参数未知的离散系统Q-学习优化状态估计与控制

李金娜， 马士凯

引用本文:
李金娜, 马士凯. 参数未知的离散系统Q-学习优化状态估计与控制[J]. 控制与决策, 2020, 35(12): 2889-2897.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0180

参数未知的离散系统Q-学习优化状态估计与控制

李金娜[1,2†]，马士凯[1]

(1. 沈阳化工大学 信息工程学院，沈阳 110142；
2. 东北大学 流程工业综合自动化国家重点实验室，沈阳 110004)

摘 要： 控制系统的应用中存在状态不能直接测量或测量成本高的实际问题,给模型参数未知的系统完全利用状态数据学习最优控制器带来挑战性难题. 为解决这一问题,首先构建具有状态观测器且系统矩阵中存在未知参数的离散线性增广系统,定义性能优化指标;然后基于分离定理、动态规划以及Q-学习方法,给出一种具有未知模型参数的非策略Q-学习算法,并设计近似最优观测器,得到完全利用可测量的系统输出和控制输入数据的非策略Q-学习算法,实现基于观测器状态反馈的系统优化控制策略,该算法的优点在于不要求系统模型参数全部已知,不要求系统状态直接可测,利用可测量数据实现指定性能指标的优化;最后,通过仿真实验验证所提出方法的有效性.

关键词： 非策略Q-学习；最优控制；状态观测器；分离定理；离散系统；近似动态规划

中图分类号： TP13 **文献标志码：** A

DOI: 10.13195/j.kzyjc.2019.0180 **开放科学(资源服务)标识码(OSID):**

引用格式： 李金娜,马士凯. 参数未知的离散系统Q-学习优化状态估计与控制 [J]. 控制与决策, 2020, 35(12): 2889-2897.

Q-learning optimal state estimation and control for discrete systems with unknown parameters

LI Jin-na[1,2†], *MA Shi-kai*[1]

(1. College of Information Engineering，Shenyang University of Chemical Technology，Shenyang 110142，China；
2. State Key Lab of Synthetical Automation for Process Industries，Northeastern University，Shenyang 110004，China)

Abstract: In the application of control systems, there is a practical problem that the state cannot be directly measured or the measurement cost is high. In order to solve this problem, a linear discrete-time augmented system with unknown parameters and a state observer is first constructed and the prescribed performance index is defined. Then, based on the separation theorem, the dynamic programming theory and the Q-learning method, a novel off-policy Q-learning algorithm is developed to approximate the optimal observer and the optimal controller for systems with unknown parameters and unmeasured states, such that the control performance is minimized using only measured data. The advantage of this algorithm is that it does not require all the system model parameters to be known and the system state to be directly measurable. Finally, the simulation experiment verifies the effectiveness of the proposed method.

Keywords: off-policy Q-learning; optimal control; state observer; separation theorem; discrete-time systems; approximate dynamic programming

0 引 言

根据系统外部变量(输入变量和输出变量)的实测值得出状态变量估计值的一类动态系统称为状态重构器. 20世纪60年代初期,为了对控制系统实现状态反馈或其他需要,人们提出了状态观测器的概念和构造方法,通过重构的途径解决了状态不能直接测量的问题. 构成状态观测器的方法依需要的不同有所差别,最简单的是开环状态观测器,开环观测器对外界干扰的抗干扰性和对参数变动的灵敏度都很差,采用闭环方式构成的状态观测器能够克服开环状态观测器的缺点[1-3]. 目前状态观测器设计普遍采用依赖系统模型信息的方法[4-6],系统模型存在未知参数,完全利用可测量数据的状态观测器设计目前还较少见到报道.

收稿日期： 2019-02-19；**修回日期：** 2019-06-17.
基金项目： 国家自然科学基金项目(61673280)；辽宁省高等学校创新人才项目(LR2017006).
责任编委： 徐胜元.
†通讯作者. E-mail: lijinna_721@126.com.

强化学习(reinforcement learning, RL)是一种从环境状态到动作映射的学习,并期望动作从环境中获得的累积奖赏最大[7-10]. 从20世纪80年代末开始,随着对强化学习的数学基础研究取得突破性进展,对强化学习的研究和应用日益开展起来,强化学习成为目前机器学习领域的研究热点之一,近年来在工程应用、模式识别、图像处理、网络优化等领域都得到了广泛应用[11]. 根据目标策略和行为策略是否一致将强化学习分为策略(on-policy)学习和非策略(off-policy)学习. 如果在学习过程中,动作选择的行为策略和学习改进的目标策略一致,则该方法称为策略学习(如Sarsa学习),否则称为非策略学习(如非策略Q-学习)[12-15].

采用强化学习方法不依赖系统模型参数信息,完全利用数据学习状态反馈控制策略、优化控制系统性能的研究成果较多,如线性系统最优二次调节、最优跟踪控制等[16-19],这些方法都假定系统的状态可以测量. 然而,在实际应用中,系统的状态可能不是物理量,不能直接测量,或是测量的成本非常高,这种情况系统无法使用基于可测量的状态信息学习得到最优控制器,并且所设计的状态反馈控制器无法实现对系统控制和优化. 对于具有未知参数的系统状态反馈最优控制,且系统状态不能直接测量的问题,目前还没有得到充分研究. 文献[17]利用系统当前和过去的可测输出和输入数据估计当前系统状态,给出基于强化学习方法的近似最优控制策略. 此种方法对于高阶系统而言计算复杂,且需要存储大量系统当前和过去的输出和输入数据.

本文针对具有未知参数的线性离散系统,考虑系统状态不可测的情况,提出一种非策略Q-学习算法,给出了基于观测器状态反馈的控制器增益和状态观测器增益学习算法. 该算法不需要系统模型参数完全已知,利用可测的当前时刻输出和输入数据优化控制系统性能. 算例仿真验证了所提出方法的有效性. 本文的创新性在于: 1)不同于模型依赖的状态观测器设计方法[4-6],本文给出了系统模型存在未知参数情况下,完全利用可测量数据的非策略Q-学习算法学习最优状态观测器增益; 2)现有的无模型控制器设计方法中往往假定系统状态直接可测[16-19],文献[17]研究了系统状态不可测情况下的无模型最优控制器学习算法,不同于文献[17],本文提出一种非策略Q-学习算法,学习基于状态观测器的状态反馈最优控制器增益.

1 具有状态观测器的最优控制问题阐述

考虑如下具有未知模型参数的线性离散系统的状态方程:

$$\begin{cases} x_{k+1} = A(\theta)x_k + B(\theta)u_k, \\ y_k = C(\theta)x_k. \end{cases} \tag{1}$$

其中: $x_k = x(k) \in R^n$、$u_k = u(k) \in R^m$、$y_k = y(k) \in R^p$ 分别为被控状态、控制输入和控制输出, $A(\theta)$、$B(\theta)$ 和 $C(\theta)$ 为适维矩阵, $k(k = 0, 1, \ldots)$ 为采样时刻, θ 为未知参数或未知向量. 假设 $(A(\theta), B(\theta))$ 能控, $(A(\theta), C(\theta))$ 能观.

鉴于实际控制系统状态很难直接测量或者只能测量部分状态信息,且测量成本较高,设计如下全维状态观测器:

$$\begin{cases} \hat{x}_{k+1} = A(\theta)\hat{x}_k + B(\theta)u_k + L(y_k - \hat{y}_k), \\ \hat{y}_k = C(\theta)\hat{x}_k. \end{cases} \tag{2}$$

其中: $\hat{x}_k$ 为 x 的重构状态或估计状态, $\hat{y}_k$ 为观测器输出, L 为观测器增益矩阵. 定义实际状态和状态估计值之差为误差向量,有

$$e_k = x_k - \hat{x}_k. \tag{3}$$

得到误差向量的动态方程为

$$e_{k+1} = (A(\theta) - LC(\theta))e_k. \tag{4}$$

选用基于观测器状态的静态反馈控制器

$$u_k = K\hat{x}_k. \tag{5}$$

控制器(5)的目标是最小化如下二次性能指标:

$$\frac{1}{2}\sum_{k=0}^{\infty} y_k^{\mathrm{T}} Q_p y_k + u_k^{\mathrm{T}} R_p u_k, \tag{6}$$

其中 Q_p 和 R_p 分别为半正定和正定矩阵. 如此,本文关注的最优控制问题可以表述如下.

问题1

$$\min_{u_k} \frac{1}{2}\sum_{k=0}^{\infty} y_k^{\mathrm{T}} Q_p y_k + u_k^{\mathrm{T}} R_p u_k;$$
$$\text{s.t. 式 (1)、(4)和(5).} \tag{7}$$

令 $\xi_k = \begin{bmatrix} x_k \\ e_k \end{bmatrix}$,由式(1)、(4)和(5)得到如下闭环增广控制系统:

$$\xi_{k+1} = \begin{bmatrix} A(\theta) + B(\theta)K & -B(\theta)K \\ 0 & A(\theta) - LC(\theta) \end{bmatrix} \xi_k. \tag{8}$$

由式(8)可知,闭环系统的极点由 $A(\theta) + B(\theta)K$ 和 $A(\theta) - LC(\theta)$ 的特征值确定,可以分别设计 L 和 K,从而配置闭环系统的极点. 因此对于优化问题1,分离定理[20-21]成立.

注1 配置 $A(\theta) + B(\theta)K$ 和 $A(\theta) - LC(\theta)$ 特征

值的方法较多,如极点配置、李雅普诺夫方法等. 本文关注的是在系统模型含有未知参数的情况下,如何完全利用可测数据分别设计L和K,保证闭环系统稳定,并优化性能指标(7).

2　最优观测器设计

本节主要给出不利用被控系统和观测器系统矩阵$A(\theta)$、$B(\theta)$和$C(\theta)$的最优观测器非策略Q-学习算法学习最优观测器. 引入观测器策略$w_k = L(y_k - \hat{y}_k)$,定义如下优化问题.

问题2

$$\min_{w_k} \sum_{k=0}^{\infty} (y_k - \hat{y}_k)^{\mathrm{T}} Q_1 (y_k - \hat{y}_k) + w_k^{\mathrm{T}} R_1 w_k; \quad (9)$$

$$\text{s.t. } e_{k+1} = A(\theta) e_k - w_k. \quad (10)$$

先给出基于系统模型的最优观测器策略w_k^*,得到不利用系统矩阵数据驱动的最优观测器策略w_k^*设计方法. 所设计的最优观测器策略w_k^*不仅要保证观测器误差收敛到零,而且能够优化性能指标(9).

2.1　依赖模型的最优观测器设计

由性能指标(9),最优值函数和最优Q函数为

$$V_o^*(e_k) = \min_{w} \sum_{i=k}^{\infty} e_i^{\mathrm{T}} \tilde{Q}_1 e_i + w_i^{\mathrm{T}} R_1 w_i; \quad (11)$$

$$Q_o^*(e_k, w_k) = e_k^{\mathrm{T}} \tilde{Q}_1 e_k + w_k^{\mathrm{T}} R_1 w_k + V_o^*(e_{k+1}). \quad (12)$$

其中

$$\tilde{Q}_1 = C(\theta)^{\mathrm{T}} Q_1 C(\theta),$$

$$V_o^*(e_{k+1}) = \min_{w} \sum_{i=k+1}^{\infty} (e_i^{\mathrm{T}} \tilde{Q}_1 e_i + w_i^{\mathrm{T}} R_1 w_i).$$

得到最优值函数和最优Q函数的关系为

$$V_o^*(e_k) = \min_{u_k} Q_o^*(e_k, w_k) = Q_o^*(e_k, w_k^*). \quad (13)$$

引理1 [18]　对于优化问题2,如果观测器策略为$w_k = L(y_k - \hat{y}_k)$,则值函数V_o^*和Q函数Q_o^*可以表示为如下二次型:

$$V_o^*(e_k) = \frac{1}{2} e_k^{\mathrm{T}} P_o e_k, \quad (14)$$

$$Q_o^*(e_k, w_k) = \frac{1}{2} \begin{bmatrix} e_k \\ w_k \end{bmatrix}^{\mathrm{T}} H_1 \begin{bmatrix} e_k \\ w_k \end{bmatrix}. \quad (15)$$

其中: P_o为正定矩阵,且有

$$H_1 = \begin{bmatrix} \tilde{Q}_1 + A(\theta)^{\mathrm{T}} P_o A(\theta) & -A(\theta)^{\mathrm{T}} P_o \\ * & R_2 + P_o \end{bmatrix} = \begin{bmatrix} H_{1,ee} & H_{1,ew} \\ H_{1,we} & H_{1,ww} \end{bmatrix}, \quad (16)$$

$$P_o = \begin{bmatrix} I \\ LC(\theta) \end{bmatrix}^{\mathrm{T}} H_1 \begin{bmatrix} I \\ LC(\theta) \end{bmatrix}. \quad (17)$$

基于动态规划,由式(12)得到基于Q函数的贝尔曼(Bellman)方程如下:

$$\begin{bmatrix} e_k \\ w_k \end{bmatrix}^{\mathrm{T}} H_1 \begin{bmatrix} e_k \\ w_k \end{bmatrix} = e_k^{\mathrm{T}} \tilde{Q}_1 e_k + w_k^{\mathrm{T}} R_1 w_k + \begin{bmatrix} e_{k+1} \\ w_{k+1} \end{bmatrix}^{\mathrm{T}} H_1 \begin{bmatrix} e_{k+1} \\ w_{k+1} \end{bmatrix} = e_k^{\mathrm{T}} \tilde{Q}_1 e_k + w_k^{\mathrm{T}} R_1 w_k + \begin{bmatrix} e_k \\ w_k \end{bmatrix}^{\mathrm{T}} \Big(\begin{bmatrix} I \\ LC(\theta) \end{bmatrix} [A(\theta) \quad -I] \Big)^{\mathrm{T}} H_1 \cdot \Big(\begin{bmatrix} I \\ LC(\theta) \end{bmatrix} [A(\theta) \quad -I] \Big) \begin{bmatrix} e_k \\ w_k \end{bmatrix}. \quad (18)$$

根据实现最优性能的必要条件,由$\partial Q_o^*/\partial w_k = 0$可得

$$\begin{cases} w_k^* = -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} e_k, \\ L^* C(\theta) = -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}}. \end{cases} \quad (19)$$

其中

$$\begin{cases} H_{1,we} = -(A(\theta)^{\mathrm{T}} P_o)^{\mathrm{T}}, \\ H_{1,ww} = R_1 + P_o. \end{cases} \quad (20)$$

将式(19)代入(18),得到代数黎卡提方程

$$\begin{bmatrix} e_k \\ -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} e_k \end{bmatrix}^{\mathrm{T}} H_1 \begin{bmatrix} e_k \\ -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} e_k \end{bmatrix} = e_k^{\mathrm{T}} \tilde{Q}_1 e_k + e_k^{\mathrm{T}} (H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}})^{\mathrm{T}} R_1 H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} e_k + \begin{bmatrix} e_k \\ -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} e_k \end{bmatrix}^{\mathrm{T}} \Big(\begin{bmatrix} I \\ -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} \end{bmatrix} \cdot [A(\theta) \quad -I] \Big)^{\mathrm{T}} H_1 \Big(\begin{bmatrix} I \\ -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} \end{bmatrix} [A(\theta) \quad -I] \Big) \cdot \begin{bmatrix} e_k \\ -H_{1,ww}^{-1} (H_{1,we})^{\mathrm{T}} e_k \end{bmatrix}. \quad (21)$$

为求解式(21)中Q函数矩阵H_1,给出如下算法.

算法1　基于模型的策略迭代算法.

step 1: 初始化. 给定保证估计误差稳定的观测器增益L^0,令$j = 0$,j为迭代指标.

step 2: 策略评估. 通过式(22)求解Q函数迭代矩阵H_1^{j+1},有

$$H_1^{j+1} = \begin{bmatrix} \tilde{Q}_1 & 0 \\ 0 & R_1 \end{bmatrix} + \Big(\begin{bmatrix} I \\ L^j C(\theta) \end{bmatrix} [A(\theta) \quad -I] \Big)^{\mathrm{T}} H_1^{j+1} \cdot \Big(\begin{bmatrix} I \\ L^j C(\theta) \end{bmatrix} [A(\theta) \quad -I] \Big). \quad (22)$$

step 3: 策略更新. 有

$$w_k^{j+1}=-(H_{1,ww}^{j+1})^{-1}(H_{1,we}^{j+1})^{\mathrm{T}}e_k. \tag{23}$$

step 4: 如果$\|H_1^j-H_1^{j+1}\|\leqslant\varepsilon(\varepsilon>0)$,则算法停止,否则令$j=j+1$,转至step 2.

注2 文献[18-19]已经证明$\lim\limits_{j\to\infty}H_1^{j+1}=H_1$,$\lim\limits_{j\to\infty}w_k^{j+1}=w_k^*$(其中$H_1$为式(21)的解).

由算法1可知,学习Q函数矩阵H_1要求系统矩阵$A(\theta)$和$C(\theta)$精确已知. 然而,实际控制系统模型中含有未知参数,算法1无法执行. 完全利用可测数据,针对具有未知参数的系统设计最优观测器是本文研究的目的. 下一节将给出一种非策略的Q-学习算法,在系统矩阵$A(\theta)$、$B(\theta)$和$C(\theta)$中含有未知参数的情况下,学习得到近似最优观测器增益.

2.2 最优观测器非策略Q-学习算法设计

引入两种操作,一种是定义一个与最优Q函数矩阵相关的虚拟Q函数矩阵$\bar{H}_1$,有

$$H_1=\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}^{\mathrm{T}}\bar{H}_1\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}; \tag{24}$$

另一种是在系统(4)中引入一个辅助变量$w_k^j=-(H_{1,ww}^j)^{-1}(H_{1,we}^j)^{\mathrm{T}}e_k$,得到

$$\begin{aligned}e_{k+1}&=A(\theta)e_k-w_k-w_k^j+w_k^j=\\&\quad(A(\theta)e_k-w_k^j)+(w_k^j-w_k).\end{aligned} \tag{25}$$

其中:w_k用于产生系统数据,称为行为策略;w_k^i为目标策略,本文目的是w_k^i收敛到最优观测器策略. 由式(16)和(24)得到

$$\begin{aligned}H_1&=\begin{bmatrix}\tilde{Q}_1+A(\theta)^{\mathrm{T}}P_oA(\theta)&-A(\theta)^{\mathrm{T}}P_o\\*&R_1+P_o\end{bmatrix}=\\&\begin{bmatrix}H_{1,ee}&H_{1,ew}\\H_{1,we}&H_{1,ww}\end{bmatrix}=\\&\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}^{\mathrm{T}}\bar{H}_1\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}=\\&\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\bar{H}_{1,11}&\bar{H}_{1,12}\\*&\bar{H}_{1,22}\end{bmatrix}\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}=\\&\begin{bmatrix}C(\theta)^{\mathrm{T}}\bar{H}_{1,11}C(\theta)&C(\theta)^{\mathrm{T}}\bar{H}_{1,12}\\*&\bar{H}_{1,22}\end{bmatrix}.\end{aligned} \tag{26}$$

沿着系统轨迹(25),结合式(22)和(26),得到

$$\begin{aligned}&\begin{bmatrix}y_k-\hat{y}_k\\w_k\end{bmatrix}^{\mathrm{T}}\bar{H}_1^{j+1}\begin{bmatrix}y_k-\hat{y}_k\\w_k\end{bmatrix}-\\&(y_{k+1}-\hat{y}_{k+1})^{\mathrm{T}}\begin{bmatrix}I\\L^j\end{bmatrix}^{\mathrm{T}}\bar{H}_1^{j+1}\begin{bmatrix}I\\L^j\end{bmatrix}(y_{k+1}-\hat{y}_{k+1})=\\&(y_k-\hat{y}_k)^{\mathrm{T}}Q_1(y_k-\hat{y}_k)+(w_k^j)^{\mathrm{T}}R_1w_k^j+\\&2(y_k^{\mathrm{T}}-\hat{y}_k^{\mathrm{T}})^{\mathrm{T}}\bar{H}_{1,12}^{j+1}(w_k^j-w_k)+\\&w_k^j(\bar{H}_{1,22}^{j+1}-R_1)(w_k^j-w_k)+\\&w_k^{\mathrm{T}}(\bar{H}_{1,22}^{j+1}-R_1)(w_k^j-w_k).\end{aligned} \tag{27}$$

令$e_{y_k}=y_k-\hat{y}_k$,式(27)改写为

$$\theta^j(k)h_o^{j+1}=\rho_k^j. \tag{28}$$

其中

$$\begin{aligned}&\rho_k^j=e_{y_k}^{\mathrm{T}}Q_1e_{y_k}+w_k^{\mathrm{T}}R_1w_k,\\&h_o^{j+1}=\\&[(\mathrm{vec}(\bar{H}_{1,11}^{j+1}))^{\mathrm{T}}\ \ (\mathrm{vec}(\bar{H}_{1,12}^{j+1}))^{\mathrm{T}}\ \ (\mathrm{vec}(\bar{H}_{1,22}^{j+1}))^{\mathrm{T}}]^{\mathrm{T}},\\&\theta^j(k)=[\theta_1^j\ \ \theta_2^j\ \ \theta_3^j],\\&\theta_1^j=e_{y_k}^{\mathrm{T}}\otimes e_{y_k}^{\mathrm{T}}-e_{y_{k+1}}^{\mathrm{T}}\otimes e_{y_{k+1}}^{\mathrm{T}},\\&\theta_2^j=4e_{y_k}^{\mathrm{T}}\otimes w_k-2e_{y_{k+1}}^{\mathrm{T}}\otimes(L^je_{y_{k+1}})^{\mathrm{T}}-\\&\qquad2e_{y_k}^{\mathrm{T}}\otimes(L^je_{y_k})^{\mathrm{T}},\\&\theta_3^j=2w_k^{\mathrm{T}}\otimes w_k^{\mathrm{T}}-(L^je_{y_{k+1}})^{\mathrm{T}}\otimes(L^je_{y_{k+1}})^{\mathrm{T}}-\\&\qquad(L^je_{y_k})^{\mathrm{T}}\otimes(L^je_{y_k})^{\mathrm{T}}.\end{aligned} \tag{29}$$

由式(26),式(23)改写为

$$\begin{aligned}w_k^{j+1}&=-(H_{1,ww}^{j+1})^{-1}(H_{1,we}^{j+1})^{\mathrm{T}}e_k=\\&\quad-(\bar{H}_{1,22}^{j+1})^{-1}(\bar{H}_{1,12}^{j+1})^{\mathrm{T}}(y_k-\hat{y}_k),\\L^{j+1}&=-(\bar{H}_{1,22}^{j+1})^{-1}(\bar{H}_{1,12}^{j+1})^{\mathrm{T}}.\end{aligned} \tag{30}$$

定理1 如果矩阵$C(\theta)C(\theta)^{\mathrm{T}}$可逆,则存在唯一矩阵$\bar{H}_1^{j+1}$,满足

$$H_1^{j+1}=\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix}^{\mathrm{T}}\bar{H}_1^{j+1}\begin{bmatrix}C(\theta)&0\\0&I\end{bmatrix} \tag{31}$$

和式(27),使得式(30)收敛到最优观测器策略,即$\lim\limits_{j\to\infty}w_k^{j+1}=w_k^*$.

证明 首先证明如果矩阵$\bar{H}_1^{j+1}$是迭代方程(27)的解,则由式(31)得到的矩阵H_1^{j+1}满足式(22). 已知$y_k-\hat{y}_k=C(\theta)e_k$和$e_k$的动态式(25),如果矩阵$\bar{H}_1^{j+1}$是迭代方程(27)的解,则$\bar{H}_1^{j+1}$保证下式成立:

$$\begin{aligned}&\begin{bmatrix}y_k-\hat{y}_k\\w_k\end{bmatrix}^{\mathrm{T}}\bar{H}_1^{j+1}\begin{bmatrix}y_k-\hat{y}_k\\w_k\end{bmatrix}-\\&((A(\theta)e_k-w_k^j)+w_k^j-w_k)^{\mathrm{T}}\cdot\\&\begin{bmatrix}C(\theta)\\L^jC(\theta)\end{bmatrix}^{\mathrm{T}}\bar{H}_1^{j+1}\begin{bmatrix}C(\theta)\\L^jC(\theta)\end{bmatrix}\cdot\\&((A(\theta)e_k-w_k^j)+w_k^j-w_k)=\\&(y_k-\hat{y}_k)^{\mathrm{T}}Q_1(y_k-\hat{y}_k)+(w_k^j)^{\mathrm{T}}R_1w_k^j-\\&2e_k^{\mathrm{T}}A(\theta)^{\mathrm{T}}P_o^{j+1}(w_k^j-w_k)+2w_k^jP_o^{j+1}(w_k^j-w_k)-\\&(w_k^j-w_k)^{\mathrm{T}}P_o^{j+1}(w_k^j-w_k).\end{aligned} \tag{32}$$

由引理1的式(17)和(32)可知,式(31)定义的矩阵H_1^{j+1}保证式(22)成立. 下面证明式(27)存在唯一解

$\bar{H}_1^{j+1}$. 假设式(27)存在两个不同的解$\bar{H}_1^{j+1}$和$\bar{W}_1^{j+1}$, 那么由式(31)能够得到矩阵H_1^{j+1}和W_1^{j+1},其中

$$W_1^{j+1}=\begin{bmatrix}C(\theta) & 0\\ 0 & I\end{bmatrix}^{\mathrm{T}}\bar{W}_1^{j+1}\begin{bmatrix}C(\theta) & 0\\ 0 & I\end{bmatrix}.$$

矩阵$C(\theta)C(\theta)^{\mathrm{T}}$可逆,有

$$\begin{bmatrix}(C(\theta)C(\theta)^{\mathrm{T}})^{-1} & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}C(\theta) & 0\\ 0 & I\end{bmatrix}W_1^{j+1}\cdot$$

$$\begin{bmatrix}C(\theta)^{\mathrm{T}} & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}(C(\theta)C(\theta)^{\mathrm{T}})^{-1} & 0\\ 0 & I\end{bmatrix}=\bar{W}_1^{j+1},$$

$$\begin{bmatrix}(C(\theta)C(\theta)^{\mathrm{T}})^{-1} & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}C(\theta) & 0\\ 0 & I\end{bmatrix}H_1^{j+1}\cdot$$

$$\begin{bmatrix}C(\theta)^{\mathrm{T}} & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}(C(\theta)C(\theta)^{\mathrm{T}})^{-1} & 0\\ 0 & I\end{bmatrix}=\bar{H}_1^{j+1}.$$

如果矩阵H_1^{j+1}与矩阵W_1^{j+1}相同,则矩阵$\bar{H}_1^{j+1}$与$\bar{W}_1^{j+1}$相等,所以矩阵H_1^{j+1}与矩阵W_1^{j+1}不相等. 然而式(22)有唯一解,根据反证法,原假设式(27)存在两个不同的解$\bar{H}_1^{j+1}$和$\bar{W}_1^{j+1}$不成立. 由式(30)有

$$\lim_{j\to\infty}w_k^{j+1}=\lim_{j\to\infty}\{-(H_{1,ww}^{j+1})^{-1}(H_{1,we}^{j+1})^{\mathrm{T}}e_k\}=w_k^*.\ \square$$

注3　定理1要求即使矩阵$C(\theta)$中含有未知参数,$C(\theta)C(\theta)^{\mathrm{T}}$也应能够判断是否可逆.

算法2　非策略Q-学习算法.

step 1: 数据收集. 给定保证估计误差稳定的观测器增益w_k,收集数据e_{y_k},存储到样本集合$\theta^j(k)$和ρ_k^j.

step 2: 初始化. 给定初始观测器增益,保证估计误差系统(4)稳定,令$j=0$.

step 3: 执行Q学习. 用最小二乘法估计式(28)中虚拟Q函数矩阵$\bar{H}_1^{j+1}$,然后利用式(30)更新观测器增益L^{j+1}.

step 4: 如果$\|L^{j+1}-L^j\|\leqslant\varepsilon(\varepsilon>0)$,则算法停止,否则令$j=j+1$,转至step 3.

注4　利用算法2可以通过非策略Q-学习方法较容易地学习最优观测器增益,算法2最重要的优点是它独立于系统矩阵、不要求系统矩阵$A(\theta)$、$B(\theta)$和$C(\theta)$完全已知. 另外,矩阵$\bar{H}_1^{j+1}$的引入使得仅利用可测的输入和输出数据来学习最优观测器增益成为可能.

3　最优控制器设计

根据分离定理,控制器增益矩阵K应保证$A(\theta)+B(\theta)K$镇定,并且优化问题1中的性能指标(7). 因此,给出如下优化问题.

问题3

$$\min_{u_k}\frac{1}{2}\sum_{k=0}^{\infty}y_k^{\mathrm{T}}Q_py_k+u_k^{\mathrm{T}}R_pu_k;\tag{33}$$

$$\text{s.t. }x_{k+1}=A(\theta)x_k+B(\theta)u_k,$$

$$u_k=Kx_k.\tag{34}$$

问题3改写为

$$\min_{u_k}\frac{1}{2}\sum_{k=0}^{\infty}x_k^{\mathrm{T}}\tilde{Q}_px_k+u_k^{\mathrm{T}}R_pu_k;$$

$$\text{s.t. 式(34).}\tag{35}$$

其中$\tilde{Q}_p=C(\theta)^{\mathrm{T}}Q_pC(\theta)$.

由性能指标(35),最优值函数和最优Q函数为

$$V_x^*(x_k)=\min_{u_k}\sum_{i=k}^{\infty}x_i^{\mathrm{T}}\tilde{Q}_px_i+u_i^{\mathrm{T}}R_pu_i,\tag{36}$$

$$Q_x^*(x_k,u_k)=x_k^{\mathrm{T}}\tilde{Q}_px_k+u_k^{\mathrm{T}}R_pu_k+V_x^*(x_{k+1}).\tag{37}$$

得到最优值函数和最优Q函数的关系为

$$V_x^*(x_k)=\min_{u_k}Q_x^*(x_k,u_k)=Q_x^*(x_k,u_k^*).\tag{38}$$

引理2 [17]　对于可镇定的控制策略$u_k=Kx_k$,最优值函数和最优Q函数可以表示为如下二次型:

$$V_x^*(x_k)=\frac{1}{2}x^{\mathrm{T}}(k)P_2x(k),\tag{39}$$

$$Q_x^*(x_k,u_k)=\frac{1}{2}\begin{bmatrix}x_k\\ u_k\end{bmatrix}^{\mathrm{T}}H_2\begin{bmatrix}x_k\\ u_k\end{bmatrix}.\tag{40}$$

其中

$$H_2=\begin{bmatrix}H_{2,xx} & H_{2,xu}\\ * & H_{2,uu}\end{bmatrix}=$$

$$\begin{bmatrix}A(\theta)^{\mathrm{T}}P_2A(\theta)+\tilde{Q}_p & A(\theta)^{\mathrm{T}}P_2B(\theta)\\ B(\theta)^{\mathrm{T}}P_2A(\theta) & B(\theta)^{\mathrm{T}}P_2B(\theta)+R_p\end{bmatrix};\tag{41}$$

$$\begin{cases}P_2=\begin{bmatrix}I\\ -K\end{bmatrix}^{\mathrm{T}}H_2\begin{bmatrix}I\\ -K\end{bmatrix},\\ P_2=P_2^{\mathrm{T}}>0.\end{cases}\tag{42}$$

在系统矩阵$A(\theta)$、$B(\theta)$和$C(\theta)$含有未知参数的情况下,给出求解问题3的不要求系统矩阵完全已知的策略Q-学习算法和非策略Q-学习算法.

3.1　策略Q-学习算法设计

采用动态规划方法,由式(37)、(38)和(40)得到基于最优Q函数的贝尔曼方程

$$\begin{bmatrix}x(k)\\ u(k)\end{bmatrix}^{\mathrm{T}}H_2\begin{bmatrix}x(k)\\ u(k)\end{bmatrix}=$$

$$x^{\mathrm{T}}(k)\tilde{Q}_px(k)+u^{\mathrm{T}}(k)R_pu(k)+$$

$$\begin{bmatrix}x(k+1)\\ u(k+1)\end{bmatrix}^{\mathrm{T}}H_2\begin{bmatrix}x(k+1)\\ u(k+1)\end{bmatrix}.\tag{43}$$

根据实现最优性能的必要条件,执行

$$\frac{\partial Q_x^k(x(k),u(k))}{\partial u(k)}=0,$$

得到

$$u^*(k)=-(H_{2,uu})^{-1}H_{2,ux}x(k),\tag{44}$$

$$K^*=-H_{2,uu}^{-1}(H_{2,ux})^{\mathrm{T}}.\tag{45}$$

下面的算法3给出了求解式(43)中最优Q函数矩阵H_2的方法,注意到该方法不需要系统矩阵已知.

算法3 依赖策略的迭代算法.

step 1: 初始化:给定可镇定控制器增益矩阵K^0,令$j=0,j$为迭代指标.

step 2: 策略评估. $u_k^j=K^jx_k$,由式(43)可以得到

$$\begin{aligned}&\begin{bmatrix}x(k)\\K^jx(k)\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}x(k)\\K^jx(k)\end{bmatrix}=\\&y(k)^{\mathrm{T}}Q_py(k)+(K^jx(k))^{\mathrm{T}}R_p(K^jx(k))+\\&\begin{bmatrix}x(k+1)\\K^jx(k+1)\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}x(k+1)\\K^jx(k+1)\end{bmatrix},\end{aligned}\tag{46}$$

其中$x_{k+1}=A(\theta)x_k+B(\theta)u_k^j$. 由式(46)可以求解最优Q函数迭代矩阵$H_2^{j+1}$.

step 3: 策略更新

$$K^{j+1}=-(H_{2,uu}^{j+1})^{-1}(H_{2,ux}^{j+1})^{\mathrm{T}}.\tag{47}$$

step 4: 如果$\|H_2^j-H_2^{j+1}\|\leqslant\varepsilon(\varepsilon>0)$,则停止迭代,否则转至step 2,继续执行算法.

注5 执行算法3, $\lim\limits_{j\to\infty}H_2^{j+1}=H_2$且$\lim\limits_{j\to\infty}K^{j+1}=K^*$[18-19],算法3要求系统添加探测噪声以满足持续激励条件. 采用策略Q-学习算法3,如果系统加入探测噪声,则式(46)的解H_2可能会有偏差. 而对于非策略Q-学习算法,即使系统加入探测噪声,也会学习得到无偏解[12-16,22-23]. 因此,在下面的第3.2节给出非策略 Q-学习算法,以便找到最优控制器增益.

3.2 非策略Q-学习方法

在系统(34)中引入辅助控制策略u_k^j,有

$$\begin{aligned}x_{k+1}&=A(\theta)x_k+B(\theta)u_k^j-B(\theta)u_k^j+B(\theta)u_k=\\&A(\theta)x_k+B(\theta)K^jx_k+B(\theta)(u_k-u_k^j)=\\&(A(\theta)+B(\theta)K^j)x_k+B(\theta)(u_k-u_k^j).\end{aligned}\tag{48}$$

其中:u_k为用于产生系统数据的行为策略;u_k^j为目标策略,算法目标是当$j\to\infty$时,u_k^j收敛到u_k^*. 沿着系统(48)的轨迹,有

$$\begin{aligned}&\begin{bmatrix}x_k\\u_k\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}x_k\\u_k\end{bmatrix}-(A(\theta)x_k+B(\theta)u_k^j)^{\mathrm{T}}\cdot\\&\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}(A(\theta)x_k+B(\theta)u_k^j)=\\&x_k^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_k-\\&(x_{k+1}-B(\theta)(u_k-u_k^j))^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\cdot\\&\begin{bmatrix}I\\K^j\end{bmatrix}(x_{k+1}-B(\theta)(u_k-u_k^j))=\\&x_k^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_k-x_{k+1}^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}\cdot\\&H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_{k+1}+2x_{k+1}^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\cdot\\&\begin{bmatrix}I\\K^j\end{bmatrix}B(\theta)(u_k-u_k^j)-(u_k-u_k^j)^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}\cdot\\&H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}B(\theta)(u_k-u_k^j).\end{aligned}\tag{49}$$

由式(41)、(42)、(46),式(49)改写为

$$\begin{aligned}&x_k^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_k-\\&x_{k+1}^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_{k+1}+\\&2x_{k+1}^{\mathrm{T}}P^{j+1}B(\theta)(u_k-u_k^j)-\\&(u_k-u_k^j)^{\mathrm{T}}P^{j+1}B(\theta)(u_k-u_k^j)=\\&y_k^{\mathrm{T}}Q_py_k+(u_k^j)^{\mathrm{T}}R_pu_k^j.\end{aligned}\tag{50}$$

进而有

$$\begin{aligned}&x_k^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_k-\\&x_{k+1}^{\mathrm{T}}\begin{bmatrix}I\\K^j\end{bmatrix}^{\mathrm{T}}H_2^{j+1}\begin{bmatrix}I\\K^j\end{bmatrix}x_{k+1}+\\&2x_k^{\mathrm{T}}H_{2,xu}^{j+1}(u_k-u_k^j)+u_k^{\mathrm{T}}(H_{2,uu}^{j+1}-R_p)(u_k-u_k^j)+\\&(u_k^j)^{\mathrm{T}}(H_{2,uu}^{j+1}-R_p)(u_k-u_k^j)=\\&y_k^{\mathrm{T}}Q_py_k+(u_k^j)^{\mathrm{T}}R_pu_k^j.\end{aligned}\tag{51}$$

进一步处理,有

$$\begin{aligned}&x_k^{\mathrm{T}}H_{2,xx}^{j+1}x_k-x_{k+1}^{\mathrm{T}}H_{2,xx}^{j+1}x_{k+1}-\\&2x_{k+1}^{\mathrm{T}}H_{2,xu}^{j+1}u_{k+1}^j-(u_{k+1}^j)^{\mathrm{T}}H_{2,uu}^{j+1}u_{k+1}^j+\\&2x_k^{\mathrm{T}}H_{2,xu}^{j+1}u_k^j+u_k^{\mathrm{T}}(H_{2,uu}^{j+1}-R_p)-\\&(u_k^j)^{\mathrm{T}}R_p(u_k-u_k^j)=\\&y_k^{\mathrm{T}}Q_py_k+(u_k^j)^{\mathrm{T}}R_pu_k^j.\end{aligned}\tag{52}$$

式(52)可以改写为

$$\varphi^j(k)h_c^{j+1}=\beta_k^j.\tag{53}$$

其中

$$u_k^j=K^jx_k,\ u_{k+1}^j=K^jx_{k+1},$$

$$K^{j+1}=-(H_{2,uu}^{j+1})^{-1}H_{2,ux}^{j+1},$$

$$\beta_k^j=y_k^{\mathrm{T}}Q_py_k+u_k^{\mathrm{T}}R_pu_k,$$

$$h_c^{j+1}=$$

$$[(\mathrm{vec}(H_{2,xx}^{j+1}))^{\mathrm{T}}\quad(\mathrm{vec}(H_{2,xu}^{j+1}))^{\mathrm{T}}\quad(\mathrm{vec}(H_{2,uu}^{j+1}))^{\mathrm{T}}]^{\mathrm{T}},$$

$$\varphi^j(k)=[\varphi_1^j(k)\quad\varphi_2^j(k)\quad\varphi_3^j(k)],$$

$$\varphi_1^j(k)=x_k^{\mathrm{T}}\otimes x_k^{\mathrm{T}}-x_{k+1}^{\mathrm{T}}\otimes x_{k+1}^{\mathrm{T}},$$

$$\varphi_2^j(k)=2x_k^{\mathrm{T}}\otimes u_k^{\mathrm{T}}-2x_{k+1}^{\mathrm{T}}\otimes(u_k^{j+1})^{\mathrm{T}},$$

$$\varphi_3^j(k)=u_k^{\mathrm{T}}\otimes u_k^{\mathrm{T}}-(u_k^{j+1})^{\mathrm{T}}\otimes(u_k^{j+1})^{\mathrm{T}}.$$

注 6 [16,22-23]　如果 H_2^{j+1} 是式(46)的解, 当且仅当 H_2^{j+1} 是式(53)的解.

注 7　利用最小二乘法求解式(53)中最优Q函数迭代矩阵 H_2^{j+1} 时, 需要利用行为控制策略 u_k 产生的状态数据 x_k. 然而, 状态数据 x_k 不可测量, 由于本文采用非策略Q-学习算法, 可以取可镇定的观测器行为策略 w_k 和控制器行为控制策略 u_k 作用系统(1)和(2). 比较 y_k 和 $\hat{y}_k$, 当 $\hat{y}_k$ 接近 y_k 时, 用 $\hat{x}_k$ 代替 x_k, 求解式(53), 得到最优Q函数迭代矩阵 H_2^{j+1}, 进而得到控制器增益 K^{j+1}.

结合算法2和算法3给出算法4, 得到近似最优控制器增益矩阵.

算法 4　基于非策略Q-学习方式的最优状态估计反馈控制器设计算法.

step 1: 选取行为控制策略 u_k 和行为观测器策略 w_k 作用系统, 收集数据 $\hat{x}_k$、u_k、$\hat{y}_k$ 和 y_k.

step 2: 给定初始观测器增益 L^0 和可镇定控制器增益 K^0, 令 $j=0$.

step 3: 执行算法2中step 3, 得到虚拟Q函数迭代矩阵 $\bar{H}_1^{j+1}$ 和观测器策略增益 L^{j+1}.

step 4: 用 $\hat{x}_k$ 代替 x_k(前提是 $\hat{y}_k$ 接近 y_k), 利用式(53)计算Q函数迭代矩阵 H_2^{j+1}, 计算

$$K^{j+1}=-(H_{2,uu}^{j+1})^{-1}(H_{2,ux}^{j+1})^{\mathrm{T}}.$$

如果 $\|\bar{H}_2^{j+1}-\bar{H}_2^j\|\leqslant\varepsilon$ 且 $\|L^{j+1}-L^j\|\leqslant\varepsilon(\varepsilon>0)$, 则停止迭代, 否则令 $j=j+1$, 转至step 3.

执行算法4, 如果 ε 足够小, 迭代指数 j 足够大, 则 L^{j+1} 无限接近 L^*, K^{j+1} 无限接近 K^*. 对于具有未知参数的离散线性系统, 算法4完全利用可测的输入、输出, 状态观测器状态学习最优观测器增益、最优控制器增益. 并且, 如果系统矩阵 $A(\theta)$、$B(\theta)$ 和 $C(\theta)$ 完全未知, 但是能够实际应用判断系统的能控性和能观性, 并且能保证矩阵 $C(\theta)^{\mathrm{T}}C(\theta)$ 可逆, 则算法4仍然适用.

注 8　文献[17]也考虑到系统状态不可测的情况, 采用强化学习方法学习最优控制策略, 需要利用系统当前和过去的输出和输入数据估计当前系统状态, 所计算的未知迭代矩阵中含有独立的未知变元的个数为 $(Nm+Np)\times(Nm+Np+1)/2(N\geqslant 1)$. 本文式(53)中未知迭代矩阵 H_2^{j+1} 中独立未知变元个数为 $m+p$.

4　仿真实验

考虑如下离散RLC电路系统:

$$\begin{bmatrix}x_{1(k+1)}\\x_{2(k+1)}\end{bmatrix}=\begin{bmatrix}1&-\dfrac{\Delta t}{N}\\\dfrac{\Delta t}{L}&1-\dfrac{r\cdot\Delta t}{L}\end{bmatrix}\begin{bmatrix}x_{1k}\\x_{2k}\end{bmatrix}+\begin{bmatrix}\dfrac{\Delta t}{N}\\0\end{bmatrix}u_k.\tag{54}$$

其中: x_{1k}、x_{2k} 分别为电容电压和感应电流; 参数 $\Delta t=0.2\,\mathrm{s}, N=0.1\,\mathrm{F}, r=3\,\Omega, L=1\,\mathrm{H}$, 假设这些参数未知, 系统能控、能观. 选取 $C=\begin{bmatrix}1&0\\0&1\end{bmatrix}$, $Q_1=Q_2=\begin{bmatrix}10&0\\0&10\end{bmatrix}$, $R_1=R_2=\begin{bmatrix}1&0\\0&1\end{bmatrix}$.

1) 基于模型求最优观测器策略和控制器策略.

如果参数 $\Delta t=0.2\,\mathrm{s}$、$N=0.1\,\mathrm{F}$、$r=3\,\Omega$、$L=1\,\mathrm{H}$ 已知, 则对于优化问题2, 利用Matlab软件“dare”命令可以获得最优Q-函数矩阵, 并由式(24)计算得到

$$H_1^*=\begin{bmatrix}75.813\,1&-14.183\,8&-12.430\,2&-26.691\,4\\-14.183\,8&72.584\,2&29.736\,8&-7.776\,5\\-12.430\,2&29.736\,8&15.646\,8&-1.108\,3\\-26.691\,4&-7.776\,5&-1.108\,3&14.899\,9\end{bmatrix}.\tag{55}$$

由式(19)计算最优观测器增益为

$$L^*=\begin{bmatrix}0.926\,2&-1.873\,4\\1.860\,3&0.382\,6\end{bmatrix}.\tag{56}$$

对于优化问题3, 利用Matlab软件中“dare”命令可以获得最优Q函数矩阵

$$H_2^*=\begin{bmatrix}151.605\,5&-133.199\,1&146.836\,5\\-133.199\,1&223.637\,7&-128.434\,1\\146.836\,5&-218.434\,1&226.274\,0\end{bmatrix}.\tag{57}$$

最优控制器增益为

$$K^*=[-0.648\,9\quad 0.965\,4].\tag{58}$$

2) 模型参数 Δt、N、r 和 L 未知情况下, 求最优观测器和控制器策略.

给定观测器增益矩阵初始值为

$$L^0=\begin{bmatrix}0.305\,6&-1.527\,8\\1&0.200\,0\end{bmatrix},\tag{59}$$

控制器增益矩阵初始值为

$$K^0 = [-0.5119 \;\; 0.9783]. \tag{60}$$

执行算法4,如图1所示,经过10次策略迭代,观测器最优Q函数迭代矩阵和迭代增益矩阵分别收敛到最优$\bar{H}_1^*$和L^*. 图2为控制器最优Q函数迭代矩阵收敛到最优H_2^*和控制器增益收敛到K^*的曲线.

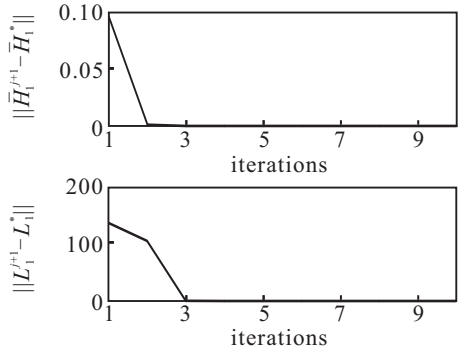

图1 $\bar{H}_1^{j+1}$和L^{j+1}收敛过程

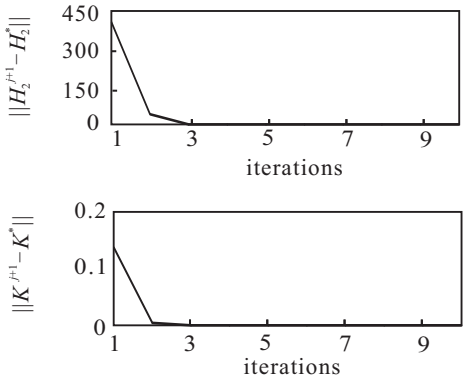

图2 H_2^{j+1}和K^{j+1}收敛过程

图3为系统响应曲线,图4为最优控制律和观测器策略曲线. 可以看出,观测器较好地估计了系统的状态. 表1给出了最优控制器和一般可镇定控制器下系统性能的比较. 由表1可以看出,在系统矩阵存在

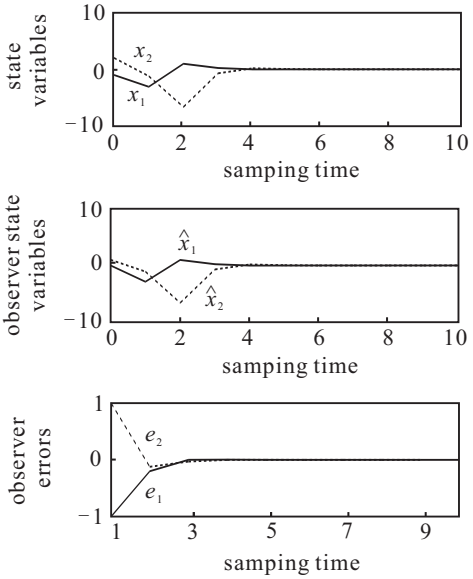

图3 系统响应曲线

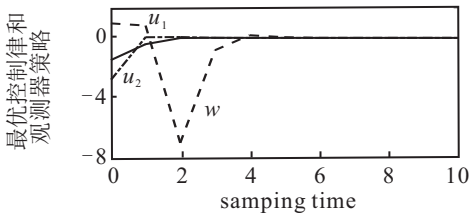

图4 最优控制律和观测器策略

参数未知的情况下,采用算法4能够利用可测的系统输入、输出数据和观测器状态学习最优控制器增益和最优观测器增益,实现基于观测器状态反馈的优化控制.

表1 性能比较

	控制器和观测器增益矩阵	最优性能
最优控制算法	$K^* = [-0.6489 \;\; 0.9654]$ $L^* = \begin{bmatrix} 0.9262 & -1.8734 \\ 1.8603 & 0.3826 \end{bmatrix}$	73.2065
可镇定控制器	$K = [-0.5119 \;\; 0.9783]$ $L^* = \begin{bmatrix} 0.9262 & -1.8734 \\ 1.8603 & 0.3826 \end{bmatrix}$	76.1342

5 结 论

本文针对系统状态不能直接测量或测量成本较高的控制系统,在系统矩阵存在未知参数的情况下,提出了一种基于非策略Q-学习方法的观测器状态反馈优化控制策略. 首先,对于具有未知参数的线性离散系统,定义基于观测器状态反馈的优化问题;然后,利用分离原理、动态规划和Q-学习算法,提出非策略Q-学习算法. 在系统矩阵$A(\theta)$、$B(\theta)$和$C(\theta)$存在未知参数的情况下,利用可测数据学习最优观测器增益和最优控制器增益. 仿真结果也验证了所提出方法的有效性.

参考文献(References)

[1] 康军, 戴冠中. 具有状态观测器的网络化控制系统的设计[J]. 控制与决策, 2010, 25(6): 943-947.
(Kang J, Dai G Z. Design of networked control system with state observer[J]. Control and Decision, 2010, 25(6): 943-947.)

[2] 邱占芝, 张庆灵. 一类基于观测器的网络控制系统鲁棒控制器设计[J]. 控制与决策, 2007, 22(10): 1165-1169.
(Qiu Z Z, Zhang Q L. Robust controller design for a class of network control system based on observer[J]. Control and Decision, 2007, 22(10): 1165-1169.)

[3] 年晓红, 曹莉. 基于微分对策的最优状态观测器和最优状态反馈控制器的设计[J]. 自动化学报, 2006,

32(5): 807-812.
(Nian X H, Cao L. Design of optimal state observer and optimal state feedback controller based on differential games[J]. Acta Automatica Sinica, 2006, 32(5): 807-812.)

[4] 陈志翔, 高钦和. 修改型扩张状态观测器:分析与实现 [J]. 控制理论与应用, 2018, 35(8): 1199-1206.
(Chen Z X, Gao Q H. Modified expansion state observer: Analysis and implementation[J]. Control Theory & Applications, 2018, 35(8): 1199-1206.)

[5] 李向阳. 迭代扩张状态观测器及其在迭代学习控制中的应用 [J]. 控制与决策, 2015, 30(3): 473-478.
(Li X Y. Iterative extended state observer and its application in iterative learning control[J]. Control and Decision, 2015, 30(3): 473-478.)

[6] 朱芳来, 侯永建, 赵旭东, 等. 非线性切换系统基于观测器的容错控制器设计 [J]. 控制与决策, 2017, 32(10): 1855-1863.
(Zhu F L, Hou Y J, Zhao X D, et al. Design of fault tolerant controller based on observer for nonlinear switching system[J]. Control and Decision, 2017, 32(10): 1855-1863.)

[7] Zhang H G, Cui X H, Luo Y H, et al. Finite-horizon H_∞ tracking control for unknown nonlinear systems with saturating actuators[J]. IEEE Transactions on Neural Networks & Learning Systems, 2018, 29(4): 1200-1213.

[8] Wei Q L, Liu D R, Shi G. A novel dual iterative Q-learning method for optimal battery management in smart residential environments[J]. IEEE Transactions on Industrial Electronics, 2015, 62(4): 2509-2518.

[9] Wang D, He H B, Liu D R. Adaptive critic nonlinear robust control: A survey[J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3429-3451.

[10] Liu Y, Yu R. Model-free optimal tracking control for discrete-time system with delays using reinforcement Q-learning[J]. Electronics Letters, 2018, 54(12): 750-752.

[11] Wei Q L, Liu D R. Data-driven neuro-optimal temperature control of water–gas shift reaction using stable iterative adaptive dynamic programming[J]. IEEE Transactions on Industrial Electronics, 2014, 61(11): 6399-6408.

[12] Li J N, Chai T Y, Lewis F L, et al. Off-policy Q-learning: Set-point design for optimizing dual-rate rougher flotation operational processes[J]. IEEE Transactions on Industrial Electronics, 2018, 65(5): 4092-4102.

[13] Precup D, Autton R S, Dasgupta S. Off-policy temporal-difference learning with function approximation[C]. Proceedings of the 18th International Conference on Machine Learning. Williamstown: IEEE, 2001: 417-424.

[14] Sutton R S, Szepesvari C, Maei H R. A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation[C]. Proceedings of the 22nd Annual Conference on Neural Information Processing Systems. Vancouver: IEEE, 2008: 1609-1616.

[15] Maei H R, Szepesvari C, Bhstngagar S, et al. Toward off-policy learning control with function approximation[C]. Proceedings of the 27th International Conference on Machine Learning. Haifa: IEEE, 2010: 719-727.

[16] Kiumarsi B, Lewis F L, Jiang Z P. H_∞ control of linear discrete-time systems: Off-policy reinforcement learning[J]. Automatica, 2017, 37(1): 144-152.

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

[18] 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_∞ control[J]. Automatica, 2007, 43(3): 473-481.

[19] Kim J H, Lewis F L. Model-free H_∞ control design for unknown linear discrete-time systems via Q-learning with LMI[J]. Automatica, 2010, 46(8): 1320-1326.

[20] Kilicaslan S, Banks S P. A separation theorem for nonlinear systems[J]. Automatica, 2009, 45(4): 928-935.

[21] Germani A, Manes C, Pepe P. Separation theorems for a class of retarded nonlinear systems[J]. IFAC Proceedings Volumes, 2010, 43(2): 21-26.

[22] Li J N, Chai T Y, Lewis F L, et al. Off-policy interleaved Q-learning: Optimal control for affine nonlinear discrete-time systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(5): 1308-1320.

[23] Luo B, Liu D R, Huang T, et al. Model-free optimal tracking control via critic-only Q-learning[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(10): 2134-2144.

作者简介

李金娜 (1977–), 女, 教授, 博士, 从事数据驱动控制、运行优化控制、强化学习、网络控制等研究, E-mail: lijinna_721@126.com;

马士凯 (1994–), 男, 硕士生, 从事强化学习、状态观测器的研究, E-mail: 2358022230@qq.com.

(责任编辑：郑晓蕾)