控制与决策
Control and Decision

丢包扰动环境下基于强化学习的最优输出调节

崔云芳, 范家璐

引用本文:
崔云芳,范家璐. 丢包扰动环境下基于强化学习的最优输出调节[J]. 控制与决策, 2023, 38(2): 403−412.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1147

丢包扰动环境下基于强化学习的最优输出调节

崔云芳，范家璐†

(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: 针对存在线性外部干扰和状态反馈过程中发生丢包的网络控制系统的跟踪控制问题,采用输出调节的思想,提出基于离轨策略强化学习的数据驱动最优输出调节控制方法,实现仅利用在线数据即可求解控制策略. 首先,对系统状态在网络传输过程存在丢包的情况,利用史密斯预估器重构系统的状态;然后基于输出调节控制框架,提出一种基于离轨策略强化学习的数据驱动最优控制算法,在系统状态发生丢包时仅利用在线数据计算反馈增益,在求解反馈增益过程中找到与求解输出调节问题的联系;接着基于求解反馈增益过程中得到的与输出调节问题中求解调节器方程相关的参数,计算前馈增益的无模型解;最后,通过仿真结果验证所提出方法的有效性.

关键词: 输出调节；强化学习；丢包；史密斯预估器；离轨策略；跟踪控制

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1147 **开放科学(资源服务)标识码(OSID):**

引用格式: 崔云芳,范家璐. 丢包扰动环境下基于强化学习的最优输出调节[J]. 控制与决策, 2023, 38(2): 403-412.

Optimal output regulation based on reinforcement learning for systems with dropouts and disturbances

CUI Yun-fang, FAN Jia-lu†

(State Key Laboratory of Synthetical Automation for Process Industries，Northeastern University，Shenyang 110004，China)

Abstract: In this paper, a data-driven optimal output regulation control method using off-policy reinforcement learning is proposed for tracking control of discrete-time networked control systems with both linear disturbance and state dropouts in the feedback process. This method uses only measured online data to calculate control policies. First, in the environment where state dropouts exist, a restructured state of the system is established by using the Smith predictor. Then, under the output regulation framework, a data-driven optimal tracking control method using off-policy reinforcement learning is developed to calculate the feedback gain using only the measured data when dropout occurs. The connection with solving the output regulation problem is found in the process of solving the feedback gain. Based on the parameters related to solving the regulator equation in the process of solving the feedback gain, a model-free solution of forward gain is calculated. Finally, simulation results demonstrate the effectiveness of the proposed approach.

Keywords: output regulation；reinforcement learning；dropout；Smith predictor；off-policy；tracking control

0 引 言

网络控制系统(networked control system, NCS)是计算机技术、网络通信技术与控制理论相融合的产物,近年来引起控制领域的广泛关注. 通信网络的使用为控制系统提供诸多优势,同时也产生一些具有挑战性的问题,例如,数据传输过程发生的时间延迟、丢包等问题[1-4]会降低系统的性能甚至会导致系统不稳定. 对于具有丢包的网络控制系统,现有的研究方法主要是基于系统模型完全已知的情况[2-4].

模型驱动控制器的性能很大程度依赖于所建立模型的精确度. 实际中建立精确的模型存在困难,因此,研究数据驱动的控制算法具有重要意义. 强化学习是一种有效的数据驱动控制算法,已成为近年来的研究热点,在求解离散系统[5-8]和连续系统[9-11]的最优控制问题中得到了广泛应用. 强化学习算法通常分为两种: 同轨策略(on-policy)方法[7]和离轨策略

收稿日期: 2021-07-01；**录用日期:** 2021-12-09.
基金项目: 辽宁省“兴辽英才计划”项目(XLYC2007135).
责任编委: 俞立.
†通讯作者. E-mail: jlfan@mail.neu.edu.cn.

(off-policy)方法[8-9]. 其中,文献[8]证明了离轨策略强化学习在满足持续激励条件时可保证所求跟踪控制策略无偏性.

实现跟踪控制的另一方法是输出调节理论,文献[12-14]和文献[15]分别研究了线性系统和非线性系统的输出调节问题. 以上方法都需要准确的系统模型信息. 考虑到模型信息未知的情况,一些学者将输出调节理论与自适应动态规划方法相结合,突破了传统的输出调节问题需要已知系统模型的束缚[16-19],但是未考虑网络传输异常对系统的影响[16-18]. 文献[19]针对存在丢包问题的离散系统设计输出调节最优控制器实现了跟踪控制,但没有考虑系统的外部干扰.

经过对现有成果的充分了解可以得出,如何在丢包的环境下对存在外部干扰以及模型未知的系统,实现无偏的数据驱动输出调节最优跟踪控制还有待解决. 为此,本文将强化学习与最优输出调节控制方法相结合,提出基于离轨策略强化学习算法的输出调节控制方法. 主要贡献在于: 1) 针对存在丢包和外界干扰的离散系统,提出基于强化学习的离轨策略输出调节方法实现跟踪控制. 一方面,本文研究对象更符合实际生产中的计算机控制环境;另一方面,同时考虑了实际控制中存在的网络丢包以及外部干扰情况. 2) 所提出方法可以实现系统状态反馈过程出现随机丢包时,仅利用测量数据即可求解控制律,能够克服控制器对模型的依赖. 3) 给出了丢包环境下线性输出调节问题的可解性和稳定性证明.

1 控制问题描述

1.1 被控对象模型

考虑如下线性离散时间系统:

$$\begin{cases} x(k+1)=Ax(k)+Bu(k)+Dv(k),\\ y(k)=Cx(k). \end{cases} \tag{1}$$

其中:$x\in \boldsymbol{R}^n$、$u\in \boldsymbol{R}^m$和$v\in \boldsymbol{R}^q$分别为系统的状态向量、输入向量和外系统的状态向量;$y\in \boldsymbol{R}^r$为系统的输出向量;状态矩阵$A\in \boldsymbol{R}^{n\times n}$,$D\in \boldsymbol{R}^{n\times q}$;输入矩阵$B\in \boldsymbol{R}^{n\times m}$;输出矩阵$C\in \boldsymbol{R}^{r\times n}$.

系统(1)的参考轨迹由外部系统提供,形式如下:

$$v(k+1)=Fv(k), \tag{2}$$

$$w(k+1)=Ev(k). \tag{3}$$

式(2)和(3)是产生系统扰动$Dv(k)$和参考信号$w(k)$的外系统. 其中:$w\in \boldsymbol{R}^r$,$v(k)$、$w(k)$均可测,状态矩阵$F\in \boldsymbol{R}^{q\times q}$,输出矩阵$E\in \boldsymbol{R}^{r\times q}$.

假设1 (A,B)是可控的,(A,C)是可观的.

1.2 无线网络环境下丢包模型

在无线网络环境下,系统的状态$x(k)$通过网络传输反馈给控制器. 假设系统在k时刻的状态信息以单包的形式传输,并且传输过程中会发生丢包,经过网络传输后,控制器可用的状态为

$$x_f(k)=x(k-\delta_{f_n}(k)), \tag{4}$$

其中$\delta_{fn}(k)$为k时刻及之前发生连续丢包的次数.

假设2 最大连续丢包次数$\delta_{f\max}$为有界值,否则控制系统相当于开环系统.

结合式(1)和(4)重建丢包系统当前的状态,有

$$\begin{aligned} x(k)=&\\ &A^{\delta_{fn}}(k)x(k-\delta_{f_n}(k))+\sum_{i=1}^{\delta_{fn}(k)}A^{i-1}Bu(k-i)+\\ &\sum_{i=1}^{\delta_{fn}(k)}A^{i-1}Dv(k-i), \end{aligned} \tag{5}$$

其中$\delta_{fn}(k)$是已知的,且由假设2可知$0\leqslant\delta_{fn}(k)\leqslant\delta_{f\max}$. 式(5)可以等价于

$$x(k)=G\eta(k). \tag{6}$$

其中

$$\begin{aligned} G=[&I \quad A \quad \dots \quad A^{\delta_{f\max}} \quad B \quad AB \quad \dots \to\\ &\leftarrow A^{\delta_{f\max}-1}B \quad D \quad AD \quad \dots \quad A^{\delta_{f\max}-1} \quad D], \end{aligned} \tag{7}$$

$\eta\in \boldsymbol{R}^z$,$z=n\times(\delta_{f\max}+1)+m\times\delta_{f\max}+q\times\delta_{f\max}$. 当$\delta_{fn}(k)=0,1,\dots,\delta_{f\max}$时,$\eta(k)$可分别表示为

$$\eta(k)=\begin{cases} \left[\underbrace{x_f^{\mathrm{T}}(k)\ 0\ \dots\ 0}_{\delta_{f\max}+1}\ \underbrace{0\ \dots\ 0}_{\delta_{f\max}}\ \underbrace{0\ \dots\ 0}_{\delta_{f\max}}\right]^{\mathrm{T}},\\ \quad \delta_{fn}(k)=0;\\ \left[0\ \underbrace{x_f^{\mathrm{T}}(k)\ \dots\ 0}_{\delta_{f\max}+1}\ \underbrace{u^{\mathrm{T}}(k-1)\ \dots\ 0}_{\delta_{f\max}}\ \underbrace{v^{\mathrm{T}}(k-1)\ \dots\ 0}_{\delta_{f\max}}\right]^{\mathrm{T}},\\ \quad \delta_{fn}(k)=1;\\ \left[\underbrace{0\,0\ \dots\ x_f^{\mathrm{T}}(k)}_{\delta_{f\max}+1}\ \underbrace{u^{\mathrm{T}}(k-1)\ \dots\ u^{\mathrm{T}}(k-\delta_{f\max})}_{\delta_{f\max}}\to\right.\\ \left.\leftarrow\ \underbrace{v^{\mathrm{T}}(k-1)\ \dots\ v^{\mathrm{T}}(k-\delta_{f\max})}_{\delta_{f\max}}\right]^{\mathrm{T}},\\ \quad \delta_{fn}(k)=\delta_{f\max}. \end{cases} \tag{8}$$

在k时刻,向量$\eta(k)$是已知的,故当G已知时,可以由式(6)重建系统发生丢包时的状态$x(k)$.

将式(6)代入(1)且由(7)可知G行满秩,存在右逆$G^*=G^{\mathrm{T}}(GG^{\mathrm{T}})^{-1}$,可得

$$\begin{cases}\eta(k+1)=G^*AG\eta(k)+G^*Bu(k)+G^*Dv(k),\\ y(k)=CG\eta(k).\end{cases}\tag{9}$$

1.3 丢包环境控制目标

本文的控制目标是在丢包环境下,对系统(1)设计数据驱动最优输出调节器,使系统的输出$y(k)$跟踪参考轨迹(3),即

$$\lim_{k\to\infty}(y(k)-w(k))=0.\tag{10}$$

根据文献[20],典型的无丢包线性输出调节问题的控制输入为

$$u(k)=-L_1x(k)+L_2v(k).\tag{11}$$

其中: $L_1\in\boldsymbol{R}^{m\times n}$为反馈增益矩阵, $L_2\in\boldsymbol{R}^{m\times q}$为前馈增益矩阵. 结合式(6),无丢包线性输出调节问题的控制输入(11)转化到丢包环境可表示为

$$\begin{aligned}u(k)=&-L_1G\eta(k)+L_2v(k)=\\&-\bar{L}_1\eta(k)+L_2v(k).\end{aligned}\tag{12}$$

其中: $\bar{L}_1\in\boldsymbol{R}^{m\times z}$, $\bar{L}_1=L_1G$. 所以, $\bar{L}_1$和L_2是本文要求解的目标量.

2 控制算法

首先分析丢包环境下线性输出调节问题的可解性,提出求解最优控制器需要求解的两个优化问题,给出控制器设计思路;然后提出模型未知、数据驱动的控制器求解方法.

2.1 丢包环境下线性输出调节问题的可解性分析

假设3 令$\sigma(F)$为式(2)中F的谱半径,λ为矩阵F的特征值,对于所有特征值$\forall\lambda\in\sigma(F)$,有

$$\operatorname{rank}\begin{bmatrix}llA-\lambda I & B\\ C & 0\end{bmatrix}=n+q.\tag{13}$$

定理1 对于满足假设1和假设3的控制系统(1),如果式(12)中的L_1使得$A-BL_1$是Schur矩阵,且存在矩阵X_c、L_1、L_2满足

$$\begin{cases}X_cF=G^*(AGX_c-BL_1GX_c+BL_2+D),\\ CGX_c-E=0,\end{cases}\tag{14}$$

其中$X_c\in\boldsymbol{R}^{z\times q}$. 则输出调节问题可解,且闭环系统全局稳定,系统的输出可以跟踪参考轨迹$w(k)$.

证明 式(14)两边同时左乘G,令

$$X=GX_c\ ,\ U=L_2-L_1X,\tag{15}$$

则式(14)可转化为

$$\begin{cases}XF=AX+BU+D,\\ CX-E=0.\end{cases}\tag{16}$$

其中: $X\in\boldsymbol{R}^{n\times q}$, $U\in\boldsymbol{R}^{m\times q}$. 根据文献[20],假设3可保证对于任意矩阵$D$和$F$,调节器方程(16)有唯一解. 因此求解满足式(14)的(X_c,L_1,L_2)问题可转化为求解(16)中的(X,U)问题. 由式(15)可知

$$L_2=U+L_1X.\tag{17}$$

定义

$$\bar{x}(k)=x(k)-Xv(k),\tag{18}$$

$$\bar{u}(k)=u(k)-Uv(k).\tag{19}$$

结合式(1)、(2)、(11)、(16)～(19),可得

$$\bar{x}(k+1)=(A-BL_1)\bar{x}(k).\tag{20}$$

由于$A-BL_1$为Schur矩阵,由式(18)～(20)可得

$$\begin{aligned}&\lim_{x\to\infty}(x(k)-Xv(k))=\lim_{x\to\infty}(G\eta(k)-Xv(k))=0,\\&\lim_{k\to\infty}(u(k)-Uv(k))=0.\end{aligned}\tag{21}$$

将式(18)左乘C,结合式(16)中$CX-E=0$,得到

$$C\bar{x}(k)=Cx(k)-Ev(k)=e(k).\tag{22}$$

进而得到

$$\lim_{k\to\infty}e(k)=0,\tag{23}$$

即式(10)成立. □

根据定理1,选取适宜的反馈增益L_1,使$A-BL_1$为Schur矩阵,利用调节器方程(16) 解出(X,U),进而利用式(17)求解前馈增益L_2,即可得到系统的控制器(11)和(12). 但是需要知道系统的模型,并且求解得到的并不是最优解. 值得注意的是,求解输出调节器方程(16)与反馈增益L_1是两个独立的过程,可以将以上求解过程转化为两个优化问题求解最优控制器.

2.2 控制器设计思路

将定理1的控制器求解问题转化为以下两个最优问题的求解.

问题1 定义如下静态优化问题:

$$\begin{aligned}&\min_{(X,U)}\ \operatorname{trace}(X^{\mathrm{T}}QX+U^{\mathrm{T}}RU);\\&\text{s.t. }XF=AX+BU+D,\\&\qquCX-E=0.\end{aligned}\tag{24}$$

其中$Q>0$和$R>0$为适当维数的对称正定矩阵.

结合式(18)～(20)和(22),系统(1)可被改写为

$$\begin{cases}\bar{x}(k+1)=A\bar{x}(k)+B\bar{u}(k),\\ e(k)=C\bar{x}(k).\end{cases}\tag{25}$$

基于系统(25)给出问题2.

问题2 求解如下动态最优问题的最优控制器$\bar{u}(k)=-L_1^*\bar{x}(k)$:

$$\begin{aligned}&\min\ J(\bar{x}(k))=\sum_{i=k}^{\infty}(\bar{x}^{\mathrm{T}}(i)\bar{Q}\bar{x}(i)+\bar{u}^{\mathrm{T}}(i)\bar{R}\bar{u}(i));\\ &\text{s.t. }\bar{x}(k+1)=A\bar{x}(k)+B\bar{u}(k),\\ &\quad e(k)=C\bar{x}(k).\end{aligned}\tag{26}$$

其中$\bar{Q}>0$和$\bar{R}>0$为适当维数的对称正定矩阵. 于是,由式(12)可知丢包时最优策略是$\bar{L}_1^*=L_1^*G$.

通过对问题1求解,可以求得(X,U),对问题2求解可以得到最优反馈增益矩阵L_1^*和$\bar{L}_1^*$,再根据式(17)求得最优的前馈增益矩阵L_2^*,即

$$L_2^*=U+L_1^*X,\tag{27}$$

从而得到式(11)和(12)中的最优控制输入

$$u^*(k)=-L_1^*x(k)+L_2^*v(k)=-\bar{L}_1^*\eta(k)+L_2^*v(k).\tag{28}$$

2.3 基于模型的最优输出调节控制器设计

给出模型已知时问题1和问题2的求解方法. 定义如下塞尔维斯特(Sylvester)映射:

$$N(X)=XF-AX.\tag{29}$$

根据文献[18],式(29)中X(即式(16)的一般解)为

$$X=X_1+\sum_{i=2}^{\tau+1}\beta_iX_i.\tag{30}$$

其中:$\beta_i\in\boldsymbol{R}$、$X_1\in\boldsymbol{R}^{n\times q}$为方程$CX_1=E$的特解,$X_i\in\boldsymbol{R}^{n\times q}(i=2,3,\ldots,\tau+1,\tau=(n-m)q)$为方程$CX=0$的基础解系. 根据式(16)、(29)和(30)可得

$$\begin{aligned}&N(X)=\\ &N(X_1)+\sum_{i=2}^{\tau+1}\beta_iN(X_i)=BU+D.\end{aligned}\tag{31}$$

于是,式(30)和(31)可表示为如下形式:

$$\Lambda\chi=\xi.\tag{32}$$

其中

$$\Lambda=\begin{bmatrix}\mathrm{vec}(N(X_2))&\ldots&\mathrm{vec}(N(X_{\tau+1}))&0&-I_q\otimes B\\ \mathrm{vec}(X_2)&\ldots&\mathrm{vec}(X_{\tau+1})&-I_{n\times q}&0\end{bmatrix}=\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\ \Lambda_{21}&\Lambda_{22}\end{bmatrix},\tag{33}$$

$$\chi=[\beta_2\ \ldots\ \beta_{\tau+1}\ \mathrm{vec}(X)^{\mathrm{T}}\ \mathrm{vec}(U)^{\mathrm{T}}]^{\mathrm{T}},\tag{34}$$

$$\xi=\begin{bmatrix}\mathrm{vec}(-N(X_1)+D)\\ -\mathrm{vec}(X_1)\end{bmatrix}=\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix},\tag{35}$$

$$\begin{aligned}&\Lambda_{11}=[\mathrm{vec}(N(X_2))\ \ \ldots\ \ \mathrm{vec}(N(X_{\tau+1}))],\\ &\Lambda_{12}=[0\ \ -I_q\otimes B],\\ &\Lambda_{21}=[\mathrm{vec}(X_2)\ \ \ldots\ \ \mathrm{vec}(X_{\tau+1})],\\ &\Lambda_{22}=[-I_{n\times q}\ \ 0],\end{aligned}\tag{36}$$

$\Lambda_{21}\in\boldsymbol{R}^{\tau\times\tau}$为非奇异矩阵.

基于以上推导以及文献[18],问题1可以写为

$$\begin{aligned}&\min\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}I_q\otimes Q&0\\ 0&I_q\otimes R\end{bmatrix}\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix};\\ &\text{s.t. }S\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}=T.\end{aligned}\tag{37}$$

其中

$$\begin{aligned}S&=-\Lambda_{11}\Lambda_{21}^{-1}\Lambda_{22}+\Lambda_{12},\\ T&=-\Lambda_{11}\Lambda_{21}^{-1}\xi_2+\xi_1.\end{aligned}$$

采用拉格朗日乘子法,引入拉格朗日乘子$\lambda\in\boldsymbol{R}^p,p=n\times q+m\times q$,式(37)中的$(X,U)$满足

$$\begin{bmatrix}2\begin{bmatrix}I_q\otimes Q&0\\ 0&I_q\otimes R\end{bmatrix}&S^{\mathrm{T}}\\ S&0\end{bmatrix}\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\\ \lambda\end{bmatrix}=\begin{bmatrix}0\\ 0\\ T\end{bmatrix}.\tag{38}$$

问题2是线性二次型调节器问题,目标是设计控制器最小化如下性能指标:

$$J(\bar{x}(k))=\sum_{i=k}^{\infty}(\bar{x}^{\mathrm{T}}(i)\bar{Q}\bar{x}(i)+\bar{u}^{\mathrm{T}}(i)\bar{R}\bar{u}(i)).\tag{39}$$

结合式(6)和(18),控制器形式为

$$\bar{u}(k)=-L_1\bar{x}(k)=-\bar{L}_1\eta(k)+L_1Xv(k).\tag{40}$$

根据最优控制参考文献[5],对于任意可稳定的反馈控制增益L_1,系统的性能指标(39)可表示为

$$J(\bar{x}(k))=\bar{x}^{\mathrm{T}}(k)P\bar{x}(k),\tag{41}$$

其中$P=P^{\mathrm{T}}>0$. 根据文献[21],能够最小化性能指标(41)的最优反馈增益(40)的解可表示为

$$L_1^*=(\bar{R}+B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA,\tag{42}$$

$$\bar{L}_1^*=(\bar{R}+B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAG,\tag{43}$$

其中P满足代数Riccati方程

$$P=\bar{Q}-A^{\mathrm{T}}PB(\bar{R}+B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA+A^{\mathrm{T}}PA,\tag{44}$$

对应的Lyapunov方程为

$$\begin{aligned}&P=\\ &\bar{Q}+L_1^{\mathrm{T}}\bar{R}L_1+(A-BL_1)^{\mathrm{T}}P(A-BL_1).\end{aligned}\tag{45}$$

于是,当系统模型已知时,可以通过式(42)～(44)求解P、L_1^*和$\bar{L}_1^*$,也可以通过文献[22]提出的迭代算法求解反馈增益. 由于该算法被用于稳定性和收敛性的分析,重述如下.

引理1[22]　令$A-BL_1^0$为稳定矩阵,利用下式求解Lyapunov方程的解$P^j={P^j}^{\mathrm{T}}>0$:

$$P^j=(A-BL_1^j)^{\mathrm{T}}P^j(A-BL_1^j)+\bar{Q}+{L_1^j}^{\mathrm{T}}\bar{R}L_1^j. \tag{46}$$

由式(42)和(43)可知

$$\begin{aligned}L_1^{j+1}&=(\bar{R}+B^{\mathrm{T}}P^jB)^{-1}B^{\mathrm{T}}P^jA,\\ \bar{L}_1^{j+1}&=(\bar{R}+B^{\mathrm{T}}P^jB)^{-1}B^{\mathrm{T}}P^jAG,\end{aligned} \tag{47}$$

$j=0,1,\ldots$为当前迭代次数,那么如下性质成立:

1) $A-BL_1^j$是稳定矩阵;

2) $P\leqslant P^{j+1}\leqslant P^j$;

3) $\lim\limits_{j\to\infty}L_1^j=L_1^*$, $\lim\limits_{j\to\infty}\bar{L}_1^j=\bar{L}_1^*$, $\lim\limits_{j\to\infty}P^j=P$.

2.4　数据驱动的最优输出调节控制器设计

考虑丢包环境下,系统模型A、B和D未知时,采用数据驱动方法求解最优输出调节控制器.

首先采用离轨策略强化学习算法求解问题2中的最优反馈增益L_1^*和$\bar{L}_1^*$,定义新的系统状态

$$\bar{x}_i(k)=x(k)-X_iv(k)=G\eta(k)-X_iv(k). \tag{48}$$

结合式(1)、(2)、(29)和(48)得到新的系统动态方程为

$$\bar{x}_i(k+1)=A\bar{x}_i(k)+Bu(k)-\pi(X_i)v(k), \tag{49}$$

其中$\pi(X_i)=N(X_i)-D$. 令$\bar{u}_i(k)=-L_1^j\bar{x}_i(k)$,其中$j=1,2,\ldots$表示当前迭代次数. 将式(49)加上并减去$B\bar{u}_i(k)$然后代入(41),结合式(46),有

$$\begin{aligned}&\bar{x}_i^{\mathrm{T}}(k+1)P^j\bar{x}_i(k+1)-\bar{x}_i^{\mathrm{T}}(k)P^j\bar{x}_i(k)=\\ &\bar{x}_i^{\mathrm{T}}(k)(-\bar{Q}-(L_1^j)^{\mathrm{T}}\bar{R}L_1^j)\bar{x}_i+\\ &2\bar{x}_i^{\mathrm{T}}(k)A^{\mathrm{T}}P^jB(L_1^j\bar{x}_i(k)+u(k))+\\ &(-L_1^j\bar{x}_i(k)+u(k))^{\mathrm{T}}B^{\mathrm{T}}P^jB(L_1^j\bar{x}_i(k)+u(k))-\\ &2\bar{x}_i^{\mathrm{T}}(k)A^{\mathrm{T}}P^j\pi(X_i)v(k)-2u^{\mathrm{T}}(k)B^{\mathrm{T}}P^j\pi(X_i)v(k)+\\ &v^{\mathrm{T}}(k)(\pi(X_i))^{\mathrm{T}}P^j\pi(X_i)v(k).\end{aligned} \tag{50}$$

在存在丢包的情况下,将式(48)预估的系统状态$\bar{x}_i(k)$代入(50),同时令$\eta(k)=[{\eta_1}^{\mathrm{T}}(k)\ {\eta_2}^{\mathrm{T}}(k)]^{\mathrm{T}}$,其中$\eta_1(k)\in\boldsymbol{R}^n$、$\eta_2(k)\in\boldsymbol{R}^\omega$分别为$\eta(k)$列向量中的前$n$个元素和余下的元素,即$\omega=z-n$. 将对应于式(7)的$G$表示为$G=[I\ \ \bar{G}]$,并利用克罗内克积(Kroneker product)展开,式(50)变为

$$\begin{aligned}&[\eta_1^{\mathrm{T}}(k+1)\otimes\eta_1^{\mathrm{T}}(k+1)-\eta_1^{\mathrm{T}}(k)\otimes\eta_1^{\mathrm{T}}(k)+\\ &2(X_iv(k))^{\mathrm{T}}\otimes\eta_1^{\mathrm{T}}(k)-(X_iv(k))^{\mathrm{T}}\otimes(X_iv(k))^{\mathrm{T}}-\\ &2(X_iv(k+1))^{\mathrm{T}}\otimes\eta_1^{\mathrm{T}}(k+1)+\\ &(X_iv(k+1))^{\mathrm{T}}\otimes(X_iv(k+1))^{\mathrm{T}}]\mathrm{vec}(P^j)+\\ &[\eta_2^{\mathrm{T}}(k+1)\otimes\eta_2^{\mathrm{T}}(k+1)-\\ &\eta_2^{\mathrm{T}}(k)\otimes\eta_2^{\mathrm{T}}(k)]\mathrm{vec}(\bar{G}^{\mathrm{T}}P^j\bar{G})+[2v^{\mathrm{T}}(k)\otimes\eta_2^{\mathrm{T}}(k)-\\ &2v^{\mathrm{T}}(k+1)\otimes\eta_2^{\mathrm{T}}(k+1)]\mathrm{vec}(\bar{G}^{\mathrm{T}}P^jX_i)-\\ &[\eta_2^{\mathrm{T}}(k)\otimes\eta_2^{\mathrm{T}}(k)]\mathrm{vec}(\bar{G}^{\mathrm{T}}(-\bar{Q}-(L_1^j)^{\mathrm{T}}\bar{R}L_1^j)\bar{G})-\\ &2[(\bar{L}_1^j\eta(k)-L_1^jX_iv(k)+u(k))^{\mathrm{T}}\otimes(\eta_1^{\mathrm{T}}(k)-\\ &(X_iv(k))^{\mathrm{T}})]\mathrm{vec}(A^{\mathrm{T}}P^jB)-\\ &[(\bar{L}_1^j\eta(k)-L_1^jX_iv(k)+u(k))^{\mathrm{T}}\otimes\\ &(L_1^jX_iv(k)-\bar{L}_1^j\eta(k)+u(k))^{\mathrm{T}}]\mathrm{vec}(B^{\mathrm{T}}P^jB)-\\ &2[(\bar{L}_1^j\eta(k)-L_1^jX_iv(k)+u(k))^{\mathrm{T}}\otimes\\ &\eta_2^{\mathrm{T}}(k)]\mathrm{vec}(\bar{G}A^{\mathrm{T}}P^jB)-\\ &[v(k)^{\mathrm{T}}\otimes v(k)^{\mathrm{T}}]\mathrm{vec}(\pi(X_i)^{\mathrm{T}}P^j\pi(X_i))+\\ &2[v(k)^{\mathrm{T}}\otimes(\eta_1^{\mathrm{T}}(k)-(X_iv(k))^{\mathrm{T}})]\mathrm{vec}(A^{\mathrm{T}}P^j\pi(X_i))+\\ &2[v(k)^{\mathrm{T}}\otimes u(k)^{\mathrm{T}}]\mathrm{vec}(B^{\mathrm{T}}P^j\pi(X_i))+\\ &2[v(k)^{\mathrm{T}}\otimes\eta_2^{\mathrm{T}}(k)]\mathrm{vec}(\bar{G}^{\mathrm{T}}(-\bar{Q}-(L_1^j)^{\mathrm{T}}\bar{R}L_1^j)X_i+\\ &\bar{G}^{\mathrm{T}}A^{\mathrm{T}}P^j\pi(X_i))=\\ &[\eta_1^{\mathrm{T}}(k)\otimes\eta_1^{\mathrm{T}}(k)-2(X_iv(k))^{\mathrm{T}}\otimes\eta_1^{\mathrm{T}}(k)+\\ &(X_iv(k))^{\mathrm{T}}\otimes(X_iv(k))^{\mathrm{T}}]\mathrm{vec}(-\bar{Q}-(L_1^j)^{\mathrm{T}}\bar{R}L_1^j).\end{aligned} \tag{51}$$

由式(51),定义

$$\begin{aligned}&V_i^{j+1}=\\ &[\mathrm{vec}(P^j)^{\mathrm{T}},\mathrm{vec}(\bar{G}^{\mathrm{T}}P^j\bar{G})^{\mathrm{T}},\mathrm{vec}(\bar{G}^{\mathrm{T}}P^jX_i)^{\mathrm{T}},\\ &\mathrm{vec}(\bar{G}^{\mathrm{T}}(-\bar{Q}-(L_1^j)^{\mathrm{T}}\bar{R}L_1^j)\bar{G})^{\mathrm{T}},\mathrm{vec}(M_1^{j+1})^{\mathrm{T}},\\ &\mathrm{vec}(M_2^{j+1})^{\mathrm{T}},\mathrm{vec}(M_3^{j+1})^{\mathrm{T}},\mathrm{vec}(M_{4i}^{j+1})^{\mathrm{T}},\\ &\mathrm{vec}(M_{5i}^{j+1})^{\mathrm{T}},\mathrm{vec}(B^{\mathrm{T}}P^j\pi(X_i))^{\mathrm{T}},\mathrm{vec}(\bar{G}^{\mathrm{T}}(-\bar{Q}-\\ &(L_1^j)^{\mathrm{T}}\bar{R}L_1^j)X_i+\bar{G}^{\mathrm{T}}A^{\mathrm{T}}P^j\pi(X_i))^{\mathrm{T}}]^{\mathrm{T}}.\end{aligned} \tag{52}$$

其中

$$\begin{aligned}&M_1^{j+1}=A^{\mathrm{T}}P^jB,\ M_2^{j+1}=B^{\mathrm{T}}P^jB,\\ &M_3^{j+1}=\bar{G}A^{\mathrm{T}}P^jB,\ M_{4i}^{j+1}=\pi(X_i)^{\mathrm{T}}P^j\pi(X_i),\\ &M_{5i}^{j+1}=A^{\mathrm{T}}P^j\pi(X_i);\end{aligned} \tag{53}$$

$$\phi_i^j(k)=[\theta_i^j(k)\ \ \theta_i^j(k+1)\ \ \theta_i^j(k+s-1)]^{\mathrm{T}}; \tag{54}$$

$$\begin{aligned}&\theta_i^j(k)=\\ &[\eta_1^{\mathrm{T}}(k)\otimes\eta_1^{\mathrm{T}}(k)-2(X_iv(k))^{\mathrm{T}}\otimes\eta_1^{\mathrm{T}}(k)+\\ &(X_iv(k))^{\mathrm{T}}\otimes(X_iv(k))^{\mathrm{T}}]\mathrm{vec}(-\bar{Q}-(L_1^j)^{\mathrm{T}}\bar{R}L_1^j);\end{aligned} \tag{55}$$

$$\psi_i^j(k)=\begin{bmatrix}H_1(1)&H_1(2)&\ldots&H_1(11)\\ \vdots&\vdots&\ddots&\vdots\\ H_s(1)&H_s(2)&\ldots&H_s(11)\end{bmatrix}; \tag{56}$$

$$
\begin{aligned}
&H_l(1)=\\
&\eta_1^{\mathrm{T}}(k+l)\otimes\eta_1^{\mathrm{T}}(k+l)-\eta_1^{\mathrm{T}}(k+l-1)\otimes\\
&\eta_1^{\mathrm{T}}(k+l-1)-2(X_iv(k+l))^{\mathrm{T}}\otimes\eta_1^{\mathrm{T}}(k+l)+\\
&2(X_iv(k+l-1))^{\mathrm{T}}\otimes\eta_1^{\mathrm{T}}(k+l-1)+\\
&(X_iv(k+l))^{\mathrm{T}}\otimes(X_iv(k+l))^{\mathrm{T}}-\\
&(X_iv(k+l-1))^{\mathrm{T}}\otimes(X_iv(k+l-1))^{\mathrm{T}},\\
&H_l(2)=\eta_2^{\mathrm{T}}(k+l)\otimes\eta_2^{\mathrm{T}}(k+l)-\\
&\qquad\eta_2^{\mathrm{T}}(k+l-1)\otimes\eta_2^{\mathrm{T}}(k+l-1),\\
&H_l(3)=-2v^{\mathrm{T}}(k+l)\otimes\eta_2^{\mathrm{T}}(k+l)+\\
&\qquad 2v^{\mathrm{T}}(k+l-1)\otimes\eta_2^{\mathrm{T}}(k+l-1),\\
&H_l(4)=-\eta_2^{\mathrm{T}}(k+l-1)\otimes\eta_2^{\mathrm{T}}(k+l-1),\\
&H_l(5)=-2[(\bar{L}_1^j\eta(k+l-1)-L_1^jX_iv(k+l-1)+\\
&\qquad u(k+l-1))^{\mathrm{T}}\otimes(\eta_1^{\mathrm{T}}(k+l-1)-\\
&\qquad(X_iv(k+l-1))^{\mathrm{T}})],\\
&H_l(6)=-[(\bar{L}_1^j\eta(k+l-1)-L_1^jX_iv(k+l-1)+\\
&\qquad u(k+l-1))^{\mathrm{T}}\otimes(\bar{L}_1^j\eta(k+l-1)-\\
&\qquad L_1^jX_iv(k+l-1)+u(k+l-1))^{\mathrm{T}}],\\
&H_l(7)=-2[(\bar{L}_1^j\eta(k+l-1)-L_1^jX_iv(k+l-1)+\\
&\qquad u(k+l-1))^{\mathrm{T}}\otimes\eta_2^{\mathrm{T}}(k+l-1)],\\
&H_l(8)=-v^{\mathrm{T}}(k+l-1)\otimes v^{\mathrm{T}}(k+l-1),\\
&H_l(9)=2[v^{\mathrm{T}}(k+l-1)\otimes(\eta_1^{\mathrm{T}}(k+l-1)-\\
&\qquad(X_iv(k+l-1))^{\mathrm{T}})],\\
&H_l(10)=2[v^{\mathrm{T}}(k+l-1)\otimes u^{\mathrm{T}}(k+l-1)],\\
&H_l(11)=2[v^{\mathrm{T}}(k+l-1)\otimes\eta_2^{\mathrm{T}}(k+l-1)];
\end{aligned}\tag{57}
$$

$s\geqslant 0$为采样数据的组数. 基于式(52)$\sim$(57),式(51)可转化为

$$
\psi_i^j(k)V_i^{j+1}=\phi_i^j(k).\tag{58}
$$

于是可通过式(58)求得V_i^{j+1},从而由式(41)和(43),式(53)反馈增益可表示为

$$
\begin{aligned}
&L_1^{j+1}=(\bar{R}+M_2^{j+1})^{-1}(M_1^{j+1})^{\mathrm{T}},\\
&\bar{L}_1^{j+1}=(\bar{R}+M_2^{j+1})^{-1}[M_1^{j+1};\ M_3^{j+1}]^{\mathrm{T}}.
\end{aligned}\tag{59}
$$

需要注意的是,在求解式(58)时,因为式(52)中待求未知参数的个数为

$$
\begin{aligned}
&\alpha=\\
&(z-n)(z-n+m-1)+(2z-n+m)q+mn+\\
&\frac{n(n+1)+m(m+1)+q(q+1)}{2},
\end{aligned}
$$

所以采样数据的组数要不小于未知参数的个数,即$s\geqslant\alpha$. 同时可加入探测噪声使得$\mathrm{rank}(\psi_i^j(k))=\alpha$,即可用最小二乘法对式(58)进行求解,得到

$$
V_i^{j+1}=((\psi_i^j(k))^{\mathrm{T}}\psi_i^j(k))^{-1}(\psi_i^j(k))^{\mathrm{T}}\phi_i^j(k).\tag{60}
$$

注1 采用离轨策略强化学习算法求解最优反馈增益L_1^*和$\bar{L}_1^*$时,需要对控制输入$\bar{u}(k)$和外系统的状态$v(k)$加入探测噪声以达到满足$\psi_i^j(k)$列满秩的条件. 由文献[8,21]可知,离轨策略算法可保证在系统存在探测噪声的情况下所求解是无偏的.

下面利用在线数据求解问题1的解(X,U). 定义

$$
\begin{aligned}
&\bar{N}(X)=A^{\mathrm{T}}P^jN(X),\\
&\bar{\pi}(X)=A^{\mathrm{T}}P^j\pi(X).
\end{aligned}\tag{61}
$$

结合式(53)和(61)可以得到

$$
\begin{aligned}
&\bar{\pi}(X_0)=-A^{\mathrm{T}}P^jD=M_{50}^{i+1},\\
&\bar{\pi}(X_i)=-A^{\mathrm{T}}P^j\pi(X_i)=M_{5i}^{i+1},\\
&\bar{N}(X_i)=A^{\mathrm{T}}P^jN(X_i)=M_{5i}^{i+1}-M_{50}^{i+1},
\end{aligned}\tag{62}
$$

其中$X_0=0$. 从而式(37)中约束条件可写为

$$
\bar{S}\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}=\bar{T}.\tag{63}
$$

其中

$$
\begin{aligned}
&\bar{S}=-\bar{\Lambda}_{11}\bar{\Lambda}_{21}^{-1}\bar{\Lambda}_{22}+\bar{\Lambda}_{12},\\
&\bar{T}=-\bar{\Lambda}_{11}\bar{\Lambda}_{21}^{-1}\bar{\xi}_2+\bar{\xi}_1,\\
&\bar{\Lambda}=\\
&\begin{bmatrix}\mathrm{vec}(\bar{N}(X_2))&\dots&\mathrm{vec}(\bar{N}(X_{\tau+1}))&0\\ \mathrm{vec}(X_2)&\dots&\mathrm{vec}(X_{\tau+1})&-I_{n\times q}\end{bmatrix}\rightarrow\\
&\leftarrow\begin{bmatrix}-I_q\otimes A^{\mathrm{T}}P^{j+1}B\\ 0\end{bmatrix}=\\
&\begin{bmatrix}\mathrm{vec}(M_{52}^{i+1}-M_{50}^{i+1})&\dots&\mathrm{vec}(M_{5(\tau+1)}^{i+1}-M_{50}^{i+1})\\ \mathrm{vec}(X_2)&\dots&\mathrm{vec}(X_{\tau+1})\end{bmatrix}\rightarrow\\
&\leftarrow\begin{bmatrix}0&-I_q\otimes M_1^{i+1}\\ -I_{n\times q}&0\end{bmatrix}=\begin{bmatrix}\bar{\Lambda}_{11}&\bar{\Lambda}_{12}\\ \bar{\Lambda}_{21}&\bar{\Lambda}_{22}\end{bmatrix},
\end{aligned}\tag{64}
$$

$$
\begin{aligned}
&\bar{\Lambda}_{11}=\\
&[\mathrm{vec}(M_{52}^{i+1}-M_{50}^{i+1})\ \dots\ \mathrm{vec}(M_{5(\tau+1)}^{i+1}-M_{50}^{i+1})],\\
&\bar{\Lambda}_{12}=[0\ -I_q\otimes M_1^{i+1}],\\
&\bar{\Lambda}_{21}=[\mathrm{vec}(X_2)\ \dots\ \mathrm{vec}(X_{\tau+1})],\\
&\bar{\Lambda}_{22}=[-I_{n\times q}\ 0],
\end{aligned}\tag{65}
$$

$$
\bar{\xi}=\begin{bmatrix}\mathrm{vec}(-\bar{N}(X_1)+\bar{\pi}(X_0))\\ -\mathrm{vec}(X_1)\end{bmatrix}=\begin{bmatrix}\bar{\xi}_1\\ \bar{\xi}_2\end{bmatrix}.\tag{66}
$$

因为$\bar{\Lambda}_{21}=\Lambda_{21}$,所以$\bar{\Lambda}_{21}$同样为非奇异矩阵.

基于式(37)、(38)和(63)可知,式(37)中的(X,U)满足如下方程:

$$\begin{bmatrix} 2\begin{bmatrix} I_q\otimes Q & 0 \\ 0 & I_q\otimes R \end{bmatrix} & \bar{S}^{\mathrm{T}} \\ \bar{S} & 0 \end{bmatrix} \begin{bmatrix} \mathrm{vec}(X) \\ \mathrm{vec}(U) \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \bar{T} \end{bmatrix}. \quad (67)$$

由式(64)～(66)可知$\bar{S}$和$\bar{T}$是已知的,故由式(67)可计算得到(X,U).

算法1　离轨策略强化学习算法.

step 1: 计算$CX=0$的基础解系$X_i\in \boldsymbol{R}^{n\times q}(i=2,3,\dots,\tau+1,\ \tau=(n-m)q)$,以及$CX=E$的特解$X_1$,给定初始稳定的控制策略$L_1^0$和$\bar{L}_1^0$,设$i=1,j=0$,同时将$\widehat{u}=-\bar{L}_1^j\eta(k)+e_1(k)$作为控制输入信号,$\widehat{v}=v(k)+e_2(k)$作为外系统状态,$e_1(k)$和$e_2(k)$为探测噪声,采集并存储数据于$\psi_i^j(k)$.

step 2: 策略评估,由式(58)求得V_i^{j+1}.

step 3: 策略更新,由式(59)求得L_1^{j+1}和$\bar{L}_1^{j+1}$.

step 4: 令$j=j+1$,重复step 2和step 3直到满足$\|L_1^{j+1}-L_1^j\|_2\leqslant\mathcal{E}_1,\|\bar{L}_1^{j+1}-\bar{L}_1^j\|_2\leqslant\mathcal{E}_2$($\mathcal{E}_1$和$\mathcal{E}_2$为小正常数),令反馈增益$L_1^\infty=L_1^{j+1},\bar{L}_1^\infty=\bar{L}_1^{j+1}$.

step 5: 令$i=i+1,j=0$,重复step 2～step 4,求解式(58)得到M_{5i}^{j+1},直到$i=\tau+1$.

step 6: 由式(67)求得(X,U).

step 7: 由式(27)求得前馈增益L_2^∞,即

$$L_2^\infty=U+L_1^\infty X. \quad (68)$$

step 8: 由式(28)计算控制输入$u(k)$.

注2　因为式(58)由(50)转化所得,所以由式(58)和(59)求得的P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$满足(50),且由文献[23]可知$\psi_i^j(k)$列满秩保证所求得的解是唯一解.

2.5　控制器稳定性和算法收敛性分析

定理2　由算法1得到的控制器$u(k)$可以使得系统(1)稳定并实现$\lim\limits_{k\to\infty}e(k)=0$.

证明　由引理1可知,给定初始稳定的控制策略L_1^0,可通过式(46)和(47)求解P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$,且$A-BL_1^{j+1}$为稳定矩阵. 由注2可知式(58)和(59)的唯一解P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$满足式(50),且式(50)的解满足Lyapunov方程(45),所以算法1中的式(58)和(59)等价于引理1中式(46)和(47). 由引理1的性质1)可知算法1求得的反馈增益能够使得$A-BL_1^{j+1}$为稳定矩阵. 因为式(67)解得的(X,U)满足(16),由定理1可知算法1解得的控制器(28)在保证系统稳定性的同时也使得系统的输出$y(k)$跟踪参考轨迹$w(k)$. □

定理3　当满足$\mathrm{rank}(\psi_i^j(k))=\alpha$时,给定初始稳定的控制策略$L_1^0$和$\bar{L}_1^0$,由算法1求解得到的$P^j$、$L_1^{j+1}$和$\bar{L}_1^{j+1}$分别收敛至其最优值$P$、$L_1^*$和$\bar{L}_1^*$.

证明　由引理1可知,给定初始稳定的控制策略L_1^0,通过式(46)和(47)求解的P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$分别收敛至满足式(44)、(42)和(43)的最优值. 由定理2可知式(58)、(59)与(46)、(47)等价. 并且根据文献[23],$\psi_i^j(k)$列满秩保证了所求解是唯一解,所以同理在算法1中给定稳定的初始控制策略L_1^0和$\bar{L}_1^0$,由式(58)和(59)求解的P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$分别收敛至最优值P、L_1^*和$\bar{L}_1^*$. □

3　仿真实验

为了验证所提出方法的有效性,给出基于模型的计算结果以及数据驱动算法1的仿真结果.

3.1　仿真实验参数选择以及模型驱动方法计算结果

考虑如下线性离散时间系统:

$$\begin{aligned} &x(k+1)=\\ &\begin{bmatrix} -1 & 2 \\ 2.2 & 1.7 \end{bmatrix}x(k)+\begin{bmatrix} 2 \\ 1.6 \end{bmatrix}u(k)+\begin{bmatrix} 0 \\ -1 \end{bmatrix}v(k),\\ &y(k)=[1\ \ 2]x(k). \end{aligned} \quad (69)$$

系统开环极点分别为$-2.144\,5$和$2.844\,5$,它们均在单位圆外,所以开环系统是不稳定的.

选取参考轨迹(2)和(3)中的$E=1,F=-1$,并取$v(k)$的初值为15,则参考轨迹如下:

$$\begin{aligned} v(k+1)&=-v(k),\\ w(k)&=v(k). \end{aligned} \quad (70)$$

选取系统的最大连续丢包次数$\delta_{f\max}=1$,可以得到史密斯预估器的参数为

$$G=[I\ \ A\ \ B\ \ D]=\begin{bmatrix} 1 & 0 & -1 & 2 & 2 & 0 \\ 0 & 1 & 2.2 & 1.7 & 1.6 & -1 \end{bmatrix}. \quad (71)$$

选取性能指标(24)的权重为$Q=8I$, $R=1$,式(26)的权重为$\bar{Q}=8I$, $\bar{R}=1$. 基于以上参数,给出如下计算结果. 问题1的解为

$$\begin{aligned} X&=[0.272\,7;\ 0.363\,6],\\ U&=-0.363\,6. \end{aligned} \quad (72)$$

通过求解式(42)～(44),可以求得问题2中Riccati方程的解P和最优的反馈增益L_1^*、$\bar{L}_1^*$,即

$$P=\begin{bmatrix} 81.188\,0 & 2.090\,2 \\ 2.090\,2 & 9.088\,8 \end{bmatrix}, \quad (73)$$

$$\begin{aligned} L_1^*&=[-0.343\,6\ \ 1.002\,4],\\ \bar{L}_1^*&=[-0.343\,6\ \ 1.002\,4\ \ 2.548\,9\ \ 1.016\,8\rightarrow\\ &\quad\leftarrow 0.916\,6\ \ 1.002\,4]. \end{aligned} \quad (74)$$

进而由式(28)求得L_2^*,即

$$L_2^*=-0.092\,8. \quad (75)$$

3.2 数据驱动算法1仿真结果

考虑系统的最大连续丢包次数$\delta_{f\max}=1$,丢包概率4%,控制律初值选为

$$L_1^0=[-0.2\ \ 1],$$
$$\bar{L}_1^0=[-0.2\ \ 1\ \ 2.4\ \ 1.2\ \ 1.1\ \ -1]. \tag{76}$$

算法1的仿真结果如图1～图4所示. 图1显示了被控系统输出跟踪性能,可见最后得到了满意的跟踪效果. 图2为系统的控制输入轨迹. 图3显示了学习过程中P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$向式(73)和(74)中P、L_1^*和$\bar{L}_1^*$收敛的情况. 图4为连续丢包次数.

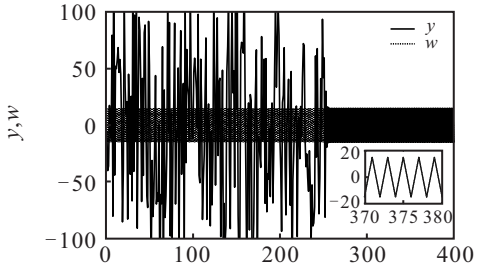

图1 系统输出跟踪性能(问题1)

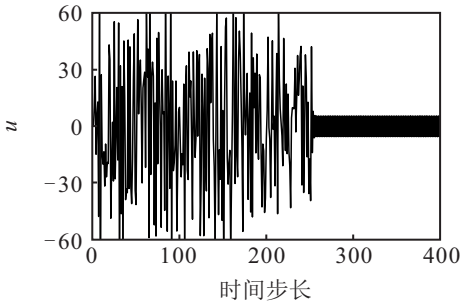

图2 系统控制输入轨迹

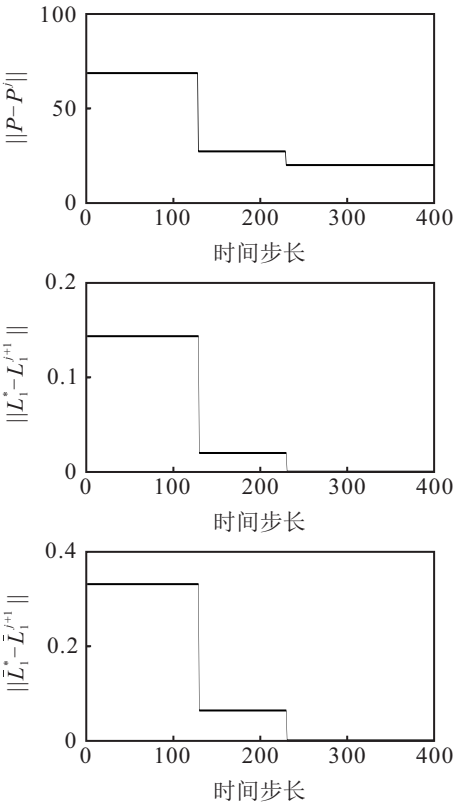

图3 学习过程中P^j、L_1^{j+1}和$\bar{L}_1^{j+1}$收敛曲线

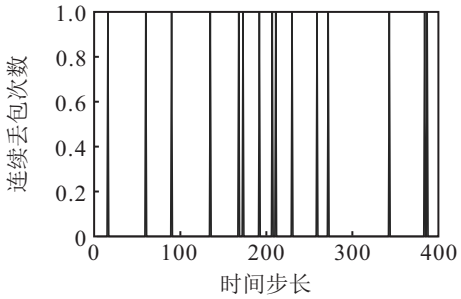

图4 4%丢包率随机丢包序列

算法1的收敛解如下所示.

问题2中黎卡提方程的收敛解为

$$P^\infty=\begin{bmatrix}81.1946 & 2.0907\\ 2.0907 & 9.0926\end{bmatrix}. \tag{77}$$

反馈控制律的收敛值L_1^∞和$\bar{L}_1^\infty$分别为

$$L_1^\infty=[-0.3438\ \ 1.0024],$$
$$\bar{L}_1^\infty=[-0.3438\ \ 1.0024\ \ 2.5491\ \ 1.0166\rightarrow$$
$$\leftarrow 0.9163\ \ 1.0024]. \tag{78}$$

问题1的解为

$$X=[0.2727;\ 0.3636],\ U=-0.3636. \tag{79}$$

进而由式(68)可得

$$L_2^\infty=-0.0928. \tag{80}$$

通过比较基于模型的计算结果(72)～(75)和算法1的仿真结果(77)～(80),可知在模型未知时,采用算法1 得到的解会收敛到根据模型计算得到的最优解.

下面考虑相同丢包概率下,最大连续丢包次数$\delta_{f\max}=2$时算法1的仿真结果. 被控系统输出跟踪性能如图5所示,可见最后得到了满意的跟踪效果. 图6为连续丢包次数.

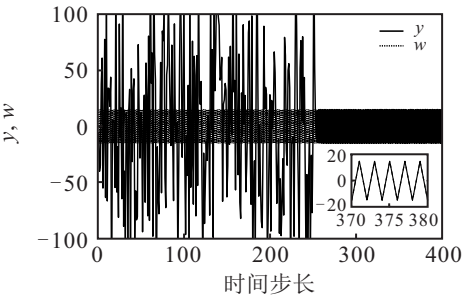

图5 系统输出跟踪性能(问题2)

3.3 对比仿真实验

为了表明低概率丢包对系统性能的影响,本节设置与第3.2节相同的丢包概率,采用不进行丢包补偿处理的数据驱动输出调节算法进行对比仿真实验.

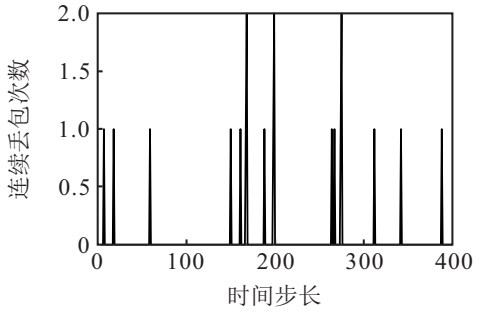

图 6　最大连续丢包次数为2的随机丢包序列

仿真结果如图7和图8所示,分别为被控系统输出跟踪性能和控制输入轨迹. 由图7可以明显看出,即使在低丢包概率下,若不采取有效的方法对丢包数据进行补偿和处理,则系统的跟踪性能也会受到影响.

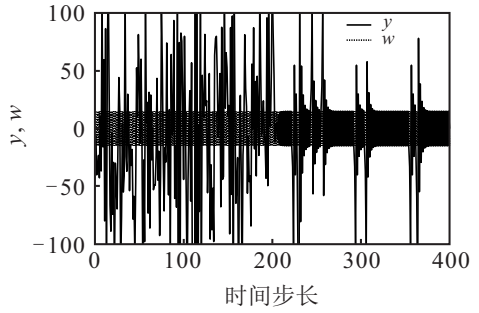

图 7　系统输出跟踪性能(对比仿真)

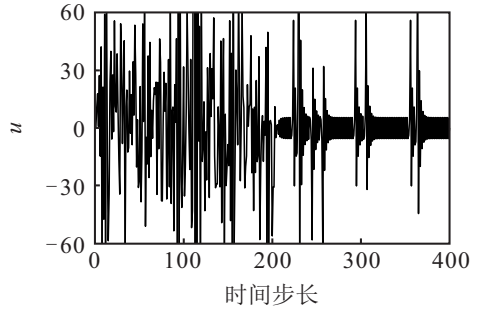

图 8　系统控制输入轨迹(对比仿真)

本文引入绝对积分误差(integral absolute error, IAE)和均方误差(mean square error, MSE)[6,11]两个指标评价低概率丢包对系统性能的影响,有

$$\mathrm{IAE}=\sum_{k=1}^{k^*}|w(k)-y(k)|,$$

$$\mathrm{MSE}=\frac{1}{k^*}\sum_{k=1}^{k^*}|w(k)-y(k)|^2. \tag{81}$$

表 1　实验评价指标

	时间范围	IAE	MSE
对比算法	$360\leqslant k\leqslant 380$	37.8843	387.5939
本文算法	$360\leqslant k\leqslant 380$	0.0766	2.4640×10^{-4}

评价结果如表1所示,可以看出本文算法在系统丢包时跟踪性能更好.

4　结　论

本文提出了基于离轨策略强化学习迭代算法的数据驱动输出调节控制方法,用于求解丢包扰动环境下控制系统的最优输出调节问题. 针对丢包问题,利用史密斯预估器的思想建立无线网络环境下系统的丢包模型,给出了系统发生丢包时状态的表达式,并分别从模型已知和模型未知两个方面对控制器进行求解. 在未知系统模型参数的情况下,采用基于离轨策略强化学习迭代算法的数据驱动算法,只利用在线数据即可求得存在丢包问题时输出调节问题的解,从而实现不依赖模型,仅通过采集的数据求解实现系统干扰抑制以及对参考轨迹渐进跟踪的控制律. 仿真结果验证了所提出方法的有效性.

参考文献(References)

[1] Xing Z R, Xia Y Q, Yan L P, et al. Multisensor distributed weighted Kalman filter fusion with network delays, stochastic uncertainties, autocorrelated, and cross-correlated noises[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 48(5): 716-726.

[2] Li C C, Han C Y, He F. Receding horizon estimation for networkde control systems with packet losses[C]. The 15th International Conference on Control, Automation, Robotics and Vision. Singapore, 2018: 1010-1015.

[3] Li Z M, Chang X H, Yu L. Robust quantized H_∞ filtering for discrete-time uncertain systems with packet dropouts[J]. Applied Mathematics and Computation, 2016, 275: 361-371.

[4] Li Z M, Chang X H, Wang Y M. Robust observer-based H_1 control for networked control systems with measurement quantization and packet dropouts[C]. The 12th IEEE Conference on Industrial Electronics and Applications. Siem Reap, 2017: 739-744.

[5] Al-Tamimi A, Lewis F L, Abu-Khalaf M. Model-free Q-learning designs for discrete-time zero-sum games with application to H_∞ control[J]. Automatica, 2007, 43(3): 473-481.

[6] Jiang Y, Fan J L, Chai T Y, et al. Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 4607-4620.

[7] Xue W Q, Fan J L, Lopez V G, et al. New methods for optimal operational control of industrial processes using reinforcement learning on two time scales[J]. IEEE

Transactions on Industrial Informatics, 2020, 16(5): 3085-3099.

[8] Kiumarsi B, Lewis F L, Jiang Z P. H_1 control of linear discrete-time systems: Off-policy reinforcement learning [J]. Automatica, 2017, 78: 144-152.

[9] Xue W Q, Fan J L, Lopez V G, et al. Off-policy reinforcement learning for tracking in continuous-time systems on two time scales[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(10): 4334-4346.

[10] Cheng W R, Xiao Z F, Li J N. Optimal tracking control of partial unknown continuous-time systems using integral reinforcement learning[C]. The 35th Youth Academic Annual Conference of Chinese Association of Automatio. Zhanjiang, 2020: 308-311.

[11] Kamalapurkar R, Dinh H, Bhasin S, et al. Approximate optimal trajectory tracking for continuous-time nonlinear systems[J]. Automatica, 2015, 51: 40-48.

[12] Liang D, Huang J. Robust output regulation of linear systems by event-triggered dynamic output feedback control[J]. IEEE Transactions on Automatic Control, 2021, 66(5): 2415-2422.

[13] Teng Y. Solution to output regulation problems for linear systems[C]. The 8th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou, 2016: 179-182.

[14] Yan Y M, Huang J. Output regulation problem for discrete-time linear time-delay systems[C]. The 34th Chinese Control Conference. Hangzhou, 2015: 5681-5686.

[15] Jiang Y, Dai J Y. Adaptive output regulation of a class of nonlinear output feedback systems with unknown high frequency gain[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(2): 568-574.

[16] Gao W N, Jiang Z P. Adaptive dynamic programming and adaptive optimal output regulation of linear systems[J]. IEEE Transactions on Automatic Control, 2016, 61(12): 4164-4169.

[17] Gao W N, Jiang Z P, Gao W N, et al. Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2614-2624.

[18] Jiang Y, Kiumarsi B, Fan J L, et al. Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning[J]. IEEE Transactions on Cybernetics, 2020, 50(7): 3147-3156.

[19] Fan J L, Wu Q, Jiang Y, et al. Model-free optimal output regulation for linear discrete-time lossy networked control systems[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(11): 4033-4042.

[20] Huang J. Nonlinear output regulation: Theory and applications[M]. Philadelphia: SIAM, 2004: 1-31.

[21] Lewis F L, Vrabie D, Syrmos V L. Optimal control[M]. The 3rd edition. NewYork: Wiley, 2012: 19-109.

[22] 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.

[23] Gao W N, Yu J, Jiang Z P, et al. Adaptive and optimal output feedback control of linear systems: An adaptive dynamic programming approach[C]. Proceeding of the 11th World Congress on Intelligent Control and Automation. Shenyang, 2014: 2085-2090.

作者简介

崔云芳 (1995−), 女, 硕士生, 从事强化学习、网络控制的研究, E-mail: yunfangcui102@163.com;

范家璐 (1985−), 女, 副教授, 博士, 从事强化学习、网络化运行控制、工业无线网络等研究, E-mail: jlfan@mail.neu.edu.cn.

(责任编辑：郑晓蕾)