文章编号：1671-0576(2024)02-0001-10

基于 Dueling DQN 的临近空间飞行器再入轨迹规划

田若岑[1]， 刘益吉[1]， 肖 涛[1]， 张顺家[1]， 陆 远[2]
（1. 上海机电工程研究所，上海 201109；2. 上海航天电子技术研究所，上海 201109）

摘 要：针对临近空间飞行器再入段禁飞区规避制导问题，构建了临近空间飞行器再入过程横侧向制导的马尔可夫决策过程（Markov decision process，MDP）模型。基于竞争深度 Q 网络（dueling deep Q network，Dueling DQN），设计了横侧向制导律及满足射程需求与禁飞区规避需求的再入过程奖励函数。经仿真验证，该横侧向制导律能够通过改变倾侧角符号实现禁飞区规避，并导引飞行器到达目标区域，具备较高精度，验证了方法的有效性。

关键词：临近空间飞行器；再入轨迹规划；竞争深度 Q 网络
中图分类号：V448.235 **文献标志码**：A
DOI：10.3969/j.issn.1671-0576.2024.02.001

Near Space Vehicle Reentry Trajectory Planning Based on Dueling DQN

TIAN Ruocen[1]， *LIU Yiji*[1]， *XIAO Tao*[1]， *ZHANG Shunjia*[1]， *LU Yuan*[2]
（1. Shanghai Electro-Mechanical Engineering Institute，Shanghai 201109，China；
2. Shanghai Aerospace Electronic Technology Institute，Shanghai 201109，China）

Abstract：Aiming at the problem of no-fly zone avoidance guidance in the reentry phase for near space vehicle，the Markov decision process（MDP）model of lateral guidance in the reentry process for near space vehicle was constructed. On the basis of dueling deep Q network（Dueling DQN），the lateral guidance law and the environmental reward feedback function to satisfy the range requirement and the no-fly zone avoidance requirement were designed. The simulation results show that the lateral guidance law can avoid the no-fly zone by changing the sign of roll angle，and guide the aircraft to the target area with high precision，which verifies the effectiveness of the method.

Key words：near space vehicle；reentry trajectory planning；Dueling DQN

收稿日期：2023-12-11
作者简介：田若岑（1997—），男，硕士，助理工程师，主要从事制导控制技术研究。

0 引言

随着科技不断发展，现代战争形势已不同于以往，飞行速度更快、反应时间更短、隐蔽性更强的临近空间飞行器有着巨大的军事、政治和经济价值[1]。临近空间飞行器由于其独特优势，已经成为世界各国竞相追逐的技术制高点，是国家最高科技水平和工业水平的象征。

临近空间飞行器的再入轨迹规划是其设计中的关键一环[2]。自 20 世纪 60 年代以来，各国学者对再入轨迹规划问题进行了大量的研究。YOUSSEF 等[3]提出了预测-校正制导方法，旨在解决再入初始条件存在的大范围散布问题；SHEN 等[4]基于拟平衡滑翔假设提出了一种有效满足多约束条件的三自由度再入轨迹在线生成方法，该方法具备较强的通用性和实时性；潘乐飞等[5]采用可变容差单纯形法求解制导参数，引入惩罚函数来解决约束问题；JOSHI 等[6]提出了一种考虑路径约束的数值预测-校正制导算法，在轨迹超出阻力边界时调整倾侧角，通过迭代计算保证满足终端约束。

近些年来，伴随着凸优化理论与群体智能优化理论的兴起，越来越多的学者开始关注再入制导过程的最优性问题。LIU 等[7]成功地将凸优化理论应用于飞行器再入制导过程中，相对于传统序列二次规划（sequential quadratic programming，SQP）算法，改进算法在实时性上取得了一定突破；LIU 等[8]应用改进后的水草算法，将再入轨迹规划问题分解为多个步骤，显著提升了群体智能算法求解该问题的实时性。

自 2016 年以来，机器学习与深度强化学习的快速发展为再入段制导的实时性与最优性矛盾提供了新的解决方案[9]。文献[10]对强化学习算法与飞行器路径规划问题的结合进行了初步探索；文献[11]采用了基于深度确定性策略梯度（deep deterministic policy gradient，DDPG）算法的纵向再入轨迹规划方案，为提升制导律实时性与精确性提供了新的思路。

本文对临近空间飞行器再入段模型进行构建，建立再入过程的马尔可夫决策过程（Markov decision process，MDP）模型，设计基于竞争深度 Q 网络（dueling deep Q network，Dueling DQN）的横侧向制导律，并通过仿真验证方法的有效性。

1 再入段模型构建

1.1 再入运动学模型构建

在弹道坐标系与速度坐标系下建立的再入飞行器的质心运动方程[12]为

$$\begin{cases} \mathrm{d}R_e/\mathrm{d}t = v\sin\theta \\ \mathrm{d}\lambda/\mathrm{d}t = v\cos\theta\sin\psi/(R_e\cos\phi) \\ \mathrm{d}\phi/\mathrm{d}t = v\cos\theta\cos\psi/R_e \\ \mathrm{d}v/\mathrm{d}t = -D/m - g\sin\theta \\ \mathrm{d}\theta/\mathrm{d}t = \dfrac{1}{v}\left[\dfrac{L\cos\sigma}{m} + \left(\dfrac{v^2}{R_e} - g\right)\cos\theta\right] \\ \mathrm{d}\psi/\mathrm{d}t = \dfrac{1}{v}\left(\dfrac{L\sin\sigma}{m\cos\theta} + \dfrac{v^2}{R_e}\cos\theta\sin\psi\tan\phi\right) \\ \mathrm{d}S_e/\mathrm{d}t = v\cos\theta/R_e \end{cases} \tag{1}$$

式中：R_e 为地心距；v 为飞行器速度；θ，ψ 为弹道倾角和弹道偏角；λ，ϕ 为经度和纬度；L，D 为气动升力和气动阻力；σ 为倾侧角；m 为飞行器质量；g 为当前高度的重力加速度；S_e 为射程对应于地心的角。

1.2 再入过程约束模型建立

再入过程本质上是一个复杂的飞行运动过程。在临近空间飞行器整个再入过程中，需要利用气动力与重力设计制导律，以对飞行轨迹进行控制，同时将多余的能量进行逸散，使飞行器以预定的速度到达指定位置。为了让飞行器安全平稳地完成再入飞行任务，需要给出相应的过程约束与终端约束。主要约束包括热流、动压、过载、平衡滑翔、控制及终端等[12]。

（1）热流约束

热流约束由驻点处的热流密度 $\dot{Q}$ 来表示，其表达式为

$$\dot{Q} = \frac{C_1}{\sqrt{R_d}}\left(\frac{\rho}{\rho_0}\right)^{0.5}\left(\frac{v}{v_c}\right)^{3.15} \leqslant \dot{Q}_{\max} \tag{2}$$

式中：C_1 为热流系数；R_d 为飞行器前缘半径；ρ 为当前海拔处的大气密度；ρ_0 为零海拔处的大气密度；$v_c=\sqrt{R_0 g_0}$ 为归一化速度，其中 R_0 为地球半径，g_0 为零海拔处的重力加速度；$\dot{Q}_{\max}$ 为允许

的最大热流密度。

(2) 动压约束

动压 q 的表达式为

$$q=\frac{1}{2}\rho v^2\leqslant q_{\max} \tag{3}$$

式中：$q_{\max}$ 为允许的最大动压。

(3) 过载约束

总过载 n 的表达式为

$$n=q\sqrt{{C_D}^2+{C_L}^2}S_{\text{ref}}/(mg)\leqslant n_{\max} \tag{4}$$

式中：C_L，C_D 为升力系数和阻力系数；S_{ref} 为气动面积；$n_{\max}$ 为允许的最大总过载。

(4) 平衡滑翔约束

平衡滑翔约束的表达式为

$$\left(g-\frac{v^2}{R_e}\right)-\frac{L}{m}\cos\sigma_{\text{QEGC}}=0 \tag{5}$$

式中：σ_{QEGC} 为平衡滑翔角。对于飞行高度为 80～85 km、飞行马赫数大于 2 的中-高升阻比飞行器，平衡滑翔约束能够成立。

(5) 控制约束

控制约束由制导系统输出的倾侧角 σ 及倾侧角变化率 $\dot{\sigma}$ 来表征，应满足

$$\begin{cases}|\sigma|\leqslant 90^\circ\\ |\dot{\sigma}|\leqslant\dot{\sigma}_{\max}\end{cases} \tag{6}$$

式中：$\dot{\sigma}_{\max}$ 为最大倾侧角变化率。

(6) 终端约束

终端约束为任务要求的终端飞行速度、高度与经纬度，其表达式为

$$\begin{cases}v(t_f)=v_f\\ h(t_f)=h_f\\ \lambda(t_f)=\lambda_f\\ \phi(t_f)=\phi_f\end{cases} \tag{7}$$

式中：t_f 为终端时刻；$v(\cdot)$，$h(\cdot)$，$\lambda(\cdot)$，$\phi(\cdot)$ 分别为实际飞行过程中的速度、高度、经纬度函数；v_f，h_f，λ_f，ϕ_f 为任务要求的终端速度、高度、经纬度。在制导律的设计中，由于横向与纵向轨迹规划过程相互独立，因而可以将式(7)中的终端经纬度约束转化为终端射程约束

$$S_e(t_f)=S_{go}=\arccos(\sin\lambda\sin\lambda_f+\cos\lambda\cos\lambda_f\cos(\phi-\phi_f)) \tag{8}$$

式中：S_{go} 为剩余飞行距离。S_{go} 为飞行器当前位置与目标点之间的最小球面圆弧距离。

(7) 禁飞区约束

禁飞区是指受雷达探测、电磁干扰与拦截，以及地形、地缘政治因素等影响而形成的再入飞行器应尽量规避的区域。为了便于设计分析，将禁飞区视作无限高圆柱体，飞行器轨迹只能从其左右两侧规避，而不考虑从其上方或下方规避的情况。设 λ_m，ϕ_m 为飞行器当前经度和纬度，λ_z，ϕ_z 为禁飞区中心所在经度和纬度，R_z 为禁飞区半径，则禁飞区路径约束应满足

$$\sqrt{(\lambda_m-\lambda_z)^2+(\phi_m-\phi_z)^2}\geqslant R_z/R_0 \tag{9}$$

1.3　飞行器参数设置

本文选取美国波音公司 1998 年设计的再入飞行器 CAV-L 为研究对象，飞行器总体参数及其最大过程约束参数如表 1 和表 2[13] 所示。

表 1　飞行器总体参数

序号	参数	参数值
1	质量 m/kg	907
2	前缘半径 R_d/m	0.1
3	热流系数 C_1	11 030
4	气动面积 S_{ref}/m^2	0.35

表 2　飞行器最大过程约束参数

序号	参数	参数值
1	最大热流密度 $\dot{Q}_{\max}$/(kW·m^{-2})	1 200
2	最大动压 $q_{\max}$/kPa	200
3	最大总过载 $n_{\max}/g_0$	4

零侧滑飞行状态下，攻角和倾侧角是制导过程中的控制量，由于调控攻角 α 的代价远高于调控倾侧角 σ，故再入过程往往采用固定攻角剖面。设最大允许攻角 $\alpha_{\max}=20^\circ$，最大升阻比对应的最小攻角 $\alpha_{\min}=8.5^\circ$，速度节点 $v_1=4\,700$ m/s，$v_2=3\,100$ m/s，则本文采用的攻角剖面可表示为

$$\alpha=\begin{cases}\alpha_{\max}, & v\geqslant v_1\\ \dfrac{(\alpha_{\max}-\alpha_{\min})(v-v_2)}{v_1-v_2}, & v_2<v<v_1\\ \alpha_{\min}, & v\leqslant v_2\end{cases} \tag{10}$$

2 深度 Q 网络算法分析

2.1 算法原理

深度 Q 网络(deep Q network,DQN)是一种经典的强化学习算法。强化学习的基本思想受到了人类学习过程的启发,其主要流程如图 1 所示。图中 s_t,a_t,r_t 分别为 t 时刻的状态、动作和奖励。

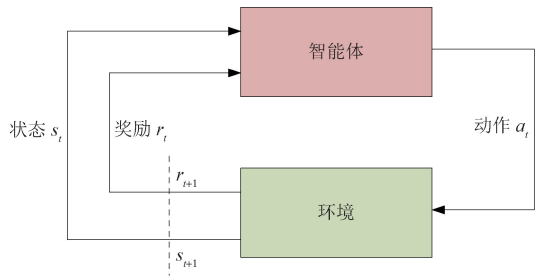

图 1 强化学习流程图

强化学习的目的是智能体在随机交互环境下,根据当前状态不断序列化选择动作,最终实现累计奖励的最大化[14]。强化学习用于处理 MDP 问题,此问题主要包含 4 个要素:

a) 状态空间 S 为所有可能状态的集合,其中 t 时刻状态 $s_t \in S$,初始状态记为 s_1;

b) 动作空间 A 为所有可能动作的集合,其中 t 时刻动作 $a_t \in A$;

c) 状态转移概率函数 $p(s_{t+1} \mid s_t, a_t)$ 满足马尔可夫性(Markov property),即 $t+1$ 时刻状态 s_{t+1} 的转移概率只与 t 时刻状态 s_t 有关,$p(s_{t+1} \mid s_t, a_t, s_{t-1}, \cdots, s_1, a_1) = p(s_{t+1} \mid s_t, a_t)$,初始状态的概率为 $p(s_1)$;

d) 奖励函数 $r_t = r(s_{t+1} \mid s_t, a_t)$ 表示由 t 时刻状态 s_t 通过执行动作 a_t 转移到 $t+1$ 时刻状态 s_{t+1} 时所获得的奖励,该函数表征了环境对行为的反馈。

在每一步的决策中,智能体根据环境状态决定所要采取的动作,而动作输出的规律为策略(policy),t 时刻的策略记为 π_t。$\pi_t(a_t \mid s_t)$ 表示在状态 s_t 下动作 a_t 的选择概率。鉴于强化学习是一个序列决策算法,因此算法可以对一个动作序列进行整体评价。将一个 t 时刻开始的动作序列的反馈奖励累积定义为收益 G_t,即

$$G_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_{t+T} \tag{11}$$

式中:T 为奖励累积的时间。

强化学习的策略决定了智能体在不同状态下的动作响应规律,策略的好坏可根据初始状态到终端状态的奖励序列进行判断。强化学习不断优化策略的过程实质是不断最大化奖励累积的过程。将一定策略下奖励累积的期望称为状态 s 的价值函数(value function),记为 $v_\pi(s)$,其表达式为

$$\begin{aligned} v_\pi(s) &= E_\pi(G_t \mid s_t = s) \\ &= E_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right) \end{aligned} \tag{12}$$

式中:$E_\pi(\cdot)$ 为策略 π 下的期望函数;k 为动作序列号;γ 为奖励反馈的折扣系数。价值函数用于对一个特定交互场景下的策略进行评价,因此价值函数本身是与策略绑定的。一个更好的策略原则上应该对应更大的价值函数值。同理,一定状态 s 下动作 a 的价值函数 $q_\pi(s,a)$ 可以定义为

$$\begin{aligned} q_\pi(s,a) &= E_\pi(G_t \mid s_t = s, a_t = a) \\ &= E_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right) \end{aligned} \tag{13}$$

在强化学习中,一个动作的好坏可以通过 $q_\pi(s,a)$ 来进行评价。强化学习的目标是学到一个最好的策略来最大化初始状态期望,即

$$\begin{aligned} J(\pi^*) &= \max(v_\pi(s_1)) \\ &= \max(E_\pi(G_t \mid s_t = s_1)) \end{aligned} \tag{14}$$

式中:$J(\cdot)$ 为目标函数;π^* 为学习到的最优策略;$\max(\cdot)$ 为最大值的取值函数。

根据状态或动作的连续性和离散性,可以将强化学习算法分为不同类型。针对再入制导轨迹横侧向规划问题,状态特性呈现为状态空间连续无限、动作空间离散有限,DQN 算法更加适配该问题。传统 DQN 算法有诸多改进版本,例如 Nature DQN,Double DQN,Prioritized Replay DQN,Dueling DQN 等。本文采用的是 Dueling DQN 算法,该算法流程如下。

步骤 1:参数初始化。初始化 Dueling DQN 中 Q_local、Q_target 的网络参数与状态。Q_local 网络的主要功能是对当前状态的各个动作价值进行评价。通过设置 Q_target 网络参数改善单网络自举导致的行为价值过估计现象,提升训练效果。

步骤 2:动作估值函数计算。在每个训练周

期内，根据当前的状态 s_t 、动作 a_t 和策略 π_t ，可确定 Q_local 网络对 t 时刻状态下动作价值的估值函数

$$Q_{\mathrm{loc}}(s_t,a_t)=\mathrm{NET}_{\pi}(s_t,a_t) \tag{15}$$

式中：$\mathrm{NET}_{\pi}(\cdot)$ 为策略 π 下的网络函数。

步骤 3：实际价值函数计算。根据当前状态及取得的最大价值动作与环境进行交互，得到奖励函数 $r(s_t,a_t,s_{t+1})$；根据环境反馈，得到当前状态与动作的实际价值函数 $Q(s_t,a_t)$ ，有

$$Q(s_t,a_t)=\begin{cases} r(s_t,a_t,s_{t+1}), & s_{t+1}=s_{\mathrm{f}} \\ r(s_t,a_t,s_{t+1})+ \\ \gamma\max(Q_{\mathrm{tar}}(s_{t+1},a_{t+1})), & s_{t+1}\neq s_{\mathrm{f}} \end{cases} \tag{16}$$

式中：s_{f} 为终端状态；$Q_{\mathrm{tar}}(\cdot)$ 为 Q_target 网络的估值函数。Q_target 网络参数通过软更新的方式实现更新，更新方程为

$$Q_{\mathrm{tar}}(s_t,a_t)\leftarrow l_{\mathrm{r}}Q_{\mathrm{loc}}(s_t,a_t)+(1-l_{\mathrm{r}})Q_{\mathrm{tar}}(s_t,a_t) \tag{17}$$

式中：l_{r} 为软更新的学习率。软更新的目的是使 Q_target 的网络参数逐渐向 Q_local 的逼近，避免了硬更新的强烈冲击。

步骤 4：Q_local 网络参数更新。更新网络参数需要计算损失函数 $L(\pi)$ ，其表达式为

$$L(\pi)=E((Q_{\mathrm{loc}}(s_t,a_t)-Q(s_t,a_t))^2/2) \tag{18}$$

式中：$E(\cdot)$ 为数学期望函数。

通过反向传播(BP)算法实现网络参数更新，$L(\pi)$ 的梯度表达式为

$$\frac{\partial L(\pi)}{\partial\pi}= E\left((Q_{\mathrm{loc}}(s_t,a_t)-Q(s_t,a_t))\frac{\partial Q_{\mathrm{loc}}(s_t,a_t)}{\partial\pi}\right) \tag{19}$$

基于梯度信息的网络优化方法有 AdaGrad、梯度下降(stochastic gradient descent，SGD)、Adam 等，其中 Adam 优化器性能最佳，是最为常用的神经网络优化方法。

Dueling DQN 在 Double DQN 基础上，对网络结构进行了改进。Dueling DQN 的价值函数不再由全连接层直接获得，而是增加了子网络结构，其中一部分子网络用于拟合当前状态价值函数 $v_{\pi}(s_t)$ ，另一部分用于拟合优势函数 $A(s_t,a_t)$。则改进的网络估值函数表达式为

$$Q_{\mathrm{NET}}(s_t,a_t)= v_{\pi}(s_t)+\left(A(s_t,a_t)-\frac{1}{N(s_t)}\sum_{a_t\in A}A(s_t,a_t)\right) \tag{20}$$

式中：$N(s_t)$ 为状态 s_t 对应的可选动作数。

通过对网络结构加以改进，可以有效提升网络的收敛速度和训练效果。

2.2　MDP 建模

(1) 横侧向制导 MDP 建模

将再入制导问题与强化学习算法相结合，首先要建立再入制导的 MDP 模型，即给出其状态空间、动作空间、奖励函数和状态转移概率函数。

根据建立的再入动力学模型，考虑横侧向制导需求，即实现禁飞区规避以及最小化终端射程误差，选取状态空间

$$S=\{h,\ \lambda,\ \phi,\ v,\ S_{\mathrm{go}},\ \psi\} \tag{21}$$

式中：h 为高度。由于状态空间中变量的单位、物理意义各不相同，若直接将这些变量输入网络，容易导致网络的局部饱和与权重不协调。为解决该问题，对各变量进行归一化处理。各变量的归一化表达式为

$$\begin{cases} \tilde{h}=[h-(h_0+h_{\mathrm{f}})/2]/(h_0-h_{\mathrm{f}}) \\ \tilde{\lambda}=[\lambda-(\lambda_0+\lambda_{\mathrm{f}})/2]/(\lambda_0-\lambda_{\mathrm{f}}) \\ \tilde{\phi}=[\phi-(\phi_0+\phi_{\mathrm{f}})/2]/(\phi_0-\phi_{\mathrm{f}}) \\ \tilde{v}=[v-(v_0+v_{\mathrm{f}})/2]/(v_0-v_{\mathrm{f}}) \\ \widetilde{S}_{\mathrm{go}}=S_{\mathrm{go}}/S_{\mathrm{need}} \\ \tilde{\psi}=\psi/2 \end{cases} \tag{22}$$

式中：下标 0 表示初始时刻状态取值；S_{need} 为根据初始状态与终端状态约束计算出的理论射程。变量归一化后的状态空间

$$S=\{\tilde{h},\ \tilde{\lambda},\ \tilde{\phi},\ \tilde{v}\ ,\widetilde{S}_{\mathrm{go}},\ \tilde{\psi}\} \tag{23}$$

经过上述处理，变量在各个状态下的取值会随时间均匀映射到[−1，1]区间内，成为无量纲变量。

动作所涉及的变量为倾侧角，由于横侧向制导律需要确定倾侧角方向，故动作空间

$$A=\{\mathrm{sign}(\sigma)\}=\{-1,+1\},\quad \sigma\neq 0 \tag{24}$$

式中：$\mathrm{sign}(\cdot)$ 为符号函数。当倾侧角 σ 的取值为正时，表明升力在水平面上的分量指向顺航向右侧，导致速度方向向右偏转，反之向左偏转。

再入轨迹规划过程是一个确定性过程，体现在再入动力学方程中，确定的输入对应确定的输出，故状态转移概率为 1。

奖励函数的设置是整个强化学习任务的关键，类似优化问题中的目标优化，二者虽直接关联，但强化学习中奖励函数的设置更为复杂。区别于优化问题直接去优化目标，强化学习需要根据策略产生的每步决策去探索追求更大的序列奖励累积，当前决策的影响具有间接性与延迟性，其正向反馈可能来自多步之后。针对再入过程横侧向制导律的两大目标，即禁飞区规避与终端射程误差最小，在再入过程中，只有靠近禁飞区并到达终端位置时，才能获取奖励，因此奖励具备较强的稀疏性。

针对上述问题，设置再入过程奖励函数

$$f_{\mathrm{rwd}}=\begin{cases}\varepsilon_{\mathrm{r}}-R_{\mathrm{ran}}, & s_t=s_{\mathrm{f}}\\ \varepsilon_{\mathrm{r}}-R_{\mathrm{nf}}, & s_t\neq s_{\mathrm{f}},s_t\in Z_{\mathrm{nf}}\\ \varepsilon_{\mathrm{r}}, & s_t\neq s_{\mathrm{f}},s_t\notin Z_{\mathrm{nf}}\end{cases}\tag{25}$$

式中：ε_{r} 为倾侧角方向维持奖励；R_{ran} 为射程误差奖励；R_{nf} 为禁飞区规避奖励；Z_{nf} 为禁飞区状态空间；$s_t\in Z_{\mathrm{nf}}$ 表示当前飞行器位置靠近禁飞区，有穿过禁飞区风险。

射程误差奖励 R_{ran} 用于促使飞行器向目标靠近，其表达式为

$$R_{\mathrm{ran}}=\begin{cases}\dfrac{S_{\mathrm{err}}}{4}, & S_{\mathrm{err}}\geqslant 1\ 000\\ 100+150\left(\dfrac{S_{\mathrm{err}}}{1\ 000}\right)^2, & 100<S_{\mathrm{err}}<1\ 000\\ \left(\dfrac{10}{5^{\log_{20}10}}\right)S_{\mathrm{err}}^{\log_{20}10}, & S_{\mathrm{err}}\leqslant 100\end{cases}\tag{26}$$

式中：S_{err} 表示当飞行器到达终端状态时的距离目标位置的射程误差。

禁飞区规避奖励 R_{nf} 用于促使飞行器远离禁飞区域，其表达式为

$$R_{\mathrm{nf}}=[R_{\mathrm{z}}/(d_{\mathrm{z}}+R_{\mathrm{z}}/10)]^2/10\tag{27}$$

式中：d_{z} 为飞行器到禁飞区中心的距离。

倾侧角方向维持奖励 ε_{r} 主要用于抑制飞行器倾侧角频繁翻转，其表达式为

$$\varepsilon_{\mathrm{r}}=\begin{cases}0.01, & \mathrm{sign}(\sigma_t)=\mathrm{sign}(\sigma_{t-1})\\ -0.10, & \mathrm{sign}(\sigma_t)\neq\mathrm{sign}(\sigma_{t-1})\end{cases}\tag{28}$$

式中：σ_t 为 t 时刻的倾侧角。

通过设置上述奖励函数，使得再入过程与终端均能获得奖励，并且通过倾侧角翻转奖励实现了再入横侧向弹道在无需规避禁飞区时的平滑。对终端奖励的设置则主要考虑训练初期终端射程误差较大的情况，使得奖励在不同阶段遵循不同规律，引导飞行器逐步从较大误差向较小误差收敛。

(2) 网络参数设置

本文价值函数估计网络采用了 Dueling DQN，其网络结构如图 2 所示。

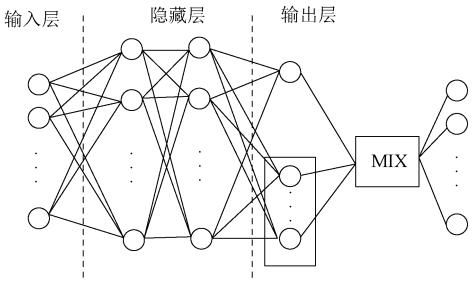

图 2 Dueling DQN 的网络结构

输入层输入的是归一化后的状态空间，如式(28)所示；隐藏层有两层，用于状态特征的转化，每层节点数均为 64，激活函数采用 ReLU 函数；输出层采用了 Dueling 设计，即分别拟合 $v_\pi(s_t)$ 与 $A(s_t,a_t)$，通过混合器(MIX)，实现了状态价值函数与优势函数的混合；最终输出为每个动作的价值估计。

针对再入制导轨迹规划与强化学习算法相结合过程中出现的两个问题进行分析。

一是算法的稳定性与收敛性问题。由于 Dueling DQN 算法中的训练数据是由 Q_local 网络自举产生的，所以样本具备强关联性，不满足强化学习所要求的样本独立同分布条件，这会影响算法的稳定性与收敛速度。

二是整个再入过程奖励的稀疏性问题。当飞行器穿越禁飞区时会获得较大的负向奖励；当飞行器到达目标区域后，获得的奖励需根据终端射程误差确定，射程误差越小，获得的奖励越大。在整个运动学方程积分解算过程中，其他时刻只有保留倾侧角不翻转的微小奖励，这种奖励的稀疏性降低了训练的稳定性和收敛速度。

为消除样本数据的关联性，并有效改善样本奖励的稀疏性，在 Dueling DQN 算法中采用了经

验池(Replay Buffer)技术。Replay Buffer 是一个有限长度的样本数据集，数据集中存储最新的样本数据。在算法学习过程中，Dueling DQN 不是根据最新产生的样本数据，而是从 Replay Buffer 中随机取出一小批数据(minibatch)作为学习样本数据。鉴于 DQN 本身是一个异策略(off-policy)算法，因此 Replay Buffer 应尽可能大，以达到消除数据关联的目的。

上述机制的实现流程为：网络初始化后先与环境进行交互，把产生的一条数据 $\{s_t, a_t, f_{\text{rwd}}, s_{t+1}\}$ 存入 Replay Buffer 中，此时不进行网络的更新；继续交互直至 Replay Buffer 中数据的数量达到上限 N_{buf}，从中随机抽取出 N_{bch} 条数据，以抽取后的数据为样本开始训练；与环境交互产生的新数据以队列形式不断存入 Replay Buffer 中，较早存入其中的旧数据被剔除，实现 Replay Buffer 中数据的滚动更新。

一般而言，N_{bch} 要远小于 N_{buf}，故可近似认为所抽取的训练样本是独立同分布的。通过该方法可以有效降低样本的关联性，提升训练稳定性与收敛速度。通过选取合适的 N_{bch}，可使平均梯度的计算具备较强的抗干扰能力，有效抑制扰动带来的数据奇异，并可使得梯度计算过程向量化，降低计算的资源耗费，提高训练速度。

由于通过网络构建的优化面是非凸的，沿着某一方向优化易陷入局部极值。此时经过 Adam 优化器的自适应调整，学习率已修正为较小值，探索率也已衰减到较小值，难以跳出可能存在的局部极值。当训练一定轮数后，应对优化器的学习率与动作选取的探索率进行重置，促使网络在优化过程中跳出局部极值。若优化收敛之处并非局部极值，衰减后的学习率和探索率也不会导引网络收敛到其他局部极值处。设第 q 阶段初始学习率和初始探索率为 α_{L}^{q} 与 ε^{q}，经过 E_k 轮训练后，重置网络与智能体的学习率和探索率。则第 $q+1$ 阶段的学习率和探索率的表达式为

$$\begin{cases} \alpha_{\text{L}}^{q+1} = \eta_{\alpha}\alpha_{\text{L}}^{q} \\ \varepsilon^{q+1} = \eta_{\varepsilon}\varepsilon^{q} \end{cases} \tag{29}$$

式中：η_{α}，η_{ε} 为初始学习率和探索率的衰减系数。

设置网络训练过程的仿真参数：初始学习率 $\alpha_{\text{L}} = 10^{-5}$；初始探索率 $\varepsilon = 0.2$，即训练过程中有 $1-\varepsilon$ 的概率选择当前最大价值动作，而有 ε 的概率随机选取一个动作；探索率的衰减率 $\eta = 0.999\ 97$，每次训练完成后对探索率进行衰减，这是对强化学习的探索(exploration)与利用(exploitation)的权衡折中；η_{α}，η_{ε} 分别为 0.1，0.5，$E_k = 100$；在计算网络目标状态动作对的价值时，奖励反馈的折扣系数 $\gamma = 0.9$；Replay Buffer 的 $N_{\text{buf}} = 100\ 000$，$N_{\text{bch}} = 4\ 096$；优化器采用 Adam，可自适应调节更新步长与方向；训练数据源于飞行器在仿真环境中交互所产生的轨迹数据，仿真的时间步长为 1 s。

基于 Dueling DQN 的禁飞区规避再入制导系统结构如图 3 所示。纵向制导采用跨周期迭代预测校正制导律[11]，横侧向采用基于 Dueling DQN 的深度强化学习制导律。

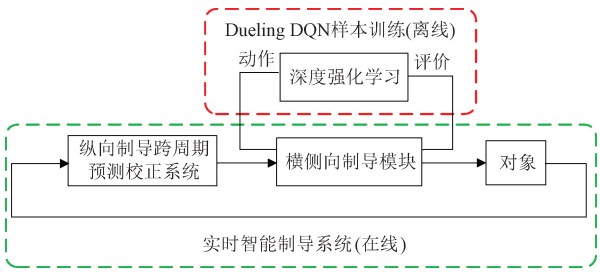

图 3　基于 Dueling DQN 的禁飞区规避再入制导系统结构示意图

3　仿真验证

在 Dueling DQN 算法的网络训练过程中，连续统计了 300 个再入任务的训练收益，累积奖励随训练轮数的变化如图 4 所示。在训练过程中，累积奖励不断增加，且随着网络的不断探索与更新，在训练后期再入制导的累积奖励波动逐渐减小，这表明深度强化学习网络具备一定的稳定性。

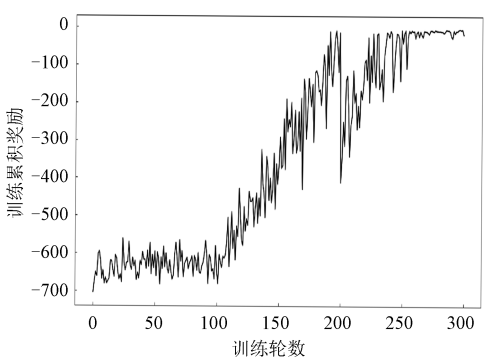

图 4　训练累积奖励随训练轮数变化图

按初始经纬度不同，设置 4 种仿真初始条件，对图 3 所示的再入制导系统进行数值仿真，验证 Dueling DQN 算法对不同初始条件的适应能力。仿真初始条件如表 3 所示。

表 3 仿真初始条件

仿真条件	λ / (°)	ϕ / (°)	禁飞区		
			λ_z /(°)	ϕ_z /(°)	R_z/km
Case 1	160	5	200	34	500
Case 2	160	−5	200	34	500
Case 3	170	5	200	34	500
Case 4	170	−5	200	34	500

在 4 种仿真初始条件下的终端射程误差如表 4 所示。其中预期射程为根据初始条件与终端条件计算得到的直线射程，而实际射程为飞行器规避禁飞区、变换倾侧角方向、改变速度水平方向后的真实射程。在 4 种仿真条件下，仿真结束时刻的射程误差均小于 3 km，可见该算法精度满足中制导转向末制导的交班要求。

表 4 终端射程误差仿真结果

仿真条件	预期射程/km	实际射程/km	射程误差/km
Case 1	8 296.77	8 490.53	2.19
Case 2	9 226.22	9 635.15	2.53
Case 3	7 690.50	8 009.54	1.97
Case 4	8 641.84	9 005.98	2.65

采用 Dueling DQN 算法，根据设置的 4 种仿真初始条件，通过大量弹道积分，与再入环境交互获取训练样本，高度-速度（HV）再入轨迹如图 5 所示，再入禁飞区规避轨迹如图 6 所示。仿真结果表明：基于 Dueling DQN 的横侧向制导律与基于跨周期迭代预测校正的纵向制导律相结合后，生成的再入轨迹较为平滑；在复合倾侧角约束条件下，再入轨迹均能满足终端高度约束，对于设定的禁飞区均具备规避能力。

在 4 种仿真条件下，倾侧角大小随速度变化的仿真曲线如图 7 所示，倾侧角符号随速度变化的仿真曲线如图 8 所示。

再入任务倾侧角的大小由纵向制导律确定。由图 7 可知：在 4 种仿真条件下，倾侧角均经历了由小变大，再逐渐变小的过程，其中初期维持较小

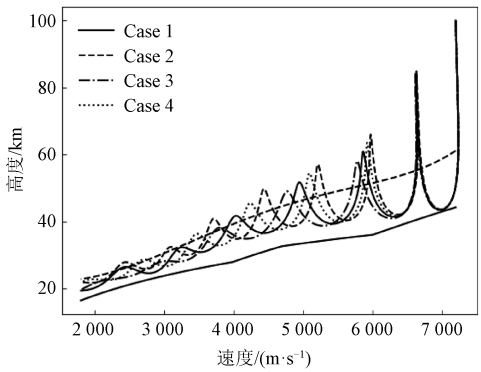

图 5 HV 再入轨迹

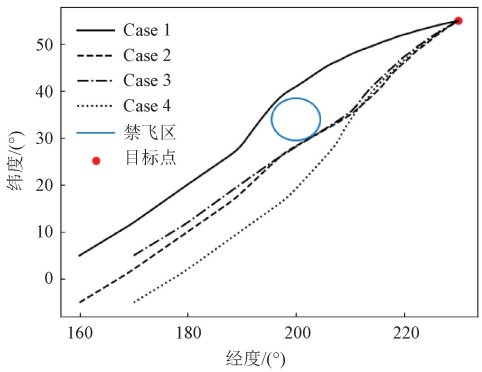

图 6 再入禁飞区规避轨迹

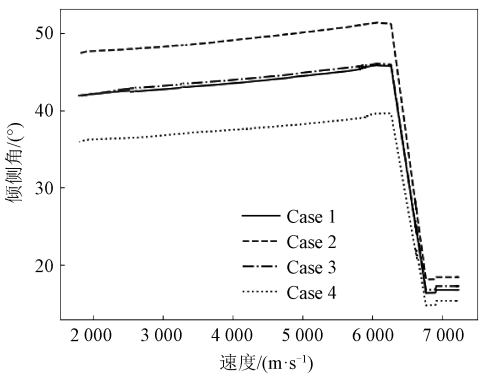

图 7 倾侧角大小随速度变化仿真曲线

倾侧角是为了保证再入初段满足热流约束条件，之后则为同时满足射程要求与过程约束开展设计，具体设计过程见文献[11]。

再入任务倾侧角的符号由横侧向制导律确定。由图 8 可知：倾侧角符号变化较为剧烈，这体现了横侧向制导律对飞行器进行航向调整，使其规避禁飞区、到达目标区域的过程。具体原因分析如下：

a) 倾侧角的符号由 Dueling DQN 预测的动作价值确定，具备一定的不可解释性，且在建立横

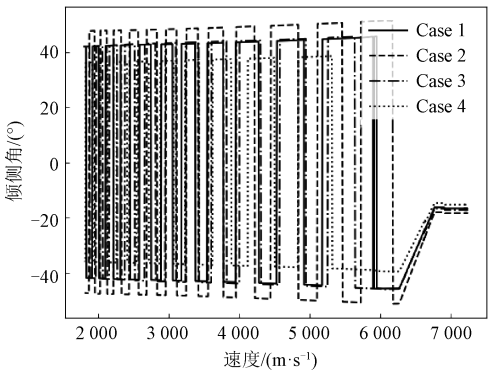

图 8　倾侧角符号随速度变化仿真曲线

侧向再入的 MDP 模型时，存在着较多的人为干扰因素，状态空间选取与奖励函数设置有待进一步优化，尤其是奖励函数的设置极大影响了强化学习算法效果；

b）网络训练过程存在较大随机性，训练至较为良好状态所需时间严重依赖于超参数设置，且再入任务参数设置也会影响训练，整个训练过程有待进一步细化提升。

设置再入任务的 50 个随机初始位置，通过大量仿真验证 Dueling DQN 算法禁飞区规避能力的鲁棒性与自适应性。随机初始位置的再入轨迹如图 9 所示。可知，该算法能够较好地满足终端高度约束，没有违反 HV 走廊下界。

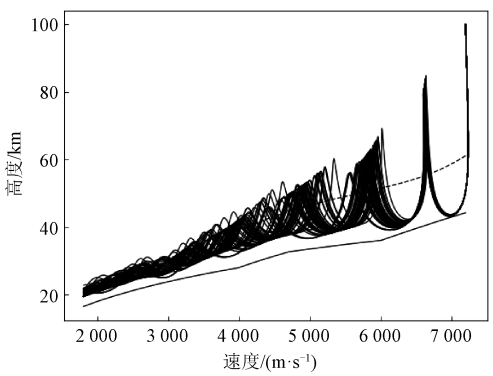

图 9　随机初始位置的再入轨迹

不同初始位置的再入禁飞区规避轨迹如图 10 所示。可知，大部分情况下算法能够实现禁飞区规避与目标点到达，但在 50 次随机初始条件仿真中，仍然有 6 次出现了违反禁飞区约束的情况，并且有 1 次仿真任务射程误差较大。

在仿真过程中，存在违反禁飞区约束和射程误差较大等问题的原因分析如下。

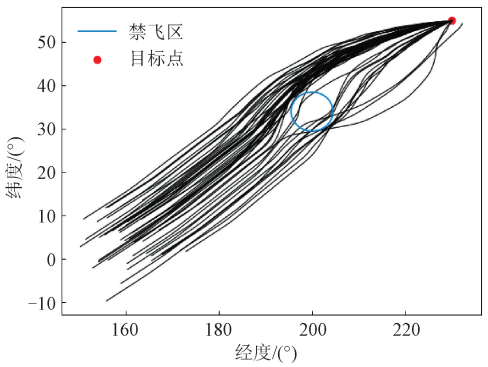

图 10　随机初始位置的再入禁飞区规避轨迹

再入飞行器机动能力有限。由于再入过程将倾侧角作为控制量，通过改变倾侧角的符号，从而改变升力水平分量的方向与大小来实现航向改变。该方式所能提供的侧向机动性有限，因此存在某些极端情况下无法完全绕开禁飞区的现象。

网络泛化能力有限。网络是针对经度为 160°、纬度为 5°的初始条件训练的，通过设置随机初始条件进行仿真，能在一定程度上验证该网络对不同初始条件的适应性。但网络不具备无限的泛化能力，不可能适用于所有情况。

奖励函数设置受主观人为因素影响。奖励函数设置是强化学习算法的核心，本文通过设置飞行器再入过程的倾侧角方向维持奖励、禁飞区规避奖励与射程误差奖励来实现对再入飞行器倾侧角符号的决策引导。如何设置更加合理的禁飞区规避奖励与射程误差奖励，以及如何合理平衡禁飞区规避奖励与射程误差奖励，有待进一步深入研究。本质上这也是多目标强化学习及多目标优化所要解决的关键问题之一。

4　结论

本文首先分析了当前再入制导问题相关模型与约束，介绍了 Dueling DQN 算法原理与实现过程。然后在此基础上建立了再入制导过程的 MDP 模型，对横侧向制导相关状态进行了归一化处理，建立了 Dueling DQN 的网络模型，并引入 Replay Buffer 机制与迭代训练机制，自适应调整学习率与探索率以增加跳出局部极值的概率。最后建立数值仿真模型，验证了纵向与横侧向制导律能够在满足过程约束条件下，导引飞行器规避

禁飞区，飞向目标区域，且具备较高精度，证明了本文所提方法的有效性。

参考文献

[1] 熊俊辉，李克勇，刘燚，等. 临近空间防御技术发展态势及突防策略[J]. 空天防御，2021，4(2)：82-86.

[2] 周蓓蓓，刘珏. 智能化技术在精确打击体系中的应用[J]. 空天防御，2019，2(3)：77-83.

[3] YOUSSEF H，CHOWDHRY RS，LEE H，et al. Predictor-corrector entry guidance for reusable launch vehicles [C]// AIAA Guidance，Navigation，and Control Conference and Exhibit，Montreal，Canada. Reston，VA：AIAA，2001：4043.

[4] SHEN Z J，LU P. On-board entry trajectory planning expanded to sub-orbital flight [C]// AIAA Guidance，Navigation，and Control Conference and Exhibit，Austin. Reston，VA：AIAA，2003：5736.

[5] 潘乐飞，李新国. 可重复使用运载器预测-校正再入制导研究[J]. 飞行力学，2007，25(1)：55-58.

[6] JOSHI A，SIVAN K，AMMA S S. Predictor-corrector reentry guidance algorithm with path constraints for atmospheric entry vehicles [J]. Journal of Guidance，Control，and Dynamics，2007，30(5)：1307-1318.

[7] LIU X D，CHENG L，ZHANG Q Z，et al. Entry trajectory optimization for hypersonic vehicle based on time-scales separation guidance with waterweeds algorithm[C]// Proceedings of 2016 Chinese Guidance，Navigation and Control Conference (CGNCC). Piscataway，NJ：IEEE Press，2016：209-215.

[8] LIU X F，SHEN Z J，LU P. Closed-loop optimization of guidance gain for constrained impact[J]. Journal of Guidance，Control，and Dynamics，2017，40(2)：453-460.

[9] 张赵寰宇. 基于深度强化学习的高超声速飞行器智能反拦截方法研究[D]. 哈尔滨：哈尔滨工业大学，2021.

[10] 吕雅丽. 基于Q学习的高超声速飞行器路径规划方法研究[D]. 成都：电子科技大学，2018.

[11] 程林. 高超声速飞行器实时最优闭环再入制导技术研究[D]. 北京：北京航空航天大学，2017.

[12] 田若岑，张庆振，郭云鹤，等. 基于禁飞区规避的高超声速飞行器再入制导律设计[J]. 空天防御，2022，5(2)：65-74.

[13] PHILLIPS T H. A common aero vehicle(CAV) model，description，and employment guide [R/OL]. (2003-01-27)[2023-12-01]. https：//www. researchgate. net/publication/272494034_.

[14] 周来，靳晓伟，郑益凯. 基于深度强化学习的作战辅助决策研究[J]. 空天防御，2018，1(1)：31-35.