doi:10.11887/j.cn.202503006 http://journal.nudt.edu.cn

针对运载火箭上升段考虑大风区减载的智能姿态控制方法

周 首,杨 豪,张士峰*,白锡斌,王 峰
(国防科技大学 空天科学学院,湖南 长沙 410073)

摘 要:针对运载火箭在上升段遭遇大风区的减载需求,提出自适应学习率的智能姿态控制方法。以运载火箭为研究对象,建立了其俯仰平面的动力学模型。基于柔性动作-评价构建了适用于运载火箭上升段飞行控制的深度强化学习框架,设计了一种综合考虑姿态跟踪精度和稳定性以及减载效果的奖励函数。在此基础上,基于步长学习率调度器实现了学习率自适应迭代,以期在快速提升控制器收敛性的基础上找到最优解。并设计了一种早停机制实现了训练过程的自动停止,以提升训练效率。仿真结果表明,所提出的方法在保证姿态跟踪精度和稳定性的前提下能够有效实现运载火箭的减载效果,并且对随机阵风干扰具有较强的鲁棒性和适应能力。

关键词:运载火箭;减载;姿态控制;深度强化学习;自适应学习率;阵风干扰
中图分类号:V249.1 **文献标志码**:A **文章编号**:1001-2486(2025)03-051-13

论文拓展

Intelligent attitude control method of launch vehicle during ascending phase considering load reduction in high wind zone

ZHOU Shou, YANG Hao, ZHANG Shifeng, BAI Xibin, WANG Feng*
(College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: To address the aerodynamic load reduction requirement when the launch vehicle flying in high wind zone during the ascending phase, an intelligent attitude control method with adaptive learning rate was proposed. Taking a certain type of launch vehicle as the research object, the dynamic model in the pitch plane was established. A deep reinforcement learning framework suitable for flight control of the launch vehicle during the ascending phase was developed based on soft actor-critic, and a reward function that comprehensively considers attitude tacking accuracy and stability, and load reduction effectiveness was designed. On this basis, an adaptive iteration of learning rate was implemented based on a step-size learning rate scheduler to quickly improve the convergence velocity and find the optimal solution of the controller. Besides, an early stopping mechanism which can automatically end the training process was designed to enhance the training efficiency. Simulations show that the proposed method can effectively achieve load reduction of the launch vehicle while ensuring attitude tracking accuracy and stability. Additionally, it has strong robustness and adaptability to random wind disturbance.

Keywords: launch vehicle; load reduction; attitude control; deep reinforcement learning; adaptive learning rate; wind disturbance

众所周知,运载火箭在上升段往往要经过稠密大气,受到大风区的影响会形成较大的气动攻角,从而形成较大的附加弯矩,这将严重影响飞行过程中的安全性和稳定性,这对运载火箭上升段飞行而言是非常不利的。因此,面对较大的阵风干扰,如何实现有效的减载显得尤为重要。运载火箭的减载技术通常是使用弹道修正或主动控制的方法,降低其在经过大风区时的气动载荷,减小其所带来的附加弯矩,有效提升箭体结构强度的可靠性并提高运载能力[1]。

总的来说,运载火箭的减载控制大致可分为两类:弹道修正补偿和主动减载控制。弹道修正补偿需要在发射前获得准确的风场信息,将其引入控制系统,离线补偿风攻角带来的影响。文献[2]以德尔塔运载火箭为研究对象设计了一种精确风载模型,提高了减载精度。但此类方法严

收稿日期:2024-12-13
基金项目:国家自然科学基金资助项目(U21B2028)
第一作者:周首(1995—),男,北京人,博士研究生,E-mail:zhoushou163@163.com
*通信作者:张士峰(1971—),男,河南新乡人,教授,博士,博士生导师,E-mail:zhang_shifeng@hotmail.com
引用格式:周首,杨豪,张士峰,等. 针对运载火箭上升段考虑大风区减载的智能姿态控制方法[J]. 国防科技大学学报,2025,47(3):51-63.
Citation: ZHOU S, YANG H, ZHANG S F, et al. Intelligent attitude control method of launch vehicle during ascending phase considering load reduction in high wind zone[J]. Journal of National University of Defense Technology, 2025, 47(3): 51-63.

重依赖于所装订的风场信息的精度,对随机阵风干扰的适应性和泛化性明显不足。因此,主动减载技术成为当前更为普遍的方法。主动减载技术是指通过实时测量或计算攻角信息,在控制系统设计时引入减载控制补偿,使得火箭在遭遇高空风时能够朝着来流方向飞行,从而减小风过载。文献[3]介绍了被动减载与主动减载机理,并给出了主动减载控制法的建模与仿真分析。文献[4]对攻角估算反馈控制与加速度反馈控制的减载效果进行了对比分析,发现加速度反馈控制由于采用了自抗扰技术,具备更强的抗干扰能力,因此在当前的减载控制方法中更为常用。文献[5]则提出了一种创新的测量方法,该方法首先通过信号辨识技术确定箭体绕质心的角加速度以及惯性测量单元相对于质心的位置,随后进一步计算得到去除了因箭体旋转而产生的线加速度影响的箭体质心处的视加速度,这一方法满足了主动减载系统对精确测量信号的需求。文献[6]研究了一种自适应姿态开环减载技术,该技术集成了加速度计和速率陀螺的实时监测,通过姿态开环控制策略,精确追踪零法向和横向加速度指令,从而调整火箭姿态以对准来流方向,有效减少气动载荷。上述主动减载控制均是传统基于过载反馈的在线减载方法,该方法面临的最大问题是跟踪精度、稳定性和减载之间的平衡问题。基于传统的主动减载控制方法,减载与跟踪这一对"矛盾体"通常与所设计的控制参数紧密相关,而面对不同的随机风型如何实现广域的适用、平衡是目前有待解决的一大难题。因此,急需探索一个能够在不确定性阵风干扰的场景下有效实现主动减载目标的控制方法。

近年来,强化学习的飞速发展为解决航天飞行器实际飞行过程中所遇到的不确定性因素提供了新的突破口[7],通过智能控制技术赋能,增强飞行器的主动适应及自主决策能力。目前,强化学习算法在飞行控制领域已得到较为广泛的应用。文献[8]提出了一种基于强化学习的误差卷积输入神经网络用于设计混合式无人机控制系统。文献[9]针对非合作目标抓捕时组合体姿态的稳定问题,利用强化学习技术对组合体的参数进行在线识别,实现卫星姿态的重新稳定。文献[10]提出了一种基于模仿强化学习的固定翼飞机姿态控制方法。文献[11]对其进行了改进,能够实现不同初始条件下飞机姿态角的快速响应。

以上研究表明,强化学习技术对外界干扰等不确定性因素下的航天器控制问题具有一定的适应性,但大多数特征状态需要人工设定,在面对高维数据所表示的复杂环境时,难以找到合适的特征表达方法,容易陷入维数灾难问题,而且传统的强化学习有着一定局限性,其动作空间和状态空间大多都是离散的,然而实际的飞行器控制中,状态空间和动作空间都是连续的[12]。因此,由强化学习定义任务的模型目标及优化的方向,深度学习给出表征问题以及解决问题的方式,能够更好地解决以上问题[13]。截至目前,深度强化学习在导弹的制导[14]和控制[15]、无人机的轨迹跟踪[16]、深空探测器的自主导航[17]等领域得到了广泛应用。对于纯航天器姿态控制方面的研究,文献[18]和文献[19]分别基于邻近策略优化(proximal policy optimization, PPO)和双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)两种深度强化学习算法设计了一套针对卫星的自适应连续姿态控制方法,实现了在轨卫星姿态的自主可控。

以上研究表明,深度强化学习对外界干扰等不确定性因素下的航天器导航、制导以及控制问题具有较强的适应性,但大多都是预先设定好学习率以后对智能体进行训练。学习率的大小决定了训练的收敛速度、性能以及稳定性:较高的学习率可以使训练过程更快地收敛,然而过高的学习率可能会导致训练过程中错过最优解;较低的学习率可能更容易找到精确的最优解,然而过低的学习率可能导致训练过程在有限时间内无法收敛。因此,找到合理的学习率对控制器的训练至关重要。然而,上述根据人工经验调节学习率的方法既耗时耗力,又无法适应复杂多变的训练环境。Dias 等[20]针对飞行器的容错控制问题,提出了一种基于监督器触发的在线自适应学习率的控制方法。文献[21]则提出了一种基于梯度下降法的在线调整强化学习动作网络学习率的算法。

受以上研究的启发,针对运载火箭上升段遇大风区的减载控制问题,本研究基于深度强化学习框架,设计了一种融合学习率自适应策略和早停机制的智能姿态控制方法。首先,构建了包含风场的运载火箭姿态动力学模型,而后将运载火箭姿态控制问题描述为马尔可夫决策过程。其中,创新地提出了一种多目标协同的奖励函数,综合考虑跟踪精度、稳定性、减载效果以及训练效率等因素,在标准柔性动作-评价(soft actor-critic, SAC)框架基础上进行了改进,最终设计出一种具备在线学习能力的智能减载姿态控制器。设计采

用基于步长学习率的自适应调节策略,动态调整网络参数的更新步长,有效提升了算法在复杂风扰下的训练稳定性。同时引入早停机制对训练过程中的收敛性能进行动态评估,当控制器性能连续迭代一定周期还未呈现显著提升时自动终止训练。为验证方法的有效性,研究通过数学仿真实验对比分析了所提方法与基于在线反馈的传统减载控制方法在解决跟踪与减载之间的平衡问题上的性能差异。

1　运载火箭动力学建模

根据文献[22]建立运载火箭的动力学模型。该型火箭是一款重型且细长的载人航天器,之所以选择该型火箭作为研究对象是因为其细长体的结构在上升段经历大风区时如形成较大的气动弯矩极容易造成空气动力学不稳定的情况,从而在设计控制器时必须考虑减载需求,这对于本研究具有重要的参考意义。这里,为了方便控制器的设计,可以作出如下假设。

假设 1:不考虑运载火箭的弹性振动模型。

假设 2:将地球看作均质圆球,忽略自转。

假设 3:只考虑火箭在俯仰平面内的运动。

假设 4:只考虑风场的风速切变,风向恒定。

1.1　坐标系转换

通常,在描述运载火箭运动的时候,用到的坐标系主要是发射惯性系、箭体坐标系和速度坐标系,上述三种坐标系可以分别用下标 I、B、V 表示。由于作用在运载火箭上的推力及其力矩和气动力及其力矩是由不同的原因产生的,在建模过程中合理地选择坐标系来分析其受力和所受力矩有助于后续控制算法的设计。下面将给出三个坐标系之间具体的转换关系。

这里需要定义如下的坐标转换矩阵:

$$\boldsymbol{R}=\begin{bmatrix}\cos\psi & -\sin\psi\\ \sin\psi & \cos\psi\end{bmatrix} \tag{1}$$

式中,ψ 代表二维平面内一个坐标系变换到另外一个坐标系的旋转角,定义逆时针旋转为正。

那么,发射惯性系到箭体坐标系的转换矩阵可以表示为:

$$\boldsymbol{C}_{\mathrm{B/I}}=-\boldsymbol{C}_{\mathrm{I/B}}=\begin{bmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{bmatrix} \tag{2}$$

同理,速度坐标系到箭体坐标系的转化矩阵可以表示为:

$$\boldsymbol{C}_{\mathrm{B/V}}=-\boldsymbol{C}_{\mathrm{V/B}}=\begin{bmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{bmatrix} \tag{3}$$

式中,φ 和 α 分别代表火箭的俯仰角和攻角。

这里需要注意的是,在后续的研究中,为了方便控制器设计,将选择箭体坐标系来描述运载火箭受到的气动力和气动力矩以及推力和推力矩。

1.2　动力学模型

运载火箭的动力学方程由质心运动方程和绕质心运动方程组成。运载火箭的质心运动方程可简单地由牛顿第二定律得出,即:

$$m\dot{\boldsymbol{V}}=\boldsymbol{F} \tag{4}$$

式中,m 是火箭的质量,$\dot{\boldsymbol{V}}$ 是火箭的加速度矢量,$\boldsymbol{F}$ 是作用在火箭质心上的合外力。

另,参考刚体绕质心运动的欧拉方程:

$$\dot{\boldsymbol{H}}=\boldsymbol{M} \tag{5}$$

式中,$\boldsymbol{H}$ 是刚体角动量矢量,$\boldsymbol{M}$ 是作用在刚体质心上的合外力矩。

角动量矢量 $\boldsymbol{H}$ 可表示为:

$$\boldsymbol{H}=\hat{\boldsymbol{I}}\cdot\boldsymbol{\omega} \tag{6}$$

式中,$\boldsymbol{\omega}$ 是刚体角速度矢量,$\hat{\boldsymbol{I}}$ 是刚体关于质心的惯性张量。

这里把运载火箭考虑为一个刚体,那么其绕质心运动方程可以表示为:

$$\hat{\boldsymbol{I}}\cdot\dot{\boldsymbol{\omega}}+\boldsymbol{\omega}\times\hat{\boldsymbol{I}}\cdot\boldsymbol{\omega}=\boldsymbol{M} \tag{7}$$

式中,$\dot{\boldsymbol{\omega}}$ 是火箭的角加速度矢量。

1.2.1　气动力及气动力矩

气动力和气动力矩取决于火箭相对于周围空气的速度,称为来流速度。针对运载火箭上升段遇大风区的实际情况,这里考虑随机阵风作为外界干扰因素,由此可以得到惯性坐标系下的来流速度矢量:

$$\boldsymbol{V}_{\mathrm{m}}=\boldsymbol{V}-\boldsymbol{V}_{\mathrm{w}} \tag{8}$$

式中,$\boldsymbol{V}$ 是运载火箭的惯性速度矢量,$\boldsymbol{V}_{\mathrm{w}}$ 是局部阵风的扰动风速矢量。

为了得到箭体坐标系下的来流速度,需要用到式(2)表示的转换矩阵,即 $\boldsymbol{V}_{\mathrm{mb}}=\boldsymbol{C}_{\mathrm{B/I}}\cdot\boldsymbol{V}_{\mathrm{m}}$。

下面分析气动力。气动力在速度系当中的两个轴向分量可以分别表示为:

$$\begin{cases}F_{\mathrm{aero},x\mathrm{V}}=F_{\mathrm{base}}-C_{\mathrm{A}}Sq=-D\\ F_{\mathrm{aero},z\mathrm{V}}=-C_{\mathrm{N}\alpha}Sq\alpha=-N\end{cases} \tag{9}$$

其中,基准力 F_{base} 是高度 h 的函数,气动力系数 C_{A} 和 $C_{\mathrm{N}\alpha}$ 是马赫数 M 的函数,均可以通过插值得到。S 代表火箭的参考横截面积,D 和 N 分别代表阻力和升力。动压 q、马赫数 M 以及攻角 α 可表示为:

$$M=\frac{V_{\mathrm{m}}}{a} \tag{10}$$

$$q = \frac{1}{2}\rho V_{\mathrm{m}}^2 \tag{11}$$

$$\alpha = \arctan\frac{V_{\mathrm{m,zb}}}{V_{\mathrm{m,xb}}} \tag{12}$$

其中，$V_{\mathrm{m,xb}}$ 和 $V_{\mathrm{m,zb}}$ 是 $\boldsymbol{V}_{\mathrm{mb}}$ 在箭体坐标系两个轴向上的分量，声速大小 a 和空气密度 ρ 是高度 h 的函数，均可以通过插值得到。

因此，在箭体坐标系下的气动力可表示为：

$$\begin{aligned}\boldsymbol{F}_{\mathrm{aero,b}} &= \begin{bmatrix} F_{\mathrm{aero,xb}} \\ F_{\mathrm{aero,zb}} \end{bmatrix} = \boldsymbol{C}_{\mathrm{B/V}} \cdot \boldsymbol{F}_{\mathrm{aero,V}} \\ &= \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} F_{\mathrm{aero},x\mathrm{V}} \\ F_{\mathrm{aero},z\mathrm{V}} \end{bmatrix}\end{aligned} \tag{13}$$

那么在箭体坐标系下，作用在火箭质心上的气动力矩大小可表示为：

$$M_{\mathrm{aero,b}} = C_{Mp\alpha} q \alpha S b + F_{\mathrm{aero,zb}} \cdot x_{\mathrm{a}} \tag{14}$$

式中，x_{a} 是气动参考点相对火箭质心位置的距离，b 是火箭的参考长度。另外，气动力矩系数 $C_{Mp\alpha}$ 是马赫数 M 的函数，也可以通过插值得到。

1.2.2 推力及推力矩

根据图1，可将火箭的推力大小简单概括为：

$$T = T_0 + (p_{\mathrm{e}} - p_0) A_{\mathrm{e}} \tag{15}$$

式中，T 是总推力，$T_0 = |\dot{m}| V_{\mathrm{e}}$ 是射流推力，$\dot{m}$ 是推进剂质量流率，V_{e} 是出口射流速度，p_{e} 是喷嘴出口压强，p_0 是局部大气压强，它也是高度 h 的函数，可以通过插值得到，A_{e} 是喷嘴出口面积。

火箭的质量为 $m = m_0 - |\dot{m}| t$，其中，m_0 为火箭的初始质量。如果将大气上方真空中的推力称为 T_∞，那么大气中任何高度的推力大小可表示为：

$$T = T_\infty - p_0 A_{\mathrm{e}} \tag{16}$$

式中，$T_\infty = T_0 + p_{\mathrm{e}} A_{\mathrm{e}}$，可以通过插值得到。

若发动机摆动喷角 δ 满足 $|\delta| \leqslant 10°$，则在箭体坐标系下的推力矢量可表示为：

$$\boldsymbol{F}_{\mathrm{rkt,b}} = \begin{bmatrix} F_{\mathrm{rkt,xb}} \\ F_{\mathrm{rkt,zb}} \end{bmatrix} = \begin{bmatrix} T\cos\delta \\ T\sin\delta \end{bmatrix} \tag{17}$$

其所产生的推力矩大小为：

$$M_{\mathrm{rkt,b}} = F_{\mathrm{rkt,zb}} \cdot x_{\mathrm{g}} \tag{18}$$

式中，x_{g} 是推力作用点相对火箭质心位置的距离。

根据以上推导，该型号运载火箭俯仰通道姿态控制动力学方程由式(7)推导可得：

$$I \cdot \frac{\mathrm{d}\omega}{\mathrm{d}t} = M_{\mathrm{aero}} + M_{\mathrm{rkt}} \tag{19}$$

式中，ω 为箭体俯仰角速度大小，I 为箭体的转动惯量大小。

根据假设2可知，地球为一均质圆球，不考虑非正球体所涉及的J4二阶模型。因此，运载火箭所受到的引力加速度 $\boldsymbol{g}$ 恒定，方向指向地心，所以火箭的引力可以表达为 $\boldsymbol{G} = m\boldsymbol{g}$。

根据以上分析及推导，可以得到简化的运载火箭纵平面受力分析图，如图1所示。

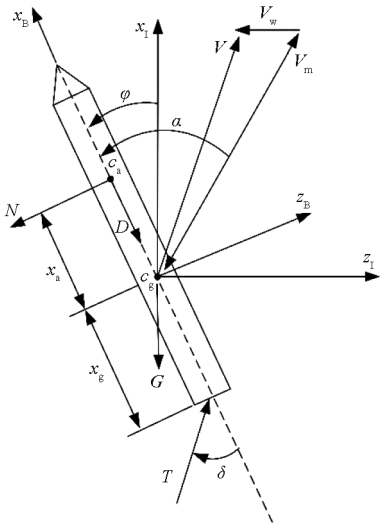


图1 运载火箭受力分析

Fig. 1 Force analysis of launch vehicle

图1中，(x,z) 代表的是垂直惯性坐标系，x 轴为垂直参考轴，z 轴为水平参考轴，$(x_{\mathrm{b}},z_{\mathrm{b}})$ 代表的是箭体坐标系，c_{a} 代表气动力作用点，而 c_{g} 代表火箭质心位置。

2 基于SAC的运载火箭姿态控制方法

2.1 深度强化学习理论基础

对于强化学习而言，智能体与环境是十分关键的两个要素，强化学习的核心机制就是通过智能体与环境不断交互、积累经验、更新策略，从而最终训练得到一个最优策略。

强化学习的本质其实是解决一个马尔可夫决策过程(Markov decision process, MDP)，该过程通常由一个四元组(S,A,P,R)组成。其中，S 代表状态空间，A 代表动作空间，P 代表状态转移概率，R 代表奖励函数。在训练过程中，智能体根据当前环境所处状态 $s_t \in S$，采取动作 $a_t \in A$，使得环境依概率 P 由状态 s_t 转移到 s_{t+1}，同时得到一个奖励值 $r_t \in R$。以上即为强化学习的基本原理，该原理如图2所示。

上述过程有两个非常重要的环节，一个是决策动作的策略 Π，以及对策略进行评价的函数 Q。训练过程中的任意时刻 t，智能体通过策略 Π 进行决策，根据环境所处的状态 s_t 输出动作 a_t。对智能体进行训练的目标是找到一个使长期累积的奖励值最大的策略 Π^*，即最优策略。这里定义 t 时刻所能带来的累积奖励为：

$$R_t = r_{t+1} + \gamma r_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (20)$$

式中,γ 表示折扣因子,满足 $\gamma \in [0,1]$。

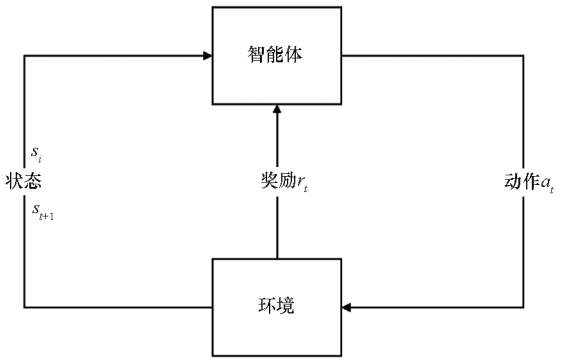


图 2　强化学习基本原理图

Fig. 2　Basic principle diagram of reinforcement learning

动作 - 评价(Actor-Critic,AC)是一种强化学习架构,该架构通过 Actor 和 Critic 两种神经网络分别逼近策略与价值函数[7]。AC 架构可基于浅层或深度网络实现,但基于此前的描述,由强化学习定义任务的模型目标,深度学习给出表征问题的方式,结合二者特点能够更好地解决连续控制问题,因此,本研究通过深度神经网络来实现 AC 框架。其中,Actor 网络负责决策,它会根据当前的状态决定每一步应该采取哪些动作,对于确定性策略,则直接输出具体的动作值,而对于随机性策略,则输出一个动作的概率分布,随后从这个分布中采样得到实际执行的动作。Critic 网络则负责评估,它会根据当前的状态和执行上述动作后获得的奖励值来预测未来可能累积的总奖励,该网络输出的是对当前状态 - 动作对的 Q 值估计,这个值反映了从当前状态开始,按照当前策略执行动作所能获得的长期累积奖励的预期。这里,Actor 网络和 Critic 网络的参数分别用 θ 和 ϕ 表示,在后续智能体的训练过程中,正是通过不断更新这两个网络参数来实现更高效的学习和控制性能[23]。

2.2　基于 SAC 算法的深度强化学习框架

SAC 算法是一种适用于连续动作空间的基于随机性策略的算法,该算法通过离策略的方法对随机策略进行优化,搭建起了随机性策略算法与深度确定性策略梯度(deep deterministic policy gradient, DDPG)和 TD3 等确定性策略算法之间的桥梁。该算法的中心特征是熵正则,目的是通过增加熵来增加策略的探索性,从而加速学习,其优化目标则是期望长期累积的奖励值与信息熵之间达到折中。因此,将信息熵融入式(20)所表示的长期累积奖励可得:

$$\hat{R}_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} + \xi H(\Pi[\cdot \mid s_t)] \quad (21)$$

式中:ξ 为熵权重系数,决定了信息熵相对于奖励的重要性,从而控制策略的随机程度;$H[\Pi(\cdot \mid s_t)]$代表信息熵,可用于度量策略的随机性,其可定义为所有动作的平均不确定性。

$$H[\Pi(\cdot \mid s_t)] = E[-\lg P(a_t \mid s_t)] = -\sum_{a_t \in A} P(a_t \mid s_t) \lg P(a_t \mid s_t) \quad (22)$$

这里需要说明的是,动作的生成是一个熵减的过程,因此策略的信息熵总是负定的。追求信息熵最大化意味着让各动作选择概率均衡,避免集中,以增强策略的探索性和鲁棒性,从而提高控制系统的稳定性和适应性。在熵正则强化学习中,智能体在每一时刻将获得正比于信息熵的额外奖励,因此,最优策略可以表示为:

$$\Pi^* = \arg\max_{\Pi} E\left(\sum_{t=0}^{\infty} \gamma^t \{ r + \alpha H[\Pi(\cdot \mid s_t)] \} \right) \quad (23)$$

式(23)表明,在状态 s_t 下选择的动作越多则能够获取的总奖励值越大。

这里,将 SAC 算法具体的训练流程总结归纳如下。

步骤 1:初始化 Actor 和 Critic 当前网络的参数 θ、ϕ_1 和 ϕ_2,并清空经验缓存。

步骤 2:将 Critic 当前网络的参数拷给对应的目标网络 $\phi_{\text{targ},1} \leftarrow \phi_1$ 和 $\phi_{\text{targ},2} \leftarrow \phi_2$。

步骤 3:重复以下步骤直到全部训练完成或者达到预期奖励值:

1)获得环境的初始状态 s_0 并根据随机策略选择一个动作 $a_t \sim \Pi(\cdot \mid s)$;

2)对训练环境执行动作 a_t,获得奖励 r_t 并根据状态转移函数 $P(s_{t+1} \mid s_t, a_t)$ 获得下一时刻的状态 s_{t+1};

3)将四元组 (s_t, a_t, r_t, s_{t+1}) 存入经验缓存并对其信息进行更新;

4)判断 s_{t+1} 是否达到终止状态,若达到则重置环境状态回到 1),若未达到则继续下面的步骤;

5)从经验缓存中,随机采样 n 组数据组成数据集 $N = \{(s_1, a_1, r_1, s_2), (s_2, a_2, r_2, s_3), \cdots, (s_t, a_t, r_t, s_{t+1})\}$,将其作为当前网络的训练数据集;

6)更新 Critic 网络参数,即

$$\phi_i \leftarrow \phi_i - \lambda_Q \nabla_{\phi_i} J_Q(\phi_i) \quad i = 1,2 \quad (24)$$

7)更新 Actor 网络参数,即

$$\theta \leftarrow \theta - \lambda_\Pi \nabla_\theta J_\Pi(\theta) \quad (25)$$

8)更新熵权重系数,即

$$\xi \leftarrow \xi - \lambda \nabla_{\xi} J(\xi) \tag{26}$$

9)更新目标网络的参数,即

$$\phi_{\text{targ},i} \leftarrow \sigma \phi_{\text{targ},i} + (1-\sigma)\phi_i \quad i=1,2 \tag{27}$$

步骤 4:输出最优网络参数 θ^*,ϕ_1^*,ϕ_2^*。

其中:$J_Q(\phi_i)$,$J_\Pi(\theta)$,$J(\xi)$分别代表 Critic 网络、Actor 网络和熵权重系数调节的代价函数;η 代表学习率;σ 代表更新系数。

此前很多研究中将熵权重系数 ξ 看作一个定值,限制了策略探索的随机性,根据式(26)可知,本研究将其考虑为一个迭代更新的变量,对其实现动态调整,平衡了探索与利用,这样做可以通过增加策略的探索性,有效避免局部最优的问题,促进全局收敛。

综上,SAC 算法的训练流程如图 3 所示。

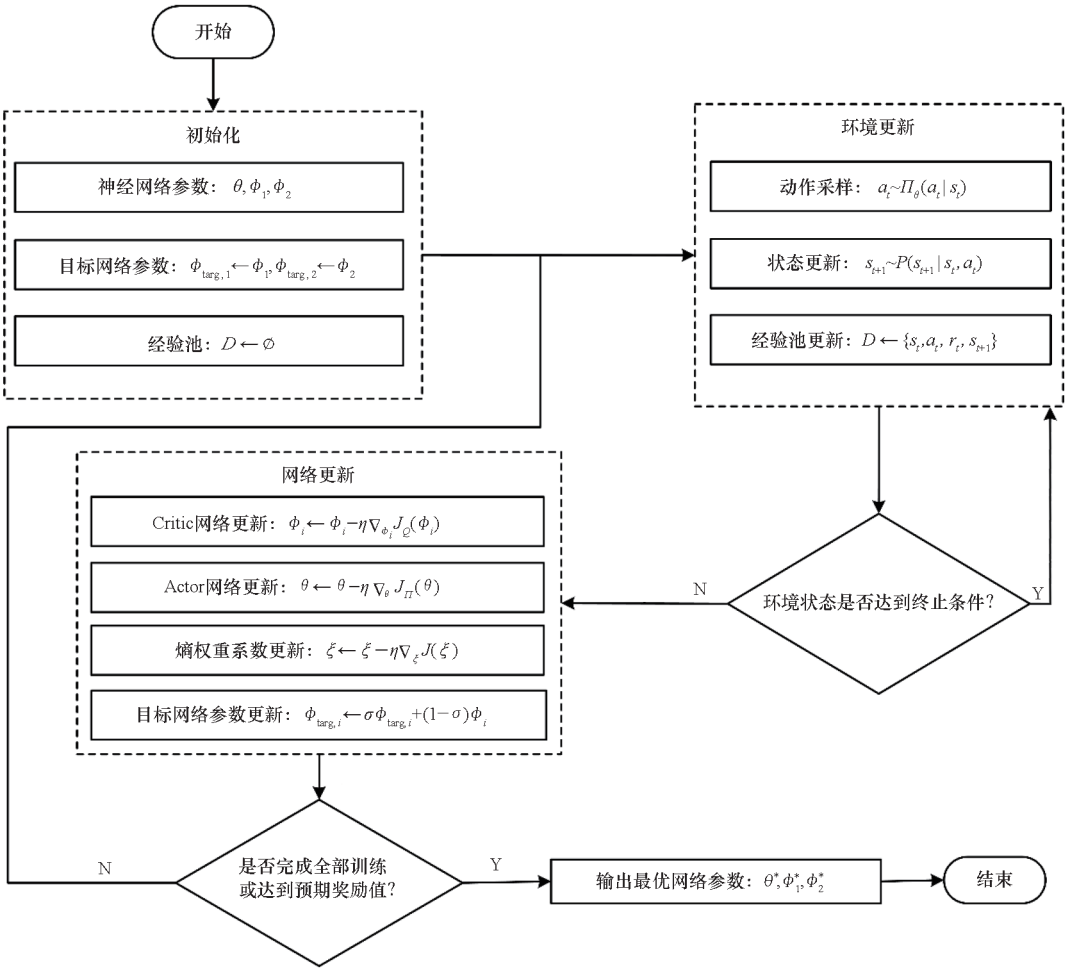


图 3　SAC 算法的训练流程图

Fig. 3　Training flow chart of the SAC algorithm

2.3　考虑大风区减载的姿态控制器设计

本研究拟采用此前构建的基于 SAC 算法的深度强化学习框架来设计针对运载火箭上升段考虑大风区减载的姿态控制器,值得一提的是本研究在原本的框架基础上引入了学习率自适应策略以及能够实现自动终止训练的早停机制,因此称之为改进后的 SAC 算法。由于本研究仅考虑火箭俯仰平面内的姿态动力学模型,因此设计控制器时仅考虑俯仰通道的姿态控制,其他两个通道的姿态控制器可以参考本研究的设计思路。

2.3.1　状态空间选择

由式(19)所表征的运载火箭姿态动力学模型可知,与火箭姿态控制系统相关的状态参数包括转动惯量 I、气动力矩 M_{aero}、推进力矩 M_{rkt}、俯仰角速度 ω、俯仰角 φ、结构参数 x_{a} 和 x_{g} 等。其他运动参数如位置、速度、加速度等均不会直接作用于姿态控制系统,遂不予考虑。上述直接作用于姿态控制系统的各运动参数中,俯仰角速度 ω 可通过箭上陀螺仪敏感得到,俯仰角 φ 可通过姿态角速度积分得到,其余运动参数均无法通过测量获取,故不能作为状态量;简体结构参数 x_{a}、x_{g} 在飞行过程中可认为是常量,对同一对象在飞行过程中不会改变,也可通过神经网络中的偏置项进行表达,故不选作状态量。针对箭体经历风过载的情况,通常使用参数 $|q\alpha|$ 来表征箭体所受的气动载荷,因此这里将该参数作为本研究在减载上

应用的状态输入,具体的实现方法采用的是风攻角的在线辨识。这一方法在工程中,通常使用大气数据测量装置[24]获得准确的来流信息,从而得到真实攻角,或者基于箭上设备如加速度表、陀螺等测量装置,根据已有信息实时计算攻角[1]。针对攻角在线辨识的研究现已有大量研究成果作为支撑,该部分内容不作为本研究的重点在此进行讨论。根据上述分析,结合控制系统的姿态跟踪目标,选择状态空间:

$$s_t = (\Delta\varphi_t, \omega, |q\alpha|) \tag{28}$$

式中,$\Delta\varphi_t$ 为俯仰角跟踪误差。

$$\Delta\varphi_t = \varphi_t - \varphi_c \tag{29}$$

其中,φ_t 为 t 时刻运载火箭的俯仰角,φ_c 为程序俯仰角指令。

2.3.2　动作空间选择

本研究中,运载火箭的姿态变化是通过固体火箭发动机改变摆动喷角来实现的,很显然,这是连续动作空间。根据此前的描述,SAC 算法是基于随机性策略的深度强化学习算法,输出层输出的是动作的概率分布,根据其分布来选择动作量。因此,在智能体的训练中,直接选择发动机的等效摆角作为智能体的动作量,即

$$a_t = \delta_z \tag{30}$$

2.3.3　奖励函数设计

奖励函数的设计是解决深度强化学习问题的关键因素,直接关系到算法能否收敛到最优解。此处设计的奖励函数包含 4 部分,分别为跟踪精度奖励、跟踪稳定性奖励、飞行时长奖励和减载效果奖励,下面将给出每一种奖励的形式。

1)跟踪精度奖励。这里,根据运载火箭对程序俯仰角指令的跟踪精度做出奖励:

$$r_1 = \frac{\Delta\varphi_{\max} - |\Delta\varphi_t|}{\Delta\varphi_{\max}} - \tau \tag{31}$$

式中,$\Delta\varphi_{\max}$ 表示最大跟踪误差,$\tau \in (0,1)$ 为控制参数。不难看出,式中加入最大跟踪误差的限制后,智能体的训练过程中当跟踪误差 $\Delta\varphi_t > \Delta\varphi_{\max}$ 时,该回合(episode)训练终止。倘若在训练过程中姿态角的跟踪误差不设限制,则会存在环境状态空间过大的问题,导致收敛非常困难,同时过大的姿态角跟踪误差在实际应用中也是不被允许的。控制参数 τ 为一常数,满足 $\tau \in (0,1)$,从设计中可得 $(\Delta\varphi_{\max} - |\Delta\varphi_t|)/\Delta\varphi_{\max} > 0$,但并不是在误差限制内所有跟踪误差都是可以被接受的,故设计了控制参数 τ,仅有 $(\Delta\varphi_{\max} - |\Delta\varphi_t|)/\Delta\varphi_{\max} > \tau$ 区间内的 $\Delta\varphi_t$ 可获得奖励,否则将具有惩罚性质。

2)跟踪稳定性奖励。跟踪稳定性评价奖励是对火箭飞行过程姿态跟踪的稳定性做出评价的指标。倘若没有设置该项奖励函数,尽管姿态角跟踪的精度很高,但跟踪的稳定性会比较差,可能存在小幅的抖动。在此,以姿态角速度作为评价指标,针对此问题设计跟踪稳定性评价奖励:

$$r_2 = -\frac{|\omega|}{\mu} \tag{32}$$

式中,$\mu \in (0,1)$ 为控制参数。可见 $r_2 \leqslant 0$,当角速度越大时则认为跟踪越不稳定。控制参数 μ 可用于调节跟踪稳定性评价奖励在总奖励中的占比。

3)飞行时长奖励。在训练过程中设置每回合训练的终止条件为 $|\Delta\varphi_t| > \Delta\varphi_{\max}$ 或 $t > t_{\text{end}}$,t_{end} 为设定的每回合终止时间。由于前期对策略的学习还处于探索阶段,这样将导致每回合训练的时长很短就终止了,距离设定的终止时间相差甚远。为了提高训练效率,故设计飞行时长奖励项:

$$r_3 = \frac{t}{t_{\text{end}}}\nu \tag{33}$$

式中,$\nu \in (0,1)$ 为控制参数,用于调节飞行时长奖励项在总奖励中的占比。

4)减载效果奖励。运载火箭在上升段通常要穿越大风区,遭遇较为强烈的高空风作用,形成较大的气动攻角,这一附加攻角通常会使箭体受到的弯矩大幅增加。因此通常在大风区通过主动减载来消除这一影响,降低运载火箭在高空风作用下的气动载荷 $|q\alpha|$,减小箭体的结构负载,提升结构强度的可靠性以及运载能力,故设计减载效果奖励:

$$r_4 = \frac{1\,000 - |q\alpha|}{3\,500}\varepsilon \tag{34}$$

式中,$\varepsilon \in (0,1)$ 为控制参数,用于调节减载效果奖励在总奖励中的占比。

综上所述,本研究所设计的总奖励函数为:

$$r_t = \sum_{i=1}^{4} r_i \tag{35}$$

根据上述分析,总奖励函数中的“飞行时长奖励”可以鼓励智能体快速完成控制任务,同时“跟踪稳定性奖励”可以抑制系统的高频振荡,确保短周期内的平滑响应。

2.3.4　学习率自适应策略

根据此前介绍,固定学习率可能导致模型在训练过程中陷入局部最优解或产生震荡,且无法适应本研究复杂且不确定的训练环境。自适应学习率能够根据实情自动调整学习率,减少这种不稳定现象,使模型更稳定地逼近全局最优解。

这里拟采用一种最直观的学习率自适应策略,即使用步长学习率调度器,这是一种在模型训

练过程中动态调整学习率的机制，旨在提高模型的训练效率和性能。具体过程可以描述为：给定一个学习率初值 η_0，然后每训练 m 轮（epoch）自动衰减一定的比例 k，那么训练 n 轮以后学习率 η 可以表达如下：

$$\eta = \eta_0 (1-k)^{\frac{n}{m}} \tag{36}$$

此外，在 Critic 网络参数更新过程中，引入梯度裁剪（gradient clipping）技术，避免因梯度爆炸导致的训练不稳定。该技术可以表示为：

$$g_{\text{clipped}} = \begin{cases} g_0 & \|g_0\| \leqslant \beta \\ \beta \cdot \dfrac{g_0}{\|g\|} & \|g_0\| > \beta \end{cases} \tag{37}$$

其中，g_0 为原始梯度，g_{clipped} 为剪裁后的梯度，β 为裁剪阈值。

2.3.5 自动终止训练的早停机制

早停（early stopping）法是一种在深度学习领域被广泛使用的正则化方法，在很多情况下比其他正则化方法更简单高效。其基本含义是在模型训练时关注模型在验证集上的表现，当模型在验证集上的表现开始下降时，系统将停止训练[25]。将早停法应用于 SAC 算法时，需要特别注意 SAC 算法的特性，因为 SAC 算法是一个基于策略梯度和价值函数的深度强化学习算法。在 SAC 算法中，学习率的调整通常不是基于单个 epoch 的验证集性能，而是基于更复杂的评判机制，比如奖励的累积回报、策略的稳定性等。

这里拟采用在 SAC 算法的训练过程中实现一种基于验证集性能的早停机制。这意味着需要在训练环境中划分出一个“验证环境”，用于在每个 epoch 的训练步骤后评估当前策略的性能。定义验证性能指标为当前 epoch 所有回合（episode）的累积奖励均值，可表达为：

$$J = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{38}$$

式中，R_i 为第 i 个回合的累积奖励，N 为回合数。

设定一个阈值 κ，用于判断验证集性能是否有显著提升。在每个 epoch 结束后，计算当前累积奖励均值与上一个 epoch 的累积奖励均值之间的差值 ΔJ，如果 $\Delta J < \kappa$，则认为验证集性能没有显著提升。设定一个耐心值（patience），表示在验证集性能不再提升之前可以容忍的训练轮数。如果在连续 patience 个 epoch 内，验证集性能都没有显著提升，甚至开始下降，则表明剩下的训练可能已经属于无效训练，继续训练可能出现越学越差的结果。因此，此时可以考虑停止训练。

3 系统仿真与分析

首先，根据上述构建的运载火箭姿态动力学模型基于 PyTorch 框架搭建智能体的训练环境，再根据上述设计的智能姿态控制方法搭建深度强化学习框架。然后，利用改进后的 SAC 算法对框架中的智能体进行训练。需要说明的是，该训练是在一台配备 Intel Core i9 - 14900HX CPU、NVIDIA RTX 4070 GPU 以及 64 GB 内存的计算机上进行的，该机器运行系统为 Windows 11。

3.1 智能体训练

3.1.1 训练目标

该仿真是在一轨道倾角为 51.6°的国际空间站交汇任务的背景下开展进行的[22]。仿真拟采用火箭清离发射台时刻的状态作为其初始状态，该时刻即为 $t=0$ 时刻，具体的仿真参数和初始条件请参考文献[22]。根据该文献的仿真条件以及制导方法对该型号运载火箭进行弹道仿真，可以得到其在上升段发射纵平面内的标称轨迹以及程序俯仰角的图像，如图 4 ~ 5 所示。不难看出，运载火箭在上升段的标称轨迹非常稳定，程序俯仰角也无异常波动。因此，在本节中智能体的主要训练目标就是使该型运载火箭在上升段飞行过程中对其程序俯仰角指令进行稳定跟踪。

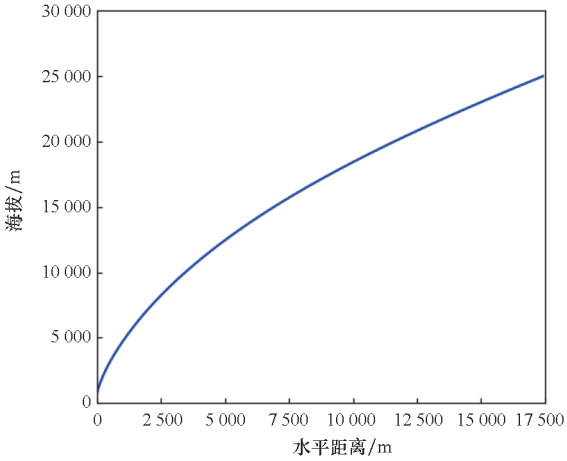


图 4 运载火箭在发射纵平面内的标称轨迹

Fig. 4 Nominal trajectory of the launch vehicle in the vertical plane

3.1.2 训练场景

根据此前的描述，火箭在上升段要经历大风区的干扰，因此，大风区减载即为本研究的训练场景。这里，根据某发射场一年四季关于风场的实测数据，随机生成其他 28 种风型，将这 32 种风型的阵风干扰引入被控对象环境。

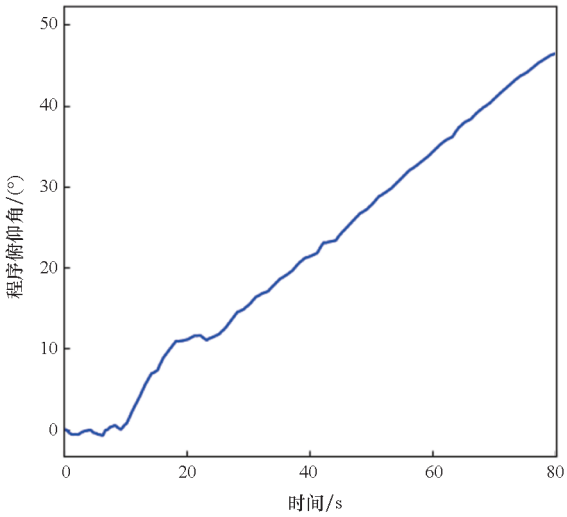


图 5　运载火箭的程序俯仰角

Fig. 5　Programmed pitch angle for the launch vehicle

由假设 4 可得,阵风干扰的风向保持恒定,仅考虑风速的垂直切变,即风速随海拔的变化,如图 6所示。需要注意的是,图中仅标记了该发射场四个季度的原始风场,其他风型由于是随机产生,因此不再逐一标注。

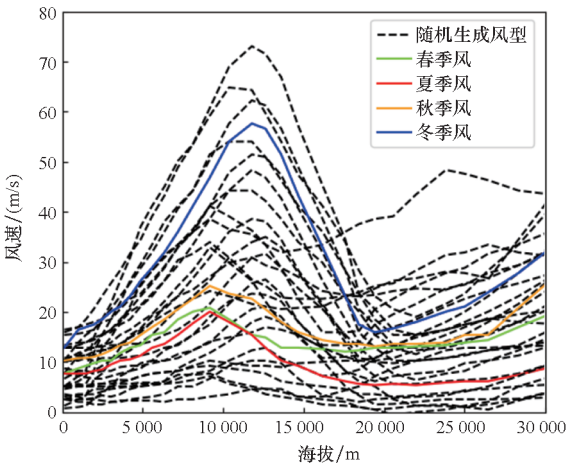


图 6　风速的垂直切变曲线

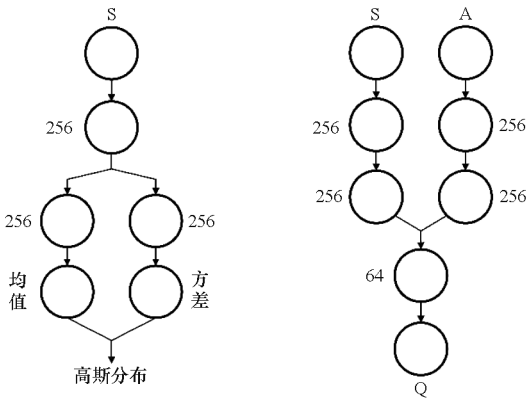
Fig. 6　Vertical shear curve of wind speed

3.1.3　训练参数

训练将采用回合制,仿真将采用四阶龙格 - 库塔法对运载火箭动力学模型进行积分,假设积分步长和控制步长保持一致,均为 0.02 s,设定训练单回合终止时间为 90 s,跟踪精度的最大允许误差为 3°,大风区的减载效果要求高于 10%,给定控制参数为 $\tau=0.7$,$\mu=0.15$,$\nu=0.1$,$\varepsilon=0.2$。

需要强调的是,之所以使用高频率控制步长(0.02 s),目的是匹配火箭动力学模型的快速变化需求,有效应对短周期扰动下的快速发散风险,有效解决此类快响应的控制问题。

这里为了验证所提方法的有效性,分别用标准的 SAC 算法和改进后的 SAC 算法训练智能体以得到基于两种算法下的控制器。两种算法均包含 1 个Actor 网络和 2 个 Critic 网络以及 1 个目标 Critic 网络。根据此前的描述,SAC 算法是基于随机性策略的深度强化学习算法,因此 AC 架构全部使用深度神经网络。其中,Actor 网络的状态路径 S 具有 1 个含有 256 个神经元的公共隐含层,均值和方差 2 条路径各具有 1 个含有 256 个神经元的隐含层。Critic 网络中,状态路径 S 和动作路径 A 各具有 2 个隐含层,每个隐含层都有 256 个神经元,2 条支路通过 1 个含有 64 个神经元的公共隐含层连接到价值路径 Q 上。两种算法中 Actor 网络和 Critic 网络的具体结构参数如图 7 所示。



(a) Actor 网络结构　(b) Critic 网络结构

(a) Actor network structure　(b) Critic network structure

图 7　两种算法神经网络结构参数示意图

Fig. 7　Schematic diagram of neural network structure parameters for two algorithms

标准 SAC 算法的训练超参数如表 1 所示,该算法将学习率固定,并且不融入早停机制。而改进后的 SAC 算法通过所设计的学习率自适应策略对固定学习率进行动态的更新迭代,并且融入早停机制以实现训练的自动停止,相关的核心参数如表 2 所示,其他超参数与表 1 保持一致。

表 1　标准 SAC 算法超参数

Tab. 1　Hyperparameters of the standard SAC algorithm

训练超参数	取值
学习率	3×10^{-4}
折扣率	0.99
经验池大小	1×10^{6}
批采样大小	100
熵权重系数	0.5
目标熵	-3
最大训练轮数	250

表 2 改进后的 SAC 算法核心参数

Tab. 2 Core parameters of the improved SAC algorithm

相关参数	取值
学习率初值	1×10^{-3}
梯度剪裁阈值	1
学习率调节步长	10
学习率衰减系数	0.1
耐心值	50

3.1.4 训练结果

在标准 SAC 算法的作用下，智能体共经历了完整的 250 轮训练后停止，总共经历了 1 000 个训练回合，总计用时 4 h 43 min 37 s。然而，在改进的 SAC 算法作用下，智能体只经历了 175 轮训练后就自动停止，共经历了 700 个训练回合，总计用时 3 h 19 min 47 s。由此可见，改进的 SAC 算法训练效率显著提升，这是加入了早停机制的缘故，在智能体性能经历多轮没有显著提升时系统自动终止了训练。

在两种算法作用下的智能体训练过程中所获得的奖励均值曲线如图 8 所示。由图 8 可见，在标准 SAC 算法的作用下，奖励均值在 80 轮左右才接近收敛，而改进后的 SAC 算法仅需 20 轮左右就已经收敛，这是因为后者在前期使用的了较高的学习率大大提升了系统的收敛速度。另外，在奖励均值的上升阶段，标准 SAC 算法的抖动较为严重，而后者非常平滑，这也是学习率自适应迭代所导致的优化结果。在改进的 SAC 算法下，学习率通过自适应策略实现了 $1\times10^{-3}\rightarrow1.7\times10^{-4}$ 的迭代。这也印证了此前的说法，过小的学习率导致系统收敛速度慢，而过大的学习率容易造成错过最优解。

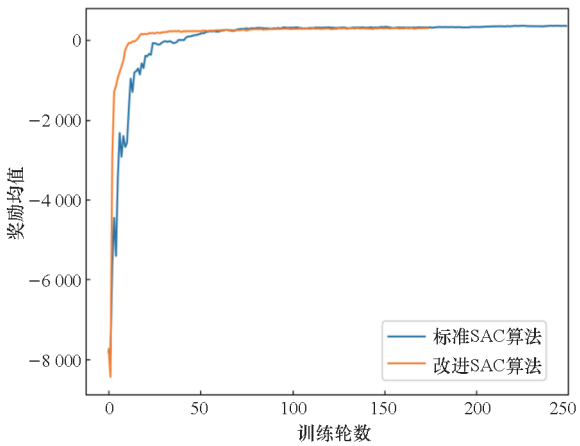


图 8 训练过程的奖励均值变化曲线

Fig. 8 Mean rewards velocity curve during the training process

因此，固定的学习率不适合此类短周期、快响应的问题，进而分析得出，改进 SAC 算法针对此类问题更有优势。

3.2 智能控制器性能测试

训练完毕后，将改进后的 SAC 控制器部署于运载火箭被控对象环境进行仿真测试。在测试中，将所改进后的 SAC 控制器与传统基于在线过载反馈的主动减载控制器进行对比，以验证所提出的控制器的有效性和优势。

测试时，在图 6 给出的 32 种风型中选择风速峰值最大的风型作为干扰，并将干扰的时机向后偏移 10 s，即从火箭起飞时刻 $t=0$ 开始算，$t=10$ s时加入该阵风干扰，以测试所提出的控制器对于随机干扰的适应能力以及对于阵风干扰时机的鲁棒性。该风型可称为最大风型，其风速随时间变化曲线如图 9 所示。

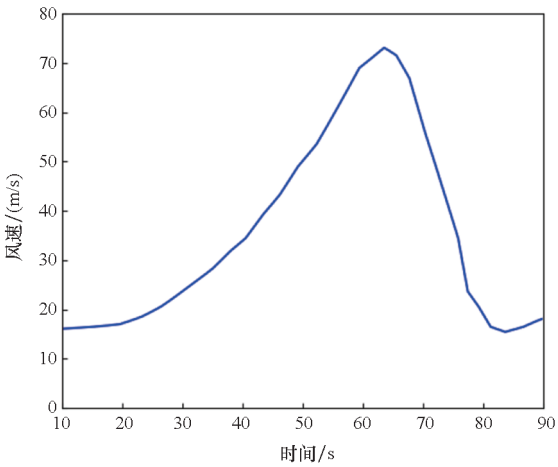
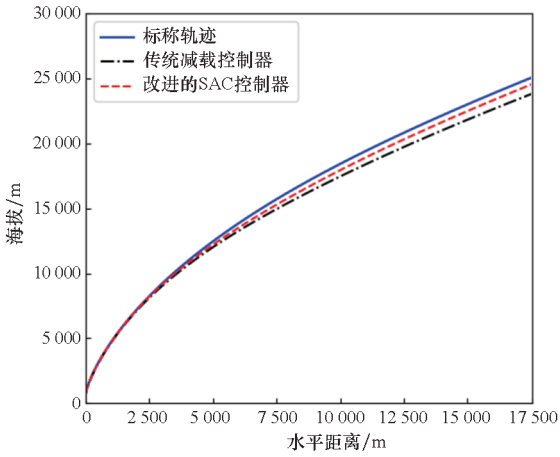


图 9 最大风型阵风干扰的风速变化曲线

Fig. 9 Wind velocity curve of maximum wind disturbance

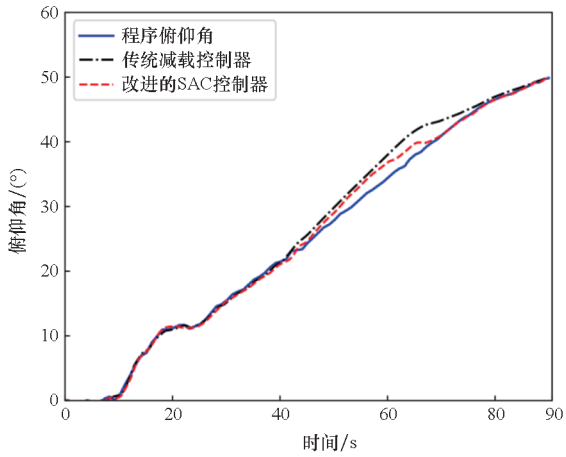
图 10 给出在传统减载控制器和改进后的 SAC 控制器的作用下运载火箭关键参数的对比效果，图 10(a) ~ (f)分别展示了运载火箭在飞行过程中的位置、俯仰角、俯仰角误差、俯仰角速度、发动机摆角以及气动载荷的曲线。为了更直观地分析测试结果，将图 10 中的关键数据用表格的形式呈现。表 3 展现了在两种控制器的作用下运载火箭的俯仰角误差、俯仰角速度、发动机摆角以及气动载荷的绝对值的最大值。

很显然，两种控制器之间的性能对比非常明显，由图 10(a)可见，运载火箭在风场的作用下飞行轨迹都发生了一定的偏差，但在本研究所提出的控制器作用下火箭的飞行轨迹相比传统减载控制器而言距离标称轨迹偏差较小。由图 10(b)和图 10(c)可知，火箭的俯仰角在两种控制器的作



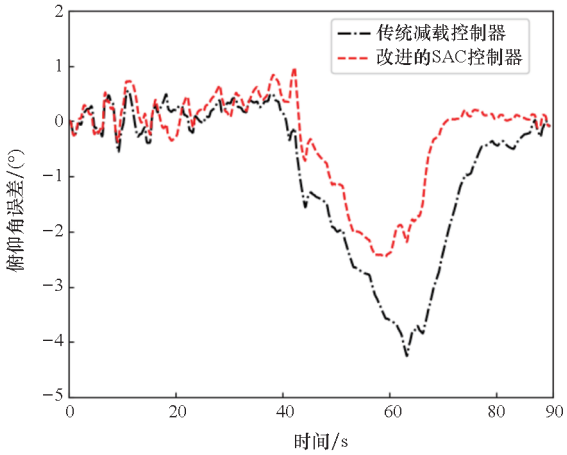
(a) 位置

(a) Position



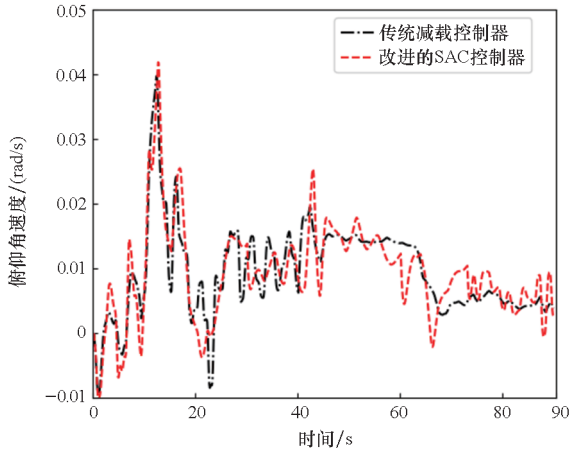
(b) 俯仰角

(b) Pitch angle



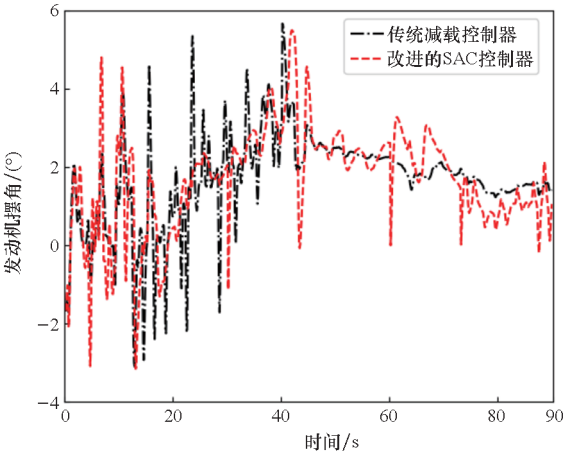
(c) 俯仰角误差

(c) Pitch angle error



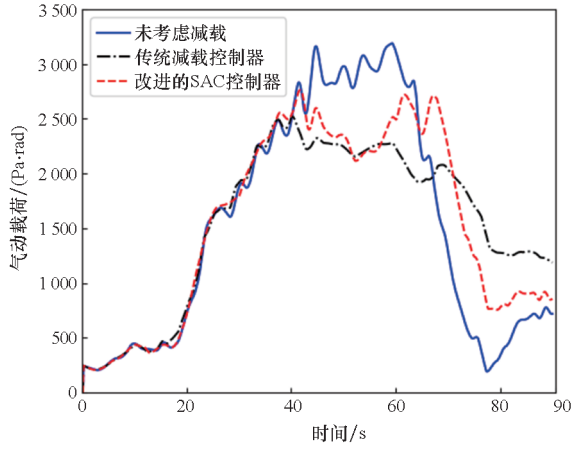
(d) 俯仰角速度

(d) Pitch angle velocity



(e) 发动机摆角

(e) Engine gimbal angle



(f) 气动载荷

(f) Aerodynamic load

图 10　两种控制器在最大风型阵风干扰下的测试结果

Fig. 10　Test results of the two controllers under the maximum wind disturbance

用下跟踪效果都表现良好,但是在中间阶段随着风场风速不断增加,在经历大风区的时候跟踪精度都有所下降,在传统减载控制器作用下最大俯仰角误差为 4.25°,在改进后的 SAC 控制器作用下最大俯仰角误差仅为 2.45°,可见前者的跟踪误差已经超过了可允许的误差范围,不满足指标要求,而后者的跟踪误差满足要求。通过图 10(d)和图 10(e)可知,两种控制器作用下,姿

表3 最大风型干扰下两种控制器的关键仿真数据对比

Tab. 3 Comparison of key simulation data between the two controllers under the maximum wind disturbance

仿真数据	传统减载控制器	改进的SAC控制器
最大俯仰角误差/(°)	4.25	2.45
最大俯仰角速度/(rad/s)	0.04	0.042
最大发动机摆角/(°)	5.67	5.49
最大气动载荷/(Pa·rad)	2 523	2 762

态跟踪的稳定性都很高,且对火箭发动机这种执行机构而言,发动机摆角也都在安全、合理的执行范围之内,由此可知跟踪稳定性奖励的设计是有效的。由图10(f)可知,在前期风速不大的阶段,两种控制器作用下的运载火箭的气动载荷相差不大,但在中间阶段经历大风区时,传统减载控制器的减载效果略优于改进后的SAC控制器,前者的最大气动载荷为2 523 Pa·rad,相比未减载情况下的3 190 Pa·rad,减载效果达到了20.1%,而后者的最大气动载荷为2 762 Pa·rad,减载效果为13.4%。可见传统减载控制器相比改进后的SAC控制器在面对大风区时的减载效果更优,但是在两者同样满足减载效果10%以上的指标要求的基础上,前者明显牺牲了更多的跟踪精度,且已超出了可允许的误差范围,这是不能接受的。

通过分析可得,姿控回路的跟踪精度和减载效果本身就是一对"矛盾体",具有此消彼长的特征。传统的减载方法大多都像上述控制器一样是以牺牲较高的跟踪精度来强化减载效果的。然而,在运载火箭控制器设计时,通常将跟踪效果作为第一指标,减载仅仅是为了保护箭体结构强度的可靠性而增添的附加指标,故在设计控制器时,只要火箭过载能够在合理的安全范围内,应当以第一指标跟踪精度作为主要评价标准。

基于以上仿真结果和分析,综合考虑跟踪精度、稳定性和减载效果之间的平衡关系,本研究所提出的改进后的SAC控制器的性能高于传统的减载控制器。

4 结论

本研究以运载火箭为研究对象。首先,在俯仰平面内建立其动力学模型,而后根据动力学模型搭建被控对象环境,并将运载火箭上升段的姿态控制问题描述为马尔可夫决策过程,构建了适用该问题的深度强化学习框架。然后,综合考虑火箭的姿态跟踪精度和稳定性以及经历大风区时的减载效果,提出了一种全新的奖励函数。此外,在原本的SAC算法框架中加入了一种学习率自适应策略以及一种能够自动终止训练的早停机制,基于这种改进后的SAC算法对智能体进行了多轮训练,在训练过程中循环改变阵风干扰的风型,最终得到一种对随机阵风干扰具备较强鲁棒性和适应能力的智能姿态控制器。最终,通过仿真测试得出本研究所设计的控制器能够在保持系统稳定性的情况下对该型号运载火箭上升段飞行中的姿态角进行准确跟踪,满足技术指标要求,并且能够使其在经历大风区时克服气动载荷过大的问题,有效实现了指标要求的减载效果,验证了该方法的有效性。仿真实验将该控制器与目前本领域较为常用的基于在线过载反馈的传统减载控制器进行对比,发现改进后的SAC控制器在解决姿态跟踪和减载效果这对"矛盾体"的平衡性上更有优势。

参考文献(References)

[1] 宋征宇. 运载火箭飞行减载控制技术[J]. 航天控制, 2013, 31(5): 3-7, 18.
SONG Z Y. Load control technology in launch vehicle[J]. Aerospace Control, 2013, 31(5): 3-7, 18. (in Chinese)

[2] BLANCHET P, BARTOS B. An improved load relief wind model for the Delta launch vehicle[C]//Proceedings of the Aerospace Sciences Meeting and Exhibit, 2001.

[3] 丁秀峰. 运载火箭减载控制技术研究[J]. 飞控与探测, 2018, 1(1): 55-58.
DING X F. Study on load control technology of launch vehicle[J]. Flight Control & Detection, 2018, 1(1), 55-58. (in Chinese)

[4] 杨伟奇, 许志, 唐硕, 等. 基于自抗扰的运载火箭主动减载控制技术[J]. 北京航空航天大学学报, 2016, 42(1): 130-138.
YANG W Q, XU Z, TANG S, et al. Active disturbance rejection control method on load relief system for launch vehicles[J]. Journal of Beijing University of Aeronautics and Astronautics, 2016, 42(1): 130-138. (in Chinese)

[5] 李效明, 许北辰, 陈存芸. 一种运载火箭减载控制工程方法[J]. 上海航天, 2004(6): 7-9, 14.
LI X M, XU B C, CHEN C Y. An engineering method on the control of decreasing load for a launch vehicle[J]. Aerospace Shanghai, 2004(6): 7-9, 14. (in Chinese)

[6] 张卫东, 贺从园, 周静, 等. 基于信号辨识的运载火箭实时减载控制技术[J]. 航天控制, 2018, 36(3): 3-8, 14.
ZHANG W D, HE C Y, ZHOU J, et al. The in-flight load relief of launch vehicles based on the signal identification[J]. Aerospace Control, 2018, 36(3): 3-8, 14. (in Chinese)

[7] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[8] XU J, DU T, FOSHEY M, et al. Learning to fly: computational controller design for hybrid UAVs with reinforcement learning[J]. ACM Transactions on Graphics, 2019, 38(4): 1 – 12.

[9] MA Z, WANG Y J, YANG Y D, et al. Reinforcement learning-based satellite attitude stabilization method for non-cooperative target capturing[J]. Sensors, 2018, 18(12): 4331.

[10] 付宇鹏, 邓向阳, 朱子强, 等. 基于模仿强化学习的固定翼飞机姿态控制器[J]. 海军航空大学学报, 2022, 37(5): 393 – 399.
FU Y P, DENG X Y, ZHU Z Q, et al. Imitation reinforcement learning based attitude controller for fixed-wing aircraft[J]. Journal of Naval Aviation University, 2022, 37(5): 393 – 399. (in Chinese)

[11] 付宇鹏, 邓向阳, 何明, 等. 基于强化学习的固定翼飞机姿态控制方法[J]. 控制与决策, 2023, 38(9): 2505 – 2510.
FU Y P, DENG X Y, HE M, et al. Reinforcement learning based attitude controller design[J]. Control and Decision, 2023, 38(9): 2505 – 2510. (in Chinese)

[12] 赵克刚, 石翠铎, 梁志豪, 等. 基于柔性演员 – 评论家算法的自适应巡航控制研究[J]. 汽车技术, 2023(3): 26 – 34.
ZHAO K G, SHI C D, LIANG Z H, et al. Research on adaptive cruise control based on soft actor-critic algorithm[J]. Automobile Technology, 2023(3): 26 – 34. (in Chinese)

[13] 多南讯, 吕强, 林辉灿, 等. 迈进高维连续空间: 深度强化学习在机器人领域中的应用[J]. 机器人, 2019, 41(2): 276 – 288.
DUO N X, LYU Q, LIN H C, et al. Step into high-dimensional and continuous action space: a survey on applications of deep reinforcement learning to robotics[J]. Robot, 2019, 41(2): 276 – 288. (in Chinese)

[14] 邱潇颀, 高长生, 荆武兴. 拦截大气层内机动目标的深度强化学习制导律[J]. 宇航学报, 2022, 43(5): 685 – 695.
QIU X Q, GAO C S, JING W X. Deep reinforcement learning guidance law for intercepting endo-atmospheric maneuvering targets[J]. Journal of Astronautics, 2022, 43(5): 685 – 695. (in Chinese)

[15] 裴培, 何绍溟, 王江, 等. 一种深度强化学习制导控制一体化算法[J]. 宇航学报, 2021, 42(10): 1293 – 1304.
PEI P, HE S M, WANG J, et al. Integrated guidance and control for missile using deep reinforcement learning[J]. Journal of Astronautics, 2021, 42(10): 1293 – 1304. (in Chinese)

[16] 宋欣屿, 王英勋, 蔡志浩, 等. 基于深度强化学习的无人机着陆轨迹跟踪控制[J]. 航空科学技术, 2020, 31(1): 68 – 75.
SONG X Y, WANG Y X, CAI Z H, et al. Landing trajectory tracking control of unmanned aerial vehicle by deep reinforcement learning[J]. Aeronautical Science & Technology, 2020, 31(1): 68 – 75. (in Chinese)

[17] 王鑫, 赵清杰, 于重重, 等. 多节点探测器软着陆的路径规划方法[J]. 宇航学报, 2022, 43(3): 366 – 373.
WANG X, ZHAO Q J, YU C C, et al. Path planning method of soft landing for multi-node probe[J]. Journal of Astronautics, 2022, 43(3): 366 – 373. (in Chinese)

[18] ELKINS J G, SOOD R, RUMPF C. Autonomous spacecraft attitude control using deep reinforcement learning[C]// Proceedings of the 71st International Astronautical Congress (IAC), 2020.

[19] ELKINS J G, SOOD R, RUMPF C. Adaptive continuous control of spacecraft attitude using deep reinforcement learning[C]//Proceedings of the AAS/AIAA Astrodynamics Specialist Conference, 2020: 420 – 475.

[20] DIAS P M, ZHOU Y, VAN KAMPEN E J. Intelligent nonlinear adaptive flight control using incremental approximate dynamic programming[C]//Proceedings of the AIAA Scitech Forum, 2019.

[21] 刘俊辉, 单家元, 荣吉利, 等. 自适应学习率的增量强化学习飞行控制[J]. 宇航学报, 2022, 43(1): 111 – 121.
LIU J H, SHAN J Y, RONG J L, et al. Incremental reinforcement learning flight control with adaptive learning rate[J]. Journal of Astronautics, 2022, 43(1): 111 – 121. (in Chinese)

[22] DU W. Dynamic modeling and ascent flight control of Ares-Ⅰ Crew Launch Vehicle[D]. Iowa: Iowa State University, 2010.

[23] 郑鹤鸣, 翟光, 孙一勇. 面向在轨加注的组合体姿态 SAC 智能控制[J]. 宇航学报, 2023, 44(7): 1020 – 1033.
ZHENG H M, ZHAI G, SUN Y Y. SAC intelligent attitude control method for on-orbit refueling combination[J]. Journal of Astronautics, 2023, 44(7): 1020 – 1033. (in Chinese)

[24] 程川, 刘阳. 运载火箭嵌入式大气数据测量系统[J]. 气体物理, 2023, 8(4): 55 – 62.
CHENG C, LIU Y. Flush air data sensing system for launch vehicle[J]. Physics of Gases, 2023, 8(4): 55 – 62. (in Chinese)

[25] 魏守鑫. 基于改进贝叶斯优化的超参数优化方法的研究与实现[D]. 西安: 西安电子科技大学, 2022.
WEI S X. Research and implementation of hyperparameter optimization method based on improved Bayesian optimization[D]. Xi′an: Xidian University, 2022. (in Chinese)