DOI:10.11918/j.issn.0367-6234.201808104

飞行时间约束下的再入制导律

方 科,张庆振,倪 昆,崔朗福

(北京航空航天大学 自动化科学与电气工程学院,北京 100083)

摘 要: 为应对现代战场的信息化与集群化发展趋势,从多高超声速飞行器饱和打击任务需求出发,针对其中的再入飞行时间约束条件进行研究,提出一套基于 Deep Q-learning Network(DQN)的时间可控再入制导律. 该制导律工作流程为首先纵向轨迹预测-校正模块根据当前飞行状态和攻角-速度剖面规划出倾侧角幅值;然后在线约束强化管理模块对其进行安全限幅处理;最后倾侧角符号规划模块以调节再入飞行时间为目标,在对横向飞行状态进行马尔科夫决策过程建模的基础上,设计相应的深度神经网络进行离线训练以在线生成倾侧角符号,进而与幅值信息共同组成最终的倾侧角指令. 多组仿真的对比分析结果表明:在标称环境下的多任务仿真中,时间可控再入制导律能够自主地进行倾侧角符号的在线规划,在不影响制导精度的前提下,对再入飞行时间进行调整以满足不同的任务需求;在参数拉偏的蒙特卡洛仿真中,其在保证再入飞行安全、稳定的同时,仍然能将时间误差控制在合理的范围之内. 从而验证了相对于传统方法而言,本研究所设计的再入制导律在任务适应性、鲁棒性与时间可控性等方面均具有良好表现,能够有效地满足飞行时间约束下的再入任务需求.

关键词: 高超声速飞行器;再入制导;预测-校正制导;深度强化学习;DQN

中图分类号: V448.235 **文献标志码:** A **文章编号:** 0367-6234(2019)10-0090-08

Reentry guidance law with flight time constraint

FANG Ke, ZHANG Qingzhen, NI Kun, CUI Langfu

(School of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China)

Abstract: In order to cope with the development trend of informationization and clustering in modern battlefields, this paper studies the reentry flight time constraint in the mission of multi-hypersonic vehicles' saturation attack, and proposes a time-controllable reentry guidance law based on the Deep Q-learning Network(DQN). Its workflow is mainly divided into three parts. First, according to the current flight state and angle of attack vs. velocity profile, the amplitude of bank angle is planned based on the prediction-correction module of longitudinal trajectory. Then the online constraint enhanced management module performs safety limiting processing the amplitude of bank angle. Finally, based on Markov Decision Process modeling of lateral flight state, the symbol planning module aims to adjust reentry flight time and to designs the corresponding deep neural network for offline training to generate the symbol of bank angle online, and then amplitude information is combined to form the final bank angle command. The comparative analysis of mutiple simulations shows that in multi-mission simulation under nominal environment, the time-controllable reentry guidance law can independently plan bank angle's symbol online and adjust reentry flight time to meet different mission requirements without affecting guidance precision. In the Monte Carlo simulations with biased parameters, the time error can still be controlled within a reasonable range while ensuring safe and stable reentry flight. Therefore, compared with the traditional method, the reentry guidance law designed in this paper has good performance in terms of task adaptability, robustness, and time controllability, and it can effectively meet reentry mission requirements with the flight-time constraint.

Keywords: hypersonic vehicle; reentry guidance; predictor-corrector guidance; deep reinforcement learning; deep Q-learning network

随着科学技术的发展,现代战争的作战形式发生了巨大变化,其中以高超声速滑翔式再入飞行器为代表的远程快速精确静默打击受到了越来越多国家的关注[1]. 与此同时,为了应对其大跨空域、高超声速飞行所带来的威胁,各国相继研发出一系列反导系统来降低高超声速飞行器的作战效能. 在这一系列技术博弈推动下,发展多高超声速飞行器协同饱和打击技术成为未来的一个必然趋势[2].

由于长距离飞行和跨空域的复杂环境变化特点,高超声速飞行器协同饱和打击任务要求各飞行器能够在一定程度上自由调节自身飞行时间,进而实现在指定时间范围内对目标的打击. 而在这一过程中,无动力滑翔再入段占据了90%以上,是影响

收稿日期: 2018-08-22
作者简介: 方 科(1994—),男,硕士研究生
通信作者: 张庆振,zhangqz@buaa.edu.cn

最终时间和饱和打击效果的决定性因素之一[2-3]，因而针对该阶段的飞行时间可控性要求进行相应的再入制导律设计具有一定的必要性与紧迫性.

现阶段再入制导方法主要有标称轨迹制导和预测-校正制导，二者各有特点并已成功运用于工程实践之中[4-6]. 但其均没有考虑飞行时间约束，仅以到达指定状态为目标，加之再入运动的复杂性，时间可控再入制导律的发展较为缓慢. 王少平等[2]利用离线轨迹优化方法对其可控性进行了验证，但无法解决实际环境中的参数摄动问题；方科等[3]利用神经网络对剩余飞行误差进行在线预估与校正，总体上具有一定的可行性但存在较多的人为限定因素. 从已有的研究成果中可知，高超声速飞行器指定飞行时间的再入制导问题的主要难点与特点为：1）大跨空域、高空高速飞行使得整个过程中存在大量相互耦合的不确定性因素；2）各类干扰对于飞行时间的影响不明确且相互耦合，难以进行详细分析；3）无动力再入过程总能量有限且控制过程不可逆，难以实现类似文献[7]中的显式制导律设计，需要进行方法创新；4）高超声速飞行器再入轨迹具有较强的横纵向解耦特性[5]. 在到达目标点的前提下，飞行时间主要与横向制导律有关，其通过对倾侧角符号的规划来改变横向机动范围，实现再入飞行时间调整[3].

由此可知，时间可控再入制导可归结为横向上的倾侧角符号规划问题，但由于其复杂性和制导实时性要求，常规方法难以应用. 而强化学习类方法在一些决策问题上的出色表现为本文供了可行的探索方向，并且其“离线训练 + 在线使用”模式具有较强的适应性与实时性[8-9]. 所以本文首先给出总体制导律结构，之后引入深度 Q-学习网络（deep Q-learning network，DQN）进行倾侧角符号规划策略设计，实现指定飞行时间的再入过程，最后进行相应的仿真验证.

1　指定飞行时间的再入问题描述

高超声速飞行器的三自由度无动力滑翔再入过程（忽略地球自转）运动方程为[5]

$$\begin{cases}\dot r = v\sin\theta,\\ \dot\lambda = \dfrac{v\cos\theta\sin\psi}{r\cos\varphi},\\ \dot\varphi = \dfrac{v\cos\theta\cos\psi}{r},\\ \dot S_{\mathrm{e}} = \dfrac{v\cos\theta}{r},\\ \dot v = -\dfrac{D}{m} - g\sin\theta,\\ \dot\theta = \dfrac{1}{v}\left[\dfrac{L\cos\sigma}{m} + \left(\dfrac{v^2}{r} - g\right)\cos\theta\right],\\ \dot\psi = \dfrac{1}{v}\left[\dfrac{L\sin\sigma}{m\cos\theta} + \dfrac{v^2}{r}\cos\theta\sin\psi\tan\varphi\right].\end{cases}$$

式中：地心距 $r = R_0 + h$，其中 $R_0 = 6\ 378$ km 为地球半径，h 为飞行高度；v 为飞行速度；θ 为速度倾角，水平为 0 且向上为正；S_{e} 为射程角；λ、φ 分别为经纬度；ψ 为弹道偏航角，正北方为 0 且顺时针为正；σ 为倾侧角，右偏为正；m 为质量；g 为当前高度下地球重力加速度；L、D 分别为升力和阻力.

再入飞行过程是一个复杂的多约束运动学问题. 对于制导律而言需要在线生成制导指令，实现多余能量的安全逸散和引导飞行器到达指定状态（高度、速度、位置等）. 具体约束项分别为：

$$\dot Q = \frac{11\ 030}{\sqrt{0.1}}\left(\frac{v}{V_{\mathrm{c}}}\right)^{3.15}\left(\frac{\rho}{\rho_0}\right)^{0.5} \leqslant \dot Q_{\max}, \tag{1}$$

$$n = \sqrt{L^2 + D^2}/mg_0 \leqslant n_{\max}, \tag{2}$$

$$q = 0.5\rho v^2 \leqslant q_{\max}, \tag{3}$$

$$(g - v^2/r) - L\cos\sigma_{\mathrm{EQ}}/m = 0, \tag{4}$$

$$|\sigma_{\mathrm{cmd}}| \leqslant 90°, |\dot\sigma_{\mathrm{cmd}}| \leqslant \dot\sigma_{\max}, \tag{5}$$

$$h(t_{\mathrm{f}}) = h_{\mathrm{f}}, v(t_{\mathrm{f}}) = v_{\mathrm{f}}, \lambda(t_{\mathrm{f}}) = \lambda_{\mathrm{f}}, \varphi(t_{\mathrm{f}}) = \varphi_{\mathrm{f}}. \tag{6}$$

式中：$\dot Q$、n、q 分别为驻点处热流密度、过载和动压，均是不可违反的硬性过程约束；式（4）为可违反的拟平衡滑翔软约束，其仅提供 0°倾侧角下的参考飞行轨迹；σ_{cmd}、$\dot\sigma_{\mathrm{cmd}}$分别为制导系统输出的倾侧角和在同一符号下的变化率；t_{f} 为终端时刻；$V_{\mathrm{c}} = \sqrt{g_0 R_0}$，$g_0 = 9.81\ \mathrm{m/s^2}$；$\rho = \rho_0 \mathrm{e}^{-h/7\,200}$为当前高度下大气密度，$\rho_0 = 1.225\ \mathrm{kg/m^3}$；$\sigma_{\mathrm{EQ}}$为拟平衡滑翔倾侧角[5].

除此之外，飞行时间约束条件为

$$t_{\mathrm{need}} = \int_{t_0}^{t_{\mathrm{f}}} \mathrm{d}t = t_{\mathrm{f}} - t_0.$$

式中：t_{need}为期望的再入飞行时间，t_0 为起始时刻.

本文以 CAV-H 为对象进行研究与仿真分析，具体参数设置见文献[10].

2　制导系统总体结构

本文所设计的制导律总体如图 1 所示，其主要包括纵向轨迹预测-校正、在线约束强化管理、倾侧角符号规划 3 个模块.

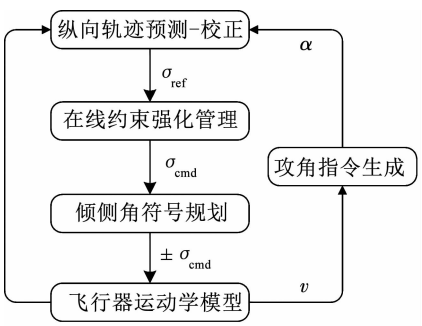

图 1　制导系统总体结构

Fig. 1　Overall scheme of reentry guidance system

在每个制导周期内,攻角指令生成模块根据如下的速度-攻角剖面[10]生成 α.

$$\alpha=\begin{cases}35°, v\geqslant 10.0Ma;\\ [35-0.45(v-10)^2]°, 2.5\ Ma\leqslant v<10.0Ma.\end{cases}$$

式中:Ma 为马赫数. 随后纵向轨迹预测-校正模块进行纵向轨迹规划并输出无符号倾侧角 σ_{ref};之后在线约束强化管理模块对其进行限幅,输出无符号的 σ_{cmd};最后倾侧角符号规划模块利用训练好的 DQN,根据当前飞行状态进行决策,输出有符号的倾侧角指令 $\pm\sigma_{\text{cmd}}$ 并作用于运动模型之中.

3 指定飞行时间的再入制导律

由于再入制导问题的复杂性和狭窄的可行解范围,首先需要进行约束降维与简化,即构建如图 2、3 所示的可行高度-速度走廊和倾侧角走廊.

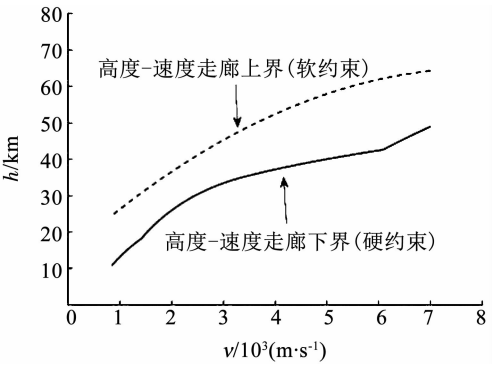

图 2 可行高度-速度走廊

Fig. 2 Feasible altitude-velocity corridor

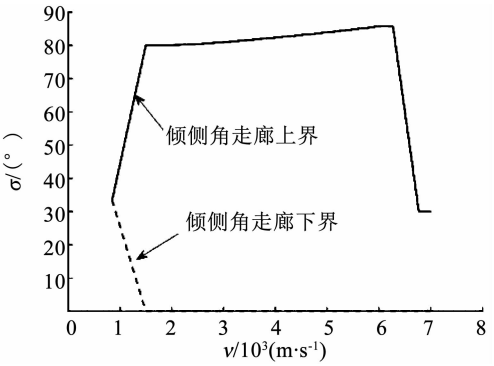

图 3 倾侧角-速度走廊

Fig. 3 Bank angle-velocity corridor

相关研究表明,再入飞行时间主要与横向机动范围、参数不确定性等因素有关[3]. 其中参数不确定性是造成时间不可控的主要被动因素,需要通过对主动因素——即横向机动范围的调节进行抑制. 而再入过程具有明显的横纵向解耦特性[5],横向机动范围只与横向制导律——即倾侧角符号规划方法有关;同时因篇幅有限,本文将重点介绍倾侧角符号规划模块的设计方法.

3.1 纵向轨迹预测-校正

根据地球任意两经纬度之间最短圆弧距离计算公式和射程角定义,可将终端经纬度约束转换为如下的射程角约束:

$$S_{\text{e}}(t_{\text{f}})=\arccos[\sin\varphi\sin\varphi_{\text{f}}+\cos\varphi\cos\varphi_{\text{f}}\cos(\lambda-\lambda_{\text{f}})]=S_{\text{go}},$$

式中 S_{go} 为所需的剩余射程角. 参数化之后的倾侧角规划剖面为

$$\sigma_{\text{design}}(v)=\omega\sigma_{\min}(v)+(1-\omega)\sigma_{\max}(v). \tag{7}$$

式中:ω 为权重系数;$\sigma_{\text{design}}(v)$ 为规划的倾侧角剖面;$\sigma_{\min}(v)$、$\sigma_{\max}(v)$ 分别为倾侧角-速度走廊的下、上界. 由此将轨迹规划转化为 ω 的单参数规划问题,而 ω 与 S_{e} 存在如下的单调递增隐函数关系[6]:

$$R_0S_{\text{e}}=f(\omega), \tag{8}$$

式(8)可在小范围内近似为[3,6]

$$S_{\text{p}}(\omega)\approx c\omega+d.$$

式中:$S_{\text{p}}(\omega)$ 为预估的剩余射程角;c、d 为待估计参数. 通过引入最小二乘方法对其进行在线参数迭代估计,使得在每个制导周期内仅需一次全弹道积分便可完成 ω 与 $\sigma_{\text{design}}(v)$ 的校正,具体计算流程如下[3,6]:

$$\begin{cases}\hat{\boldsymbol{x}}_k=[\hat{c}_k\ \hat{d}_k]^{\text{T}},\\ \hat{\boldsymbol{x}}_{k+1}=\hat{\boldsymbol{x}}_k+\boldsymbol{K}_{k+1}(z_{k+1}-\boldsymbol{h}_{k+1}^{\text{T}}\hat{\boldsymbol{x}}_k),\\ \boldsymbol{h}_k=[\omega_k\ 1]^{\text{T}},\\ z_k=S_{\text{p}}(\omega_k),\\ \boldsymbol{K}_{k+1}=\boldsymbol{P}_k\boldsymbol{h}_{k+1}(R_{k+1}+\boldsymbol{h}_{k+1}^{\text{T}}\boldsymbol{P}_k\boldsymbol{h}_{k+1})^{-1},\\ \boldsymbol{P}_{k+1}=(\boldsymbol{I}-\boldsymbol{K}_{k+1}\boldsymbol{h}_{k+1}^{\text{T}})\boldsymbol{P}_k,\\ \omega_{k+1}=\omega_k+[S_{\text{go}}-S_{\text{p}}(\omega_k)]/\hat{c}_{k+1}.\end{cases}$$

式中:下标 k 为第 k 个周期;$\hat{c}$、$\hat{d}$ 为估计值. 之后结合式(7)与校正后的 ω 得到倾侧角规划剖面 $\sigma_{\text{design}}(v)$,并根据当前速度 v 插值得到本制导周期的倾侧角指令 σ_{ref}.

3.2 在线约束强化管理

在线约束强化管理模块如图 4 所示[3,6].

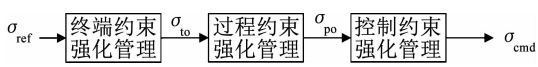

图 4 在线约束强化管理

Fig. 4 Online constraint enhanced management

其中,终端约束强化管理过程为

$$\begin{cases}\Delta\theta=\theta-\theta_{\text{f}},\\ \Delta\cos\sigma_{\text{coe}}=A_3vm\Delta\theta/L,\\ \sigma_{\text{to}}=\arccos(\cos\sigma_{\text{ref}}+\Delta\cos\sigma_{\text{coe}}),\end{cases}$$

式中常数 $A_3<0$. 过程约束强化管理过程为

$$\begin{cases}\Delta h_{\mathrm{d}}=h-h_{\mathrm{down}},\\ \cos\sigma_{\mathrm{comin}}=\dfrac{m}{L}\left[\dfrac{\Delta\ddot{h}_{\mathrm{d}}-2\lambda_h\Delta\dot{h}_{\mathrm{d}}-\lambda_h^2\Delta h_{\mathrm{d}}}{\cos\theta}-\left(g-\dfrac{v^2}{r}\right)\cos\theta\right],\\ \sigma_{\mathrm{po}}=\min(\sigma_{\mathrm{to}},\sigma_{\mathrm{comin}}).\end{cases}$$

式中：$\lambda_h\in[0,1]$ 为高度下降速率限制因子；Δh_{d} 为剩余可行高度，h_{down} 为当前速度 v 在可行高度-速度走廊上所对应的高度下界.

控制约束强化管理过程为

$$\sigma_{\mathrm{cmd}}^{k}=(1-k_{\mathrm{pf}})\sigma_{\mathrm{cmd}}^{k-1}+k_{\mathrm{pf}}\sigma_{\mathrm{po}}^{k}.$$

式中：上标 k 为第 k 个制导周期；$k_{\mathrm{pf}}\in[0,1]$. 经过以上限幅处理最终得到倾侧角指令 σ_{cmd}.

3.3　倾侧角符号规划

在传统再入制导律设计中，通常只考虑制导精度、实时性、鲁棒性等，然而饱和打击任务的出现，使得再入飞行时间的可控性也成为一个值得考虑的重要因素[1].

现阶段研究表明，高超声速飞行器无动力滑翔再入过程具有一定的时间可控性，并且主要与横向制导律有关[3]. 其通过对倾侧角符号的规划来改变横向机动范围，进而调节再入飞行时间. 倾侧角符号规划本质上而言是一个典型的二值决策问题，即根据当前状态和目标给出“+”或“-”，但受困于问题的复杂性，难以使用传统方法进行分析与设计. 然而强化学习类方法所具有的设计流程通用性、自学习自适应能力、泛化能力，为本文的横向制导律设计提供了可行的探索方向.

3.3.1　DQN 基本流程

强化学习类方法的目标是通过大量数据积累和参数寻优，得到一套能最大化如下所示的期望累计收益 $J_{\pi^\theta}(s_0,s_{\mathrm{f}})$ 的策略 $\boldsymbol{\pi}^{\theta}$[11].

$$J_{\pi^\theta}(s_0,s_{\mathrm{f}})=\boldsymbol{E}_{\pi^\theta}\left[\sum_{t=0}^{n}\eta^t P(s_t,a_t,s_{t+1})f_{\mathrm{R}}(s_t,a_t,s_{t+1})\right],\tag{9}$$

式中：$J_{\pi^\theta}(s_0,s_{\mathrm{f}})$ 为在 $\boldsymbol{\pi}^{\theta}$ 作用下，从初状态 s_0 经过 n 步到达终端状态 s_{f} 的期望累计收益；$\boldsymbol{E}[\cdot]$ 为求取期望值的符号函数；s_t 为第 t 步的状态；a_t 为第 t 步的动作；$f_{\mathrm{R}}(s_t,a_t,s_{t+1})$ 为状态 s_t 下选择动作 a_t 后到达状态 s_{t+1} 的收益函数；$P(s_t,a_t,s_{t+1})$ 为状态转移概率，表示状态 s_t 下采用 a_t 后成功转移到 s_{t+1} 的概率；$\eta\in[0,1]$ 为衰减系数. 与此同时需要将原问题建模成如下的马尔可夫决策过程（markov decision process，MDP）.

$$\mathrm{MDP}=(\boldsymbol{S},\boldsymbol{A},P,f_{\mathrm{R}},\eta).$$

式中：$\boldsymbol{S}$ 为所有状态空间的集合且 $\forall s\in\boldsymbol{S}$；$\boldsymbol{A}$ 为所有动作空间的集合且 $\forall a\in\boldsymbol{A}$. 针对本文所研究问题具有状态空间无限、动作空间有限的特点，选择 DQN 进行横向制导律设计，其基本流程如图 5 所示.

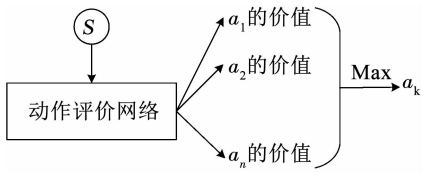

图 5　DQN 基本流程

Fig. 5　Fundamental process of DQN

在每个训练周期中，DQN 根据内部参数 $\boldsymbol{\pi}^{v}$ 的动作评价网络对当前状态 s 下所有能够采取的动作 $a_i(i=1,2,\cdots,n)$ 的价值 $Q^{\pi^v}(s,a_i)$ 进行估计，输出其中具有最大价值的动作 a_k；之后根据 a_k 的实际价值与估计值之间的误差对 $\boldsymbol{\pi}^{v}$ 进行更新. 为保证网络具有一定探索能力，定义以 η_ε 进行指数衰减的探索率 ε，使得每次输出均存在概率 ε 进行随机动作选择. 假设在第 t 步时状态为 s_t，输出的动作 a_t，收益为 R_t，则该状态-动作对的实际价值 $Q_{\mathrm{T}}^{\pi^v}(s_t,a_t)$ 为

$$Q_{\mathrm{T}}^{\pi^v}(s_t,a_t)=\begin{cases}R_t,s_{t+1}=s_{\mathrm{f}};\\ R_t+\eta\max Q^{\pi^v}(s_{t+1},a_{t+1}),s_{t+1}\neq s_{\mathrm{f}}.\end{cases}$$

动作评价网络对于该状态-动作对的估计价值为 $Q^{\pi^v}(s_t,a_t)$，可得误差函数 $L(\boldsymbol{\pi}^{v})$ 为

$$L(\boldsymbol{\pi}^{v})=\boldsymbol{E}_{\pi^\theta}\left[\frac{1}{2}(Q_{\mathrm{T}}^{\pi^v}(s_t,a_t)-Q^{\pi^v}(s_t,a_t))^2\right].$$

$L(\boldsymbol{\pi}^{v})$ 相对于各网络参数 $\boldsymbol{\pi}^{v}$ 的梯度为

$$\frac{\partial L(\boldsymbol{\pi}^{v})}{\partial\boldsymbol{\pi}^{v}}=\boldsymbol{E}_{\pi^\theta}\left[(Q_{\mathrm{T}}^{\pi^v}(s_t,a_t)-Q^{\pi^v}(s_t,a_t))\frac{\partial Q^{\pi^v}(s_t,a_t)}{\partial\boldsymbol{\pi}^{v}}\right],\tag{10}$$

根据式(10)对网络参数 $\boldsymbol{\pi}^{v}$ 进行调整与更新，并不断重复上述流程，最终在理论上可以获得一套能够根据当前状态与目标对倾侧角符号进行规划决策的网络参数 $\boldsymbol{\pi}^{v}$.

3.3.2　马尔可夫决策过程建模

为了将 DQN 应用到横向制导律设计中，首先对原问题进行马尔可夫决策过程建模. 考虑到再入飞行时间主要与横向飞行状态有关[3]，并综合射程、落点误差和再入飞行时间要求，构建如下的状态空间集合 S 为

$$\boldsymbol{S}=[h\quad v\quad\lambda\quad\varphi\quad S_{\mathrm{go}}\quad\psi\quad\Delta t]^{\mathrm{T}},$$

式中 $\Delta t=t_{\mathrm{need}}-t_{\mathrm{pre}}$ 为剩余飞行时间误差，其中 t_{pre} 为当前飞行状态下改变倾侧角符号后，剩余飞行时间的估计量，其来源于纵向轨迹预测-校正模块中的全弹道积分预测结果. 其中由于再入飞行的横纵向解耦特性，倾侧角符号的改变只会影响飞行轨迹的横向分布而不会造成总射程的变化，既不会影响到纵向轨迹规划结果. 所以只需要改变原轨迹预测过程中的倾侧角符号即可得到在倾侧角改变后的剩余时间

t_{pre},并且该过程也不会增加额外的制导周期耗时.

由于各状态量之间数量级与物理意义的差异,需要进行如下的归一化过程:

$$\tilde{h}=\frac{h-(h_0+h_{\mathrm{f}})/2}{h_0-h_{\mathrm{f}}},\tilde{v}=\frac{v-(v_0+v_{\mathrm{f}})/2}{v_0-v_{\mathrm{f}}},$$

$$\tilde{\lambda}=\frac{\lambda-(\lambda_{\mathrm{f}}+\lambda_0)/2}{\lambda_{\mathrm{f}}-\lambda_0},\tilde{\varphi}=\frac{\varphi-(\varphi_{\mathrm{f}}+\varphi_0)/2}{\varphi_{\mathrm{f}}-\varphi_0},$$

$$\tilde{S}_{\mathrm{go}}=\frac{S_{\mathrm{go}}}{S_0},\tilde{\psi}=\frac{\psi}{2},\Delta\tilde{t}=\frac{t_{\mathrm{need}}-t_{\mathrm{p}}-t_{\mathrm{pre}}}{t_{\mathrm{need}}},$$

其中归一化之后的状态空间集合 S 为

$$\boldsymbol{S}=\left[\tilde{h}\quad\tilde{v}\quad\tilde{\lambda}\quad\tilde{\varphi}\quad\tilde{S}_{\mathrm{go}}\quad\tilde{\psi}\quad\Delta\tilde{t}\right]^{\mathrm{T}}.\tag{11}$$

式中:下标 0、f 分别为初始时刻与终端时刻;S_0 为初始射程角;t_{p} 为已飞行时间.

由于倾侧角符号仅存在正负两种可能,所以动作空间集合 $\boldsymbol{A}$ 为

$$\boldsymbol{A}=\mathrm{sign}(\sigma)=\{1,-1\},\tag{12}$$

式中 sign(·)为求取正负号函数.

收益函数 f_{R} 的设计较为复杂,主要原因在再入飞行存在难以量化的多任务目标特点. 常规的强化学习类问题往往只有一个优化目标,例如最短时间等;然而指定飞行时间的再入过程存在落点误差和飞行时间两个指标,并且需要尽可能地减少倾侧角符号的变化次数. 所以结合混合奖励函数设计方法[12],将 f_{R} 设置为

$$f_{\mathrm{R}}(s_t,a_t,s_{t+1})=\begin{cases}\varepsilon_{\mathrm{r}}-0.1(\Delta S+|\Delta t_{\mathrm{f}}|),s_{t+1}=s_f;\\ \varepsilon_{\mathrm{r}},s_{t+1}\neq s_f.\end{cases}$$

式中:$\Delta S=20\ (S_{\mathrm{error}}/100)^2$,其中 S_{error} 为终端横向距离误差;$\Delta t_{\mathrm{f}}=2\ (t_{\mathrm{error}}/10)^2$,其中 t_{error} 为最终飞行时间误差;倾侧角符号奖励函数 ε_{r} 为

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

ε_{r} 的设置引导决策网络在训练中更加倾向于以更少次数的符号变化来完成给定的飞行任务,提高横向制导律的稳定性与工程可行性. 由于飞行器的运动为确定性事件并存在终端截止条件,式(9)可简化为

$$J=J_{\pi^\theta}(s_0,s_{\mathrm{f}})=\sum_{t=0}^{n}f_{\mathrm{R}}(s_t,a_t,s_{t+1}).$$

3.3.3 阶段性迭代训练方法

在实际训练过程中,由于问题较为复杂且奖励函数稀疏,可行解范围狭窄,训练过程不稳定且收敛速度慢;并且由运动模型自举产生的训练数据集内部存在强关联性,降低了网络的稳定性与泛化能力. 所以为了改善网络的稳定性、训练速度和泛化能力,需要加入以下两项改进措施:

1)经验池和小批量梯度下降. 由于再入飞行的长周期性,通过在线自举产生的训练集破坏了样本之间的相互独立性要求,所以需要引入经验池和小批量梯度下降方法. 称每次自举产生的样本(s_t,a_t,R_t,s_{t+1})为一条经验,将其放入容量大小为 V_{MP} 的经验池中,在每次训练中随机选择其中的 V_{MB} 条经验组成训练数据集. 由于 $V_{\mathrm{MP}}\gg V_{\mathrm{MB}}$,因而每次训练中使用的绝大多数数据来源于不同的再入过程,大大地降低了数据间的关联性. 并且通过小批量梯度下降方法求取的加权平均梯度能够有效抑制数据自身扰动,使得训练过程更加平稳和减少资源占用率.

2)阶段性迭代训练方法. 从凸函数上的任意一点出发,沿着梯度下降的方向移动,必然能够到达全局最小值附近,但该结论在函数非凸的情况下并不成立. 强化学习类方法中所使用的多层神经网络恰恰是非凸函数,所以其理论上存在多个局部极小值,并且随着网络复杂性与层数的增加而增多,这是造成实际训练过程不稳定和发散的主要原因之一[11,13]. 同时,问题可行域的狭小也是造成以上现象的原因之一,并且难以通过简单地设置学习率进行解决. 针对以上问题,本文结合自适应思想和逐步缩小可行域方法,设计了一套阶段性迭代训练方法.

设网络初始学习率、探索率为 α_{L} 和 ε. 在 E_{d} 个训练周期后,每个周期的总收益 J 在一定范围内振荡或出现持续衰减,则说明在该学习率下,网络训练已到达饱和或存在陷入局部极值点的风险,需要减小相关训练参数. 定义 E_{d} 为阶段训练周期,并称此时网络完成了一个阶段的学习,保存此时网络参数并将学习率改为

$$\alpha'_{\mathrm{L}}=\delta_\alpha\alpha_{\mathrm{L}},$$

式中衰减率 $\delta_\alpha\in[0.01,0.10]$. 探索率重置为

$$\varepsilon'_0=\delta_\varepsilon\varepsilon_0.$$

式中:衰减率 $\delta_\varepsilon\in[0.25,1.00]$;$\varepsilon_0$ 为上一个阶段的初始探索率.

在新的学习率 α'_{L} 和探索率 ε'_0 基础上继续训练网络约 E_{d} 个周期,若每个周期的总收益 J 满足要求则结束训练,反之重复以上的训练参数衰减过程直至 J 符合要求或不再提高.

最后综合以上给出本文所设计的横向制导律基本流程,具体如图 6 所示. 其中离线训练模块(含阶段性迭代训练方法等)仅用于离线训练过程,并不出现在实际应用中. 通过离线训练得到一套较好的动作评价网络参数 $\boldsymbol{\pi}^v$,在实际的每个制导周期内,各状态量组成式(11)的状态向量输入到动作评价

网络之中. 该网络输出式(12)两个动作的价值，进而选择具有最大价值的动作作为倾侧角的符号输出，最后与 σ_{cmd} 共同作用于实际再入运动模型.

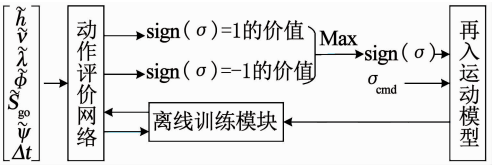

图 6　基于 DQN 的横向制导律基本流程

Fig. 6　Basic process of lateral guidance law based on DQN

4　仿真验证

4.1　网络参数设置

本文 DQN 中的动作评价网络采用如图 7 所示的全连接结构，输入层为式(11)的状态向量；两个隐藏层神经元节点数均为 500，采用 ReLU 激活函数；输出层为式(12)中两个动作的估计价值，采用线性激活函数. 在人为经验[14]和多次尝试的基础上，网络参数设定为：经验池大小 $V_{MP}=10^5$，每次抽取样本数 $V_{MB}=2^{12}$. 初始学习率 $\alpha_L=10^{-5}$，初始探索率 $\varepsilon=0.2$，衰减系数 $\eta_\varepsilon=0.999\ 97$；阶段性训练周期 $E_d=10^2$，衰减系数 $\delta_\alpha=10^{-2}$、$\delta_\varepsilon=0.5$；仿真步长 $\Delta T=1$ s. 由于飞行器在再入过程中的机动能力十分有限，时间调节能力较小，飞行任务(即飞行距离与再入时间之间的合理性)的设置较为严苛，所以为避免在训练过程中出现不合理的设置，本文再入任务的生成与可行性验证来源于文献[3]. 最后定义一次再入任务及其网络训练的完成为一个训练周期.

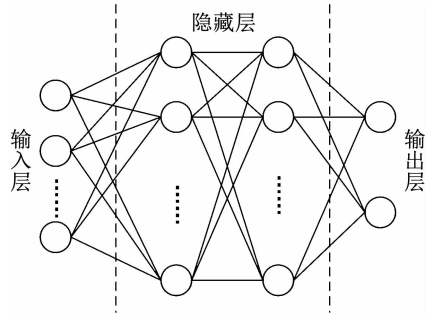

图 7　全连接 DQN 结构

Fig. 7　Structure of full-connected DQN

为验证阶段性迭代训练方法的有效性，在相同软硬件条件下进行 300 个训练周期的对比实验，每个周期总收益 J 的变化如图 8 所示.

图 8 表明在固定学习率下，所训练的网络大约在 170 个训练周期后出现总收益 J 急剧下降的现象，说明网络陷入局部极值陷阱. 虽然在 230 个训练周期后网络跳出了该陷阱，但受限于过大的学习率，网络参数接近饱和，J 在 -100 ~ -50 之间震荡. 与之相对的在加入了阶段性迭代训练方法后，网络能够成功地避开极值陷阱并一直保持较高的增长率. 这说明了本文所设计的阶段性迭代训练方法具有一定的有效性，其能够在一定程度上解决网络在长时间训练后陷入局部极值陷阱和过早饱和现象的发生.

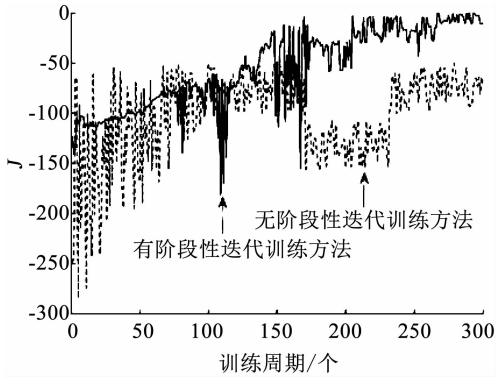

图 8　训练过程中总收益 J 的变化对比

Fig. 8　Comparison of total reward J during training

4.2　多任务标称环境仿真

为验证时间可控再入制导律的有效性和任务适应性，首先进行标称环境下的多任务指定再入时间仿真. 各再入任务初始点见表 1[10,15]，目标点的经、纬度坐标均为(215°，25°)，仿真结果如图 9 ~ 11 所示，统计结果见表 2.

表 1　再入任务初始点设置

Tab. 1　Setting of reentry missions' initial points

再入任务 i	经度 λ/(°)	纬度 φ/(°)	期望飞行时间 t_{need}/s
1	140	0	1 907
2	160	5	1 401
3	150	5	1 623
4	155	−5	1 657

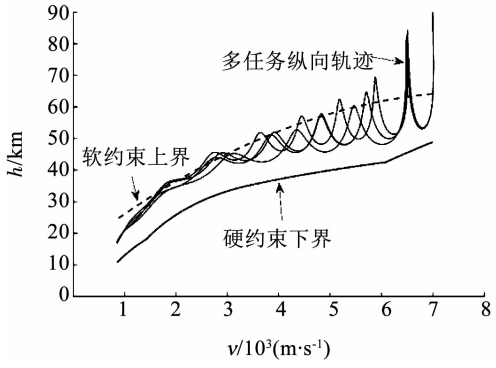

图 9　多任务纵向轨迹

Fig. 9　Longitudinal trajectory of multi-mission

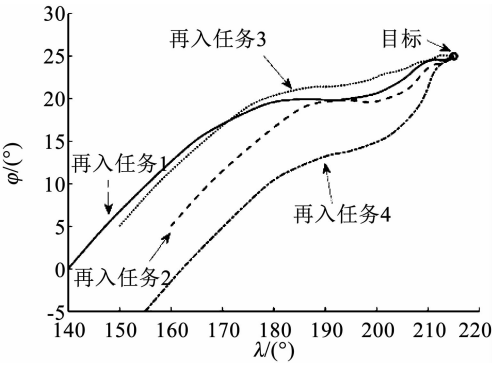

图 10 多任务横向轨迹

Fig. 10 Horizontal trajectory of multi-mission

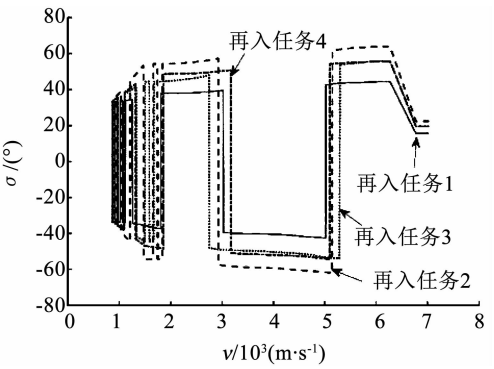

图 11 多任务倾侧角剖面

Fig. 11 Bank angle profile of multi-mission

表 2 多任务仿真结果

Tab. 2 Results of multi-mission simulation

再入任务 i	横向距离误差 S_{error}/km	飞行时间 t/s
1	1.81	1 906
2	0.98	1 399
3	1.51	1 625
4	1.32	1 661

由仿真结果可知,各任务横向飞行轨迹较为平滑,纵向轨迹均在 H-V 走廊之上,倾侧角符号的变化次数在 4 ~ 10 次之间,各模块之间的相互配合能够实现指定时间的再入飞行过程.

4.3 蒙特卡洛仿真

本文通过蒙特卡罗仿真进一步验证制导律的鲁棒性与稳定性. 仿真对象为任务 1,参数拉偏见表 3,拉偏方式均为正态分布. 随机选择 3 个可行时间并分别进行 500 次对比仿真,结果见表 4,其中"传统方法"代指文献[5]. 以期望时间为 1 891 s 为例,飞行轨迹分布如图 12、13 所示,飞行时间误差分布直方图如图 14 所示.

表 3 蒙特卡洛参数拉偏表

Tab. 3 Monte Carlo parameters deviation

拉偏参数	质量/kg	速度倾角/(°)	弹道倾角/(°)
最大偏差	40	0.2	2
拉偏参数	初始速度/(m·s^{-1})	初始高度/km	初始经、纬度/(°)
最大偏差	50	5	5
拉偏参数	升力系数/%	阻力系数/%	大气密度/%
最大偏差	10	10	10

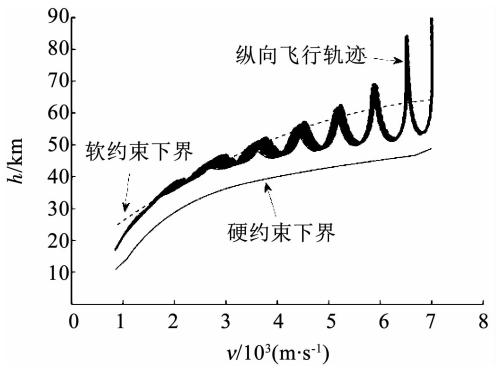

图 12 纵向轨迹分布

Fig. 12 Longitudinal trajectory distribution

表 4 蒙特卡洛仿真结果对比

Tab. 4 Comparison of Monte Carlo simulation results

期望飞行时间 t_{need}/s	平均横向误差 $\overline{S}_{error}$/km		平均终端高度误差 $\overline{\Delta h}$/km		平均时间误差 $\overline{t}_{error}$/s		时间误差方差 $D(t_{error})$/s^2	
	本文方法	传统方法	本文方法	传统方法	本文方法	传统方法	本文方法	传统方法
1 876	1.41	2.73	0.67	1.78	-1.14	3.97	2.98	27.85
1 891	1.17	1.94	0.81	1.22	-0.75	-1.12	2.69	34.43
1 918	1.32	3.14	0.78	1.91	-1.52	-4.89	3.11	31.39

由图 12、13 可知,在各类不确定性参数的干扰下,高超声速飞行器能够到达指定的终端状态并且全程没有违反过程约束条件. 表 4 说明在不同期望时间作用下,本文方法相对于传统方法而言,其具有较强的时间调节能力和控制能力. 除此之外,其制导精度和终端高度管理能力也优于传统方法. 图 14 表明最终飞行时间误差总体上呈正态分布趋势,并且在各类不确定性干扰下依然能够保证最终飞行时间误差在 ±10 s 以内,满足指定时间再入飞行的任务需求[3],具有较强的稳定性与鲁棒性.

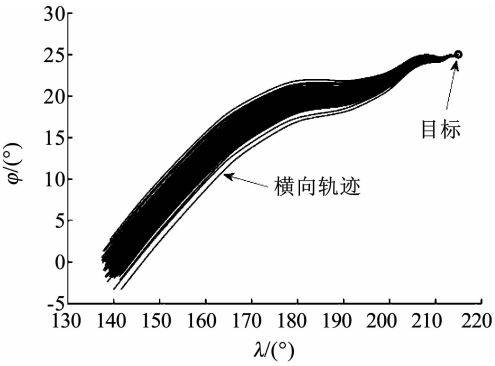

图 13　横向轨迹分布

Fig. 13　Horizontal trajectory distribution

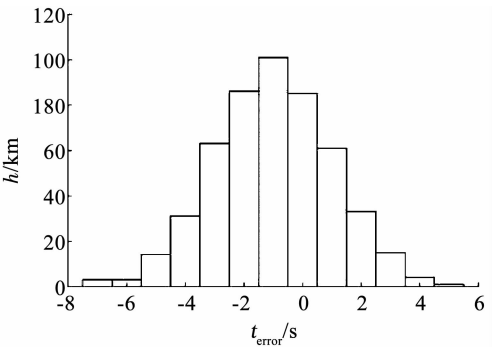

图 14　飞行时间误差直方图

Fig. 14　Histogram of flight time error

5　结　论

1)本文针对高超声速飞行器的再入飞行时间可控性要求,提出基于 DQN 的时间可控再入制导律.

2)由于再入飞行的横纵向解耦特性,飞行时间主要与倾侧角的符号规划——即横向制导律设计有关.但由于再入过程较为复杂,无法使用常规方法进行问题建模与分析,所以引入 DQN 进行制导律设计;同时针对网络训练过程中的收敛速度过慢与不稳定问题,设计了一套阶段性迭代训练方法.

3)基于 DQN 的时间可控再入制导律在线根据飞行状态进行倾侧角符号规划以调节再入时间,仿真验证了其具有良好的任务适应性、时间可控性、稳定性和鲁棒性.

参考文献

[1]黄志澄. 高超声速武器及其未来战争的影响[J]. 战术导弹技术, 2018(3): 1
HUANG Zhideng. Hypersonic weapons and its influence on future war[J]. Tactical Missile Technology, 2018(3): 1. DOI: 10.16358/j. issn. 1009 - 1300. 2018. 8. 501

[2]王少平, 董受全, 李晓阳, 等. 助推滑翔高超声速反舰导弹多方向协同突防可行性研究[J]. 指挥控制与仿真, 2017, 39(2): 55
WANG Shaoping, DONG Shouquan, LI Xiaoyang, et al. Feasibility study of multi-direction coordinated penetration of the boost-glide hypersonic anti-ship missile[J]. Command Control and Simulation, 2019, 39(2): 55. DOI: 10. 3969/j. issn. 1673 - 3819. 2017. 02. 012

[3]方科, 张庆振, 倪昆, 等. 高超声速飞行器时间协同再入制导[J]. 航空学报, 2018, 39(5): 202
FANG Ke, ZHANG Qingzhen, NI Kun, et al. Time-coordination reentry guidance law for hypersonic vehicle[J]. Acta Aeronautica et Astronautica Sinica, 2018, 39(5): 202. DOI: 10. 7527/S1000 - 6893. 2018. 21958

[4]SHEN Zuojun, LU Ping. Onboard generation of three-dimensional constrained entry trajectory[J]. Journal of Guidance, Control & Dynamics, 2003, 26(1): 111. DOI: 10. 2514/2. 5021

[5]LU Ping. Entry guidance: A unified method[J]. Journal of Guidance Control & Dynamics, 2014, 37(3): 713. DOI: 10. 2514/1. 62605

[6]CHENG Lin, WANG Zhenbo, CHENG Yang, et al. Multi constrained predictor-corrector reentry guidance for hypersonic vehicles[J]. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2017, 232(16): 0954410017724185. DOI: 10. 1177/0954410017724185

[7]JEON I S, LEE J I, TAHK M J. Impact-time-control guidance law for anti-ship missiles[J]. IEEE Transactions on Control Systems Technology, 2006, 14(2): 26. DOI: 10. 1109/tcst. 2005. 863655

[8]周锐, 陈宗基. 强化学习在导弹制导中的应用[J]. 控制理论与应用, 2001, 18(5): 748
ZHOU Rui, CHEN Zongji. Application of reinforcement learning in missile guidance[J]. Control Theory and Applications, 2001, 18(5): 748. DOI: 10. 3969/j. issn. 1000 - 8152. 2001. 05. 023

[9]JIANG X, FURFARO R, LI S. Integrated guidance for Mars entry and powered descent using reinforcement learning and Gauss pseudospectral method[EB/OL]. (2018-5-23). https://www.researchgate. net/publication/324982576_ Integrated_Guidance_for_Mars_Entry_and_ Powered_Descent_Using_Reinforcement_Learning_and_Gauss_Pseudospectral_Method

[10]王光伦. 高超声速飞行器再入段预测校正制导研究[D]. 哈尔滨: 哈尔滨工业大学, 2010
WANG Guanglun. Predictor-corrector reentry guidance for hypersonic vehicles[D]. Harbin: Harbin Institute of Technology, 2010

[11]SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. Cambridge, MA: MIT Press, 2005

[12]SEIJEN H V, FATEMI M, ROMOFF J, et al. Hybrid reward architecture for reinforcement learning[EB/OL]. (2017-06-13)[2017-11-28]. http://cn. arxiv. org/abs/1706. 04208

[13]MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB/OL]. (2013-12-19). https://arxiv. org/abs/1312. 5602

[14]CASTANO A P. Practical artificial intelligence[M]. Berkeley, CA: Apress, 2018

[15]PHILLIPS T H. A common aero vehicle (CAV) model, description, and employment guide[R]. Albuqerque: New Mexico Schafer Corp-oration for AFRL and AFSPC, 2003

(编辑　张　红)