基于深度 Q 学习的增益调度：四旋翼非线性动力学

Hossein Rastgoftar 和 Muhammad J. H. Zahed

arXiv:2603.03127v1 [eess.SY] 3 Mar 2026

摘要—**本文提出了一种基于深度 Q 网络（DQN）的增益调度框架，用于安全关键型四旋翼飞行器轨迹跟踪。该方法不直接学习控制输入，而是从有限个经过预认证的稳定增益向量中进行选择，使强化学习能够在结构化且保持稳定性的控制架构中运行。利用平移动力学的各向同性结构，各空间轴共享反馈增益，在保持性能的同时降低维度。学习策略实时调整反馈强度：在大幅瞬态响应期间施加较强控制，并在接近收敛时降低增益，以限制控制 effort。采用高保真四旋翼非线性模型进行的仿真结果表明，该方法能够实现精确的轨迹跟踪、将姿态偏移限制在有界范围内、在最终时刻后平稳过渡至悬停，并持续提升奖励，从而验证了所提出的基于学习的增益调度策略的有效性和鲁棒性。**

索引词—**强化学习、深度 Q 网络（DQN）、增益调度、自适应控制、非线性控制、四旋翼动力学、无人机轨迹跟踪**

I. 引言

自主四旋翼飞行器稳定化与悬停控制已通过非线性几何方法及基于李雅普诺夫的表述得到广泛研究，这些方法可在悬停平衡点附近提供严格的稳定性保证 [1]–[4]。此外，微分平坦性与最小快拍轨迹生成能够实现动力学可行的参考轨迹跟踪，其中悬停可视为恒定平坦输出的特例 [5]。近年来，基于平坦性的Snap控制架构已被提出，用以扩大吸引域并提升跟踪性能，同时保留非线性结构 [6]。尽管取得了这些进展，大多数高性能控制器仍依赖固定反馈增益，而这些增益在不同瞬态工况下可能偏于保守或并非最优。相反，无模型强化学习（强化学习）方法，例如深度 Q 网络 [7]，，虽具有适应性，但通常会牺牲可解释性以及安全关键航空航天系统所必需的结构稳定性保证。本文研究如何在不损害稳定性结构的前提下，将自适应强化学习与经过严格验证的基于平坦性的控制架构相结合。我们提出一种叠加于Snap控制器之上的基于 DQN 的监督式增益调度框架，从而在保留经认证的非线性控制特性的同时，实现自适应反馈控制力度调节。由此得到的方法在强化学习与结构化非线性飞行控制之间建立了一座原则性桥梁。

作者就职于亚利桑那大学航空航天与机械工程系，图森，AZ 85719，美国 {hrastgoftar, mjhz}@arizona.edu

A. 相关工作

四旋翼飞行器轨迹跟踪通常由非线性控制器处理，这些控制器可在建模假设下提供稳定性保证。反馈线性化与滑模设计提高了对不确定性与噪声的鲁棒性 [8]。在特殊欧几里得群 SE(3)上的几何跟踪可产生全局一致的姿态误差，并具有较强的跟踪性能[9]。在已部署系统中，增益调度十分常见：增益针对不同工况或故障工况进行整定，并在线切换。代表性示例包括面向四旋翼无人机的增益调度PID容错控制 [10]，以及用于Qball-X路径跟踪与容错的、基于参数相关稳定性的增益调度PID4 [11]。

与上述方法相比，基于学习的方法旨在减少手动调参并补偿未建模动力学。一种广泛使用的架构保留稳定反馈回路，同时增加一个学习得到的前馈／参考整形模块。深度神经网络已被提出作为附加模块，用于即兴四旋翼轨迹跟踪 [12]。基于逆模型的学习将这一思路扩展到具有挑战性的动力学，同时显式地推理输入输出行为与跟踪瞬态 [13]。在线变体将离线训练与飞行过程中的持续自适应相结合，并融入专家知识，以应对变化的动力学与运行不确定性 [14]。其他工作则将学习更深入地融入控制器：DATT学习自适应组件，以便在大扰动下跟踪任意（甚至不可行的）轨迹[15]。由最优轨迹数据训练的端到端神经控制器直接输出旋翼指令，并通过自适应未建模力矩来应对现实差距效应 [16]。

强化学习也已应用于无人机低层控制。GymFC 在高保真仿真中评估了用于四旋翼姿态控制的策略梯度强化学习，并探讨了鲁棒性和奖励设计挑战 [17]。通过在仿真中结合基于示范的训练学习感觉运动策略，强化学习实现了敏捷特技飞行 [18]。较早的学习型特技飞行研究包括通过策略梯度改进参数化四旋翼连续翻转 [19]。然而，大多数强化学习飞行控制器学习的是连续控制动作（或连续调节的增益），而稳定性／约束保证通常是间接实现的。安全强化学习表述通过约束策略优化 [20] 或基于李雅普诺夫的安全证书 [21],来处理约束，而深度Q学习则是用于离散动作选择的典型价值方法 [22]。

以往的增益调度通常依赖启发式方法或事件触发机制，缺乏数据驱动的状态依赖自适应；而基于强化学习的飞行控制往往直接学习连续控制

而未嵌入经过认证的稳定性。我们的方法通过使用深度Q网络策略，在有限的预先认证的稳定增益向量库中进行调度，弥补了这一差距；该方法在在线调整控制激进程度的同时，从设计上保证稳定性。通过在各轴之间共享平移增益（利用各向同性），我们进一步降低了学习维度，同时不牺牲跟踪质量。与采用连续增益的强化学习相比，这种结构化动作空间提升了探索效率，也简化了验证。

B. 贡献

本文提出了一种用于非线性四旋翼飞行器轨迹跟踪的结构化强化学习框架，该框架增强而非替代经过认证的非线性控制架构。所提出的方法将深度 Q 网络（DQN）与此前在 [6],[23], 中介绍的 Snap平坦性控制器相结合，其性能曾与广泛采用的 Mellinger最小Snap控制器 [5] 进行基准比较。不同于将状态直接映射为推力和力矩的端到端强化学习方法，所提出的框架保留了反馈线性化和微分平坦性的结构特性，同时在监督层面引入自适应增益调度。

本文的主要贡献如下：

1) **结构保持型强化学习架构**。我们提出了一种监督式强化学习方案，其中学习智能体为经过验证的基于平坦性的控制器选择稳定化增益向量，而不是直接生成控制输入。如图 1 所示，所提出的方法在实现自适应性能的同时，保留了稳定性结构和可解释性。该方法为完全无模型的策略 [7], 提供了一种替代方案，将学习嵌入控制理论骨干之中。

2) **具备认证边界的安全性导向的混合增益调度**。动作空间限定为经过预认证的 Snap 控制器稳定增益配置 [6],，切换则受驻留时间约束调节。该方法确保自适应调度不会违反稳定性保证。

3)**在不损失稳定性结构的前提下自适应调节反馈控制能力**。仿真结果表明，学习策略能够自主调节反馈控制能力——在大幅瞬态变化期间采用激进增益，在接近收敛时采用保守增益——同时保留基于平坦性的控制的理论稳定性性质。这弥合了最小Snap轨迹跟踪 [5] 与自适应强化学习之间的差距。

C. 论文结构

第二节回顾了本文研究的问题。第三节介绍了确保目标悬停状态稳定的四旋翼控制。第四节介绍了基于深度Q网络的强化学习增益调度。第五节展示了仿真结果，最后在第六节给出总结。

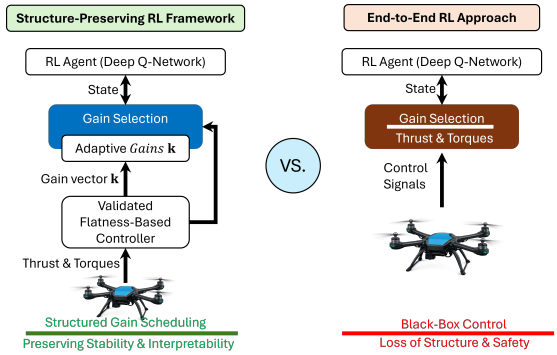


图1. 保持结构的强化学习与端到端无模型控制的比较。在所提出的架构（左）中，深度Q网络监督增益选择，以控制经过验证的基于平坦性的控制器，从而保持稳定性结构和可解释性。相比之下，端到端强化学习（右）直接将状态映射为控制输入，可能牺牲结构性保证。

II. 问题陈述

我们考虑由以下方程描述的四旋翼飞行器非线性动力学

$$\begin{cases} \dot{\mathbf{r}} = \mathbf{v}, \\ \dot{\mathbf{v}} = -g\hat{\mathbf{e}}_3 + \dfrac{mg+T}{m}\mathbf{R}(\boldsymbol{\eta})\hat{\mathbf{e}}_3, \\ \dot{\boldsymbol{\eta}} = \mathbf{E}(\phi,\theta)\boldsymbol{\omega}, \\ \mathbf{I}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}\times(\mathbf{I}\boldsymbol{\omega}) = \boldsymbol{\tau}, \\ \dot{T} = \dot{T}, \\ \ddot{T} = u_T, \end{cases} \tag{1}$$

其中$\hat{\mathbf{e}}_3 = [0\ 0\ 1]^\top$、$\mathbf{r},\mathbf{v} \in \mathbb{R}^3$ 表示惯性位置和速度，$\boldsymbol{\eta} = (\phi,\theta,\psi)$为3-2-1欧拉角，$\boldsymbol{\omega} = (p,q,r)$为机体角速度，$T$表示相对悬停的推力偏差，$\boldsymbol{\tau} = (\tau_\phi,\tau_\theta,\tau_\psi)^\top$ 为控制力矩，且$\mathbf{E}(\phi,\theta)$为欧拉运动学矩阵。

控制输入 $\boldsymbol{\tau}$ 和 u_T 由增益向量参数化

$$\mathbf{k} = [k_1,\ldots,k_{14}]^\top,$$

其中每个分量满足

$$k_i \in [k_{i,\min},k_{i,\max}], \quad i = 1,\ldots,14.$$

这些界限的选取旨在防止振荡行为，并确保渐近收敛至悬停平衡点。为保持稳定性，增益向量被限制在一个离散动作集内

$$\mathcal{A} \subset [k_{1,\min},k_{1,\max}]\times\cdots\times[k_{14,\min},k_{14,\max}].$$

在此参数化下，闭环动力学可表示为

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{k}, \tag{2}$$

其中，$\mathbf{x}\in\mathbb{R}^{14}$ 表示完整状态。

在实际应用中，非线性映射$\mathbf{f}$ 和$\mathbf{G}$会因建模误差、空气动力学扰动以及

参数变化而存在不确定性。给定一个不确定的动力学模型和离散增益集 $\mathcal{A}$，设计一种增益选择策略

$$\pi : \mathbf{x} \mapsto \mathbf{k} \in \mathcal{A}$$

从而保证安全地渐近收敛至期望的悬停平衡点$\mathbf{x}^*$，同时保持瞬态响应有界并避免振荡。

该问题被建模为一个强化学习任务，其中增益调度被视为离散动作控制问题。通过使用深度 Q 网络（DQN）近似相应的价值函数，获得最优策略，从而能够在存在不确定性的情况下进行无模型学习。

III. 四旋翼飞行器控制稳定性

本节将四旋翼动力学改写为式（2）的形式，并设计一种反馈控制律，以保证期望悬停平衡点的渐近稳定性。首先，在第三节 A 中给出四旋翼动力学的非线性状态空间表示。接着，在第三节 B 中构造稳定化反馈控制律。最后，第三节 C 推导面向控制的离散时间模型，该模型将用于构建强化学习环境。

A. 非线性状态空间动力学

对运动学关系求导

$$\dot{\boldsymbol{\eta}} = \mathbf{E}(\phi,\theta)\boldsymbol{\omega}$$

得到

$$\ddot{\boldsymbol{\eta}} = \mathbf{E}(\phi,\theta)\dot{\boldsymbol{\omega}} + \dot{\mathbf{E}}(\phi,\theta)\boldsymbol{\omega}.$$

定义虚拟控制输入

$$\mathbf{u}_\eta \triangleq \begin{bmatrix} u_\phi \\ u_\theta \\ u_\psi \end{bmatrix} := \ddot{\boldsymbol{\eta}}.$$

求解 $\dot{\boldsymbol{\omega}}$ 得到

$$\dot{\boldsymbol{\omega}} = \mathbf{E}^{-1}(\phi,\theta)\Big(\mathbf{u}_\eta - \dot{\mathbf{E}}(\phi,\theta)\boldsymbol{\omega}\Big). \tag{3}$$

将其代入刚体旋转动力学，得到所需的控制力矩

$$\boldsymbol{\tau} = \mathbf{I}\mathbf{E}^{-1}(\phi,\theta)\mathbf{u}_\eta - \mathbf{I}\mathbf{E}^{-1}(\phi,\theta)\dot{\mathbf{E}}(\phi,\theta)\boldsymbol{\omega} + \boldsymbol{\omega}\times(\mathbf{I}\boldsymbol{\omega}).$$

通过定义

$$\mathbf{x} = \begin{bmatrix}\mathbf{r}^\top & \mathbf{v}^\top & \boldsymbol{\eta}^\top & \dot{\boldsymbol{\eta}}^\top & T & \dot{T}\end{bmatrix}^\top \in \mathbb{R}^{14},$$
$$\mathbf{u} = \begin{bmatrix} u_T & \mathbf{u}_\eta^\top \end{bmatrix}^\top \in \mathbb{R}^4,$$

并利用 $\boldsymbol{\omega} = \mathbf{E}^{-1}(\phi,\ \theta)\,\dot{\boldsymbol{\eta}}$，四旋翼动力学可表示为控制仿射形式

$$\dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}) + \mathbf{G}_0(\mathbf{x})\,\mathbf{u}. \tag{4}$$

其中

$$\mathbf{f}_0(\mathbf{x}) = \begin{bmatrix} \mathbf{v}_I \\ -g\hat{\mathbf{e}}_3 + \dfrac{mg+T}{m}\mathbf{R}(\boldsymbol{\eta})\hat{\mathbf{e}}_3 \\ \dot{\boldsymbol{\eta}} \\ \mathbf{0}_3 \\ \dot{T} \\ 0 \end{bmatrix}, \tag{5}$$

$$\mathbf{G}_0(\mathbf{x}) = \begin{bmatrix} \mathbf{0}_{9\times1} & \mathbf{0}_{9\times3} \\ \mathbf{0}_{3\times1} & \mathbf{I}_3 \\ 0 & \mathbf{0}_{1\times3} \\ 1 & \mathbf{0}_{1\times3} \end{bmatrix}. \tag{6}$$

B. 四旋翼控制

我们采用 [24]中提出的 snap 控制框架，实现形式如下的期望悬停平衡状态

$$\mathbf{x}^* = \begin{bmatrix} \mathbf{r}_I^* \\ \mathbf{0}_{11\times1} \end{bmatrix},$$

其中$\mathbf{r}_I^*$ 表示期望悬停位置。为确保从任意初始位置$\mathbf{r}_0$ 平滑过渡到$\mathbf{r}_I^*$，我们定义期望轨迹

$$\mathbf{r}_d(t) = \begin{cases} (1-\beta(t))\mathbf{r}_0 + \beta(t)\mathbf{r}_I^*, & t \in [0, T_f], \\ \mathbf{r}_I^*, & t \geq T_f, \end{cases} \tag{7}$$

其中 $\beta(t)$是满足以下条件的五次多项式

$$\beta(0) = 0, \quad \beta(T_f) = 1, \quad \dot{\beta}(T_f) = 0,$$
$$\ddot{\beta}(T_f) = 0, \quad \dddot{\beta}(T_f) = 0, \quad \ddddot{\beta}(T_f) = 0.$$

这些条件确保在 post-T_f 恒定保持下， $t = T_f$ 处的 C^4 连续性。

定义$\mathbf{e}_r = \mathbf{r} - \mathbf{r}_d$、 $\mathbf{e}_v = \mathbf{v} - \dot{\mathbf{r}}_d$、 $\mathbf{e}_a = \mathbf{a} - \ddot{\mathbf{r}}_d$、 $\mathbf{e}_j = \mathbf{j} - \dddot{\mathbf{r}}_d$，外部（悬停误差）状态向量表示为

$$\mathbf{z}(\mathbf{x}) \triangleq \begin{bmatrix}\mathbf{e}_r^\top & \mathbf{e}_v^\top & \mathbf{e}_a^\top & \mathbf{e}_j^\top & \psi & \dot{\psi}\end{bmatrix}^\top \in \mathbb{R}^{14}.$$

令外部输入为

$$\mathbf{s} \triangleq \begin{bmatrix} \mathbf{s}_r \\ s_\psi \end{bmatrix} = \begin{bmatrix} \dddot{\mathbf{j}} \\ \ddot{\psi} \end{bmatrix} \in \mathbb{R}^4,$$

其中$\mathbf{s}_r$ 是平移 snap 输入， s_ψ 是偏航角加速度。于是，外部动力学的形式为

$$\dot{\mathbf{z}} = \mathbf{A}_{\mathrm{EXT}}\mathbf{z} + \mathbf{B}_{\mathrm{EXT}}\mathbf{s} + \mathbf{E}_{\mathrm{EXT}}\mathbf{r}_d^{(4)}(t), \tag{8}$$

其中$\mathbf{r}_d^{(4)}$ (t) 表示期望 snap，且

$$\mathbf{A}_{\mathrm{EXT}} = \begin{bmatrix} \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & 0 & 1 \\ \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & 0 & 0 \end{bmatrix},$$

$$\mathbf{B}_{\mathrm{EXT}}=\begin{bmatrix}\mathbf{0}_{3\times3} & \mathbf{0}_{3\times1}\\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times1}\\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times1}\\ \mathbf{I}_3 & \mathbf{0}_{3\times1}\\ \mathbf{0}_{1\times3} & 0\\ \mathbf{0}_{1\times3} & 1\end{bmatrix},\qquad \mathbf{E}_{\mathrm{EXT}}=\begin{bmatrix}\mathbf{0}_{3\times3}\\ \mathbf{0}_{3\times3}\\ \mathbf{0}_{3\times3}\\ -\mathbf{I}_3\\ \mathbf{0}_{1\times3}\\ \mathbf{0}_{1\times3}\end{bmatrix}.$$

因此，平动子系统构成一个由加加加速度驱动的四阶积分器链，而偏航动力学则构成一个由 s_ψ驱动的二阶子系统。我们选择

$$\begin{aligned}\mathbf{s}=&-k_{13}\dot{\psi}-k_{14}\psi+\mathbf{K}_j\left(\dddot{\mathbf{r}}_d-\mathbf{j}\right)+\mathbf{K}_a\left(\ddot{\mathbf{r}}_d-\mathbf{a}\right)\\&+\mathbf{K}_v\left(\dot{\mathbf{r}}_d-\mathbf{v}\right)+\mathbf{K}_p\left(\mathbf{r}_d-\mathbf{r}\right)\end{aligned}\tag{9}$$

采用正定对角增益矩阵

$$\begin{aligned}\mathbf{K}_j=&\mathrm{diag}(k_1,k_2,k_3),\\ \mathbf{K}_a=&\mathrm{diag}(k_4,k_5,k_6),\\ \mathbf{K}_v=&\mathrm{diag}(k_7,k_8,k_9),\\ \mathbf{K}_p=&\mathrm{diag}(k_{10},k_{11},k_{12}),\end{aligned}$$

以及偏航增益 k_{13}，$k_{14}>0$，其中

$$k_i=[k_{i,min},k_{i,max}],\qquad i=1,\cdots,14.\tag{10}$$

请注意，表 I 中列出的界限 $k_{i,\ \min}$ 和 $k_{i,\ \max}$ 经过选择，使得所得闭环外部动力学矩阵具有互不相同且严格位于左半平面的特征值，从而保证渐近稳定性。

外部输入 s 与四旋翼飞行器虚拟输入 u 之间的关系。定义

$$\hat{\mathbf{k}}_b=\mathbf{R}(\boldsymbol{\eta})\hat{\mathbf{e}}_3\tag{11}$$

作为四旋翼飞行器平面的单位法向量，并与推力方向对齐，$\hat{}\mathbf{e}_3$ 表示 $\mathbb{R}^3$ 中的第三个标准基向量。对平移动力学求两次导数，并利用运动学关系

$$\dot{\boldsymbol{\eta}}=\mathbf{E}(\phi,\theta)\boldsymbol{\omega},\qquad \boldsymbol{\omega}=\mathbf{E}^{-1}(\phi,\theta)\dot{\boldsymbol{\eta}},$$

外部输入 $\mathbf{s}$ 可表示为虚拟输入$\mathbf{u}$的仿射函数:

$$\mathbf{s}=\mathbf{M}(\mathbf{x})\,\mathbf{u}+\mathbf{n}(\mathbf{x}),\tag{12}$$

其中

$$\mathbf{M}(\mathbf{x})=\begin{bmatrix}\mathbf{M}'(\mathbf{x})\\ 0\ \ 0\ \ 0\ \ 1\end{bmatrix}\in\mathbb{R}^{4\times4},\qquad \mathbf{n}(\mathbf{x})=\begin{bmatrix}\mathbf{n}'(\mathbf{x})\\ 0\end{bmatrix}\in\mathbb{R}^4.\tag{13}$$

矩阵 $\mathbf{M}'(\mathbf{x})$ 可显式表示为

$$\mathbf{M}'(\mathbf{x})=\left[\frac{1}{m}\hat{\mathbf{k}}_b\quad f\left[\mathbf{E}^{-1}\hat{\mathbf{e}}_1\right]_\times\hat{\mathbf{k}}_b\quad f\left[\mathbf{E}^{-1}\hat{\mathbf{e}}_2\right]_\times\hat{\mathbf{k}}_b\quad f\left[\mathbf{E}^{-1}\hat{\mathbf{e}}_3\right]_\times\hat{\mathbf{k}}_b\right],\tag{14}$$

漂移项$\mathbf{n}'(\mathbf{x})$ 为

$$\mathbf{n}'(\mathbf{x})=-f\left[\mathbf{E}^{-1}\dot{\mathbf{E}}\boldsymbol{\omega}\right]_\times\hat{\mathbf{k}}_b+2\frac{\dot{T}}{m}\left[\boldsymbol{\omega}\right]_\times\hat{\mathbf{k}}_b+f\left[\boldsymbol{\omega}\right]_\times\left[\boldsymbol{\omega}\right]_\times\hat{\mathbf{k}}_b,\tag{15}$$

其中

$$f=\frac{mg+T}{m}.\tag{16}$$

这里，$[\cdot]_\times$ 表示与叉积对应的反对称矩阵，$\hat{}\mathbf{e}_i$ 表示 $\mathbb{R}^3$中的第 i个标准基向量，即，

$$\hat{\mathbf{e}}_1=[1\ 0\ 0]^\top,\qquad \hat{\mathbf{e}}_2=[0\ 1\ 0]^\top,\qquad \hat{\mathbf{e}}_3=[0\ 0\ 1]^\top.$$

C. 控制友好的离散时间动力学

外部输入可简洁地写为

$$\mathbf{s}=\mathbf{H}(\mathbf{x})\mathbf{k},\tag{17}$$

其中 $\mathbf{H}$： $\mathbb{R}^{14}\to\mathbb{R}^{4\times14}$。根据式（13），虚拟控制输入为

$$\mathbf{u}=\mathbf{M}^{-1}(\mathbf{H}\mathbf{k}-\mathbf{n}),\tag{18}$$

代入式 (4) 后，得到控制仿射形式

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})+\mathbf{G}(\mathbf{x})\mathbf{k},\tag{19}$$

with

$$\mathbf{f}=\mathbf{f}_0-\mathbf{G}_0\mathbf{M}^{-1}\mathbf{n},\qquad \mathbf{G}=\mathbf{G}_0\mathbf{M}^{-1}\mathbf{H}.\tag{20}$$

在采样时间为 Δt的零阶保持下，离散时间动力学为

$$\mathbf{x}_{t+1}=\Phi(\mathbf{x}_t,\mathbf{k}_t),\tag{21}$$

其中，Φ 通过对式 (19) 应用四阶 Runge–Kutta 方法进行数值计算。

IV. 基于DQN的增益调度

我们将离散时间系统上的增益调度问题表述为一个马尔可夫决策过程（MDP）

$$\mathbf{x}_{t+1}=\Phi(\mathbf{x}_t,a_t),\quad a_t\in\{1,\dots,N\},\tag{22}$$

其中，每个动作从有限的稳定化集合中选择一个增益向量

$$\mathcal{A}=\{\mathbf{k}^{(1)},\dots,\mathbf{k}^{(N)}\}\subset\mathcal{K}_1\times\cdots\times\mathcal{K}_{14},\tag{23}$$

其中，$\mathcal{K}_i\subset[k_{i,min},k_{i,max}]$ 是用于指定 k_i可取值的离散集合。每个$\mathbf{k}\ \in\mathcal{A}$都会使标称误差动力学渐近稳定。为防止稳定化增益向量之间发生具有破坏稳定作用的快速切换，系统施加了离散驻留时间约束。具体而言，每当选择新的增益向量$\mathbf{k}_t$ 时，必须将其保持不变至少 N_d个采样间隔，即:

$$\mathbf{k}_{t+i}=\mathbf{k}_t,\quad i=1,\dots,N_d.\tag{24}$$

此约束可防止高频切换，并保持所得切换闭环系统的稳定性。

A. 马尔可夫决策过程建模

增益调度问题被表述为由以下元组定义的马尔可夫决策过程

$$(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}),$$

其中，状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、转移核 $\mathcal{P}$ 和奖励函数 $\mathcal{R}$ 定义如下。

状态空间 $\mathcal{S}$： 时刻 t的状态（观测）为

$$\mathbf{o}_t = \begin{bmatrix} \mathbf{x}_t \\ \phi_t \end{bmatrix} \in \mathcal{S}, \tag{25}$$

其中，$\mathbf{x}_t \in \mathbb{R}^{14}$ 表示四旋翼飞行器的状态，且

$$\phi_t = \min(t/T_f, 1)$$

是一个相位变量，用于编码参考轨迹的进展情况。纳入 ϕ_t 可在参考值随时间变化的跟踪任务中保留马尔可夫性质。

动作空间 $\mathcal{A}$： 动作是一个离散索引

$$a_t \in \{1, \dots, N\}, \tag{26}$$

选择增益向量 $\mathbf{k}^{(a_t)} \in \mathcal{A}$，其中

$$\mathcal{A} = \{\mathbf{k}^{(1)}, \dots, \mathbf{k}^{(N)}\} \subset \mathcal{K}_1 \times \cdots \times \mathcal{K}_{14}.$$

每个$\mathbf{k} \in \mathcal{A}$ 都会使标称误差动力学渐近稳定。为防止快速切换导致系统不稳定，本文施加了式（24）给出的驻留时间约束。

转移核 $\mathcal{P}$： 状态转移概率由离散时间闭环动力学（22）决定，其中Φ 通过在零阶保持下对控制仿射系统进行数值积分得到。建模不确定性和未建模的空气动力学效应使转移核 $\mathcal{P}$未知。

奖励函数 $\mathcal{R}$： 阶段奖励对跟踪误差、姿态偏差、角速度、控制 effort 和增益切换进行惩罚：

$$\begin{aligned} r_t = -\Big(& w_r \|\mathbf{e}_r\|^2 + w_v \|\mathbf{e}_v\|^2 + w_\eta \|\boldsymbol{\eta}\|^2 + w_\omega \|\boldsymbol{\omega}\|^2 \Big) \\ & - w_u \|\mathbf{u}_t\|^2 - w_s \mathbb{I}[a_t \neq a_{t-1}]. \end{aligned} \tag{27}$$

其中 w_r、 w_v、 w_η、 w_ω、 w_u和 w_s 均为正的缩放权重。

B. 深度Q网络近似

令 $Q^\pi(\mathbf{o}\ a)$ 表示动作价值函数。最优 Q 函数满足贝尔曼方程

$$Q^\star(\mathbf{o}_t, a_t) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q^\star(\mathbf{o}_{t+1}, a') \right]. \tag{28}$$

为近似 $Q^\star$，我们采用深度 Q 网络（DQN）

$$Q_\theta : \mathbb{R}^{15} \to \mathbb{R}^{|\mathcal{A}|}.$$

该网络由两个带有 ReLU 激活函数的全连接隐藏层组成，后接一个线性输出层。这种架构在保持计算效率的同时，具备足够的表达能力来近似非线性动作价值函数。

结构化增益参数化： 通过在一个小范围网格上缩放标称极点位置，并纳入多个偏航极点对，构造出一个鲁棒的离散动作集。由此得到的增益表定义了一组稳定控制器；表 I 中的上下界 $k_{i,\min}$ 和 $k_{i,\max}$ 分别是该表各分量的最小值和最大值。按此构造，每个允许的动作都对应一个在保持不变时能够稳定标称误差动力学的控制器，同时为在不确定性下进行基于强化学习的性能自适应提供了足够的变化范围。

为降低增益选择问题的维度，我们利用平移误差动力学中对称的积分器链结构，并在每个导数阶次上对 x、 y 和 z 通道的增益施加相等约束：

$$k_{(3l-1)+1} = k_{(3l-1)+2} = k_{3l}, \qquad l = 1, 2, 3, 4. \tag{29}$$

这些增益分别对应加加速度级、加速度级、速度级和位置级反馈增益。由于偏航动力学具有不同的二阶特性，偏航增益 (k_1, k_2) 保持独立且固定。

对于每个导数阶次，选取五个可用的增益值，从而得到一个基数为

$$|\mathcal{A}| = 5^4 = 625.$$

因此，深度Q网络输出一个包含 625 个 Q 值的向量，其中第 a 个条目表示选择相应增益向量的价值。

损失函数： 通过最小化时序差分损失来更新网络参数 θ

$$\mathcal{L}(\theta) = \mathbb{E}\Big[\big(Q_\theta(\mathbf{o}_t, a_t) - y_t \big)^2 \Big], \tag{30}$$

其中

$$y_t = r_t + \gamma \max_{a'} Q_{\bar{\theta}}(\mathbf{o}_{t+1}, a'). \tag{31}$$

此处 $Q_{\bar{\theta}}$ 表示一个非周期更新的目标网络。采用经验回放来解除样本之间的相关性并提高训练稳定性。

C. 学习得到的增益调度策略

最终得到的增益调度策略为

$$\pi(\mathbf{o}_t) = \arg\max_a Q_\theta(\mathbf{o}_t, a). \tag{32}$$

由于每个动作都对应一个稳定增益向量，且通过驻留时间对切换进行调节，因此，学习得到的调度器在保持标称稳定性的同时，能够在线调整增益，以补偿模型不确定性和瞬态动力学。

V. 仿真结果

我们在四旋翼飞行器的高保真黑箱仿真中评估所提出的基于深度Q网络的增益调度控制器，该飞行器采用ZYX欧拉姿态和推力/力矩执行。物理状态为 $\mathbf{x} = [\mathbf{r}^\top,\ \mathbf{v}^\top,\ \boldsymbol{\eta}^\top,\ \boldsymbol{\omega}^\top,\ T_{\text{dev}},\ \dot{T}]^\top \in \mathbb{R}^{14}$，其中 $\mathbf{r}$，$\mathbf{v} \in \mathbb{R}^3$ 分别为惯性位置/速度，$\boldsymbol{\eta} = [\phi, \theta, \psi]^\top$ 为欧拉角，$\boldsymbol{\omega} \in \mathbb{R}^3$ 为机体角速度，而T_{dev} 和 T 用于描述推力偏差动力学。仿真中飞行器参数固定为：质量 $m = 1.5$ kg、重力加速度 $g = 9.81$ $\mathrm{m/s^2}$,，惯性矩阵为

表 I 训练环境所用稳定增益表推导出的增益界限。

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_{11}	k_{12}	k_{13}	k_{14}
$k_{i,\min}$	9.8304	24.1920	49.1520	25.6000	47.6160	78.8480	22.4000	32.9600	45.4400	8.0000	9.6000	11.2000	12.0000	8.0000
$k_{i,\max}$	49.7664	122.4720	248.8320	86.4000	160.7040	266.1120	50.4000	74.1600	102.2400	12.0000	14.4000	16.8000	32.0000	12.0000

$I = \mathrm{diag}(0.02,\ 0.02,\ 0.04)\ \mathrm{kgm}^2$。积分步长为 $\Delta t = 0.01\ \mathrm{s}$，每个回合时长为 10 s。参考轨迹通过在 $t \in [0, T_f]$ 上进行平滑的五次时间缩放生成，$T_f = 5\ \mathrm{s}$；对于 $t > T_f$，期望位置保持不变，为$\mathbf{r}_d(T_f)$，期望速度、加速度、加加速度和加加加速度均设为零。

深度Q网络的观测由 14 维物理状态和一个标量相位变量 $\min(t/T_f, 1)$ 拼接而成，构成 15 维输入。在每个步骤中，智能体选择一个离散动作，该动作对应于预先计算的稳定增益向量 $\mathbf{k} = [k_1, \ldots, k_{14}]^\top$（来自有限表）。为缩减动作空间并利用平移通道的各向同性假设，在每个导数阶次（位置/速度/加速度/加加速度）上，各轴共用增益；而由于二阶偏航动力学不同，偏航增益则保持独立。驻留时间约束规定所选动作必须保持固定步数，以防止快速切换导致系统失稳。

图 2 展示了训练后的深度 Q 网络在一个具有代表性的运行轨迹中选择的增益（仅绘制共享平动增益）。策略最初选择较激进的反馈，以快速抑制较大的初始跟踪误差；随后，当状态接近参考值且回合进入 post-T_f 保持阶段时，策略切换至低增益模式。

图 3 展示了由此产生的外部误差状态演化，其中 (x, y, z) 中位置、速度、加速度和加加速度层级的误差分量迅速衰减至原点附近，偏航通道也保持良好调节。这种表现符合增益调度的预期作用：远离目标时施加较强的控制作用，接近收敛时则降低控制 effort。

为将学习到的调度与实际运动联系起来，图 4 对比了惯性位置 $\mathbf{r}(t)$ 和期望轨迹 $\mathbf{r}_d(t)$，并在 T_f 处标出垂直标记。四旋翼飞行器沿五次多项式轨迹段跟踪，并在 T_f 后平稳地稳定在最终悬停位置。图 5 给出了欧拉角，证实瞬态期间姿态偏移较小，并衰减至接近零的稳态。

图 6 展示了仿真¨中使用的物理控制信号：推力二阶导数指令 T 和机体力矩 $\boldsymbol{\tau} = [\tau_x, \tau_y, \tau_z]^\top$。正如预期，最大控制幅值出现在回合¨初期，此时跟踪误差最大；随着轨迹跟踪到位，控制幅值随之减小。最后，图 7 中的每步奖励从初始修正阶段的较大负值提升至接近零，表明随着状态收敛，策略减少了不必要的控制 effort，成功实现了基于所学习调度的稳定化与跟踪。

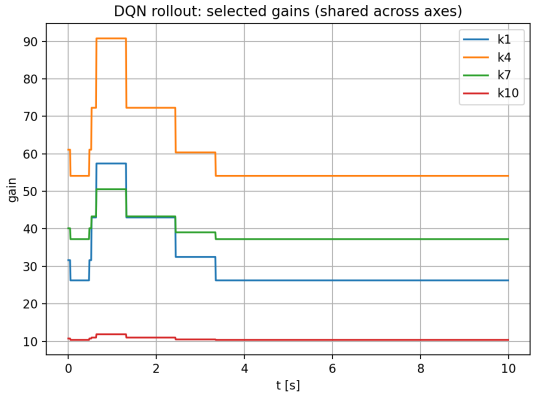


图2. 深度Q网络运行轨迹：各轴共享的选定平动反馈增益（展示代表性增益）。策略在初始瞬态期间提高增益，并随着跟踪误差减小而降低增益。

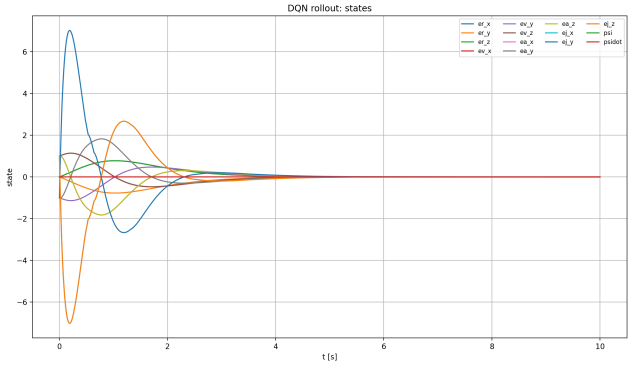


图3. 深度Q网络运行轨迹：外部误差状态。平动误差分量（位置/速度/加速度/加加速度）逐渐收敛至原点，同时偏航通道保持受控，表明在学习得到的增益调度下，闭环行为稳定。

VI. 结论

一种用于四旋翼飞行器轨迹跟踪的基于深度Q网络的增益选择策略已开发完成，并在非线性仿真中进行了评估。通过将动作空间限制为稳定化增益向量并强制采用共享平动增益，该方法在保持与底层动力学结构一致性的同时实现了降维。学习策略展现出自适应行为，在跟踪误差较大时提高反馈控制能力，并在系统收敛时降低增益，从而在性能与控制 effort 之间取得平衡。仿真结果证实了稳定闭环运行、精确的参考轨迹跟踪、有界的控制输入，以及终止时间后平滑的悬停稳定。这些结果表明，强化学习可以与结构化反馈架构相结合，以实现自适应、注重安全的

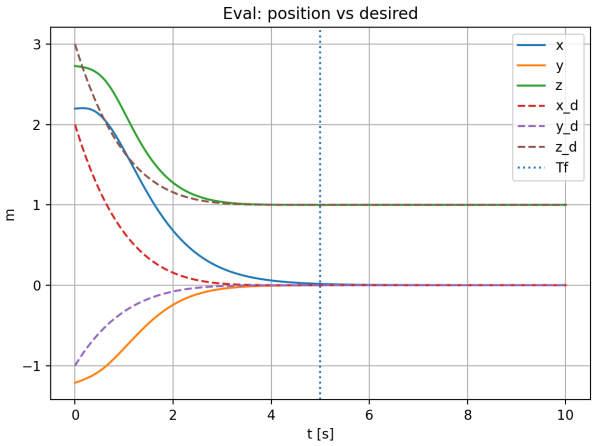


图 4. 物理评估：惯性位置与期望位置。虚线标示 T_f；对于 $t > T_f$，参考值保持在$\mathbf{r}_d(T_f)$，四旋翼飞行器最终悬停。

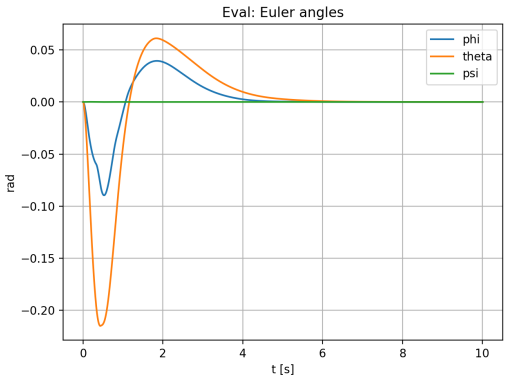


图 5. 物理评估：欧拉角（ϕ，θ，ψ）。随着跟踪逐渐收敛，姿态偏移始终较小，并逐渐衰减至接近零。

性能，同时不牺牲可解释性或稳定性保证。

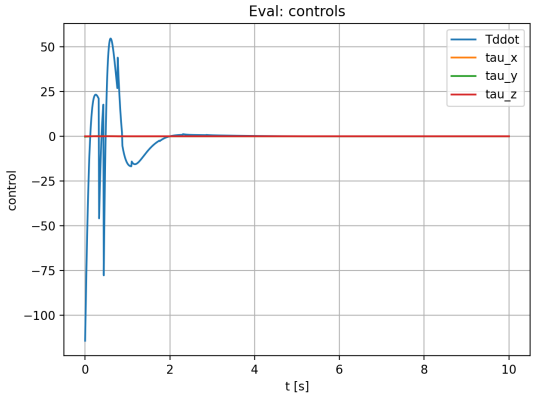


图 6. 物理评估：控制输入。推力二阶导数 $\ddot{T}$ 指令 T 和机体力矩 $\boldsymbol{\tau}$ 在初始瞬态期间最大，并随着状态接近参考值而减小。

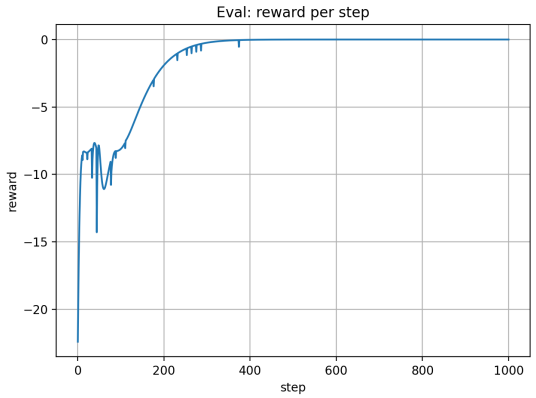


图 7. 物理评估：每步奖励。随着回合中跟踪误差和控制 effort 逐渐减小，奖励趋近于零。

参考文献

[1] T. Lee、M. Leok 和 N. H. McClamroch，“基于几何的四旋翼无人机在 SE(3) 上的跟踪控制”，载于第*49*届 *IEEE* 决策与控制会议（*CDC*），2010年，第 5420–5425 页。[2] S. Bouabdallah、P. Murrieri 和 R. Siegwart，“室内微型四旋翼无人机的设计与控制”，载于*IEEE* 国际机器人与自动化会议（*ICRA*），2004年，第 4393–4398 页。[3] R. Mahony、V. Kumar 和 P. Corke，“多旋翼飞行器：四旋翼无人机的建模、估计与控制”，*IEEE*《机器人与自动化杂志》，第 19 卷，第 3 期，第 20–32 页，2012年。[4] A. Tayebi 和 S. McGilvray，“垂直起降四旋翼飞机的姿态稳定”，*IEEE*《控制系统技术汇刊》，第 14 卷，第 3 期，第 562–571 页，2006年。[5] D. Mellinger 和 V. Kumar，“四旋翼无人机的最小加加速度轨迹生成与控制”，载于*IEEE* 国际机器人与自动化会议（*ICRA*），2011年，第 2520–2525 页。[6] A. E. Asslouj 和 H. Rastgoftar，“采用无欧拉角的基于平坦性的控制实现四旋翼飞行器跟踪”，载于欧洲控制会议论文集（*ECC*）。布加勒斯特，罗马尼亚：EUCA，2023年。[7] V. Mnih、K. Kavukcuoglu、D. Silver等，“通过深度强化学习实现人类水平控制”，《自然》，第 518 卷，533，2015年 7540，529 no. pp. – 。[8] D. Lee、H. Jin Kim 和 S. Sastry，“四旋翼直升机的反馈线性化与自适应滑模控制”，《国际控制、自动化与系统杂志》，第 7 卷，第 3 期，第 419–428 页，2009年。

[9] T. Lee、M. Leok 和 N. H. McClamroch，“基于几何的四旋翼无人机在 SE(3) 上的跟踪控制”，载于第*49*届 *IEEE* 决策与控制会议（*CDC*），2010年，第 5420–5425 页。[10] A. Milhim 、 Y. Zhang 和 C.-A. Rabbath，“ 用于四旋翼无人机容错控制的基于增益调度的 PID 控制器”，载于*AIAA infotech@ aerospace 2010*年，2010年，第 3530 页。[11] J. Qiao、Z. Liu 和 Y. Zhang，“ 用于四旋翼无人机路径跟踪与容错控制的基于增益调度的 PID 控制方法”，《国际机械工程与机器人研究期刊》，第 7 卷，第 4 期，第 401–408 页，2018年。[12] Q. Li、J. Qian、Z. Zhu、X. Bao、M. K. Helwa 和 A. P. Schoellig，“用于改善四旋翼无人机即兴轨迹跟踪的深度神经网络”，载于*2017*年 *IEEE* 国际机器人与自动化会议（*ICRA*），2017年，第 5183–5189 页。[13] S. Zhou、M. K. Helwa 和 A. P. Schoellig，“一种基于逆模型的学习方法，用于改善具有非最小相位动力学的机器人即兴轨迹跟踪”，《IEEE 机器人与自动化快报》，第 3 卷，第 3 期，第 1663–1670 页，2018年。[14] A. Sarabakha 和 E. Kayacan，“利用专家知识在线深度学习，以改善无人机轨迹跟踪”，载于*2019*年国际机器人与自动化会议（*ICRA*），2019年，第 7727–7733 页。[15] K. Huang、R. Rana、A. Spitzer、G. Shi 和 B. Boots，“DATT：用于四旋翼控制的深度自适应轨迹跟踪”，2023年。 [在线]。网址：https://arxiv.org/abs/2310 09053 。R. Ferede、G. de Croon、C. [16] De Wagter 和 D. Izzo，“基于端到端神经网络的四旋翼飞行器最优控制”，《机器人与自主系统》，第 172 卷，第 104588 页，2024年。 [在线]。网址：https://www.sciencedirect.com/science/article/pii/S0921889023002270 [17] W. Koch、R. Mancuso、R. West 和 A. Bestavros，“用于无人机姿态控制的强化学习”，第 3 卷，第 2 期，2019年2月。

[18] E. Kaufmann、A. Loquercio、R. Ranftl、M. M¨uller、V. Koltun 和 D. Scaramuzza，“深度无人机特技机器人：科学与系统（RSS）”，美国俄勒冈州，第 12–17 页，2020年。[19] S. Lupashin、A. Sch¨ollig、M. Sherback 和 R. D’Andrea，“高速四旋翼多次翻转的简单学习策略”，载于*2010*年 *IEEE* 国际机器人与自动化会议，2010年，第 1642–1648 页。[20]J. Achiam、D. Held、A. Tamar 和 P. Abbeel，“约束策略优化”，载于第*34*届国际机器学习会议论文集，《机器学习研究论文集》系列，D. Precup 和 Y. W. Teh 编，第 70 卷。PMLR，2017年8月6–11日，第 22–31 页。 [在线]。网址：https: //proceedings.mlr.press/v70/achiam17a.html [21] Y. Chow、O. Nachum、E. Duenez-Guzman 和 M. Ghavamzadeh，“一种基于李雅普诺夫的方法实现安全强化学习”，神经信息处理系统进展，第 31 卷，2018年。[22] V. Mnih、K. Kavukcuoglu、D. Silver、A. A. Rusu、J. Veness、M. G. Bellemare、A. Graves、M. Riedmiller、A. K. Fidjeland、G. Ostrovski 等，“通过深度强化学习实现人类水平控制”，《自然》，第 518 卷，第 7540 期，第 529–533 页，2015年。[23] H. Rastgoftar 和 I. V. Kolmanovsky，“大规模多四旋翼系统的基于安全仿射变换的制导”，*IEEE*网络系统控制汇刊，第 8 卷，第 2 期，第 640–653 页，2021年。[24] A. El Asslouj 和 H. Rastgoftar，“使用无欧拉角、基于平坦性的控制进行四旋翼跟踪”，载于*2023*年欧洲控制会议（*ECC*）。IEEE，2023年，第 1–6 页。