控制与决策
Control and Decision

基于增量式Q学习的固定翼无人机跟踪控制性能优化

赵振根, 程磊

引用本文:
赵振根,程磊. 基于增量式Q学习的固定翼无人机跟踪控制性能优化[J]. 控制与决策, 2024, 39(2): 391−400.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0708

基于增量式Q学习的固定翼无人机跟踪控制性能优化

赵振根†, 程 磊

(南京航空航天大学 自动化学院，南京 211106)

摘 要: 针对固定翼无人机纵向控制的高性能需求,提出一种控制系统性能优化结构. 该结构包括一个使系统稳定的标称控制器和一个参与性能优化的增量式控制器. 控制系统增量式的实现不会改变原有的控制系统,而是仅对标称控制系统做控制输入的补偿与控制性能的优化. 基于Q学习理论进行增量式控制器设计,针对状态信息完全可获得的系统,设计一种基于状态反馈的增量式Q学习算法. 当状态信息不能完全获得时,利用系统输入、输出和参考信号数据,设计一种基于输出反馈的增量式Q学习算法. 两种增量式控制器均是在数据驱动环境下自适应学习增量式控制律,无需提前知道系统动力学模型以及标称控制器的控制增益. 此外,证明了增量式Q学习方法在满足持续激励条件的激励噪声下,对Q函数贝尔曼方程的求解没有偏差. 最后,通过对F-16飞行器纵向模型实例的仿真验证该方法的有效性.

关键词: 强化学习；Q学习；增量式控制；性能优化；跟踪控制；无人机

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0708

引用格式: 赵振根,程磊. 基于增量式Q学习的固定翼无人机跟踪控制性能优化[J]. 控制与决策, 2024, 39(2): 391-400.

Performance optimization for tracking control of fixed-wing UAV with incremental Q-learning

ZHAO Zhen-gen†, CHENG Lei

(College of Automation Engineering，Nanjing University of Aeronautics and Astronautics，Nanjing 211106，China)

Abstract: Aiming at the high performance requirements of longitudinal control of a fixed-wing unmanned aerial vehicle (UAV), a performance optimization structure of the control system is proposed. This structure includes a nominal controller that stabilizes the system and an incremental controller that participates in performance optimization. The incremental implementation of the control system does not change the original control system, but compensates the control input and optimizes the control performance for the nominal control system exclusively. Based on the Q-learning theory, the incremental controller is designed. For the system with completely available state information, an incremental Q-learning algorithm based on state feedback is developed. When the state information cannot be obtained completely, an incremental Q-learning algorithm based on output feedback is designed by using the system input, output and reference trajectory data. Both incremental controllers learn incremental control laws adaptively in the data-driven environment without the need for system dynamics model and the control gain of the nominal controller. In addition, it is proved that the incremental Q-learning method has no bias in solving the Q-function Bellman equation under the excitation noise. Finally, the effectiveness of the method is verified by the simulation of an example of the longitudinal model of the F-16 aircraft.

Keywords: reinforcement learning; Q-learning; incremental control; performance optimization; tracking control; UAV

0 引 言

近些年来,无人机逐步成为当今航空航天的研究热点. 无人机的应用领域广泛、型号多样,在人类生活生产中都得到了极大的应用与普及,例如航拍摄影、无人投递、侦察探测等. 但随着无人机的任务多样化,结构复杂化,无人机也面临越来越多的挑战. 飞行控

收稿日期: 2022-04-27；**录用日期:** 2022-10-10.

基金项目: 国家自然科学基金项目(62003161)；江苏省自然科学基金项目(BK20190399)；中国博士后科学基金项目(2021M701701).

责任编委: 侯忠生.

†通讯作者. E-mail: zhaozhengen@nuaa.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

制系统作为无人机的“大脑”,不仅需要提高对环境的适应性,还要能够处理自身复杂的动力学模型和自身模型的不确定性,从而使无人机表现出一定的自主性与智能性. 目前,飞行控制系统已经得到了广泛的研究,并取得了许多成果.

传统控制方法发展已相对成熟,例如PID控制、切换控制、动态逆控制、鲁棒控制和自适应控制等[1]. 文献[2]针对无人机飞行过程中出现的参数摄动和外部扰动的影响,提出在经典PID控制的基础上加入H_∞鲁棒控制的混合控制方式. 文献[3]将非线性动态逆控制与反步法相结合,对高超音速飞机纵向运动模型设计飞行控制系统. 该系统以非线性动态逆控制作为控制内环,以反步法作为控制外环从而保证系统的全局稳定以及抑制不确定参数的扰动.

传统无人机控制设计方法大多依赖于系统动力学模型,无模型的PID控制的参数整定一般需要用试凑法反复调试才能实现,存在调参繁琐等问题. 强化学习对无模型控制问题提供了一种解决思路. 文献[4]提出了一种利用强化学习技术训练的神经网络来控制四旋翼的方法,所使用策略网络对阶跃响应能够相对准确地做出反应. 文献[5]将基于表格的Q学习应用于实际的四旋翼无人机以实现悬停控制. 文献[6]提出了一种基于深度强化学习算法的鲁棒控制方案,该方案利用神经网络实现了状态到控制指令的端对端设计. 对于无人机非线性系统的处理,代尔夫特理工大学研究团队提供了一些思路,文献[7-8]利用增量式近似动态规划处理非线性系统,它结合了线性近似动态规划方法和增量控制技术,通过系统辨识得到系统的增量模型,并在线设计最优控制器,最后对导弹俯仰平面,某固定翼飞机的高度控制以及航天器姿态控制的仿真验证了该方法的有效性.

上述关于无人机强化学习控制的文献大多利用系统的输入输出数据,在系统模型未知的情况下学习整体控制器. 但是,针对无人机系统存在标称控制器的情形,如何利用数据驱动的方法实现控制性能在线优化的研究较少. 本文将Q学习方法[9-10]和增量式控制结构相结合,提出一种增量式Q学习的无人机跟踪控制性能优化方法,在不修改预先设计的控制系统情况下,实现无人机控制性能优化. 预先设计的控制系统为针对标称系统设计的控制器,由于模型误差和工作点变化等原因,其控制性能不一定最优,增量式控制器的作用为补偿控制器增益从而优化系统性能.

本文主要工作如下: 1) 提出一种无人机控制系统性能优化结构,不改变原有无人机控制器,利用Q学习方法设计增量式控制器对标称控制器进行性能优化; 2) 分别针对无人机状态可知和状态不完全可知情况,设计基于状态反馈的增量式Q学习控制器与基于输出反馈的增量式Q学习控制器,并给出相应的策略迭代算法; 3) 在激励噪声满足持续激励条件下,证明了增量式Q学习方法进行参数估计的无偏性.

1 无人机建模与控制系统描述

1.1 无人机动力学模型

本文的仿真对象为F-16型固定翼飞行器. 假设飞行器的质量恒视为常数,并且地面坐标系视为惯性系. 在水平无侧滑和滚转的条件下,在地面坐标系中建立无人机的动力学方程,F-16的纵向平面模型描述[11]为

$$\begin{cases}\dot V=\dfrac{T\cos\alpha-D}{m}+g(-\cos\alpha\sin\theta+\sin\alpha\cos\theta),\\ \dot\alpha=\dfrac{-T\sin\alpha-L}{mV}+q+\dfrac{g(\sin\alpha\sin\theta+\cos\alpha\cos\theta)}{V},\\ \dot\theta=q,\\ \dot q=\dfrac{M}{I_y}.\end{cases}\tag{1}$$

其中: V、α、θ和q分别表示无人机的速度、迎角、俯仰角和俯仰角速率; m、g、T和I_y分别表示无人机的质量、重力加速度、发动机推力和转动惯量; M、D和L分别为俯仰力矩、阻力和升力,表达式如下:

$$\begin{cases}M=\bar qS\bar cC_m,\\ D=\bar qSC_x,\\ L=\bar qSC_z,\end{cases}\tag{2}$$

$\bar q$、S和$\bar c$分别表示动压、机翼面积和机翼平均气动弦长, C_x、C_z和C_m分别表示阻力系数、升力系数和俯仰力矩系数.

本文使用F-16飞行器的低保真空气动力学模型. 低保真模型忽略了飞机的前缘襟翼和速度制动器的作用. 纵向模型的力和力矩系数计算如下:

$$\begin{cases}C_x=C_x(\alpha,\delta_e)+\dfrac{\bar cq}{2V}C_{xq},\\ C_z=C_z(\alpha,\delta_e)+\dfrac{\bar cq}{2V}C_{zq},\\ C_m=C_m(\alpha,\delta_e)+\dfrac{\bar cq}{2V}C_{mq}+C_z(x_{cgr}-x_{cg}),\end{cases}\tag{3}$$

其中$C_x(\alpha,\delta_e)$, $C_z(\alpha,\delta_e)$, $C_m(\alpha,\delta_e)$, C_{xq}, C_{zq}和C_{mq}为气动参数并以插值表[12]形式给出. 这些空气动力学数据在$-10^\circ\leqslant\alpha\leqslant45^\circ$, $0.1\leqslant M_a\leqslant0.6$飞行范围有效, M_a为马赫数. x_{cgr}, x_{cg}和δ_e分别为参考重心位

置、重心位置和升降舵偏转角. F-16飞行器的相关参数参见文献[12].

1.2　控制系统描述

将飞行器纵向模型(1)配平线性化后,考虑以下线性离散动力学系统:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k. \end{aligned} \tag{4}$$

其中: $x_k = [V\ \ \alpha\ \ \theta\ \ q]^{\mathrm{T}} \in \boldsymbol{R}^n$ 为系统状态, $u_k = \delta_e \in \boldsymbol{R}^m$ 为控制输入, $y_k = \alpha \in \boldsymbol{R}^p$ 为系统输出. 向量维数 $n = 4$, $m = 1$, $p = 1$. 参数矩阵 $A \in \boldsymbol{R}^{n\times n}$, $B \in \boldsymbol{R}^{n\times m}$, $C \in \boldsymbol{R}^{p\times n}$. 假设 (A, B) 可控, (A, C) 可观. 参考信号可由以下动力学系统生成:

$$\begin{aligned} x_{k+1}^r &= Fx_k^r, \\ y_k^r &= C_r x_k^r. \end{aligned} \tag{5}$$

其中: $x_k^r \in \boldsymbol{R}^{n_1}$ 为参考系统状态, $y_k^r \in \boldsymbol{R}^p$ 为参考输出,参数矩阵 $F \in \boldsymbol{R}^{n_1\times n_1}$, $C_r \in \boldsymbol{R}^{p\times n_1}$. 结合式(4)和(5)可得增广系统[13]如下:

$$X_{k+1} = TX_k + B_1 u_k. \tag{6}$$

其中: $X_k = \begin{bmatrix} x_k \\ x_k^r \end{bmatrix} \in \boldsymbol{R}^{n+n_1}$, $T = \begin{bmatrix} A & 0 \\ 0 & F \end{bmatrix} \in \boldsymbol{R}^{(n+n_1)\times(n+n_1)}$, $B_1 = \begin{bmatrix} B \\ 0 \end{bmatrix} \in \boldsymbol{R}^{(n+n_1)\times m}$. 跟踪问题是将原系统与参考信号系统结合为增广系统,对增广系统进行研究并设计控制律.

为了优化系统性能,设计二次型形式的性能指标如下:

$$\begin{aligned} V(k) = \sum_{i=k}^{\infty} \gamma^{i-k}[(y_i - y_i^r)^{\mathrm{T}} Q (y_i - y_i^r) + u_i^{\mathrm{T}} R u_i], \\ Q \geqslant 0,\ R > 0. \end{aligned} \tag{7}$$

本文采用控制系统结构如图1所示,控制器包括预先设计的用来负责系统稳定的标称控制器和负责性能优化的增量式控制器. 增量式的性能优化不需改变标称控制器,而是产生一个增量式控制信号加入到控制输入端. 增量式控制器采用Q学习得到,该方法无需系统动力学模型,也无需预知预先设计的标称控制器增益大小. 在优化后,系统预先设计控制律与补偿控制律总和达到最优. 设计方法有如下两种:

1) 在系统全状态信息可获得的情况下,设计基于状态反馈的增量式Q学习控制器来完成系统对参考信号跟踪问题的性能优化.

2) 在系统全状态信息不能够获得的情况下,利用系统输入、输出数据信息重构状态变量,设计基于输出反馈的增量式Q学习控制器优化跟踪性能.

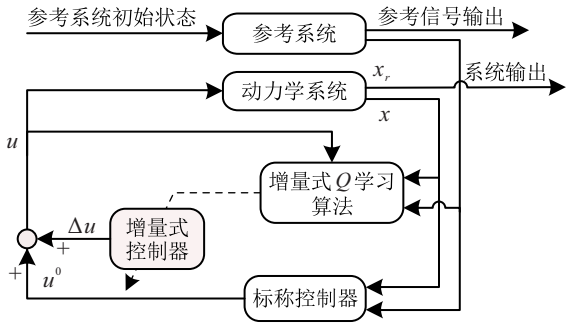

图 1　总体控制框架

两种补偿控制器设计方法均基于Q学习理论,无需知道系统动力学模型和预先设计的控制器增益大小. 在优化后,标称控制律与补偿控制律总和达到最优.

2　增量式Q学习状态反馈控制器设计

本节将Q学习方法与增量式控制相结合,介绍一种增量式Q学习方法. 通过Q学习去补偿现有控制器增益,从而对模型未知系统进行性能优化.

2.1　增量式Q学习与性能优化

考虑一种控制器结构,该控制器由预先设计的标称控制器和增量式控制器组成[14],具体表示如下:

$$\begin{cases} u_k = u_k^0 + \Delta u_k, \\ u_k^0 = K_0 X_k, \\ \Delta u_k = \Delta K X_k. \end{cases} \tag{8}$$

其中: $K_0 \in \boldsymbol{R}^{m\times(n+n_1)}$ 表示预先设计的控制器增益,它的作用是使系统稳定,但在给定的性能指标下是非最优控制; $u_k^0 \in \boldsymbol{R}^m$ 表示标称控制器, $\Delta u_k \in \boldsymbol{R}^m$ 表示补偿控制器; $\Delta K \in \boldsymbol{R}^{m\times(n+n_1)}$ 表示需要设计的增量式控制器增益. 性能函数(7)等价为

$$V(X_k) = \sum_{i=k}^{\infty} \gamma^{i-k}(X_i^{\mathrm{T}} Q_1 X_i + u_i^{\mathrm{T}} R u_i). \tag{9}$$

其中: $Q_1 = \begin{bmatrix} C^{\mathrm{T}}QC & -C^{\mathrm{T}}QC_r \\ -C_r^{\mathrm{T}}QC & C_r^{\mathrm{T}}QC_r \end{bmatrix}$, γ 为折扣因子且 $0 < \gamma \leqslant 1$. $V(X_k)$ 的贝尔曼方程形式如下:

$$V(X_k) = X_k^{\mathrm{T}} Q_1 X_k + u_k^{\mathrm{T}} R u_k + \gamma V(X_{k+1}). \tag{10}$$

对于线性二次型问题, $V(X_k)$ 的二次型形式为

$$V(X_k) = X_k^{\mathrm{T}} P X_k. \tag{11}$$

其中: P 为黎卡提方程的解,有

$$\begin{aligned} &P = Q_1 + \bar{T}^{\mathrm{T}} P \bar{T} - \bar{T}^{\mathrm{T}} P \bar{B}_1 (R + \bar{B}_1^{\mathrm{T}} P \bar{B}_1)^{-1} \bar{B}_1^{\mathrm{T}} P \bar{T}; \\ &\bar{T} = \gamma^{1/2} T,\ \bar{B}_1 = \gamma^{1/2} B_1. \end{aligned} \tag{12}$$

式(12)具有唯一解的充分必要条件是动力学系统 $(\gamma^{1/2}T, \gamma^{1/2}B_1)$ 是稳定的.

定义Q函数如下:

$$Q(X_k,u_k)=X_k^{\mathrm{T}}Q_1X_k+u_k^{\mathrm{T}}Ru_k+\gamma V(X_{k+1}). \tag{13}$$

Q函数是依赖于状态和控制的函数.

将式(11)代入(13),有

$$\begin{aligned}&Q(X_k,u_k)=\\&X_k^{\mathrm{T}}Q_1X_k+u_k^{\mathrm{T}}Ru_k+\gamma X_{k+1}^{\mathrm{T}}PX_{k+1}=\\&X_k^{\mathrm{T}}Q_1X_k+u_k^{\mathrm{T}}Ru_k+\gamma(TX_k+B_1u_k)^{\mathrm{T}}P\times\\&(TX_k+B_1u_k)=\\&X_k^{\mathrm{T}}Q_1X_k+(K_0X_k+\Delta u_k)^{\mathrm{T}}R(K_0X_k+\Delta u_k)+\\&\gamma((T+B_1K_0)X_k+B_1\Delta u_k)^{\mathrm{T}}P\times\\&((T+B_1K_0)X_k+B_1\Delta u_k):=Q(X_k,\Delta u_k).\end{aligned}$$

整理得

$$Q(X_k,\Delta u_k)=\begin{bmatrix}X_k\\ \Delta u_k\end{bmatrix}^{\mathrm{T}}H\begin{bmatrix}X_k\\ \Delta u_k\end{bmatrix}. \tag{14}$$

其中

$$T_F=T+B_1K_0,$$

$$H=\begin{bmatrix}Q_1+{K_0}^{\mathrm{T}}RK_0+\gamma{T_F}^{\mathrm{T}}PT_F & {K_0}^{\mathrm{T}}R+\gamma{T_F}^{\mathrm{T}}PB_1\\ RK_0+\gamma{B_1}^{\mathrm{T}}PT_F & R+\gamma{B_1}^{\mathrm{T}}PB_1\end{bmatrix}.$$

为了使Q函数(13)最小化,执行$\dfrac{\partial Q(X_k,\Delta u_k)}{\partial\Delta u_k}=0$,可得最优控制律为

$$\begin{aligned}&\Delta u_k=\\&-(R+\gamma{B_1}^{\mathrm{T}}PB_1)^{-1}(RK_0+\gamma{B_1}^{\mathrm{T}}PT_F)X_k.\end{aligned} \tag{15}$$

式(15)等价于

$$\begin{aligned}&\Delta u_k=\\&-(R+\gamma{B_1}^{\mathrm{T}}PB_1)^{-1}(RK_0+\\&\gamma{B_1}^{\mathrm{T}}PT+\gamma{B_1}^{\mathrm{T}}PB_1K_0)X_k=\\&-\left(\frac{R}{\gamma}+{B_1}^{\mathrm{T}}PB_1\right)^{-1}{B_1}^{\mathrm{T}}PTX_k-K_0X_k.\end{aligned} \tag{16}$$

即总的控制输入为

$$\begin{aligned}&u_k=K_0X_k+\Delta u_k=\\&-\left(\frac{R}{\gamma}+{B_1}^{\mathrm{T}}PB_1\right)^{-1}{B_1}^{\mathrm{T}}PTX_k.\end{aligned} \tag{17}$$

结合式(14),将核矩阵H表示为如下形式(即增量式控制器的无模型实现方式得以实现):

$$Q(X_k,\Delta u_k)=\begin{bmatrix}X_k\\ \Delta u_k\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}H_{XX} & H_{Xu}\\ H_{uX} & H_{uu}\end{bmatrix}\begin{bmatrix}X_k\\ \Delta u_k\end{bmatrix}.$$

其中:$H=H^{\mathrm{T}}\in\boldsymbol{R}^{l\times l}$, $l=n+n_1+m$; $H_{XX}=Q_1+{K_0}^{\mathrm{T}}RK_0+\gamma{T_F}^{\mathrm{T}}PT_F\in\boldsymbol{R}^{(n+n_1)\times(n+n_1)}$, $H_{Xu}={K_0}^{\mathrm{T}}R+\gamma{T_F}^{\mathrm{T}}PB_1\in\boldsymbol{R}^{m\times(n+n_1)}$, $H_{uu}=R+\gamma{B_1}^{\mathrm{T}}PB_1\in\boldsymbol{R}^{m\times m}$. 则对应的无模型的增量式控制律为

$$\Delta u_k=-{H_{uu}}^{-1}H_{uX}X_k. \tag{18}$$

2.2 数据驱动实现与算法

对于式(8)所示控制器结构,Q函数对应的贝尔曼方程为

$$\begin{aligned}&z_k^{\mathrm{T}}Hz_k=\\&X_k^{\mathrm{T}}Q_1X_k+u_k^{\mathrm{T}}Ru_k+\gamma z_{k+1}^{\mathrm{T}}Hz_{k+1}=\\&X_k^{\mathrm{T}}(Q_1+K_0^{\mathrm{T}}RK_0)X_k+\Delta u_k^{\mathrm{T}}R\Delta u_k+\\&\Delta u_k^{\mathrm{T}}RK_0X_k+X_k^{\mathrm{T}}K_0^{\mathrm{T}}R\Delta u_k+\gamma z_{k+1}^{\mathrm{T}}Hz_{k+1}=\\&z_k^{\mathrm{T}}S_0z_k+\gamma z_{k+1}^{\mathrm{T}}Hz_{k+1}.\end{aligned} \tag{19}$$

其中

$$z_k=[{X_k}^{\mathrm{T}}\ \ {\Delta u_k}^{\mathrm{T}}]^{\mathrm{T}}\in\boldsymbol{R}^l,$$

$$S_0=\begin{bmatrix}Q_1+\gamma{K_0}^{\mathrm{T}}PK_0 & {K_0}^{\mathrm{T}}R\\ RK_0 & R\end{bmatrix}.$$

Q函数矩阵H是未知的,需要通过Q学习算法学习得到,可通过参数化式(14)来分离H,即

$$Q(X_k,\Delta u_k)=z_k^{\mathrm{T}}Hz_k=\bar{H}^{\mathrm{T}}\bar{z}_k. \tag{20}$$

其中:$\bar{H}\in\boldsymbol{R}^{l(l+1)/2}$, $\bar{z}_k\in\boldsymbol{R}^{l(l+1)/2}$,且满足

$$\begin{aligned}&\bar{H}\triangleq\\&[h_{11},2h_{12},\dots,2h_{1l},h_{22},2h_{23},\dots,2h_{2l},\dots,h_{ll}]^{\mathrm{T}},\\&\bar{z}_k\triangleq[z_1^2,z_1z_2,\dots,z_1z_l,z_2^2,z_2z_3,\dots,z_2z_l,\dots,z_l^2]^{\mathrm{T}}.\end{aligned}$$

通过参数化式(19)后得到如下方程:

$$\bar{H}^{\mathrm{T}}\bar{z}_k=z_k^{\mathrm{T}}S_0z_k+\gamma\bar{H}^{\mathrm{T}}\bar{z}_{k+1}. \tag{21}$$

上述线性方程可以写成

$$\Phi^{\mathrm{T}}\bar{H}=\Omega. \tag{22}$$

其中

$$\begin{aligned}&\Phi=[\bar{z}_k-\gamma\bar{z}_{k+1},\bar{z}_{k+1}-\gamma\bar{z}_{k+2},\dots,\bar{z}_{k+L-1}-\gamma\bar{z}_{k+L}],\\&\Omega=[r_k,r_{k+1},\dots,r_{k+L-1}]^{\mathrm{T}},\ r_k=z_k^{\mathrm{T}}S_0z_k,\\&\Phi\in R^{l(l+1)/2\times L},\ \Omega\in R^L.\end{aligned}$$

式(22)的解为

$$\bar{H}=(\Phi\Phi^{\mathrm{T}})^{-1}\Phi\Omega. \tag{23}$$

在式(21)中,$\bar{H}$是未知的向量,为求解线性方程,需要的样本数据需满足$L\geqslant l(l+1)/2$. 下面提出策略迭代的强化学习算法来学习状态反馈增量式Q学习的补偿控制律.

算法1 状态反馈增量式Q学习策略迭代算法.

1) 初始化. 初始化H^0,选取使系统稳定的标称

控制器u^0,选择稳定的初始控制策略Δu_0,对$j(j=1,2,3,\ldots)$执行如下.

2) 策略评估.收集足够步长的数据解Q函数贝尔曼方程

$$(\bar{H}^j)^{\mathrm{T}}(\bar{z}_k-\gamma\bar{z}_{k+1})=$$
$$(y_k-y_k^r)^{\mathrm{T}}Q(y_k-y_k^r)+(u_k^j)^{\mathrm{T}}Ru_k^j.$$

3) 策略更新.在策略评估得到$\bar{H}$后构造H,计算改进后的控制策略

$$\Delta u_k^{j+1}=-(H_{uu}^j)^{-1}H_{uX}^jX_k.$$

4) 终止条件.当达到收敛条件$\|H^j-H^{j-1}\|\leqslant\varepsilon$时停止迭代,其中$\varepsilon$为一个预先设计的正常数,否则执行$j=j+1$继续迭代.

算法1中策略评估需要收集$X_k,\Delta u_k,X_{k+1}$和Δu_{k+1}的数据求解贝尔曼方程(21),$\bar{H}$的解如式(23)所示. 在策略更新步骤中,通过最小化第j个策略的Q函数,得到一个更好的策略Δu_k^{j+1}. 在策略评估过程中,需要加入激励噪声使得贝尔曼方程(21)有解,激励噪声的作用阐述如下.

注1　通过在控制端加入满足持续激励条件的噪声,使得向量$\bar{z}_k-\gamma\bar{z}_{k+1}$线性无关,从而使得线性方程(22)中的矩阵$\Phi$满秩[15]. 利用最小二乘法,加入持续激励条件的噪声保证了式(23)解的唯一性.

定理1　激励噪声在求解Q函数贝尔曼方程时,参数估计不产生偏差.

证明过程略.

为了保证式(19)的唯一解,在控制输入端加入的激励噪声对Q函数的参数估计没有偏差作用,学习到的增量式最优控制律不存在偏差. 算法1表明,控制性能的优化可以通过增量的形式来实现,而无需变动预先设计好的控制器增益.

注2　通过增量结构可使控制系统的控制器设计更加有效,预先设计的标称控制器可能表现的是非最优性能,它的主要作用是使系统稳定. 而增量式的控制器负责性能优化,它可补偿地产生控制作用,而无需更改原有控制系统.

3　增量式Q学习输出反馈控制器设计

3.1　增量式Q学习与性能优化

在算法1中,需要系统的全状态信息来设计状态反馈控制器,但状态信息在实际情况下可能获取不到. 本节将利用系统的输入、输出和参考信号序列的观测数据改进算法1,改进的策略迭代算法利用观测数据对Q函数进行策略评估与与策略更新. 由文献[16]可知,系统的状态可通过过去时刻测量到的输入、输出和参考信号数据来重构.

$$X_k=[M_u\ \ M_y\ \ M_r]\begin{bmatrix}\bar{u}_{k-1,k-N}\\ \bar{y}_{k-1,k-N}\\ x_{k-N}^r\end{bmatrix}=M\tau_k. \tag{24}$$

其中

$$\bar{u}_{k-1,k-N}=[u_{k-1}^{\mathrm{T}}\ \ u_{k-2}^{\mathrm{T}}\ \ \cdots\ \ u_{k-N}^{\mathrm{T}}]^{\mathrm{T}},$$
$$\bar{y}_{k-1,k-N}=[y_{k-1}^{\mathrm{T}}\ \ y_{k-2}^{\mathrm{T}}\ \ \cdots\ \ y_{k-N}^{\mathrm{T}}]^{\mathrm{T}}.$$

状态重构性表明,在满足系统可观性条件$\mathrm{rank}(V_N)=n$下,增广系统状态X_k可以用过去时刻系统$[k-N,k-1]$段的输入、输出以及参考信号数据来表示. 因此,输出反馈下的控制器设计为

$$\begin{cases}u_k=u_k^0+\Delta u_k,\\ u_k^0=F_0\tau_k,\\ \Delta u_k=\Delta F\tau_k,\end{cases} \tag{25}$$

其中$F_0=[F_u\ \ F_y\ \ F_r]\in\boldsymbol{R}^{m\times(mN+pN+n_1)}$. 利用该增广系统状态$X_k$的新表示形式和增量下输出的形式,可以描述输入输出数据方案下的Q函数,并在此基础上开发数据驱动的增量式Q学习算法. 为此,记$Z_k=[\bar{u}_{k-1,k-N}^{\mathrm{T}}\ \ \bar{y}_{k-1,k-N}^{\mathrm{T}}\ \ (x_{k-N}^r)^{\mathrm{T}}\ \ \Delta u_k^{\mathrm{T}}]\in\boldsymbol{R}^{l_1}$, $l_1=mN+pN+n_1+m$. 将Z_k代入式(13)中,Q函数表达式如下:

$$Q(Z_k)=Z_k^{\mathrm{T}}H_1Z_k. \tag{26}$$

$H_1=H_1^{\mathrm{T}}\in R^{l_1\times l_1}$可以分块为

$$H_1=\begin{bmatrix}H_{\bar{u}\bar{u}}&H_{\bar{u}\bar{y}}&H_{\bar{u}r}&H_{\bar{u}u}\\ H_{\bar{y}\bar{u}}&H_{\bar{y}\bar{y}}&H_{\bar{y}r}&H_{\bar{y}u}\\ H_{r\bar{u}}&H_{r\bar{y}}&H_{rr}&H_{ru}\\ H_{u\bar{u}}&H_{u\bar{y}}&H_{ur}&H_{uu}\end{bmatrix}.$$

其中

$$H_{\bar{u}\bar{u}}=M_u{}^{\mathrm{T}}Q_1M_u+F_u{}^{\mathrm{T}}RF_u+\gamma G_u{}^{\mathrm{T}}PG_u,$$
$$H_{\bar{u}\bar{y}}=M_u{}^{\mathrm{T}}Q_1M_y+F_u{}^{\mathrm{T}}RF_y+\gamma G_u{}^{\mathrm{T}}PG_y,$$
$$H_{\bar{u}r}=M_u{}^{\mathrm{T}}Q_1M_r+F_u{}^{\mathrm{T}}RF_r+\gamma G_u{}^{\mathrm{T}}PG_r,$$
$$H_{\bar{u}u}=F_u{}^{\mathrm{T}}R+\gamma G_u{}^{\mathrm{T}}PB_1,$$
$$H_{\bar{y}\bar{y}}=M_y{}^{\mathrm{T}}Q_1M_y+F_y{}^{\mathrm{T}}RF_y+\gamma G_y{}^{\mathrm{T}}PG_y,$$
$$H_{\bar{y}r}=M_y{}^{\mathrm{T}}Q_1M_r+F_y{}^{\mathrm{T}}RF_r+\gamma G_y{}^{\mathrm{T}}PG_r,$$
$$H_{\bar{y}u}=F_y{}^{\mathrm{T}}R+\gamma G_y{}^{\mathrm{T}}PB_1,$$
$$H_{rr}=M_r{}^{\mathrm{T}}Q_1M_r+F_r{}^{\mathrm{T}}RF_r+\gamma G_r{}^{\mathrm{T}}PG_r,$$
$$H_{ru}=F_r{}^{\mathrm{T}}R+\gamma G_r{}^{\mathrm{T}}PB_1,$$
$$H_{uu}=R+\gamma B_1{}^{\mathrm{T}}PB_1;$$

$H_{\bar{u}\bar{u}}\in\boldsymbol{R}^{mN\times mN}$, $H_{\bar{u}\bar{y}}\in\boldsymbol{R}^{mN\times pN}$, $H_{\bar{u}r}\in\boldsymbol{R}^{mN\times n_1}$,
$H_{\bar{u}u}\in\boldsymbol{R}^{mN\times m}$, $H_{\bar{y}\bar{y}}\in\boldsymbol{R}^{pN\times pN}$, $H_{\bar{y}r}\in\boldsymbol{R}^{pN\times n_1}$,
$H_{\bar{y}u}\in\boldsymbol{R}^{mN\times m}$, $H_{rr}\in\boldsymbol{R}^{n_1\times n_1}$, $H_{ru}\in\boldsymbol{R}^{n_1\times m}$,

$H_{uu}\in \boldsymbol{R}^{m\times m}$, $G_u=TM_u+B_1F_u$,

$G_y=TM_y+B_1F_y$, $G_r=TM_r+B_1F_r$.

式(26)将Q函数表示为输入、输出、参考信号和增量式输入的新形式,最优化控制策略通过$\partial Q(Z_k)/\partial\Delta u_k=0$获得. 利用输入、输出和参考信号序列,可以得到最优的增量式控制器为

$$\Delta u_k=-{H_{uu}}^{-1}[H_{u\bar{u}}\ \ H_{u\bar{y}}\ \ H_{ur}]\begin{bmatrix}\bar{u}_{k-1,k-N}\\ \bar{y}_{k-1,k-N}\\ x^r_{k-N}\end{bmatrix}=$$

$$\Delta F^*\begin{bmatrix}\bar{u}_{k-1,k-N}\\ \bar{y}_{k-1,k-N}\\ x^r_{k-N}\end{bmatrix}. \tag{27}$$

3.2 数据驱动实现与算法

在本节中,根据测量数据的Q函数-贝尔曼方程来评估控制策略

$$Z_k^{\mathrm{T}}H_1Z_k=(y_k-y_k^r)^{\mathrm{T}}Q(y_k-y_k^r)+u_k^{\mathrm{T}}Ru_k+\gamma Z_{k+1}^{\mathrm{T}}H_1Z_{k+1}. \tag{28}$$

为了求解H_1来实现Q学习算法,现将Q函数(26)参数化为

$$Q(Z_k)=Z_k^{\mathrm{T}}H_1Z_k=\bar{H}_1^{\mathrm{T}}\bar{Z}_k. \tag{29}$$

其中

$\bar{H}_1\triangleq$

$[h_{11},2h_{12},\dots,2h_{1l_1},h_{22},2h_{23},\dots,2h_{2l_1},\dots,h_{l_1l_1}]^{\mathrm{T}}$,

$\bar{Z}\triangleq$

$[Z_1^2,Z_1Z_2,\dots,Z_1Z_{l_1},Z_2^2,Z_2Z_3,\dots,Z_2Z_{l_1},\dots,Z_{l_1}^2]^{\mathrm{T}}$.

上述线性方程可以写成

$$\Phi_1{}^{\mathrm{T}}\bar{\boldsymbol{H}}_1=\Omega_1. \tag{30}$$

其中

$$\Phi_1=[\bar{Z}_k-\gamma\bar{Z}_{k+1},\bar{Z}_{k+1}-\gamma\bar{Z}_{k+2},\dots,\bar{Z}_{k+L-1}-\gamma\bar{Z}_{k+L}],$$

$$\Omega_1=[r_k,r_{k+1},\dots,r_{k+L-1}]^{\mathrm{T}},$$

$$r_k=(y_k-y_k^r)^{\mathrm{T}}Q(y_k-y_k^r)+{u_k}^{\mathrm{T}}Ru_k,$$

$\Phi_1\in R^{l_1(l_1+1)/2\times L}$, $\Omega_1\in R^L$.

式(30)的解为

$$\bar{H}_1=(\Phi_1{\Phi_1}^{\mathrm{T}})^{-1}\Phi_1\Omega_1. \tag{31}$$

在式(29)中,$\bar{H}_1$是未知的向量,为了解线性方程,需要的样本数据需满足$L\geqslant l_1(l_1+1)/2$. 下面提出策略迭代的强化学习算法来学习输出反馈增量式Q学习的补偿控制律.

算法2 输出反馈增量式Q学习策略迭代算法.

1) 初始化. 初始化$\bar{H}_1^0$,选取使系统稳定的标称控制器u^0,选择稳定的初始控制策略Δu_0,对$j(j=1,2,3,\dots)$执行如下操作.

2) 策略评估. 收集足够步长的数据解Q函数贝尔曼方程

$$(\bar{H}_1^j)^{\mathrm{T}}(\bar{Z}_k-\gamma\bar{Z}_{k+1})=(y_k-y_k^r)^{\mathrm{T}}Q(y_k-y_k^r)+(u_k^j)^{\mathrm{T}}Ru_k^j.$$

3) 策略更新. 在策略评估得到$\bar{H}_1$后构造H_1,计算改进后的控制策略

$$\Delta u_k^{j+1}=-(H_{uu}^j)^{-1}(H_{u\bar{u}}^j\bar{u}_{k-1,k-N}+H_{u\bar{y}}^j\bar{y}_{k-1,k-N}+H_{ur}^jx^r_{k-N}).$$

4) 终止条件. 当达到收敛条件$\|H_1^j-H_1^{j-1}\|\leqslant\varepsilon$时停止迭代,其中$\varepsilon$为一个预先设计的正常数,否则执行$j=j+1$继续迭代.

注3 算法1和算法2所示的增量式控制律是一种无模型的方法,只需获得辨识核矩阵的样本数据,便可在数据驱动的环境下自适应地学习到该增量控制律. 由于系统的标称控制律使系统稳定,在算法中使用策略迭代算法来求解H和H_1.

4 仿真分析

本文使用F-16飞行器模型进行算法验证. 选择使用低保真空气动力模型,在高度为5 000 ft和速度为300 ft/s的平稳飞行状态下配平. 离散时间取为0.5 s,离散化后F-16纵向模型的线性离散系统描述如下:

$$x_{k+1}=\begin{bmatrix}0.985\,4 & 0.652\,8 & -15.967\,7 & -4.526\,1\\ -0.000\,3 & 0.789\,1 & 0.002\,6 & 0.330\,5\\ 0 & 0.033\,2 & 1 & 0.415\,3\\ 0 & 0.119\,1 & 0.000\,1 & 0.691\,2\end{bmatrix}x_k+\begin{bmatrix}-0.048\,3\\ 0.005\,7\\ 0.006\,3\\ 0.023\,7\end{bmatrix}u_k,$$

$$y_k=[0\ \ 57.3\ \ 0\ \ 0]x_k. \tag{32}$$

其中:$x_k=[V\ \ \alpha\ \ \theta\ \ q]^{\mathrm{T}}$,$u=\delta_e$,$\delta_e$的偏转范围为$\pm25.0^\circ$.

参考信息由以下动力学系统[17]生成:

$$x_{k+1}^r=\begin{bmatrix}0.995\,4 & -0.095\,4\\ 0.095\,4 & 0.995\,4\end{bmatrix}x_k^r,\ y_k^r=[5\ \ 0]x_k^r. \tag{33}$$

4.1　增量式状态反馈

各个状态的初始值设置为$V = 10\,\mathrm{ft/s}$, $\alpha = 10°$, $\theta = -10°$, $q = 10°/\mathrm{s}$,参考轨迹系统初始状态设置为$[1,-1]^{\mathrm{T}}$,代价函数中的权重矩阵$R = 1$, $Q = 20$,折扣因子$\gamma = 0.3$, $H^0 = 0.01I$. 由于系统(32)是不稳定的,需要预先设计标称控制器使其稳定,标称控制器增益设计如下:

$$K_0 = [-0.0061\ \ 6\ \ 0.007\,94\ \ 3\ \ -1\ \ 0.178\,4]. \quad (34)$$

在使用算法1学习最优增量式控制器时,需要辨识的H维度为7×7,则辨识一次需要满足采样数据长度$L \geqslant 28$,这里L取28,通过算法1可得

$$H_{uX} = [0\ \ -0.010\,5\ \ -0.000\,1\ \ -0.005\,3],$$
$$H_{ur} = [0.001\,1\ \ -0.000\,1],\ \ H_{uu} = 0.000\,2. \quad (35)$$

将式(35)代入(18),可得最优增量式控制律为

$$\Delta K = [-0.020\,3\ \ 47.461\,0\ \ 0.254\,6\ \ \rightarrow$$
$$\leftarrow 23.966\,7\ \ -5.157\,9\ \ 0.603\,4]. \quad (36)$$

将标称控制律(34)与学习到的增量式控制律(36)相加可得总控制律为

$$K_{\mathrm{sum}} = [-0.026\,4\ \ 53.461\,0\ \ 0.334\,3\ \ \rightarrow$$
$$\leftarrow 26.966\,7\ \ -6.157\,9\ \ 0.781\,8]. \quad (37)$$

通过非增量式Q学习得到的控制律[18]为

$$K^*_{\mathrm{sum}} = [-0.026\,4\ \ 53.469\,2\ \ 0.332\,3\ \ \rightarrow$$
$$\leftarrow 26.967\,8\ \ -6.158\,8\ \ 0.775\,2]. \quad (38)$$

比较式(37)与(38)可知,增量式Q学习能够在系统预先有一个标称控制器的情况下,自适应地补偿标称控制律,使得总的控制律达到最优.

状态反馈仿真如图2～图6所示. 图2表示迎角的变化曲线,可以看出,在标称控制器下,迎角虽然不发散,但对参考信号跟踪效果不理想,在算法1的控制律补偿下,迎角在第40步能够跟踪上参考信号且跟踪效果显著优于未优化时的效果. 图3为状态反馈下的升降舵偏转角曲线,其范围为$[-4°, 10°]$. 无人机配平点的升降舵偏转角为$-4.238\,3°$,因此,总的升降

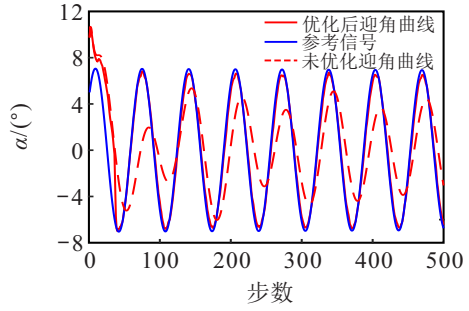

图2　状态反馈下迎角变化曲线

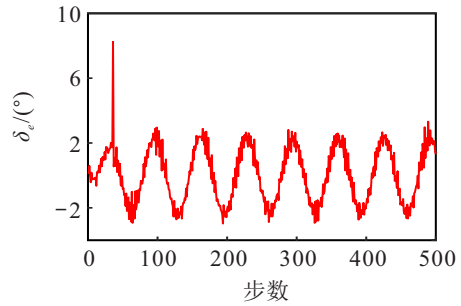

图3　状态反馈下升降舵偏转角

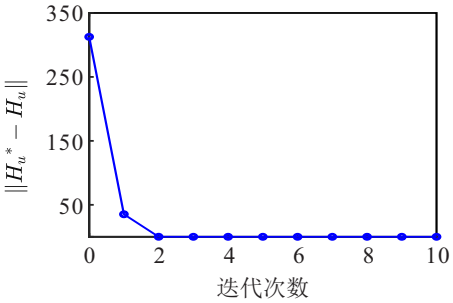

图4　状态反馈下H的收敛性

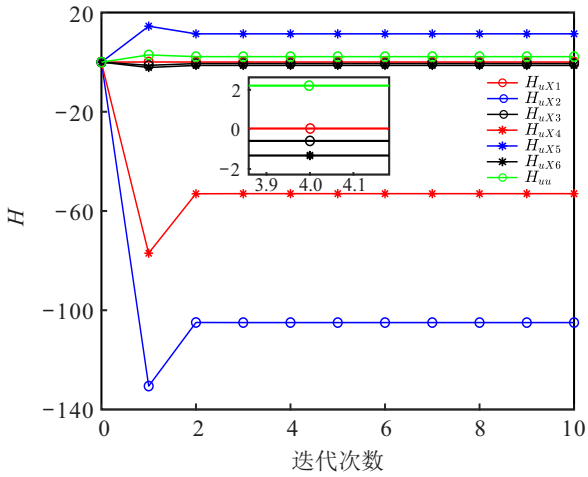

图5　状态反馈下H元素的收敛性

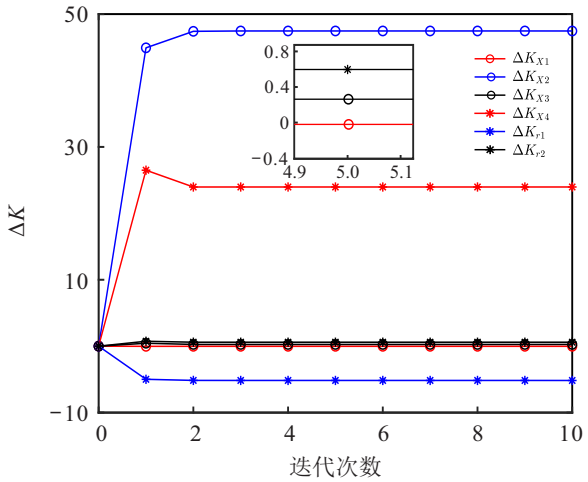

图6　状态反馈下控制律增益收敛性

舵偏转角在偏转角约束范围$\pm 25.0°$内. 图4和图5表示H中与增量式控制律有关的各元素收敛性,在第2次迭代中即可学习到最优的增量式控制律. 图6为状

态反馈下增量式控制器增益收敛性.

4.2 增量式输出反馈

各个状态的初始值设置为$V=10\,\text{ft/s}, \alpha=10°$, $\theta=-10°$, $q=10°/\text{s}$,参考轨迹系统初始状态设置为$[1,-1]^{\mathrm{T}}$,代价函数中的权重矩阵$R=1$, $Q=7$,折扣因子$\gamma=0.5$, $H_1^0=0.01I$. 预先设计如下标称控制器使系统(32)稳定:

$$\begin{aligned}&F_0=[F_u\ \ F_y\ \ F_r],\\&F_u=[-0.217\,4\ \ 0.521\,7\ \ 0.060\,9\ \ -0.252\,2],\\&F_y=[2.434\,8\ \ -4.217\,4\ \ 2.652\,2\ \ -0.573\,9],\\&F_r=[-1.321\,7\ \ 0.704\,3].\end{aligned}\tag{39}$$

在使用算法2学习最优增量式控制器时,需要辨识的H_1维度为11×11,则辨识一次需要满足采样数据长度$L\geqslant 66$,这里L取70,通过算法2可得

$$\begin{aligned}&H_{u\bar{u}}=[1.244\,4\ \ -4.561\,3\ \ -0.193\,2\ \ 2.232\,2],\\&H_{uy}=[-18.245\,1\ \ 34.663\,4\ \ -22.789\,1\ \ 5.079\,1],\\&H_{ur}=[6.371\,1\ \ -3.898\,7],\ H_{ur}=2.110\,7.\end{aligned}\tag{40}$$

将式(40)代入(27),可得最优增量式控制律为

$$\begin{aligned}&\Delta F_u=[-0.589\,6\ \ 2.161\,0\ \ 0.091\,5\ \ -1.057\,5],\\&\Delta F_y=[8.643\,9\ \ -16.318\,2\ \ 10.796\,8\ \ -2.406\,3],\\&\Delta F_r=[-3.018\,4\ \ 1.847\,1].\end{aligned}\tag{41}$$

将标称控制律(40)与学习到的增量式控制律(41)相加,可得总控制律为

$$\begin{aligned}&F_{u\_\text{sum}}=[-0.807\,0\ \ 2.682\,7\ \ 0.152\,4\ \ -1.309\,7],\\&F_{y\_\text{sum}}=[11.078\,7\ \ -20.535\,6\ \ 13.449\,0\ \ -2.980\,2],\\&F_{r\_\text{sum}}=[-4.340\,1\ \ 2.551\,4].\end{aligned}\tag{42}$$

通过非增量式Q学习而学习到的控制律[18]为

$$\begin{aligned}&F_u^*=[-0.806\,9\ \ 2.686\,9\ \ 0.151\,1\ \ -1.312\,6],\\&F_y^*=[11.091\,4\ \ -20.567\,3\ \ 13.474\,3\ -2.986\,7],\\&F_r^*=[-4.340\,4\ \ 2.551\,2].\end{aligned}\tag{43}$$

比较式(42)与(43)可知,输出反馈增量式Q学习能够在系统预先有一个标称控制器的情况下,利用输入输出数据设计增量式控制律,自适应地补偿标称控制律,使得总的控制律达到最优.

输出反馈仿真如图7～图11所示. 图7表示迎角的变化曲线,可以看出,标称控制器对参考信号跟踪效果不理想,在算法2的控制律补偿下,迎角在第100步能够跟踪上参考信号且跟踪效果显著优于未优化时的效果,输出反馈迎角跟踪上参考信号的时间长于状态反馈,是由于辨识一次H所需的样本数据步长不同,算法2所需数据步长明显多于算法1. 图8为输出反馈下的升降舵偏转角,其范围为$[-8°,6°]$. 根据无人机配平点的升降舵偏转角为$-4.238\,3°$,因此,总的升降舵偏转角在偏转角约束范围$\pm 25.0°$内. 图9和图10表示H_1中与增量式控制律有关的各元素收敛性,在第2次迭代中即可学习到最优的增量式控制律. 图11表示增量式控制器增益收敛性.

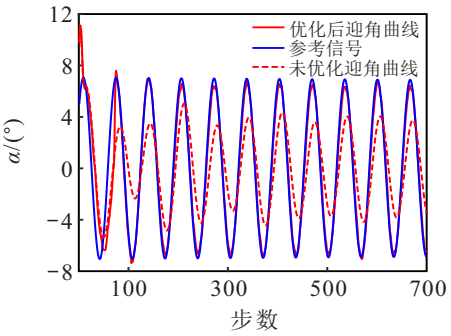

图 7 输出反馈下迎角变化曲线

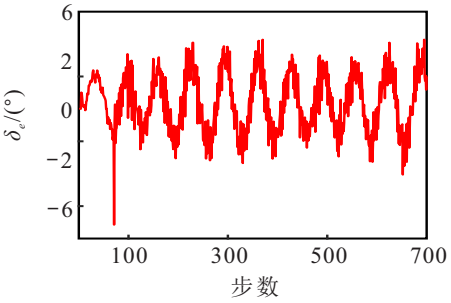

图 8 输出反馈下升降舵偏转角

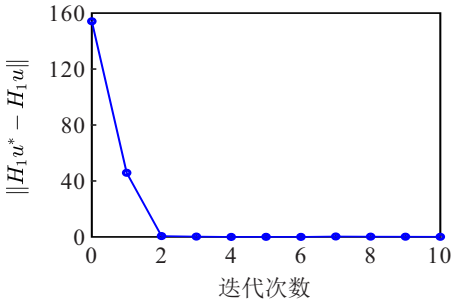

图 9 输出反馈 H_1 的收敛性

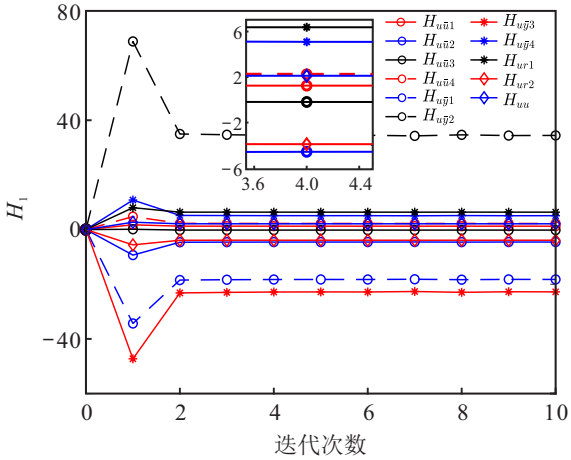

图 10 输出反馈 H_1 元素的收敛性

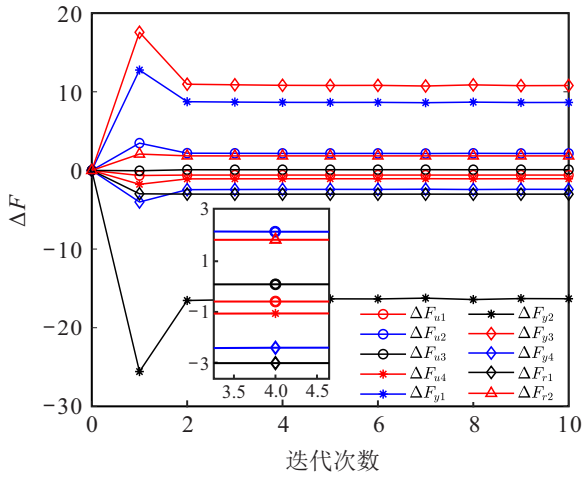

图 11　输出反馈控制律增益收敛性

注 4　当激励噪声满足持续激励条件时, 数据驱动的策略迭代算法的收敛性和利用李雅普诺夫方程迭代求解黎卡提方程的收敛性一致[14]. 由于采用迭代算法的控制器增益满足二次收敛特性[19], 本文采用的Q学习算法也具有二次收敛特性, 其收敛速度较快. 在二次收敛特性下, Q学习算法的参数设置主要包括权重矩阵Q, R, 折扣因子γ, 初始矩阵H^0和每步策略评估的数据长度N. 当数据长度N满足矩阵满秩条件时, 增大每步的数据长度不影响收敛速度. 因此, Q学习的收敛速度主要与参数Q, R, γ, H^0的设置有关.

4.3　数字PID控制器性能对比

对于本文的迎角跟踪, PID控制系统包含迎角偏差控制通道, 迎角偏差控制通道控制器为

$$e_k = y_k - y_k^r$$

$$u_k = K_p e_k + K_i \sum_{j=0}^{k} e_j + K_d(e_k - e_{k-1}). \qquad (44)$$

通过调试, PID控制参数为$K_{\rm p} = 0.5$, $K_{\rm i} = 0.34$, $K_{\rm d} = 0.25$. 将PID控制与本文方法进行对比, 如图12和图13所示. 图12是增量式Q学习状态反馈优化控制与PID控制性能对比图, 图13是增量式Q学习输出反馈优化控制与PID控制性能对比图, 三者都能实现迎角对参考信号的跟踪, 但本文提出的增量式Q学习算法比PID控制跟踪误差较小, 且在参考信号跟踪初

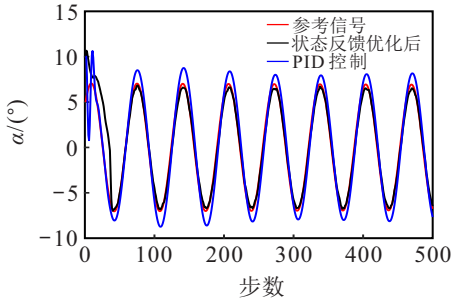

图 12　状态反馈与PID性能对比

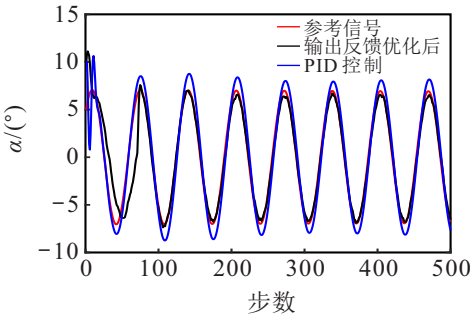

图 13　输出反馈与PID性能对比

始段比PID控制的响应更为平稳. 更重要的是, 当无人机的工作点发生变化时, PID控制不能自适应地调整, 其控制性能也会随着下降, 而增量式Q学习控制可以根据工作点的变化自适应地学习一个新的控制器, 实现无人机控制性能的实时优化. 因此, 增量式Q学习控制器相比PID控制具有更优良的性能.

5　结　论

本文针对固定翼无人机纵向控制的高性能需求, 提出一种控制系统性能优化结构. 该控制结构在不修改预先设计的控制系统的情况下, 通过优化一个增量式的控制系统增益来优化性能. 增量式控制器的设计分为两种: 1) 在无人机状态信息可获得情况下, 设计基于状态反馈的增量式Q学习的补偿控制律; 2) 当无人机状态信息不能完全获得时, 利用输入、输出和参考信号数据设计基于输出反馈的增量式Q学习的补偿控制律. 两种设计方法均为数据驱动下的无模型方法, 且激励噪声对Q函数贝尔曼方程的参数估计没有偏差. 通过对选用的F-16纵向模型的仿真可知, 增量式Q学习算法会自适应地补偿标称控制增益使总的控制输入达到最优. 与标称控制性能相比, 增量式Q学习控制器的跟踪控制性能得到了较好的优化.

参考文献(References)

[1]　王美仙, 李明, 张子军. 飞行器控制律设计方法发展综述[J]. 飞行力学, 2007, 25(2): 1-4.
(Wang M X, Li M, Zhang Z J. Developing status of control law design methods for flight[J]. Flight Dynamics, 2007, 25(2): 1-4.)

[2]　江琼, 陈怀民, 吴佳楠. H_∞鲁棒控制与PID控制相结合的无人机飞行控制研究[J]. 宇航学报, 2006, 27(2): 192-195.
(Jiang Q, Chen H M, Wu J N. Research on UAV flight control based on PID control and H_∞ robust control[J]. Journal of Astronautics, 2006, 27(2): 192-195.)

[3]　刘燕斌, 陆宇平. 基于反步法的高超音速飞机纵向逆飞行控制[J]. 控制与决策, 2007, 22(3): 313-317.

(Liu Y B, Lu Y P. Longitudinal inversion flight control based on backstepping for hypersonic vehicle[J]. Control and Decision, 2007, 22(3): 313-317.)

[4] Hwangbo J, Sa I, Siegwart R, et al. Control of a quadrotor with reinforcement learning[J]. IEEE Robotics and Automation Letters, 2017, 2(4): 2096-2103.

[5] Sugimoto T, Gouko M. Acquisition of hovering by actual UAV using reinforcement learning[C]. The 3rd International Conference on Information Science and Control Engineering. Beijing: IEEE, 2016: 148-152.

[6] Wang Y D, Sun J, He H B, et al. Deterministic policy gradient with integral compensator for robust quadrotor control[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(10): 3713-3725.

[7] Zhou Y, Kampen E J V, Chu Q P. Nonlinear adaptive flight control using incremental approximate dynamic programming and output feedback[J]. Journal of Guidance, Control, and Dynamics, 2016, 40(2): 493-496.

[8] de Alvear Cárdenas J I, Sun B, Van Kampen E J. Intelligent adaptive control using LADP and IADP applied to F-16 aircraft with imperfect measurements[C]. AIAA Scitech 2021 Forum. Reston: AIAA, 2021: 1119.

[9] Kiumarsi B, Lewis F L, Modares H, et al. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics[J]. Automatica, 2014, 50(4): 1167-1175.

[10] Kiumarsi B, Lewis F L, Naghibi-Sistani M B, et al. Optimal tracking control of unknown discrete-time linear systems using input-output measured data[J]. IEEE Transactions on Cybernetics, 2015, 45(12): 2770-2779.

[11] Nguyen L T. Simulator study of stall/post-stall characteristics of a fighter airplane with longitudinal static stability[M]. Washington, DC: National Aeronautics and Space Administration, Scientific and Technical Information Branch, 1979: 36-43.

[12] Stevens B L, Lewis F L, Johnson E N. Aircraft control and simulation: Dynamics, controls design, and autonomous systems[M]. The 3rd edition. Hoboken: Wiley, 2015: 584-592.

[13] Peng Y J, Chen Q, Sun W J. Reinforcement Q-learning algorithm for H_∞ tracking control of unknown discrete-time linear systems[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(11): 4109-4122.

[14] Xu Y S, Zhao Z G, Yin S. Performance optimization and fault-tolerance of highly dynamic systems via Q-learning with an incrementally attached controller gain system[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022 (99): 1-11.

[15] Lewis F L, Vamvoudakis K G. Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data[J]. IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics, 2011, 41(1): 14-25.

[16] Rizvi S A A, Lin Z L. Output feedback Q-learning for discrete-time linear zero-sum games with application to the H_∞ control[J]. Automatica, 2018, 95: 213-221.

[17] Peng Y J, Meng Q Q, Sun W J. Adaptive output-feedback quadratic tracking control of continuous-time systems via value iteration with its application[J]. IET Control Theory & Applications, 2020, 14(20): 3621-3631.

[18] Sun W J, Zhao G Y, Peng Y J. Adaptive optimal output feedback tracking control for unknown discrete-time linear systems using a combined reinforcement Q-learning and internal model method[J]. IET Control Theory & Applications, 2019, 13(18): 3075-3086.

[19] Hewer G. An iterative technique for the computation of the steady state gains for the discrete optimal regulator[J]. IEEE Transactions on Automatic Control, 1971, 16(4): 382-384.

作者简介

赵振根 (1989−), 男, 讲师, 博士, 从事信息物理系统安全、无人机智能控制等研究, E-mail: zhaozhengen@nuaa.edu.cn;

程磊 (1997−), 男, 硕士生, 从事无人机智能控制的研究, E-mail: chenglei@nuaa.edu.cn.