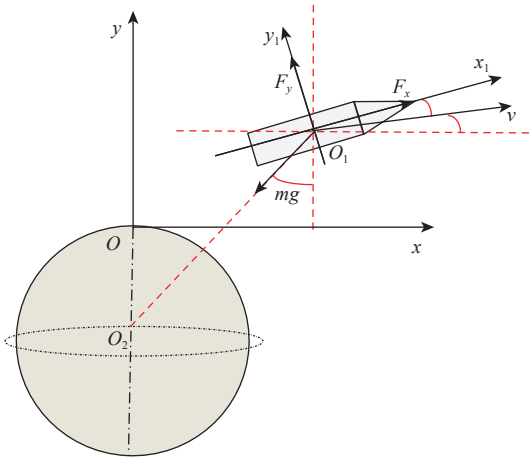


图 2　导弹受力分析

Fig. 2　Force analysis of missile

式中：S 为导弹特征面积；C_x 为沿 O_1x_1 轴的气动系数，C_y 为沿 O_1y_1 轴的气动系数，C_x 和 C_y 通过插值函数从已知的气动系数表中求得；ρ 为大气密度。

1.2　参数设置

1）航程设定。本文如何通过控制合适的导弹攻角，使其航程最大化。在导弹跳跃飞行段，导弹水平飞行距离远大于垂直下降距离，所以其航程可由横坐标 x 取值近似表示。

2）初始参数。导弹跳跃飞行阶段的参数设置如表 1 所示。

表 1　导弹跳跃飞行仿真参数

Table 1　Simulation parameters of missile skipping flight

参数	数值
初始位置/m	(0, 45 000)
初速度大小/(m·s^{-1})	2 224
初始速度倾角/(°)	0
质量/kg	574
特征面积/m^2	0.203
高度约束/km	25～45

3）高度约束。导弹在海拔 25～45 km 范围内跳跃飞行过程的轨迹优化，该过程跳跃飞行示意图如图 3 所示。海拔高度恰好为大气平流层范围，平流层大气温度随海拔上升而升高，气体不易产生对流，大气相对稳定。此外，25 km 以上也是目前防空及其平台难以企及的高度，所以在此高度采用攻角控制策略优化航程有一定的工程实际应用背景。

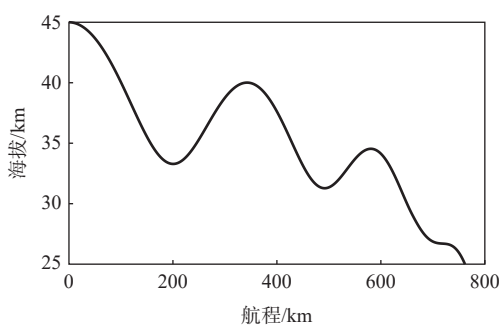


图 3　导弹跳跃飞行示意图

Fig. 3　Diagram of skipping flight of missile

4）过程约束。导弹在跳跃飞行过程中须满足驻点热流密度、动压和过载约束，设导弹允许最大驻点热流密度、最大动压和最大过载分别为 $\dot{Q}_{\max}$ = 300 W/cm^2 和 $q_{\max}$ = 35 kPa 和 $n_{\max}$ = 2g，则有

$$\dot{Q} = k\rho^{0.5}v^{3.15} \leqslant \dot{Q}_{\max} \tag{5}$$

$$q = \frac{1}{2}\rho v^2 \leqslant q_{\max} \tag{6}$$

$$n = \frac{\sqrt{F_x^{\,2} + F_y^{\,2}}}{mg_0} \leqslant n_{\max} \tag{7}$$

式中：$\dot{Q}$ 为驻点热流密度；k 为与导弹相关的常数；q 为动压；n 为过载；g_0 为海平面重力加速度。

5）末速度约束。导弹的速度性能意味着导弹的突防性能，故在优化导弹航程的同时，需要保证导弹的速度性能不受影响。OCAOA 方案中，导弹的末速度为 715 m/s，故本文设计了不小于 700 m/s 的末速度约束。

6）控制量约束。在无动力情况下，为控制导弹飞行稳定性和最大过载，对攻角幅值的约束为 [0,30°]；考虑到导弹控制系统具有良好快速性，导弹飞行动态品质良好，并未设置攻角的变化率约束。本文综合考虑寻优空间的大小与寻优时间的开销，在攻角幅值约束下，控制攻角取值为 0～30°之间的整数。

2　基于 NEO-DDQN 的轨迹优化

2.1　DDQN 算法

Q-learning 算法是强化学习算法中基于值函数的经典算法，该算法利用 Q 表来存储每一个状态动作对的评估价值，即 Q 值。然而，当问题涉及大量状态输入时，将出现“维数灾难”问题，为解决该问题，可考虑利用一个函数来拟合 Q 表所表征的状态动作对与 Q 值的映射关系，而由于该函数的形式化表达十分困难，所以可利用几乎能逼近任何非线性函数关系的深度神经网络来拟合，该深度神经网络被称为深度 Q 网络（deep Q-network, DQN）[24-25]。

DQN 算法是 Q-learning 算法的改进算法，图 4 为 DQN（DDQN）算法的训练模型，展示了 DQN 算法训练的全过程，最终得到的目标网络作用与 Q 表类似，即为待求的动作选择策略。图 4 包含了 DQN 算法所做的 3 方面的改进：拟合 Q 表的深度神经网络、评估网络和目标网络、记忆经验池。

1）深度神经网络。DQN 解决“维数灾难”的

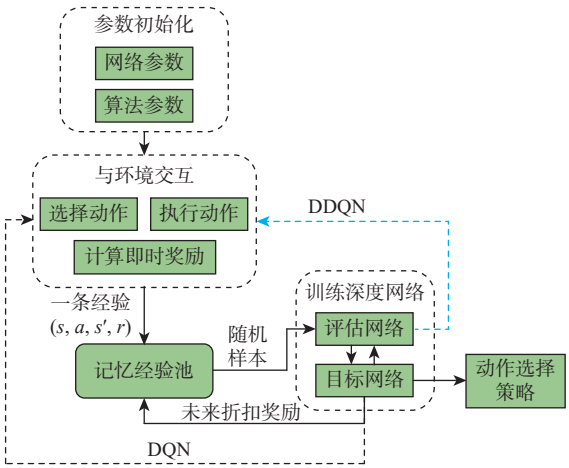


图 4 DQN(DDQN)算法的训练模型

Fig. 4 Training model of DQN (DDQN) algorithm

基本思想为：使用深度神经网络 $f(s,a,\theta)$来近似代替 Q 表所表示的状态动作值函数$Q(s,a)$：

$$f(s,a,\theta)\approx Q(s,a) \tag{8}$$

式中：θ为网络参数；$f(\cdot)$为神经网络函数，可接收无限多个(s,a)状态动作对的输入，并输出相应的 Q 值来近似状态动作值函数$Q*(s,a)$的输出。

2）评估网络和目标网络。DQN 算法中，存在 2 个结构完全相同的网络，分别为评估神经网络和目标神经网络，它们的不同在于网络参数更新的频率不同。每隔一定步数，便利用训练数据训练评估神经网络，评估网络的参数实时更新，而训练一定次数的评估网络后，便将评估网络中的参数完全复制到目标网络中。DQN 算法中利用时间差分方法更新评估网络参数，设评估网络的输出为$Q(s,a,\theta)$，目标网络的输出为$Q(s,a,\theta^-)$，则评估网络的参数更新为

$$\begin{aligned}Q(s,a,\theta)=\alpha(r(s,a)+\gamma\max_{a'}(Q(s',a',\theta^-)))+\\(1-\alpha)Q(s,a,\theta)\end{aligned} \tag{9}$$

式中：α为学习率；γ为奖励衰减因子（折扣因子）；$r(s,a)$为从状态 s 执行动作 a 转移到状态 s'所获得的奖励；$\max_{a'}(Q(s',a',\theta^-))$为假定从状态 s'进行状态转移，根据目标网络$Q(s,a,\theta^-)$选择所有可能动作所能获得的最大 Q 值。

3）记忆经验池。DQN 算法的训练样本通过实时仿真得到，数据样本之间并不独立，存在很强的相关性，为解决该问题，DQN 算法采用经验回放机制，即构建一个经验池来存储最近得到的训练样本。在训练评估网络时，将随机地从经验池中选择一定量的样本进行训练，这样就可以打破相邻训练样本间的相关性，避免模型陷入局部最优。

4）DDQN 算法。本文在选择强化学习算法时，实际使用的是规避了 DQN 算法过估计问题的 DDQN 算法，DDQN 算法被证明可以估计出更准确的状态动作值，智能体可以学到更好的决策策略[26]。如图 4 所示，DDQN 算法与 DQN 算法的不同在于其将动作的选择和目标价值的计算分开以解决过估计问题，即利用评估网络选择动作而使用目标网络计算未来折扣奖励。DDQN 算法依据式 (10) 进行参数的更新：

$$\begin{aligned}Q(s,a,\theta)=\alpha(r+\gamma Q(s',\max_{a'}(Q(s',a',\theta),\theta^-))+\\(1-\alpha)Q(s,a,\theta)\end{aligned} \tag{10}$$

2.2 网络优选策略

在 DDQN 算法训练过程中，训练数据由仿真程序实时仿真得到，在较短的训练时间内，利用新产生的数据训练后可能得到比之前性能更差的网络。所以，在训练深度神经网络时，并不能保证在达到设定的最大训练回合数时得到的网络训练效果就是训练过程中性能最佳的网络，即在训练的全过程中某个时刻的网络性能可能比最终得到的网络性能更好。基于上述原因，本文提出 NEO 策略，具体运行逻辑分为 2 步：

1）攻角记忆。维护一个最佳攻角序列的变量。在所有回合的训练中记录下每个回合的攻角序列及其对应的航程，以航程最大化为指标不断更新所维护的变量。

2）NEO 策略。记忆全程最佳攻角序列的同时，记录此时目标网络的参数。所以，训练结束将得到 2 个目标网络，一是利用评估网络不断更新得到的目标网络 1(ε-greedy)，二是依据全程最佳攻角序列、利用网络 1(ε-greedy)不断更新得到的目标网络 2(ε-greedy)。

基于 DDQN 算法，加入 NEO 策略，本文提出 NEO 双深度 Q 网络（network optimization double deep Q-network, NEO-DDQN）算法。图 5 为 NEO-DDQN 算

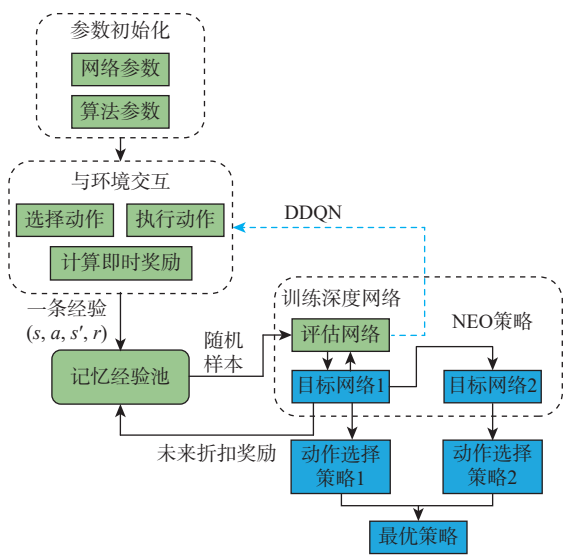


图 5 NEO-DDQN 算法的训练模型

Fig. 5 Training model of NEO-DDQN algorithm

法的训练模型, 图 5 中目标网络 1 代表 DDQN 算法的动作选择策略, 而目标网络 2 则为 NEO-DDQN 算法的动作选择策略, 二者对比得到最优的策略。

2.3　基于 NEO-DDQN 的轨迹优化

导弹在无动力跳跃飞行阶段, 通过选择合适的攻角以获得一定的气动力, 可对其飞行轨迹进行优化, 如达到航程最大、气动加热最小等。本文导弹轨迹优化问题的优化目标为飞行航程最大, 该问题可转换为状态空间连续、动作空间离散的强化学习决策问题, 而深度强化学习 DDQN 算法适合于解决这类问题。

基于 DDQN 算法, 本文算法主要分为 3 部分: 动作空间设计、状态空间设计和奖励函数设计。

1) 动作空间设计。攻角可通过改变气动力来控制跳跃导弹的飞行过程, 所以在设计本文算法时, 将攻角取值集合作为待选的动作集合。攻角取值范围为 0～30°之间的整数, 故可选的动作空间为 0～30°之间的 31 个整数值。

2) 状态空间设计。导弹当前所处的位置、速度和当前所选择的攻角作为网络的输入, 当前的状态动作值, 即 Q 值作为网络的输出。通过在指定的大气环境下进行飞行仿真实验, 实验中利用ε-greedy 贪心策略选择动作, 从当前状态执行该动作后转移到下一状态, 将该状态转换对及此次状态转移所获得的奖励存储到经验回放池中, 这样就可以采集到大量的学习训练样本。

3) 奖励函数设计。在强化学习算法中, 奖励函数的设计显著影响最终的训练效果。在本文跳跃导弹轨迹优化任务中, 优化目标为最大化导弹的航程, 所以可直接利用导弹所处状态的航程信息设计奖励函数。实验中, 本文希望导弹飞行不超过海拔 45 km, 同时希望延迟导弹到达 25 km 下界的时间以增加飞行航程, 所以当导弹海拔 H 高于 45 km 或低于 25 km 时将获得−50 的惩罚值, 其他情况获得奖励值:

$$r(s,a)=\begin{cases}-50 & H>45\ \text{km},\ H<25\ \text{km}\\ \dfrac{x^2}{10^{10}} & \text{其他}\end{cases} \tag{11}$$

其他情况的奖励值为航程的二次方除以 10^{10}, 目的是将奖励值压缩至 1～100 之间, 实验表明该设计有利于引导网络学习到更好的结果。

基于 NEO-DDQN 算法的跳跃导弹轨迹优化算法执行流程, 其中输出目标网络θ_1、θ_2分别为 DDQN 算法、NEO-DDQN 算法的训练结果。

NEO-DDQN 算法:

输入: 导弹初始位置、速度等, 算法参数学习率, 折扣率等。

初始化记忆经验池 D, 容量为 M
随机初始化评估网络的参数θ
随机初始化目标网络的参数$\theta_1=\theta$, $\theta_2=\theta_1$
repeat
　导弹初始状态s
　　repeat
　　　ε-greedy贪心策略选择动作a, 即控制攻角的值
　　　执行动作a, 得到奖励r和下一飞行状态s'
　　　将 (s,a,r,s') 存入经验池 D
　　　从经验池 D 随机采样 (s_j,a_j,r_j,s_{j+1})
　　　计算
$$y=\begin{cases}r_j & s_{j+1}\text{是末状态}\\ r_j+\gamma\max_{a'}Q(s_{j+1},a',\theta^-) & s_{j+1}\text{不是末状态}\end{cases}$$
　　　以 $((y-Q(s_j,a_j,\theta))^2)$ 为损失训练评估网络
　　　$s=s'$
　　　每 L 步迭代后, 更新目标网络$\theta_1=\theta$
　　　until s为末状态
　　　依据最佳攻角序列, 更新目标网络$\theta_2=\theta_1$
until 训练期结束

输出: 目标网络θ_1、θ_2。

3　实验验证

基于 1.2 节中所述的仿真参数设置, 本文首先采用分段控制思想; 其次, 设计网络的结构及 NEO-DDQN 算法的参数, 确定算法的学习率、贪心系数等关键参数的取值; 再次, 本文引入 NEO 策略, 通过迭代学习得到能有效拟合导弹飞行状态-最优控制攻角之间映射关系的深度网络; 最后, 为说明得到的网络的性能, 设计性能验证实验和与 GA 的对比实验。

由于需要考虑大气环境等对导弹飞行过程的影响, 故在进行分段处理时, 有均分与不均分 2 种方式。如 1.2 节中所述, 本文导弹飞行大气环境为平流层, 平流层大气相对稳定, 所以可以不考虑大气环境对分段的影响, 即可对导弹的飞行过程进行平均分段处理, 且在各段飞行时间内使用同一攻角控制导弹的飞行。

在分段处理时, 分段时间间隔对轨迹优化效果的影响很大。间隔较小时, 动作搜索范围较大, 理论上可以对飞行过程实施更精细的控制, 得到更好的优化效果, 但较大的搜索范围也将导致训练时间长、效率低等问题; 如果分段时间间隔过大, 则搜索快、训练时间短, 但对飞行过程的控制精度必然降低, 最终优化结果也较差。为得到最佳的分段时间间隔, 开展间隔分别为 10 s、20 s、40 s、50 s 的 4 组

对比实验。不同时间间隔对训练的影响如表2所示，结果表明，时间间隔为20 s、40 s时，航程的优化效果较好且2种情况下效果相差较小。然而，更小的分段间隔意味着更精细的控制，为了给后续的寻优过程留足可优化空间，本文选择20 s作为后续网络训练的时间间隔。

表2 不同时间间隔对训练的影响

Table 2 Influence of different time intervals on training

分段时间间隔/s	航程/m
10	747 878
20	776 593
40	775 804
50	768 351

3.1 网络设计与训练

本节将详细描述网络的结构和参数、NEO-DDQN算法参数的设计过程。本文的深度网络为全连接前馈神经网络，采用标准的深度神经网络结构，即包含输入层、隐含层和输出层。网络结构的具体参数如表3所示。其中，隐含层1、隐含层2均使用tanh激活函数，输出层则使用purelin线型激活函数。表4列举了经实验确定后较合适的NEO-DDQN算法参数的取值。

表3 网络结构参数

Table 3 Parameters of network structure

层数	神经元个数
输入层	5
隐含层1	40
隐含层2	40
输出层	1

表4 NEO-DDQN算法参数

Table 4 Parameters of NEO-DDQN algorithm

参数	数值
观察期回合数N_1	500
训练期回合数N_2	59 500
训练频率N_3	20
更新频率N_4	200
折扣因子γ	0.95
批大小B	32
记忆池大小M	4 000

确定网络结构及算法参数后，开始训练网络并实时绘制训练迭代图，用以表征训练效果。具体操作如下：记录每个训练回合导弹从初始状态到末状态的状态转移及对应的航程，每100个回合结束后，取当前回合数为横坐标，最近100个回合的航程平均值为纵坐标，在训练迭代图上增加一个点。以此类推，直到训练结束，得到完整的训练迭代。

网络的迭代训练依式(10)进行，即时奖励和未来折扣奖励相加得到目标Q值，目标Q值与当前预测Q值差值的α倍累加到当前Q值作为网络的输出标签。DQN算法的学习率参数α是影响训练效果的关键参数，其决定着网络训练收敛到最佳值的快慢及最终训练的效果，为保证训练的速度与精度，本文设计了学习率参数α分别取0.001、0.01、0.1、0.2、0.5的5组对比实验，并绘制了不同学习率之下的训练迭代图，如图6所示。可知，α=0.01时航程最终收敛值最大；类似地，表5的实验结果同样表明α=0.01时算法寻优性能最佳。因此，后续实验均采用学习率α=0.01进行。

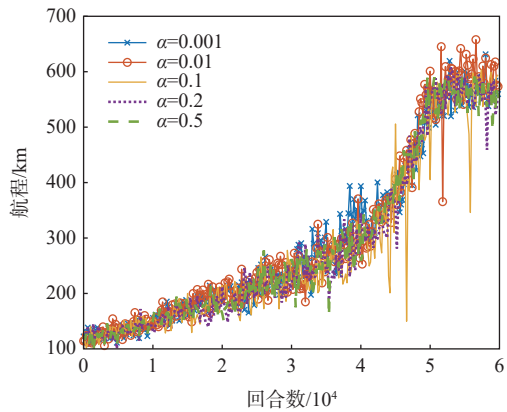


图6 5种学习率的训练迭代

Fig. 6 Training iteration for 5 learning rates

表5 5种学习率最终航程的收敛值

Table 5 Final range convergence values of 5 learning rates

学习率	收敛值/m
0.001	745 654
0.01	781 434
0.1	774 864
0.2	743 651
0.5	746 867

动作选择的探索与利用是强化学习中重要概念，二者之间的有效平衡更是直接影响最终的学习效果。较大的探索系数过于重视随机地探索，不利用学习已经得到的经验；较小的探索系数则可能局限于较少的经验，学习到局部最优策略。为平衡探索与利用，本文采用变探索系数的策略，具体实施方案为：学习初期，探索系数取值为1，算法以1的概率进行随机动作的探索，随着迭代回合数的增加，直到回合数为epoch时，探索系数从1线性递减到0.05。

为选择合适的epoch值，在相同学习率α=0.01的情况下，本文设计了epoch为10 000、20 000、30 000、

40 000、50 000 的 5 种贪心策略的对比实验，并绘制了 5 种贪心策略的训练迭代，如图 7所示，表 6 列举了 5 种贪心策略最终航程的收敛值。由图 7 可知，当回合数为 epoch 时，各组实验都基本收敛，且 epoch 取值越小算法收敛越快，然而实验表明，epoch 越小，迭代寻优时间也越长；由表 6 可知，epoch 取值为 50 000 时最终航程的收敛值也较大，故兼顾收敛速度与收敛精度，本文将选择 epoch=50 000 进行后续的实验。

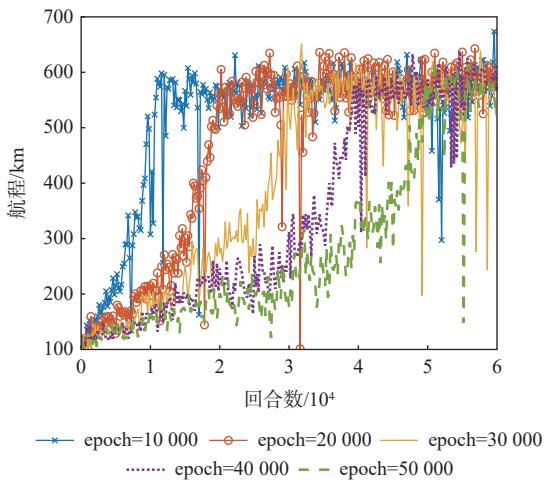


图 7　5 种贪心策略的训练迭代

Fig. 7　Training iteration under 5 greedy strategies

表 6　5 种贪心策略最终航程的收敛值

Table 6　Final range convergence values of 5 greedy strategies

epoch	收敛值/m
10 000	771 926
20 000	778 832
30 000	771 433
40 000	764 878
50 000	776 593

3.2　网络优选策略

在 DDQN 算法的训练过程中，由于训练数据实时产生，存放训练数据的记忆池被不断刷新，故利用较新的数据进行训练得到的网络性能可能劣于先前的网络；另外，在训练全过程中某个时刻的网络性能也可能比最终得到的网络性能更好。为解决此问题，本文提出加入 NEO 策略的 DDQN 算法，即 NEO-DDQN 算法。

表 7 记录了 10 次 NEO 策略对算法结果的影响，θ_1、θ_2分别为加入 NEO 策略前后的结果，分别代表 DDQN 算法和 NEO-DDQN算法，$\theta_2-\theta_1$表示加入 NEO 策略后结果的改善情况。10 次实验的算法参数相同，且均为 3.1 节中最终确定的参数值。

表 7　NEO 策略对算法结果的影响

Table 7　Influence of NEO strategy on algorithm results

编号	θ_1/m	θ_2/m	$\theta_2-\theta_1$/m
1	767 040	771 830	4 790
2	760 609	774 440	13 831
3	767 642	769 734	2 092
4	770 645	771 134	489
5	771 587	784 626	13 039
6	702 365	781 316	78 951
7	762 389	785 295	22 906
8	764 834	783 575	18 741
9	754 677	777 747	23 070
10	755 698	768 921	13 223
平均值	757 748	776 861	19 113

由表 7 中θ_1结果可知，未加入 NEO 策略时，目标网络θ_1的优化效果具有一定的不稳定性，得到的优化结果波动较大；对比 10 次实验的θ_1、θ_2结果可知，加入 NEO 策略后，得到的θ_2网络一定程度解决了这种不稳定性；而单独对比各组实验可以发现，NEO 策略的引入也让算法的迭代结果值得到一定的提升；就 10 次结果的平均值而言，NEO-DDQN 算法比 DDQN 算法航程提升了 2.52%。综上，NEO 策略不仅可以有效改善算法的不稳定性，并能一定程度提升寻优结果。

NEO 策略之所以帮助智能体学习到了更好的决策策略，是因为智能体学习经验获取知识的过程本身就是波动的，NEO 策略观察了智能体学习的全过程，记录了智能体学习到经验知识相对更多、更准确的时候，并存储了此时的决策网络，该网络拥有相对更多的先验知识，故能做出更优的决策。

3.3　优化结果分析

GA 模拟生物种群进化过程中选择、交叉、变异和优胜劣汰的机制，是典型的启发式搜索优化算法，由于其特殊的运行机制，GA 无需梯度信息即可对问题进行具有并行搜索能力的全局寻优，该特点适用于求解导弹的轨迹优化这一复杂非线性、多约束问题。

作为对比方案，本文设置了 OCAOA 方案，即导弹跳跃飞行的全过程，取恒定的控制攻角，在满足过程约束、末速度约束等约束条件下，让导弹航程最远。OCAOA 取值为 13°。

本节对比了 NEO-DDQN 算法与 OCAOA 方案、GA 的轨迹优化结果。GA 参数设置如表 8 所示。为避免偶然性，运行了 10 次 GA 的寻优实验，如表 9

所示。与NEO-DDQN对比时，取表中最好的编号为6的结果，此时优化航程为775 017 m。

表8 GA的参数设置

Table 8 Parameters setting of genetic algorithm

参数	数值
时间间隔T/s	20
攻角/(°)	0～30(整数)
种群大小N_S	10
交叉概率P_c	0.8
变异概率P_v	0.9
迭代次数I	100

表9 GA的优化结果

Table 9 Optimization results of genetic algorithm

编号	航程/m
1	741 955
2	767 656
3	752 029
4	730 913
5	721 760
6	775 017
7	764 682
8	751 243
9	744 770
10	761 691

表10为OCAOA方案、GA、NEO-DDQN算法这3种算法的优化结果(其中航程提升为相对于OCAOA的航程提升比)，图8为3种算法的热流密度、动压、过载与约束界限的关系图，图9为3种算法的航程-海拔图。由结果可知，在满足热流密度、动压、过载等过程约束及末速度等终端约束的同时，NEO-DDQN算法和GA相较于OCAOA方案，航程得到一定的提升，其中GA提升1.26%，NEO-DDQN算法提升2.61%。NEO-DDQN算法与GA对比，航程提升了1.33%，说明NEO-DDQN算法有良好的优化效果。

图10为3种算法得到的攻角序列图，由图10可知，GA和NEO-DDQN算法主要通过调整末段飞行过程的攻角取值来优化航程。NEO-DDQN算法优化的航程更大，因为NEO-DDQN算法在更大的搜索空间中，避免了GA陷入局部最优的缺点，因而成功实施了更精确的控制。

NEO-DDQN算法之所以能避免GA陷入局部最优的缺点，是因为深度强化学习的搜索机制带有记忆性，这种记忆表现为“知识”。强化学习在探索动作时，新的动作被探索、执行后，该动作的优劣性被深度神经网络感知到并以Q值的形式记忆下来，“知识”就被网络“学习”到了，随着学习的时间延长、学习样本的增加，网络有望积累越来越多的经验知识，从而做出更准确的决策；并且，由于学习

表10 3种算法的优化结果

Table 10 Optimization results of three algorithms

算法	航程/m	末速度/(m·s^{-1})	航程提升/%
OCAOA	765 343	715	0.00
GA	775 017	722	1.26
NEO-DDQN	785 295	713	2.61

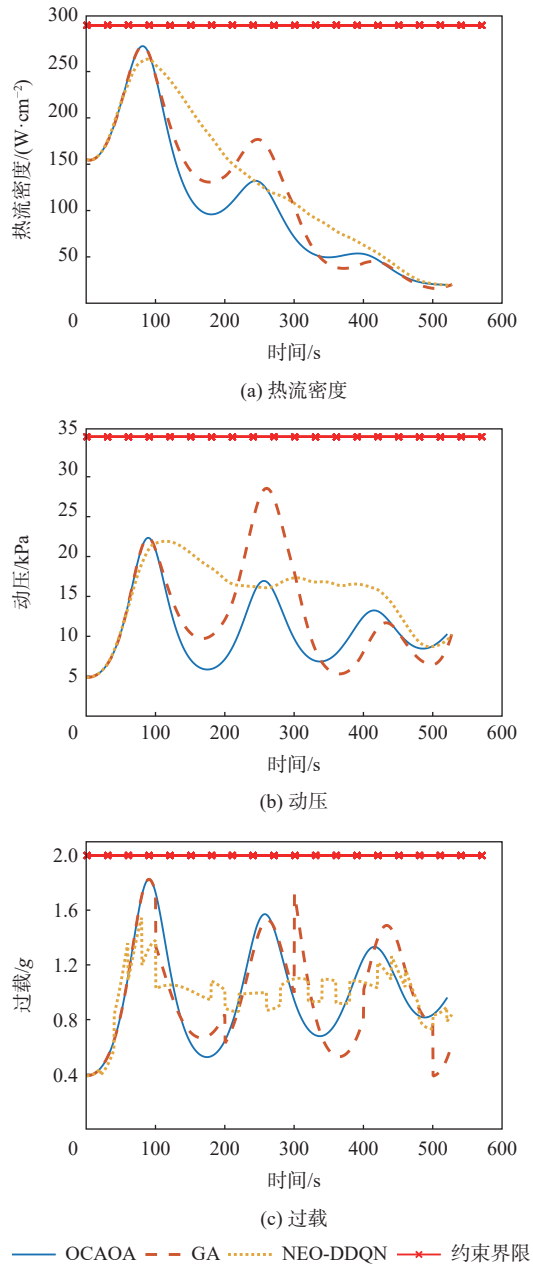


图8 3种算法的热流密度、动压、过载与约束界限的关系

Fig. 8 Relationship between heat flux, dynamic pressure, overload and restraint for three algorithms

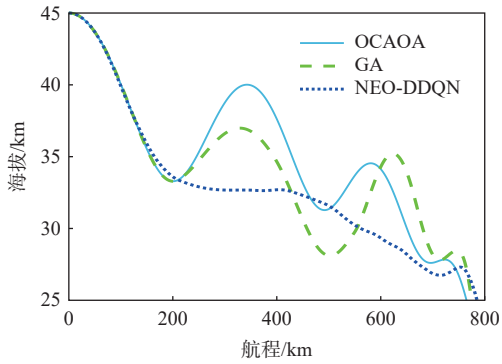


图 9　3 种算法的航程-海拔图

Fig. 9　Range-altitude diagram of three algorithms

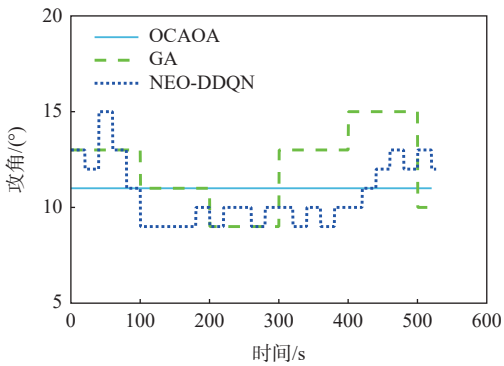
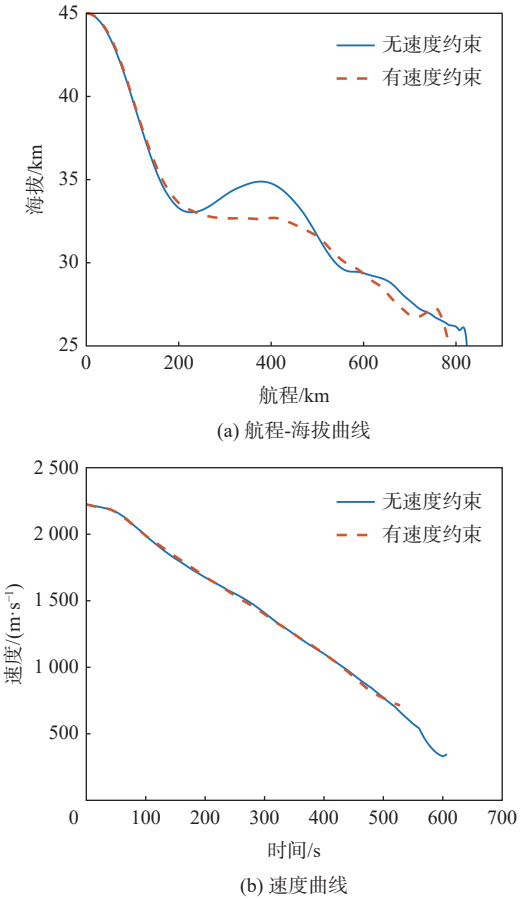


图 10　3 种算法的攻角序列图

Fig. 10　Sequence diagram of three algorithms of angle of attack



(a) 航程-海拔曲线

(b) 速度曲线

图 11　导弹跳跃飞行的性能分析

Fig. 11　Performance analysis of missile skipping flight

初期的随机性, 必然会出现大量劣质解, 而网络可利用记忆机制在之后的探索过程中回避选择这些劣质的动作。相似地, GA 同样存在探索的机制, 也在适应度函数的引导下, 探索更优的解。然而, GA 只能依靠交叉、变异等遗传学原理进行随机搜索, 这种搜索过程是无记忆的, 无法累计经验知识; 甚至, GA 不可避免地会由于重复搜索到相同的解而做大量无意义的搜索。

图 11 为导弹跳跃飞行的性能分析, 对比分析 NEO-DDQN 算法在有大于 700 m/s 末速度约束和无此约束情况下的飞行性能。由图 11(a) 可知, 无速度约束情况下, 导弹跳跃性更强, 有效延长了飞行时间, 最终的飞行航程也更大(823 577 m); 而图 11(b) 则表明, 无约束情况下更剧烈的跳跃飞行虽然使导弹获得了更大的航程, 但跳跃过程中的能量损耗导致导弹最终的速度性能大大下降(346 m/s)。所以在有速度约束的情况下, 为了在优化航程性能的同时保证末速度性能, NEO-DDQN 算法学习到了较“聪明”的策略: 前期通过减小跳跃性来维持导弹的能量, 接近结束时适当增加跳跃性以延长飞行时间, 提高飞行航程。

4　结　论

1) 与无 NEO 策略的传统 DDQN 算法相比, NEO-DDQN 算法记录了训练过程中拥有更多经验知识的网络, 不仅有效提高了算法的稳定性, 并在一定程度上提升了优化效果。

2) 与启发式搜索 GA 相比, NEO-DDQN 算法具有积累经验知识、避免重复搜索等特点, 故在相同大小的搜索空间中, NEO-DDQN 并未陷入局部最优, 取得了更好的优化效果。

3) 与传统间接法和直接法相比, 本文提出的 NEO-DDQN 算法是一种新颖的导弹轨迹优化算法, 其通过拟合导弹飞行状态-最优控制攻角之间的映射关系, 避免了直接处理复杂非线性微分方程, 为轨迹优化问题提供了一种新型的基于学习的算法。

参考文献 (References)

[1] 王在铎, 王惠, 丁楠, 等. 高超声速飞行器技术研究进展[J]. 科技导报, 2021, 39(11): 59-67.
WANG Z D, WANG H, DING N, et al. Research on the development of hypersonic vehicle technology[J]. Science & Technology

Review, 2021, 39(11): 59-67(in Chinese).

[2] 邵雷, 雷虎民, 赵锦. 临近空间高超声速飞行器轨迹预测方法研究进展[J]. 航空兵器, 2021, 28(2): 34-39.
SHAO L, LEI H M, ZHAO J. Research progress in trajectory prediction for near space hypersonic vehicle[J]. Aero Weaponry, 2021, 28(2): 34-39(in Chinese).

[3] 陈小庆, 侯中喜, 刘建霞. 高超声速滑翔式飞行器再入轨迹多目标多约束优化[J]. 国防科技大学学报, 2009, 31(6): 77-83.
CHEN X Q, HOU Z X, LIU J X. Multi-objective optimization of reentry trajectory for hypersonic glide vehicle with multi-constraints[J]. Journal of National University of Defense Technology, 2009, 31(6): 77-83(in Chinese).

[4] AN K, GUO Z Y, XU X P, et al. A framework of trajectory design and optimization for the hypersonic gliding vehicle[J]. Aerospace Science and Technology, 2020, 106: 106110.

[5] 何烈堂, 柳军, 侯中喜, 等. 无动力跳跃式跨大气层飞行的可行性研究[J]. 弹箭与制导学报, 2008, 28(2): 155-157.
HE L T, LIU J, HOU Z X, et al. Feasibility study of unpropulsive skipping trans-atmospheric flight[J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2008, 28(2): 155-157(in Chinese).

[6] 国海峰, 黄长强, 丁达理, 等. 考虑随机干扰的高超声速滑翔飞行器轨迹优化[J]. 北京航空航天大学学报, 2014, 40(9): 1281-1290.
GUO H F, HUANG C Q, DING D L, et al. Trajectory optimization for hypersonic gliding vehicle considering stochastic disturbance[J]. Journal of Beijing University of Aeronautics and Astronautics, 2014, 40(9): 1281-1290(in Chinese).

[7] GATH P F, WELL K H, MEHLEM K. Initial guess generation for rocket ascent trajectory optimization using indirect methods[J]. Journal of Spacecraft and Rockets, 2002, 39(4): 515-521.

[8] BARRON R L, CHICK C M. Improved indirect method for air-vehicle trajectory optimization[J]. Journal of Guidance, Control, and Dynamics, 2006, 29(3): 643-652.

[9] ROSA SENTINELLA M, CASALINO L. Genetic algorithm and indirect method coupling for low-thrust trajectory optimization[C]// 42nd AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit. Reston: AIAA, 2006: 4468.

[10] 李瑜, 杨志红, 崔乃刚. 洲际助推-滑翔导弹全程突防弹道优化[J]. 固体火箭技术, 2010, 33(2): 125-130.
LI Y, YANG Z H, CUI N G. Optimization of overall penetration trajectory for intercontinental boost-glide missile[J]. Journal of Solid Rocket Technology, 2010, 33(2): 125-130(in Chinese).

[11] 涂良辉, 袁建平, 岳晓奎, 等. 基于直接配点法的再入轨迹优化设计[J]. 西北工业大学学报, 2006, 24(5): 653-657.
TU L H, YUAN J P, YUE X K, et al. Improving design of reentry vehicle trajectory optimization using direct collocation method[J]. Journal of Northwestern Polytechnical University, 2006, 24(5): 653-657(in Chinese).

[12] RAO A, CLARKE K. Performance optimization of a maneuvering re-entry vehicle using a Legendre pseudospectral method[C]//AIAA Atmospheric Flight Mechanics Conference and Exhibit. Reston: AIAA, 2002: 4885.

[13] COTTRILL G C, HARMON F G. Hybrid Gauss pseudospectral and generalized polynomial chaos algorithm to solve stochastic optimal control problems[C]//AIAA Guidance, Navigation, and Control Conference. Reston: AIAA, 2011: 6572.

[14] KUMAR G N, AHMED M S, SARKAR A K, et al. Reentry trajectory optimization using gradient free algorithms[J]. IFAC-Papers OnLine, 2018, 51(1): 650-655.

[15] CHAI R Q, SAVVARIS A, TSOURDOS A, et al. Solving multi-objective aeroassisted spacecraft trajectory optimization problems using extended NSGA-II[C]// AIAA SPACE and Astronautics Forum and Exposition. Reston: AIAA, 2017: 5193.

[16] ZHAO J, ZHOU R. Particle swarm optimization applied to hypersonic reentry trajectories[J]. Chinese Journal of Aeronautics, 2015, 28(3): 822-831.

[17] SHAHZAD SANA K, HU W D. Hypersonic reentry trajectory planning by using hybrid fractional-order particle swarm optimization and gravitational search algorithm[J]. Chinese Journal of Aeronautics, 2021, 34(1): 50-67.

[18] SONG X, HAN D L, SUN J H, et al. A data-driven neural network approach to simulate pedestrian movement[J]. Physica A: Statistical Mechanics and its Applications, 2018, 509(11): 827-844.

[19] 桑晨, 郭杰, 唐胜景, 等. 基于DDPG算法的变体飞行器自主变形决策[J]. 北京航空航天大学学报, 2022, 48(5): 910-919.
SANG C, GUO J, TANG S J, et al. Autonomous deformation decision making of morphing aircraft based on DDPG algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(5): 910-919(in Chinese).

[20] 许轲, 吴凤鸽, 赵军锁. 基于深度强化学习的软件定义卫星姿态控制算法[J]. 北京航空航天大学学报, 2018, 44(12): 2651-2659.
XU K, WU F G, ZHAO J S. Software defined satellite attitude control algorithm based on deep reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(12): 2651-2659(in Chinese).

[21] NI W J, WU D, MA X P. Energy-optimal flight strategy for solar-powered aircraft using reinforcement learning with discrete actions[J]. IEEE Access, 2021, 9: 95317-95334.

[22] JIANG L, NAN Y, LI Z H. Realizing midcourse penetration with deep reinforcement learning[J]. IEEE Access, 2021, 9: 89812-89822.

[23] GAO J S, SHI X M, CHENG Z T, et al. Reentry trajectory optimization based on deep reinforcement learning[C]//2019 Chinese Control and Decision Conference (CCDC). Piscataway: IEEE Press, 2019: 2588-2592.

[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[25] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB/OL]. (2013-03-02) [2021-05-15]. http://arxiv.org/abs/1312.5602.

[26] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2016: 2094–2100.

Trajectory optimization algorithm of skipping missile based on deep reinforcement learning

GONG Kaiqi[1], WEI Hongkui[2], LI Jiawei[2], SONG Xiao[1,*], LI Yong[1], LI Yixin[2], ZHANG Yue[2]

(1. School of Cyber Science and Technology, Beihang University, Beijing 100191, China;
2. Beijing Institute of Electronic System Engineering, Beijing 100191, China)

Abstract: The skipping flight process of the skipping missile can be modeled as a set of time-varying nonlinear differential equations which cannot be solved analytically. Therefore, it brings great difficulties to optimize the trajectory optimization of the skipping missile. To solve this problem, a deep reinforcement learning trajectory optimization method based on double deep Q-network (DDQN) is proposed to maximize the range of the missile under certain constraints of heat flux, dynamic pressure, and overload. The procedure of this method is as follows. Firstly, the action space, state space, and reward function of the algorithm are designed. The appropriate greedy strategy is then determind, along with the learning rate, an important algorithmic parameter. Further, the network optimization (NEO) strategy is introduced and then the NEO-DDQN algorithm is proposed. Finally, comparison experiments with the optional constant angle of attack (OCAOA) scheme and genetic algorithm (GA) are designed. Results show that the network optimization strategy effectively improves the stability of the algorithm and increases the flight range by 2.52%. Compared with OCAOA scheme and GA, the NEO-DDQN method improves the range of skipping missiles by 2.61% and 1.33% respectively. In addition, the proposed method successfully avoids directly dealing with complex nonlinear differential equations and innovatively provides a learning-based method for the trajectory optimization of the missile.

Keywords: missile; skipping flight; trajectory optimization; deep reinforcement learning; network optimization

Received: 2021-08-02; **Accepted**: 2021-11-11; **Published Online**: 2021-12-14 12:49

URL: kns.cnki.net/kcms/detail/11.2625.V.20211213.1850.002.html

Foundation item: National Key R& D Program of China (2018YFB1702703)

* **Corresponding author**. E-mail: songxiao@buaa.edu.cn