DOI:10.13878/j.cnki.jnuist.2021.05.006

张正义[1]　赵学艳[1]

基于Q学习算法的随机离散时间系统的随机线性二次最优追踪控制

摘要

针对随机线性离散时间系统,利用Q学习算法求解无限时域的随机线性二次最优追踪控制(SLQT)问题.首先,假设通过命令生成器生成追踪所需的参考信号,并建立一个由原随机系统和参考轨迹系统组成的增广系统,把最优追踪问题转化为最优调节问题的形式.其次,为了在线求解随机系统的最优追踪问题,将随机系统转为确定性系统,并根据增广系统定义随机线性二次最优追踪控制的Q函数,在无需知道系统模型参数的情况下在线求解增广随机代数方程(GSAE).再次,证明了Q学习算法和增广随机代数方程的等价性,给出了Q学习算法实现步骤.最后,给出一个仿真实例说明Q学习算法的有效性.

关键词

随机系统;Q学习算法;最优追踪控制;随机代数方程

中图分类号 O232;TP13

文献标志码 A

收稿日期 2021-09-12

资助项目 国家自然科学基金(61873099,62073144);广东省自然科学基金(2020A1515010441);广州市科技计划(202002030158,202002030389)

作者简介

张正义,男,硕士生,研究方向为自适应动态规划、最优控制、强化学习.zhang810024143@163.com

赵学艳(通信作者),女,副教授,硕士生导师,主要从事随机系统和非线性系统的稳定性与镇定,复杂系统的建模、分析和控制的研究.auxyzhao@scut.edu.cn

1 华南理工大学 自动化科学与工程学院,广州,510640

0 引言

最优控制的目标是找到最优的控制策略,使得被控系统达到指定目标状态的同时,使系统预定义的性能指标为最小.最优控制问题主要有两个研究方向,分别是最优调节问题和最优追踪问题.对于线性系统的二次调节(Linear Quadratic Regulator,LQR)问题,传统方法通常是通过离线求解其对应的代数里卡蒂(Riccati)方程,这种方法需要完全已知系统参数的全部动力学信息[1-2].但是,在实际情况下,系统动力学信息完全已知的条件难以满足,传统方法不可能得到解析解.所以,通常需要在系统参数未知的情况下在线求解最优控制器,因此利用自适应动态规划(Adaptive Dynamic Programming,ADP)和神经网络方法求解最优控制在近些年备受关注.自适应动态规划[3]是在系统参数未知或系统参数不确定的情况下设计系统的控制器,不需要提前知道系统动力学信息,充分利用系统的状态信息在线求解最优控制.近些年来,ADP方法在离散系统和连续系统中有了广泛的应用.文献[4]针对连续时间线性系统提出了自适应动态规划方法,在系统参数矩阵部分未知的情况下得到最优控制器;文献[5]进一步针对连续时间线性系统提出了一种自适应策略迭代方法,在系统参数完全未知的情况下得到最优控制器;文献[6]针对线性离散时间系统的追踪问题使用强化Q学习方法,在系统参数完全未知的情况下求解最优控制器.

随机系统控制理论由于其自身的学术难度以及广泛的应用领域,已成为控制理论的重要组成部分与研究热点[7-8],尤其是随机系统的最优控制问题受到越来越多的关注.与确定性问题相似,随机系统的线性二次最优控制问题(Stochastic Linear Quadratic,SLQ)的可解性等价于随机代数Riccati方程的可解性,文献[9]研究了线性终端状态约束下不定随机线性二次最优控制问题,文献[10]研究了具有乘性噪声的随机离散系统的带约束线性二次最优控制问题,但是文献[9-10]需要完全已知的系统参数信息.因此,文献[11]针对随机连续时间系统在系统参数部分未知的情况下提出了策略迭代方法求解随机系统的最优控制问题,文献[12]针对系统参数完全未知的随机线性离散系统提出了使用自适应动态规划的方法求解最优控制问题,

文献[13]针对模型自由的随机线性离散系统提出了Q学习算法求解最优控制问题.相较于最优调节问题,最优追踪问题在现实中往往有更多的应用,例如文献[14]针对参数未知的随机离散系统提出了基于神经网络的自适应动态规划方法求解最优追踪控制问题.求解系统的最优控制问题,大多需要系统的完全动力学信息,使用Q学习算法的优点是不用直接求解复杂的随机代数方程,而是充分利用系统的状态信息在线求得系统的最优控制.受到文献[13-14]的启发,本文针对离散时间系统的随机线性二次最优追踪控制问题,提出了解决随机线性二次最优追踪控制的Q学习算法,给出算法的具体实现步骤,使用Q学习算法在线解决追踪控制问题而无需系统模型参数,最后给出仿真实例,表明系统输出可以有效地追踪参考轨迹.

本文的结构安排如下:第一节对问题进行描述,定义参考信号系统,将原随机系统和参考信号系统组成增广系统,把最优追踪问题转化为最优调节问题的形式;第二节对随机系统进行了问题转变,将随机系统转化为确定性系统;第三节推导了Q函数;第四节给出算法的具体实现步骤;第五节给出仿真实例;第六节对全文进行了总结.

注1 本文中:$\mathbf{R}$表示实数集;$\mathbf{R}^n$表示n维欧几里得空间;$\mathbf{R}^{n\times m}$表示全体$n\times m$维实数矩阵的集合;$\|\cdot\|$表示$\mathbf{R}^n$中的欧几里得范数;$\|\boldsymbol{P}\|_F$表示假设$\boldsymbol{P}$为$n\times m$维实矩阵,则其F范数定义为$\|\boldsymbol{P}\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$;$\boldsymbol{Z}^{\mathrm{T}}$表示矩阵或向量$\boldsymbol{Z}$的转置;$\boldsymbol{Z}^{-1}$表示方阵$\boldsymbol{Z}$的逆;$\boldsymbol{I}$表示适当维度的单位矩阵;$\mathrm{tr}(\boldsymbol{Z})$表示方阵$\boldsymbol{Z}=(z_{ij})_{n\times n}$的迹,$\mathrm{tr}(\boldsymbol{Z})=\sum_{i=1}^{n}z_{ii}$;$\mathrm{E}(\cdot)$表示数学期望;$(\Omega,\mathcal{F},\{\mathcal{F}_k\}_{k\geqslant k_0},\mathcal{P})$表示具有流$\{\mathcal{F}_k\}_{k\geqslant k_0}$的完备概率空间,其中$\Omega$是样本点的全体,$\mathcal{F}$是$\Omega$上的一个$\sigma$-代数,$\mathcal{P}$是$\mathcal{F}$上的一个概率测度,滤子族$\{\mathcal{F}_k\}_{k\geqslant k_0}$是一族单增的$\mathcal{F}$子$\sigma$-代数;$\boldsymbol{w}_k(k=0,1,2,\cdots,\boldsymbol{w}_0=0)$是定义在完备概率空间上的一维Wiener过程,又称为布朗运动;$\boldsymbol{P}>0$表示对称矩阵$\boldsymbol{P}$为正定矩阵.

1 问题描述

给定随机离散时间线性系统为

$$\begin{cases}\boldsymbol{x}_{k+1}=[\boldsymbol{A}\boldsymbol{x}_k+\boldsymbol{B}\boldsymbol{u}_k]+[\boldsymbol{C}\boldsymbol{x}_k+\boldsymbol{D}\boldsymbol{u}_k]\boldsymbol{w}_k,\\ \boldsymbol{y}_k=\boldsymbol{G}\boldsymbol{x}_k,\end{cases}\tag{1}$$

式中:$\boldsymbol{x}_k\in\mathbf{R}^n$表示系统的状态变量;$\boldsymbol{x}_0$是系统的初始状态;$\boldsymbol{u}_k\in\mathbf{R}^m$表示系统控制输入;$\boldsymbol{y}_k\in\mathbf{R}^q$表示系统输出变量;$\boldsymbol{A}\in\mathbf{R}^{n\times n},\boldsymbol{B}\in\mathbf{R}^{n\times m},\boldsymbol{C}\in\mathbf{R}^{n\times n},\boldsymbol{D}\in\mathbf{R}^{n\times m},\boldsymbol{G}\in\mathbf{R}^{q\times n}$均为常数矩阵;$\boldsymbol{w}_k$是定义在完备概率空间$(\Omega,\mathcal{F},\{\mathcal{F}_k\}_{k\geqslant k_0},\mathcal{P})$上的一维标准布朗运动,$\boldsymbol{w}_0=0$,且满足$\mathrm{E}(\boldsymbol{w}_{k+1}\mid\mathcal{F}_k)=0,\mathrm{E}(\boldsymbol{w}_{k+1}^2\mid\mathcal{F}_k)=1$.

首先,对于追踪问题,需要定义追踪问题的参考信号.假设命令生成器生成的随机线性二次最优追踪控制的参考轨迹模型为

$$\boldsymbol{r}_{k+1}=\boldsymbol{F}\boldsymbol{r}_k,\tag{2}$$

式中:$\boldsymbol{r}_k\in\mathbf{R}^q$表示参考轨迹系统的输出变量;$\boldsymbol{F}\in\mathbf{R}^{q\times q}$为常数矩阵.为了生成更为广泛的参考轨迹,假设$\boldsymbol{F}$不必是赫尔维茨的.

对于追踪控制问题,需要重点关注系统的输出与需要追踪的参考信号输出之间的误差,因此定义系统的追踪误差为

$$\boldsymbol{e}_k=\boldsymbol{y}_k-\boldsymbol{r}_k,\tag{3}$$

式中$\boldsymbol{r}_k$为参考轨迹模型(2)的输出变量.

随机线性最优追踪问题的目标是设计出最优的控制器,不但能够确保系统的输出变量$\boldsymbol{y}_k$可以稳定地追踪上给定的参考信号$\boldsymbol{r}_k$,并且最小化预定义的性能指标函数.定义性能指标函数为

$$J(\boldsymbol{x}_k,\boldsymbol{r}_k,\boldsymbol{u})=\mathrm{E}\sum_{i=k}^{\infty}U_i=\mathrm{E}\sum_{i=k}^{\infty}[(\boldsymbol{G}\boldsymbol{x}_i-\boldsymbol{r}_i)^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{G}\boldsymbol{x}_i-\boldsymbol{r}_i)+\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_i],\tag{4}$$

式中:$\boldsymbol{r}_k$为参考轨迹信号,$\boldsymbol{u}$为控制序列,U_i为$i=k$时对应的性能指标函数,$\boldsymbol{Q}\in\mathbf{R}^{q\times q}>0,\boldsymbol{R}\in\mathbf{R}^{m\times m}>0.(\boldsymbol{G}\boldsymbol{x}_i-\boldsymbol{r}_i)^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{G}\boldsymbol{x}_i-\boldsymbol{r}_i)$表示为追踪误差变量的二次型,$\boldsymbol{Q}$为正定的权重系数矩阵,因此$\boldsymbol{Q}$矩阵越大表示误差变量对性能指标函数的影响越大;$\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_i$表示为系统控制输入变量的二次型,$\boldsymbol{R}$为正定的权重系数矩阵,因此$\boldsymbol{R}$矩阵越大表示对系统控制的约束越大.

基于随机系统模型(1)和参考轨迹模型(2),建立一个由原随机系统和参考轨迹系统组成的增广系统,把最优追踪问题转化为最优调节问题的形式.

因此,随机线性离散时间系统(1)和参考轨迹系统(2)构建的增广系统如下:

$$\boldsymbol{X}_{k+1}=\begin{bmatrix}\boldsymbol{x}_{k+1}\\ \boldsymbol{r}_{k+1}\end{bmatrix}=\begin{bmatrix}\boldsymbol{A}+\boldsymbol{C}\boldsymbol{w}_k & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{F}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_k\\ \boldsymbol{r}_k\end{bmatrix}+\begin{bmatrix}\boldsymbol{B}+\boldsymbol{D}\boldsymbol{w}_k\\ \boldsymbol{O}\end{bmatrix}\boldsymbol{u}_k\equiv\boldsymbol{T}\boldsymbol{X}_k+\boldsymbol{B}_0\boldsymbol{u}_k,\tag{5}$$

式中: $\boldsymbol{T} \in \mathbf{R}^{(n+q)\times(n+q)}$, $\boldsymbol{B}_0 \in \mathbf{R}^{(n+q)\times m}$, 增广系统的状态为 $\boldsymbol{X}_k = \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{r}_k \end{bmatrix} \in \mathbf{R}^{(n+q)}$.

显然,只有当 $\boldsymbol{F}$ 为赫尔维茨(Hurwitz)时才可以使用性能指标函数(4),即要求参考轨迹系统是渐近稳定的,如果参考轨迹随着时间的推迟不趋于零,那么性能指标函数(4)将是无界的.但是在实际应用中并不总是追踪渐近稳定的信号,而通常需要追踪正弦信号等其他复杂信号.为了放宽 $\boldsymbol{F}$ 为赫尔维茨的限制,在性能指标函数(4)中引入折扣系数 γ,可以使得 $\boldsymbol{F}$ 为非赫尔维茨,因此基于(4)重新定义带折扣系数的性能指标函数为

$$J(\boldsymbol{x}_k, \boldsymbol{r}_k, \boldsymbol{u}) = \mathrm{E}\sum_{i=k}^{\infty} \gamma^{i-k} U_i = \mathrm{E}\sum_{i=k}^{\infty} \gamma^{i-k}\left[(\boldsymbol{G}\boldsymbol{x}_i - \boldsymbol{r}_i)^{\mathrm{T}} \boldsymbol{Q} (\boldsymbol{G}\boldsymbol{x}_i - \boldsymbol{r}_i) + \boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{R} \boldsymbol{u}_i\right], \tag{6}$$

式中: $0 < \gamma \leqslant 1$ 是折扣系数,注意到只有当 $\boldsymbol{F}$ 为赫尔维茨时,此时参考轨迹为渐近稳定的,γ 可以取值为 1,性能指标函数由式(6)化简为式(4).

引入折扣系数 γ 后,基于增广系统(5)和性能指标函数(6),性能指标函数可进一步重写为

$$J(\boldsymbol{X}_k, \boldsymbol{u}) = \mathrm{E}\sum_{i=k}^{\infty} \gamma^{i-k}\left[\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{Q}_1 \boldsymbol{X}_i + \boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{R} \boldsymbol{u}_i\right], \tag{7}$$

式中: $\boldsymbol{Q}_1 = \begin{bmatrix} \boldsymbol{G} & -\boldsymbol{I} \end{bmatrix}^{\mathrm{T}} \boldsymbol{Q} \begin{bmatrix} \boldsymbol{G} & -\boldsymbol{I} \end{bmatrix} \in \mathbf{R}^{(n+q)\times(n+q)}$.

本文的目标是找到最优控制增益矩阵 $\boldsymbol{K}^*$ 使得性能指标函数(7)达到最小,同时能够追踪参考信号.经过增广原系统,最优追踪问题已经转化为最优调节问题的形式,定义状态反馈线性控制器为

$$\boldsymbol{u}_k = \boldsymbol{K}\boldsymbol{X}_k, \tag{8}$$

式中: $\boldsymbol{K} \in \mathbf{R}^{m\times(n+q)}$, $\boldsymbol{K}$ 表示系统的控制增益矩阵.

同时定义性能指标函数的最优值为

$$V^*(\boldsymbol{X}_0, \boldsymbol{u}) = \min_{\boldsymbol{u}} J(\boldsymbol{X}_0, \boldsymbol{u}). \tag{9}$$

定义 1[14] 如果存在线性状态反馈控制器 $\boldsymbol{u}_k$,对于任意初始状态,使得 $\lim\limits_{k\to\infty} \mathrm{E}(\boldsymbol{e}_k^{\mathrm{T}} \boldsymbol{e}_k) = 0$,那么称 $\boldsymbol{u}_k$ 是均方稳定的.若系统(5)存在均方稳定的控制器,该系统称为均方可稳的.

定义 2[10] 如果 $-\infty < V^*(\boldsymbol{X}_0, \boldsymbol{u}) < +\infty$,则称随机线性二次最优控制问题是适定的.

定义 3[15] 如果线性反馈控制 $\boldsymbol{u}_k$ 满足:1) $\boldsymbol{u}_k$ 是关于 $\mathcal{F}_k$ 适定并且可测的随机过程;2) $\boldsymbol{u}_k$ 能够使得性能指标函数式(6)为最小值;3) $\boldsymbol{u}_k$ 是均方稳定的.那么称 $\boldsymbol{u}_k$ 是容许控制.

引理 1 如果存在均方稳定的容许控制 $\boldsymbol{u}_k = \boldsymbol{K}\boldsymbol{X}_k$,那么随机线性二次最优追踪控制是适定的,并且值函数可重写为

$$V(\boldsymbol{X}_k, \boldsymbol{K}) = \mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{P} \boldsymbol{X}_k), \tag{10}$$

式中: $\boldsymbol{P} \in \mathbf{R}^{(n+q)\times(n+q)}$, $\boldsymbol{P}$ 为正定矩阵,且为式(11)增广随机代数方程的解.

$$\begin{aligned} \boldsymbol{P} = {} & \gamma(\boldsymbol{A}_1 + \boldsymbol{B}_1\boldsymbol{K})^{\mathrm{T}} \boldsymbol{P} (\boldsymbol{A}_1 + \boldsymbol{B}_1\boldsymbol{K}) + \\ & \gamma(\boldsymbol{C}_1 + \boldsymbol{D}_1\boldsymbol{K})^{\mathrm{T}} \boldsymbol{P} (\boldsymbol{C}_1 + \boldsymbol{D}_1\boldsymbol{K}) + \boldsymbol{Q}_1 + \boldsymbol{K}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{K}, \end{aligned} \tag{11}$$

式中:

$$\boldsymbol{A}_1 = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{F} \end{bmatrix} \in \mathbf{R}^{(n+q)\times(n+q)},$$

$$\boldsymbol{B}_1 = \begin{bmatrix} \boldsymbol{B} \\ \boldsymbol{O} \end{bmatrix} \in \mathbf{R}^{(n+q)\times m},$$

$$\boldsymbol{C}_1 = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \in \mathbf{R}^{(n+q)\times(n+q)},$$

$$\boldsymbol{D}_1 = \begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{O} \end{bmatrix} \in \mathbf{R}^{(n+q)\times m}.$$

证明 由式(7)和式(8)有:

$$V(\boldsymbol{X}_k, \boldsymbol{K}) = \mathrm{E}\sum_{i=k}^{\infty} \gamma^{i-k}\left[\boldsymbol{X}_i^{\mathrm{T}} (\boldsymbol{Q}_1 + \boldsymbol{K}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{K}) \boldsymbol{X}_i\right],$$

拆开化简得:

$$V(\boldsymbol{X}_k, \boldsymbol{K}) = \mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}} (\boldsymbol{Q}_1 + \boldsymbol{K}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{K}) \boldsymbol{X}_k) + \gamma V(\boldsymbol{X}_{k+1}, \boldsymbol{K}),$$

代入式(10)可得:

$$\mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{P} \boldsymbol{X}_k) = \mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}} (\boldsymbol{Q}_1 + \boldsymbol{K}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{K}) \boldsymbol{X}_k) + \gamma \mathrm{E}(\boldsymbol{X}_{k+1}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{X}_{k+1}),$$

代入增广系统(5),替换 $\boldsymbol{X}_{k+1}$ 化简后得式(11).

对于随机线性二次最优追踪控制问题:随机线性二次最优追踪控制问题等价于求解其对应的增广随机代数 Riccati 方程的解,并且需要系统模型参数的全部信息.

2 问题转换

目前,确定性系统的最优追踪控制问题有着广泛的研究并且已经得到了很好的解决,随机系统因为随机参数的存在使得系统输出轨迹存在不确定性,且性能指标函数带有期望,在线算法无法实现期望功能.因此本节通过系统转变将随机系统转变为确定性系统,进而将随机系统的最优追踪控制问题转化为确定性的系统最优追踪控制问题.

给定容许控制 $\boldsymbol{u}_k = \boldsymbol{K}\boldsymbol{X}_k$,令 $\boldsymbol{T}_k = \mathrm{E}(\boldsymbol{X}_k \boldsymbol{X}_k^{\mathrm{T}})$,系统(5)转化为

$$\boldsymbol{T}_{k+1} = \mathrm{E}(\boldsymbol{X}_{k+1} \boldsymbol{X}_{k+1}^{\mathrm{T}}) = \mathrm{E}((\boldsymbol{T} + \boldsymbol{B}_0\boldsymbol{K}) \boldsymbol{X}_k \boldsymbol{X}_k^{\mathrm{T}} (\boldsymbol{T} + \boldsymbol{B}_0\boldsymbol{K})^{\mathrm{T}}), \tag{12}$$

代入增广系统式(5),化简得:

$$\boldsymbol{T}_{k+1}=\mathrm{E}(\boldsymbol{X}_{k+1}\boldsymbol{X}_{k+1}^{\mathrm{T}})=(\boldsymbol{A}_1+\boldsymbol{B}_1\boldsymbol{K})\boldsymbol{T}_k(\boldsymbol{A}_1+\boldsymbol{B}_1\boldsymbol{K})^{\mathrm{T}}+(\boldsymbol{C}_1+\boldsymbol{D}_1\boldsymbol{K})\boldsymbol{T}_k(\boldsymbol{C}_1+\boldsymbol{D}_1\boldsymbol{K})^{\mathrm{T}},\tag{13}$$

式中:$\boldsymbol{T}_k\in\mathbf{R}^{(n+q)\times(n+q)}$ 是确定性系统的状态.令 $\boldsymbol{T}_0=\mathrm{E}(\boldsymbol{X}_0\boldsymbol{X}_0^{\mathrm{T}})$ 为初始状态.

相应地,性能指标函数式(7)重写为

$$J(\boldsymbol{T}_k,\boldsymbol{K})=\mathrm{tr}\Big\{\sum_{i=k}^{\infty}\gamma^{i-k}(\boldsymbol{Q}_1+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K})\boldsymbol{T}_i\Big\}.\tag{14}$$

注 2 经系统转变后,随机系统转变为确定性系统.由式(13)可知,系统转变后随机参数 $\boldsymbol{w}_k$ 对系统的影响包含在系统状态中,这为下文 Q 学习算法的推导和应用做出了准备.

3 Q 学习算法的推导

为了求解随机系统最优追踪问题而无需系统参数信息,本节推导出在线求解随机系统追踪问题的 Q 学习算法.

引理 2[14] 给定容许控制 $\boldsymbol{u}_k=\boldsymbol{K}\boldsymbol{X}_k$,由引理 1 可知系统的最优值函数为

$$V^*(\boldsymbol{X}_k)=\mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{X}_k)=\mathrm{tr}(\boldsymbol{P}^*\boldsymbol{T}_k),\tag{15}$$

最优控制增益矩阵为

$$\boldsymbol{u}_k^*=\boldsymbol{K}^*\boldsymbol{X}_k=-(\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{D}_1)^{-1}\gamma(\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{A}_1+\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{C}_1)\boldsymbol{X}_k,\tag{16}$$

式中:矩阵 $\boldsymbol{P}^*$ 是式(17)增广随机代数方程的解.

$$\boldsymbol{P}=\boldsymbol{Q}_1+\gamma(\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}_1+\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}_1)-\gamma(\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B}_1+\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{D}_1)\times(\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{D}_1)^{-1}\times\gamma(\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}_1+\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}_1),\tag{17}$$

式中:$\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{D}_1>0$.

注 3 由引理 2 可知,求解 SLQ 最优追踪控制的有效方法之一就是求解增广随机代数方程.然而,求解式(17)需要完全已知系统的全部动力学信息,因此,当系统动力学参数未知时,式(17)就无法求解.

为了求解最优追踪控制无需系统的动力学信息,下文给出 Q 函数的定义和对应的 $\boldsymbol{H}$ 矩阵,基于式(7)得:

$$J(\boldsymbol{X}_k,\boldsymbol{u})=\mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{X}_k+\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_k)+\gamma\sum_{i=k+1}^{\infty}\gamma^{i-(k+1)}[\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{X}_i+\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_i],\tag{18}$$

由式(18)和贝尔曼最优性原理得:

$$V^*(\boldsymbol{X}_k)=\min_{\boldsymbol{u}_k}\{\mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{X}_k+\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_k)+\gamma V^*(\boldsymbol{X}_{k+1})\},\tag{19}$$

定义 Q 函数为

$$Q(\boldsymbol{X}_k,\boldsymbol{u}_k)=\mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{X}_k+\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_k)+\gamma\mathrm{E}(\boldsymbol{X}_{k+1}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}_{k+1}),\tag{20}$$

代入增广系统(5),式(20)变为

$$Q(\boldsymbol{X}_k,\boldsymbol{u}_k)=\mathrm{E}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{X}_k+\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_k)+\gamma\mathrm{E}((\boldsymbol{T}\boldsymbol{X}_k+\boldsymbol{B}_0\boldsymbol{u}_k)^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{T}\boldsymbol{X}_k+\boldsymbol{B}_0\boldsymbol{u}_k)),\tag{21}$$

化简式(21)可得:

$$Q(\boldsymbol{X}_k,\boldsymbol{u}_k)=\mathrm{E}\left\{\begin{bmatrix}\boldsymbol{X}_k\\\boldsymbol{u}_k\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{H}_{XX}&\boldsymbol{H}_{uX}\\\boldsymbol{H}_{Xu}&\boldsymbol{H}_{uu}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_k\\\boldsymbol{u}_k\end{bmatrix}\right\}.\tag{22}$$

进一步化简式(22):

$$Q(\boldsymbol{X}_k,\boldsymbol{u}_k)=\mathrm{E}\left\{\begin{bmatrix}\boldsymbol{X}_k\\\boldsymbol{u}_k\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{H}_{XX}&\boldsymbol{H}_{uX}\\\boldsymbol{H}_{Xu}&\boldsymbol{H}_{uu}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_k\\\boldsymbol{u}_k\end{bmatrix}\right\}=\mathrm{E}\left\{\begin{bmatrix}\boldsymbol{X}_k\\\boldsymbol{u}_k\end{bmatrix}^{\mathrm{T}}\boldsymbol{H}\begin{bmatrix}\boldsymbol{X}_k\\\boldsymbol{u}_k\end{bmatrix}\right\}.\tag{23}$$

式中:$\boldsymbol{H}=\boldsymbol{H}^{\mathrm{T}}\in\mathbf{R}^{(n+q+m)\times(n+q+m)}$,

$\boldsymbol{H}_{XX}=\gamma\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}_1+\gamma\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}_1+\boldsymbol{Q}_1\in\mathbf{R}^{(n+q)\times(n+q)}$,

$\boldsymbol{H}_{uX}=\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}_1\in\mathbf{R}^{m\times(n+q)}$,

$\boldsymbol{H}_{Xu}=\gamma\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B}_1+\gamma\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{D}_1\in\mathbf{R}^{(n+q)\times m}$,

$\boldsymbol{H}_{uu}=\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{D}_1\in\mathbf{R}^{m\times m}$.

由式(16)可知:$\boldsymbol{u}_k^*=\boldsymbol{K}^*\boldsymbol{X}_k=-(\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{D}_1)^{-1}\gamma(\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{A}_1+\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}^*\boldsymbol{C}_1)\boldsymbol{X}_k$.当控制策略 $\boldsymbol{u}_k^*$ 为最优的控制策略时,$Q^*(\boldsymbol{X}_k,\boldsymbol{u}_k)$ 和 $V^*(\boldsymbol{X}_k)$ 是等价的,由最优化一阶必要条件,通过微分 $\frac{\partial}{\partial\boldsymbol{u}_k}Q(\boldsymbol{X}_k,\boldsymbol{u}_k)=0$,可以得到:

$$\boldsymbol{K}^*=-\boldsymbol{H}_{uu}^{-1}\boldsymbol{H}_{uX}.\tag{24}$$

并且 Q 函数可重写为

$$Q(\boldsymbol{X}_k,\boldsymbol{u}_k)=\mathrm{E}\{[\boldsymbol{X}_k^{\mathrm{T}}\quad\boldsymbol{u}_k^{\mathrm{T}}]\boldsymbol{H}[\boldsymbol{X}_k^{\mathrm{T}}\quad\boldsymbol{u}_k^{\mathrm{T}}]^{\mathrm{T}}\}.\tag{25}$$

由引理 1 和式(25)可知 $\boldsymbol{P}$ 和 $\boldsymbol{H}$ 的联系为

$$\boldsymbol{P}=[\boldsymbol{I}\quad\boldsymbol{K}^{\mathrm{T}}]\boldsymbol{H}[\boldsymbol{I}\quad\boldsymbol{K}^{\mathrm{T}}]^{\mathrm{T}}.\tag{26}$$

通过式(24)可知,求解随机线性二次最优追踪控制增益矩阵摆脱了系统模型参数的限制,系统控制的最优解仅依赖于 $\boldsymbol{H}$ 矩阵,$\boldsymbol{H}$ 矩阵包含了系统的动态特性信息,如果 $\boldsymbol{H}$ 矩阵已知,就能够实现在无需系统参数的情况下在线求解系统的最优控制增益矩阵 $\boldsymbol{K}^*$.

文献[16]使用 Q 学习算法解决确定性离散系统的 H_∞ 最优控制问题,文献[13]使用 Q 学习算法解决随机离散系统的最优调节控制问题,并且均证明了Q学习算法的收敛性.本文的Q学习算法推导过程基于文献[13]和文献[16].基于值迭代方法[13],由式(25)可得:

$$\mathrm{E}\{[\boldsymbol{X}_k^{\mathrm{T}}\quad\boldsymbol{u}_k^{\mathrm{T}}]\boldsymbol{H}_{i+1}[\boldsymbol{X}_k^{\mathrm{T}}\quad\boldsymbol{u}_k^{\mathrm{T}}]^{\mathrm{T}}\}=\mathrm{E}\{\boldsymbol{X}_k^{\mathrm{T}}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_{i+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{X}_k\}=\mathrm{tr}\{[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_{i+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k\}.\tag{27}$$

同时由式(21)、(25)以及式(20)的右半部分化简为

$$\begin{aligned}&\mathrm{E}\{(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{X}_k+\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_k)+\gamma[\boldsymbol{X}_{k+1}^{\mathrm{T}}\quad\boldsymbol{u}_{k+1}^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{X}_{k+1}^{\mathrm{T}}\quad\boldsymbol{u}_{k+1}^{\mathrm{T}}]^{\mathrm{T}}\}=\\&\quad\mathrm{E}\{[\boldsymbol{X}_k^{\mathrm{T}}\quad\boldsymbol{u}_k^{\mathrm{T}}]\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}[\boldsymbol{X}_k^{\mathrm{T}}\quad\boldsymbol{u}_k^{\mathrm{T}}]^{\mathrm{T}}+\\&\quad\gamma[\boldsymbol{X}_{k+1}^{\mathrm{T}}\quad\boldsymbol{u}_{k+1}^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{X}_{k+1}^{\mathrm{T}}\quad\boldsymbol{u}_{k+1}^{\mathrm{T}}]^{\mathrm{T}}\}=\\&\quad\mathrm{E}\{\boldsymbol{X}_k^{\mathrm{T}}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{X}_k+\\&\quad\gamma\boldsymbol{X}_{k+1}^{\mathrm{T}}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{X}_{k+1}\}=\\&\quad\mathrm{tr}\{[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k+\\&\quad\gamma[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_{k+1}\}.\end{aligned}\tag{28}$$

由式(27)、(28),可以得到:

$$\begin{aligned}&\mathrm{tr}\{[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_{i+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k\}=\\&\qquad\mathrm{tr}\{[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k+\\&\qquad\gamma[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_{k+1}\}.\end{aligned}\tag{29}$$

由式(24)得控制迭代式:

$$\boldsymbol{K}_i=-(\boldsymbol{H}_{uu}^{-1}\boldsymbol{H}_{uX})_i.\tag{30}$$

式(29)、(30)构成了 Q 学习算法,可知系统参数信息 $\boldsymbol{A}\in\mathbf{R}^{n\times n},\boldsymbol{B}\in\mathbf{R}^{n\times m},\boldsymbol{C}\in\mathbf{R}^{n\times n},\boldsymbol{D}\in\mathbf{R}^{n\times m}$ 均包含在系统状态 $\boldsymbol{T}_k$ 和 $\boldsymbol{T}_{k+1}$ 之中,因此系统矩阵信息 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}$ 矩阵可以是未知的.接下来说明 Q 学习算法的等价性和收敛性.

定理 1 Q 学习算法式(29)—(30)等价于

$$\begin{aligned}\boldsymbol{P}_{i+1}=&\boldsymbol{Q}_1+\gamma(\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{A}_1+\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{C}_1)-\\&\gamma(\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{B}_1+\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{D}_1)\times\\&(\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{D}_1)^{-1}\times\\&\gamma(\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{A}_1+\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{C}_1).\end{aligned}\tag{31}$$

证明 由于

$$\begin{aligned}&\mathrm{E}\{[\boldsymbol{X}_{k+1}^{\mathrm{T}}\quad\boldsymbol{u}_i^{\mathrm{T}}(\boldsymbol{X}_{k+1})]\boldsymbol{H}_i[\boldsymbol{X}_{k+1}^{\mathrm{T}}\quad\boldsymbol{u}_i^{\mathrm{T}}(\boldsymbol{X}_{k+1})]^{\mathrm{T}}\}=\\&\quad\mathrm{E}\{\boldsymbol{X}_{k+1}^{\mathrm{T}}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{X}_{k+1}\}=\\&\quad\mathrm{E}\{(\boldsymbol{T}\boldsymbol{X}_k+\boldsymbol{B}_0\boldsymbol{u}_k)^{\mathrm{T}}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}(\boldsymbol{T}\boldsymbol{X}_k+\boldsymbol{B}_0\boldsymbol{u}_k)\},\end{aligned}\tag{32}$$

代入式(26),化简式(27)和式(28)的左半部分后,得:

$$\begin{aligned}\boldsymbol{H}_{i+1}=&\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}+\begin{bmatrix}\gamma\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{A}_1&\gamma\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{B}_1\\\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{A}_1&\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{B}_1\end{bmatrix}+\\&\begin{bmatrix}\gamma\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{C}_1&\gamma\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{D}_1\\\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{C}_1&\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{D}_1\end{bmatrix},\end{aligned}\tag{33}$$

又因为有:

$$\boldsymbol{P}_{i+1}=[\boldsymbol{I}\quad\boldsymbol{K}_{i+1}^{\mathrm{T}}]\boldsymbol{H}_{i+1}[\boldsymbol{I}\quad\boldsymbol{K}_{i+1}^{\mathrm{T}}]^{\mathrm{T}}.\tag{34}$$

结合式(30)、(33)和(34)可得:

$$\begin{aligned}\boldsymbol{P}_{i+1}=&\boldsymbol{Q}_1+\gamma(\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{A}_1+\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{C}_1)-\\&\gamma(\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{B}_1+\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{D}_1)\times\\&(\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{D}_1)^{-1}\times\\&\gamma(\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{A}_1+\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{C}_1).\end{aligned}\tag{35}$$

式中:$\boldsymbol{R}+\gamma\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B}_1+\gamma\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{P}\boldsymbol{D}_1>0$.

Q 学习算法基于值迭代算法,值迭代算法的收敛性在文献[14]已经有详细的证明,并且 Q 学习算法的迭代和增广系统的随机代数方程自身的迭代是等价的.

4 Q 学习算法的实现

本节给出 Q 学习算法的具体实现步骤.

式(29)可进一步化简为

$$\begin{aligned}&\mathrm{tr}\{\boldsymbol{H}_{i+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\}=\\&\qquad\mathrm{tr}\left\{\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]+\right.\\&\qquad\left.\gamma\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_{k+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\right\}.\end{aligned}\tag{36}$$

为了方便计算式(36),另记

$$\boldsymbol{Z}_k=[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}],\boldsymbol{G}=\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix},$$

使用最小二乘法求解式(36)之前,首先引入三个算子:

$$\begin{aligned}&\boldsymbol{H}\in\mathbf{R}^{(n+q+m)\times(n+q+m)}\rightarrow\hat{\boldsymbol{H}}\in\mathbf{R}^{\frac{1}{2}(n+q+m)\times(n+q+m+1)},\\&\boldsymbol{Z}_k\in\mathbf{R}^{(n+q+m)\times(n+q+m)}\rightarrow\hat{\boldsymbol{Z}}_k\in\mathbf{R}^{\frac{1}{2}(n+q+m)\times(n+q+m+1)},\\&\boldsymbol{G}\in\mathbf{R}^{(n+q+m)\times(n+q+m)}\rightarrow\hat{\boldsymbol{G}}\in\mathbf{R}^{\frac{1}{2}(n+q+m)\times(n+q+m+1)}.\end{aligned}$$

式中:

$$\begin{aligned}&\hat{\boldsymbol{H}}=[h_{11},2h_{12},2h_{13},\cdots,2h_{1n},h_{22},2h_{23},\cdots,2h_{(n-1)n},h_{nn}]^{\mathrm{T}},\\&\hat{\boldsymbol{Z}}_k=[z_{11},z_{12},z_{13},\cdots,z_{1n},z_{22},z_{23},\cdots,z_{(n-1)n},z_{nn}]^{\mathrm{T}},\\&\hat{\boldsymbol{G}}=[g_{11},2g_{12},2g_{13},\cdots,2g_{1n},g_{22},2g_{23},\cdots,2g_{(n-1)n},g_{nn}]^{\mathrm{T}}.\end{aligned}$$

因此,式(36)可化简为

$$\begin{aligned}&\mathrm{tr}\{\boldsymbol{H}_{i+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\}=\hat{\boldsymbol{Z}}_k^{\mathrm{T}}\hat{\boldsymbol{H}}_{i+1},\\&\mathrm{tr}\left\{\begin{bmatrix}\boldsymbol{Q}_1&\boldsymbol{O}\\\boldsymbol{O}&\boldsymbol{R}\end{bmatrix}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_k[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]+\right.\\&\qquad\left.\gamma\boldsymbol{H}_i[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{T}_{k+1}[\boldsymbol{I}\quad\boldsymbol{K}_i^{\mathrm{T}}]\right\}=\hat{\boldsymbol{G}}^{\mathrm{T}}\hat{\boldsymbol{Z}}_k+\gamma\hat{\boldsymbol{Z}}_{k+1}^{\mathrm{T}}\hat{\boldsymbol{H}}_i.\end{aligned}$$

由最小二乘法可知,至少需要收集 N 组数据,即 $N\geqslant\frac{1}{2}(n+q+m)(n+q+m+1)$.记 $\hat{\boldsymbol{Z}}_{k+N}^{\mathrm{T}}$ 为收集的 N 个状态信息构成的矩阵,$\hat{\boldsymbol{Z}}_{k+N}^{\mathrm{T}}=[\hat{\boldsymbol{Z}}_1,\hat{\boldsymbol{Z}}_2,\hat{\boldsymbol{Z}}_3,\cdots,\hat{\boldsymbol{Z}}_N]^{\mathrm{T}}$,如果矩阵 $\hat{\boldsymbol{Z}}_{k+N}^{\mathrm{T}}$ 是列满秩的,则式(36)可以由下式直接求解:

$$\hat{\boldsymbol{H}}_{i+1}=(\hat{\boldsymbol{Z}}_{k+N}\hat{\boldsymbol{Z}}_{k+N}^{\mathrm{T}})^{-1}\hat{\boldsymbol{Z}}_{k+N}(\hat{\boldsymbol{G}}^{\mathrm{T}}\hat{\boldsymbol{Z}}_{k+N}+\gamma\hat{\boldsymbol{Z}}_{k+1+N}^{\mathrm{T}}\hat{\boldsymbol{H}}_i),\quad(37)$$

Q 学习算法的流程如图 1 所示.

注 4 因为 $\boldsymbol{H}$ 是维度为 $(n+q+m)\times(n+q+m)$ 的对称矩阵,因此可知 $\boldsymbol{H}$ 有 $\frac{1}{2}(n+q+m)\times(n+q+m+1)$ 个独立元素,所以求解式(37):

1) 至少需要与 $\boldsymbol{H}$ 独立元素相同数量的方程,也就是必须保证 $\mathrm{rank}(\hat{\boldsymbol{Z}}_{k+N}^{\mathrm{T}})=\frac{1}{2}(n+q+m)\times(n+q+m+1)$,收集足够多的数据以保证式(37)可解;

2) 同时在系统(13)需要引入探测噪声,以保证能够充分探测系统参数信息.

5 系统仿真

本节给出一个仿真实例来说明 Q 学习算法的有效性.

假设乘性噪声随机线性系统动态模型如下:

$$\begin{cases}\boldsymbol{x}_{k+1}=\begin{bmatrix}0.2 & -0.8\\ 0.5 & -0.7\end{bmatrix}\boldsymbol{x}_k+\begin{bmatrix}0.03\\ -0.5\end{bmatrix}\boldsymbol{u}_k+\\ \qquad\left(\begin{bmatrix}-0.04 & 0.4\\ -0.3 & 0.13\end{bmatrix}\boldsymbol{x}_k+\begin{bmatrix}0.05\\ -0.3\end{bmatrix}\boldsymbol{u}_k\right)\boldsymbol{w}_k,\\ \boldsymbol{y}_k=\begin{bmatrix}3 & 3\end{bmatrix}\boldsymbol{x}_k.\end{cases}\quad(38)$$

假设参考轨迹的动态模型如下:

$$\boldsymbol{r}_{k+1}=-\boldsymbol{r}_k.\quad(39)$$

设置 $\boldsymbol{Q}=10$,$\boldsymbol{R}=1$,$\gamma=0.9$,增广系统的初始状态 $\boldsymbol{X}_0=\begin{bmatrix}10\\10\\1\end{bmatrix}$,参考轨迹的初始状态 $\boldsymbol{r}_0=1$.更具一般性,初始化 $\boldsymbol{H}=\boldsymbol{I}$,此时 $\boldsymbol{K}=[0\ 0\ 0]$,在每一次迭代过程中,加入探测噪声并收集 20 组数据去求解更新控制增益矩阵 $\boldsymbol{K}$.经过 Q 学习算法迭代后,控制增益矩阵曲线如图 2 所示.

为了验证 Q 学习算法迭代求解的 $\boldsymbol{K}$ 矩阵的准确性,需要把迭代计算的 $\boldsymbol{K}$ 矩阵同 SAE 方程求解出的最优控制增益矩阵 $\boldsymbol{K}^*$ 比较.首先通过式(16)、(17)的 SAE 方程求得系统的最优控制矩阵 $\boldsymbol{K}^*$.$\boldsymbol{K}$ 矩阵收敛到最优控制矩阵 $\boldsymbol{K}^*$ 的过程如图 3 所示.由图 3 可以看出,随着迭代次数的增加,控制矩阵 $\boldsymbol{K}$ 收敛到最优控制矩阵 $\boldsymbol{K}^*$,并且 $\boldsymbol{P}$ 收敛到代数方程的解 $\boldsymbol{P}^*$.$\boldsymbol{P}$ 收敛到代数方程的解 $\boldsymbol{P}^*$ 的过程如图 4 所示,验证了 Q 学习算法的有效性.

追踪问题的目标是追踪到参考信号轨迹,结果如图 5 所示,系统的期望输出 $\mathrm{E}(\boldsymbol{y})$ 追踪到了参考轨迹 $\boldsymbol{r}$.

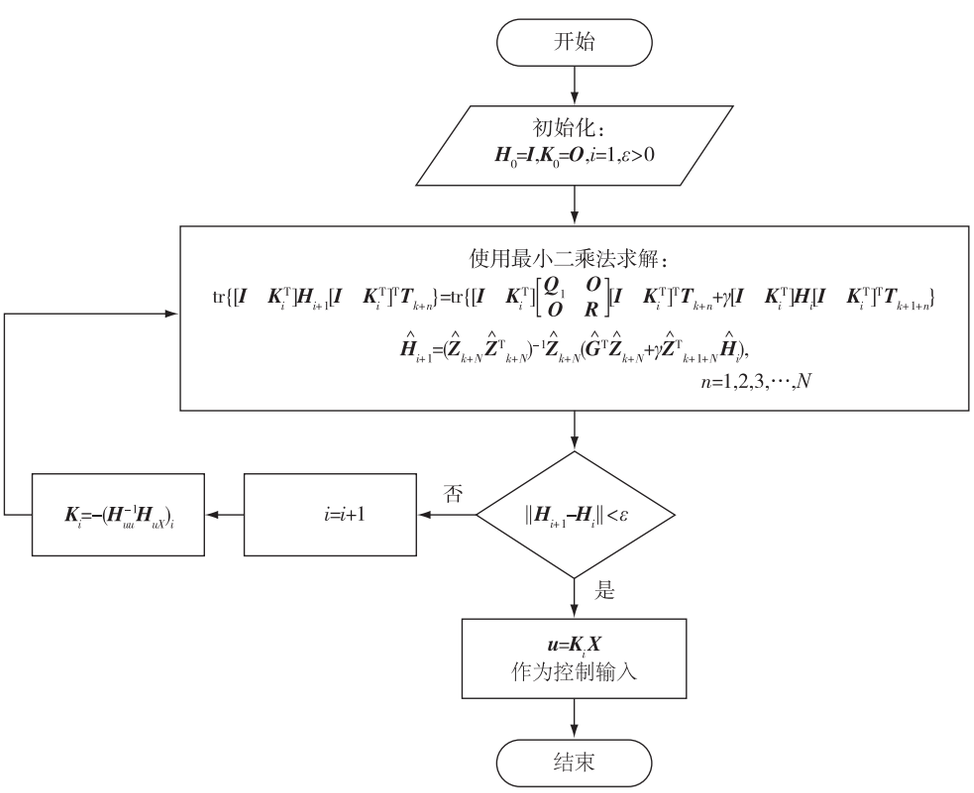

图 1 Q 学习算法流程

Fig. 1 Flowchart of Q-learning

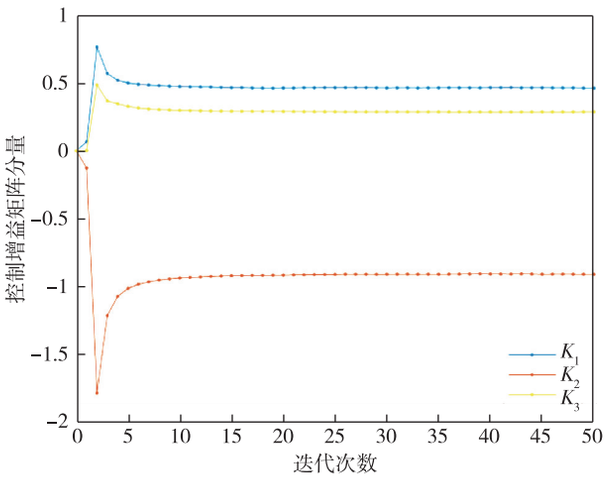

图 2　控制增益矩阵 $\boldsymbol{K}$ 曲线

Fig. 2　Curves of control gain matrix $\boldsymbol{K}$

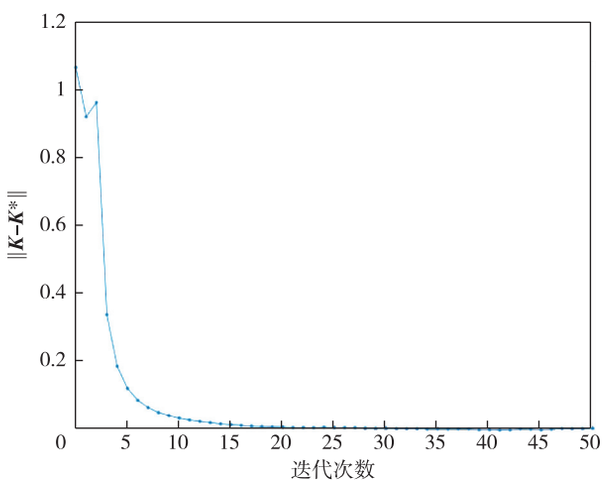

图 3　控制矩阵 $\boldsymbol{K}$ 的收敛

Fig. 3　Convergence of control gain matrix $\boldsymbol{K}$

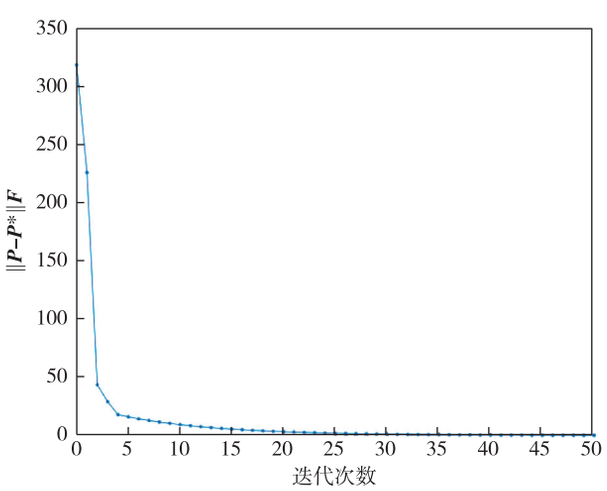

图 4　矩阵 $\boldsymbol{P}$ 的收敛

Fig. 4　Convergence of matrix $\boldsymbol{P}$

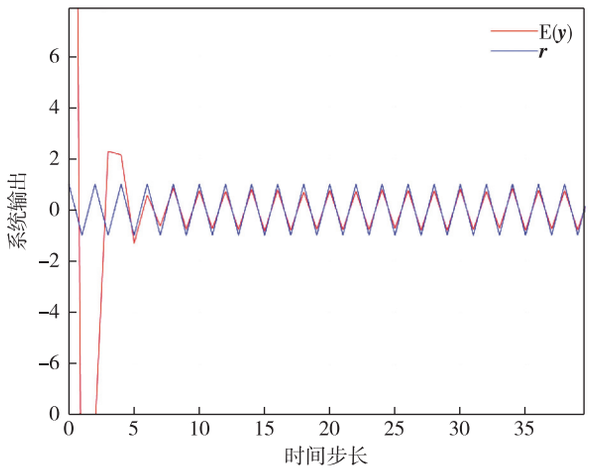

图 5　系统的期望输出曲线

Fig. 5　Curves of system output E($\boldsymbol{y}$)

6　结论

通常来说,求解随机最优追踪控制问题需要完全的系统参数信息,本文针对离散时间系统的随机线性二次最优追踪控制问题,推导了 Q 学习算法,给出算法的具体实现步骤,使用 Q 学习算法在线解决追踪控制问题而无需系统模型参数,最后给出仿真结果表明系统输出可以有效地追踪参考轨迹.

参考文献

References

[1] Byers R. Solving the algebraic Riccati equation with the matrix sign function[J]. Linear Algebra and Its Applications, 1987, 85: 267-279

[2] Kleinman, D. On an iterative technique for Riccati equation computations [J]. IEEE Transactions on Automatic Control, 1968, 13(1): 114-115

[3] Wang F Y, Zhang H G, Liu D R. Adaptive dynamic programming: an introduction[J]. IEEE Computational Intelligence Magazine, 2009, 4(2): 39-47

[4] Vrabie D, Pastravanu O, Abu-Khalaf M, et al. Adaptive optimal control for continuous-time linear systems based on policy iteration[J]. Automatica, 2009, 45(2): 477-484

[5] Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics [J]. Automatica, 2012, 48 (10): 2699-2704

[6] 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

[7] Zhao X Y, Deng F Q. Divided state feedback control of stochastic systems [J]. IEEE Transactions on Automatic Control, 2015, 60(7): 1870-1885

[8] Zhao X Y, Deng F Q. A new type of stability theorem for

stochastic systems with application to stochastic stabilization[J].IEEE Transactions on Automatic Control,2016,61(1):240-245
[9] 黄玉林,张维海.约束随机线性二次最优控制的研究[J].自动化学报,2006,32(2):246-254
HUANG Yulin, ZHANG Weihai. Study on stochastic linear quadratic optimal control with constraint[J]. Acta Automatica Sinica,2006,32(2):246-254
[10] Liu X K, Li Y, Zhang W H. Stochastic linear quadratic optimal control with constraint for discrete-time systems[J]. Applied Mathematics and Computation,2014,228:264-270
[11] 王涛,张化光.基于策略迭代的连续时间系统的随机线性二次最优控制[J].控制与决策,2015,30(9):1674-1678
WANG Tao, ZHANG Huaguang. Stochastic linear quadratic optimal control for continuous-time systems based on policy iteration[J].Control and Decision,2015,30(9):1674-1678
[12] Wang T, Zhang H G, Luo Y H. Infinite-time stochastic linear quadratic optimal control for unknown discrete-time systems using adaptive dynamic programming approach[J].Neurocomputing,2016,171:379-386
[13] Wang T,Zhang H G,Luo Y H.Stochastic linear quadratic optimal control for model-free discrete-time systems based on Q-learning algorithm[J]. Neurocomputing,2018,312:1-8
[14] Chen X, Wang F. Neural-network-based stochastic linear quadratic optimal tracking control scheme for unknown discrete-time systems using adaptive dynamic programming[J]. Control Theory and Technology,2021,19(3):315-327
[15] Xiao G Y, Zhang H G, Luo Y H, et al. Data-driven optimal tracking control for a class of affine non-linear continuous-time systems with completely unknown dynamics[J].IET Control Theory & Applications,2016,10(6):700-710
[16] Al-Tamimi A, Lewis F L, Abu-Khalaf M. Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control[J]. Automatica,2007,43(3):473-481

Stochastic linear quadratic optimal tracking control for stochastic discrete time systems based on Q-learning

ZHANG Zhengyi[1] ZHAO Xueyan[1]
1 School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640

Abstract For stochastic linear discrete time systems, a Q-learning algorithm is proposed in this paper to solve the stochastic linear quadratic optimal tracking control problem in the infinite time domain. First, it is assumed that the reference signal required for tracking is generated by the command generator, and an augmented system consisting of the original stochastic system and the reference trajectory system is established, then the optimal tracking problem is transformed into an optimal regulation problem. Second, in order to solve the optimal tracking problem online, the stochastic system is transformed into a deterministic one, the Q function of stochastic linear quadratic optimal tracking control is defined according to the augmented system, and the augmented stochastic algebraic equation is solved online without knowing the parameters of the system model. Third, the equivalence between the Q-learning algorithm and the augmented stochastic algebraic equation is proved, and the implementation steps of the Q-learning algorithm are given. Finally, a simulation example is given to illustrate the effectiveness of the proposed Q-learning algorithm.
Key words stochastic systems; Q-learning algorithm; optimal tracking control; stochastic algebraic equation