基于强化学习的部分线性离散时间系统的最优输出调节

庞文砚 [1] 范家璐 [1] 姜 艺 [1] LEWIS Frank Leroy [2]

摘 要 针对同时具有线性外部干扰与非线性不确定性下的离散时间部分线性系统的最优输出调节问题, 提出了仅利用在线数据的基于强化学习的数据驱动控制方法. 首先, 该问题可拆分为一个受约束的静态优化问题和一个动态规划问题, 第一个问题可以解出调节器方程的解. 第二个问题可以确定出控制器的最优反馈增益. 然后, 运用小增益定理证明了存在非线性不确定性离散时间部分线性系统的最优输出调节问题的稳定性. 针对传统的控制方法需要准确的系统模型参数用来解决这两个优化问题, 提出了一种数据驱动离线策略更新算法, 该算法仅使用在线数据找到动态规划问题的解. 然后, 基于动态规划问题的解, 利用在线数据为静态优化问题提供了最优解. 最后, 仿真结果验证了该方法的有效性.

关键词 输出调节, 离散时间系统, 强化学习, 非线性未知动态

引用格式 庞文砚, 范家璐, 姜艺, Lewis Frank Leroy. 基于强化学习的部分线性离散时间系统的最优输出调节. 自动化学报, 2022, **48**(9): 2242−2253

DOI 10.16383/j.aas.c190853

Optimal Output Regulation of Partially Linear Discrete-time Systems Using Reinforcement Learning

PANG Wen-Yan[1] FAN Jia-Lu[1] JIANG Yi[1] LEWIS Frank Leroy[2]

Abstract A data-driven control method only using online data based on reinforcement learning is proposed for the optimal output regulation problem of discrete-time partially linear systems with both linear disturbance and nonlinear uncertainties. First, the problem can be split into a constrained static optimization problem and a dynamic one. The solution of the first problem is corresponding to the solution of the regulator equation. The second can determine the optimal feedback gain of the controller. Then the small-gain theorem is used to prove the stability of the optimal output regulation problem of discrete-time partially linear systems with nonlinear uncertainties. The traditional control method needs the dynamics of the system to solve the two problems. But for this problem, a data-driven off-policy algorithm is proposed using only the measured data to find the solution of the dynamic optimization problem. Then, based on the solution of the dynamic one, the solution of the static optimization problem can be found only using data online. Finally, simulation results verify the effectiveness of the proposed method.

Key words Output regulation, discrete-time system, reinforcement learning, nonlinear unknown dynamics

Citation Pang Wen-Yan, Fan Jian-Lu, Jiang Yi, Lewis Frank Leroy. Optimal output regulation of partially linear discrete-time systems using reinforcement learning. *Acta Automatica Sinica*, 2022, **48**(9): 2242−2253

输出调节问题是一种对于线性和非线性动态系统, 设计反馈控制器从而使系统实现渐近跟踪和干扰抑制的问题[1−5]. 输出调节问题的显著特征则是参考输入和干扰由已知的外系统自主微分或差分方产生的[5]. 目前, 已有学者研究了连续时间系统的输出调节问题[6−8]. 文献 [5] 对线性和非线性连续时间系统的输出调节问题给出了解决框架. 文献 [6] 研究了一类加入瞬态性能概念的输出调节问题, 详细研究了可解性条件和调节器结构等问题. 而文献 [5−6] 都需要在系统的动态模型参数已知的情况下, 解决其输出调节问题.

强化学习作为一种机器学习方法, 是以目标为导向的学习工具, 其中智能体或是决策者通过与环境交互为最优化长期奖励来学习控制策略[9−11], 可主要解决控制领域中的最优控制问题, 其中包括最优调节, 最优跟踪以及最优协同问题. 最优控制问题是一类通过使得代价函数或性能指标达到最优而为动态系统寻找控制律的问题. 典型的最优控制问题是需要系统的模型参数完全已知, 问题的求解是

收稿日期 2019-12-16 录用日期 2020-04-07
Manuscript received December 16, 2019; accepted April 7, 2020
国家自然科学基金 (61533015, 61991404, 61991403) 和辽宁省兴辽英才计划 (XLYC2007135) 资助
Supported by National Natural Science Foundations of China (61533015, 61991404, 61991403) and Liaoning Revitalization Talents Program (XLYC2007135)
本文责任编委 魏庆来
Recommended by Associate Editor WEI Qing-Lai
1. 东北大学流程工业综合自动化国家重点实验室 沈阳 110819 中国 2. 德克萨斯大学阿灵顿分校 沃斯堡 76118 美国
1. State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China 2. University of Texas at Arlington, Fort Worth 76118, USA

离线的, 其不能适应动态系统中模型参数的变化和不确定性, 因此数据驱动的强化学习方法也就应运而生, 广泛应用于解决离散时间和连续时间不确定系统的最优控制问题. 文献 [12] 利用数据驱动的强化学习方法利用沿着系统的数据解决了线性系统的最优跟踪问题, 又因为系统的状态数据往往难以获得, 文献 [13] 提出仅利用输入输出数据, 利用强化学习中的策略迭代和值迭代算法在线寻得最优控制律从而实现最优跟踪. 这 2 篇文献是针对于线性系统, 文献 [14] 则针对于非线性系统, 采用基于 Actor-Critic 结构的强化学习方法数据驱动在线学习跟踪哈密顿−雅可比−贝尔曼方程 (Hamilton-Jacobi-Bellman, HJB), 从而解决最优跟踪问题. 由于 H 无穷问题也可看作是一种最优控制问题, 主要是分别找出最优反馈控制律和最优扰动控制律的一类问题, 因此强化学习也应用于该问题的解决. 针对于 H 无穷控制问题, 对于线性系统模型参数未知的文献 [15], 该文采用强化学习离线策略控制方法进行解决, 并证明了探测噪声会对在线策略迭代算法产生影响使获得参数不准确, 而则不会对离线的策略迭代算法产生影响, 同时证明了离线策略迭代算法的收敛性. 文献 [16]则对于未知的非线性系统, 采用强化学习的离线策略方法学习跟踪哈密顿−雅可比−艾萨克方程方程 (Hamilton-Jacobi-Isaac, HJI) 的解, 在不知道系统模型参数的情况下解决了 H 无穷跟踪控制问题, 并给出所提算法的收敛性. 数据驱动的强化学习方法还可应用于无线网络环境下的控制问题, 文献 [17] 就针对于离散时间的网络系统利用沿着系统轨迹的数据实现网络控制系统的最优跟踪问题. 数据驱动的强化学习方法近年来解决了线性与非线性系统、连续和离散系统、传统状态空间控制和网络控制系统、利用沿系统轨迹数据和利用输入输出数据等的最优控制问题.

前文提到传统的输出调节问题都是基于系统的模型参数即模型已知的前提下求解输出调节问题. 而文献 [7–8] 则是在系统模型参数不确定的情况下利用数据驱动的方法解决输出调节问题. 在文献 [7–8] 中, 对于连续时间系统分别采用近似动态规划和鲁棒近似动态规划的方法解决了线性系统和部分线性系统的最优输出调节问题. 由于强化学习是解决最优控制问题的有力工具, 前述也有许多学者采用了强化学习方法解决最优跟踪问题, 现在另外考虑外部系统的干扰, 把强化学习应用到解决最优输出调节问题中. 文献 [18] 将文献 [7] 中利用数据驱动方法求解线性连续时间系统的最优输出调节问题拓展到线性离散时间系统中. 本文则是针对部分线性的离散时间系统, 在具有模型参数未知的情况下, 利用基于强化学习的离线策略更新方法数据驱动求解最优输出调节问题.

本文将数据驱动的强化学习方法与最优输出调节问题相结合. 主要贡献如下: 针对于存在线性干扰和非线性不确定性的部分离散时间系统的最优输出调节问题, 提出基于强化学习的离线策略更新算法. 该方法不需要知道系统的模型参数, 只利用测量数据在线求解即可实现对最优输出调节控制律的自适应学习, 即可应对系统模型参数的变化, 且提出的方法不仅可以抑制线性的外部干扰并且对动态非线性不确定性存在鲁棒性保证渐近跟踪. 并运用了小增益定理说明了本文提出的方法可以保证闭环系统的稳定性.

本文结构如下: 第 1 节介绍离散时间部分线性系统的最优输出调节问题. 提出最优输出调节问题中的两个优化问题, 分别为静态优化问题和动态优化问题; 然后将该离散时间系统转化为误差系统, 通过证明误差系统的全局渐近稳定性以推出原系统的最优输出调节问题的可解性. 第 2 节针对具有线性外部干扰和非线性不确定性的部分线性离散时间系统, 提出离线策略更新算法利用在线数据求解动态规划问题, 并基于动态规划问题的解, 用数据驱动的方法解静态规划问题以此解决其最优输出调节问题. 第 3 节提供仿真结果验证本文方法的有效性, 并进行对比实验, 比较性能指标突显本文方法的优越性. 第 4 节为结束语.

符号说明及概念介绍. $\mathbf{R}_+$表示非负实数集, $\mathbf{R}^{n\times m}$ 表示 $n\times m$ 维矩阵, $\mathbf{R}^n$ 即 $\mathbf{R}^{n\times 1}$, $\mathbf{Z}_+$ 表示非负整数集, $\otimes$ 表示克罗内克积, vec 为矩阵的拉直运算, 把矩阵按照列的顺序一列接一列的组成一个长向量, trace 表示矩阵的迹, Id 表示恒等函数, $\circ$ 表示函数的复合运算, $f\circ g$ 表示函数 f 和 g 的复合函数, 即 $f\circ\ g(x)=f(g(x))$, $\lambda_{\max}$ ($\lambda_{\min}$) 表示矩阵的最大 (最小) 特征值, $|x|$ 表示向量 x 的欧几里得范数, $\|A\|$ 表示矩阵 A 诱导欧几里得范数, x^{T} 表示向量 x 的转置. $\|u\|$ 表示 $\sup_{k>0}|\,u\,(\,k\,)|$.

$\mathcal{K}$ 类函数[19]. 该类函数为一个严格递增连续函数 α: $\mathbf{R}_+\to\mathbf{R}_+$ 且 $\alpha(0)=0$, 其可以表示为 $\alpha\in\mathcal{K}$.

$\mathcal{K}_\infty$ 类函数[19]. 一个函数为 $\mathcal{K}$ 类函数, 当 $s\to\infty$ 时 $\alpha(s)\to\infty$, 那么该类函数是 $\mathcal{K}_\infty$ 类函数, 其可以表示为 $\alpha\in\mathcal{K}_\infty$.

$\mathcal{KL}$ 类函数[19]. 一个连续函数 β: $\mathbf{R}_+\times\mathbf{R}_+\to\mathbf{R}_+$. 如果对于每个特定的 $t\in\mathbf{R}_+$, $\beta(\cdot,t)$ 均是一个 $\mathcal{K}$ 类函数, 并且对于每个特定的 $s>0$, $\beta(s,\cdot)$ 递减并满足 $\lim_{t\to\infty}\beta(s,t)=0$, 那么就称 β 为 $\mathcal{KL}$ 类函

数, 并表示为 $\beta \in \mathcal{KL}$.

1 控制问题描述

1.1 离散时间部分线性系统被控对象

考虑一组离散时间部分线性系统:

$$\zeta(k+1) = g(\zeta(k), y(k), v(k)) \tag{1}$$

$$x(k+1) = Ax(k) + B[u(k)+ \Delta(\zeta(k), y(k), v(k))] + Dv(k) \tag{2}$$

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

$$y(k) = Cx(k) \tag{4}$$

$$r(k) = -Fv(k) \tag{5}$$

$$e(k) = y(k) - r(k) \tag{6}$$

式中, k 是描述系统运行轨迹的时间步骤, $x(k) \in \mathbf{R}^n$ 为系统的状态向量, $u(k) \in \mathbf{R}^p$ 为系统的输入向量, $\zeta(k) \in \mathbf{R}^p$, $v(k) \in \mathbf{R}^q$ 是外系统的状态向量, $y(k) \in \mathbf{R}^r$ 是系统的输出向量, $r(k) \in \mathbf{R}^r$ 是参考输入向量, $e(k) \in \mathbf{R}^r$ 是跟踪误差向量, $Dv(k) \in \mathbf{R}^n$ 是系统干扰向量, $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{r\times n}$, $D \in \mathbf{R}^{n\times q}$, $E \in \mathbf{R}^{q\times q}$, $F \in \mathbf{R}^{r\times q}$ 是系统矩阵, 其中 (A, B) 是可镇定的, (A, C) 是可观测的. $g(\zeta(k), y(k), v(k)) : \mathbf{R}^p \times \mathbf{R}^r \times \mathbf{R}^q \to \mathbf{R}^p$, $\Delta(\zeta(k), y(k), v(k)) : \mathbf{R}^p \times \mathbf{R}^r \times \mathbf{R}^q \to \mathbf{R}^m$ 是充分光滑的函数, 满足 $g(0, 0, 0) = 0$, $\Delta(0, 0, 0) = 0$. 该系统中 A、B、D、g 和 Δ 是未知的.

本文控制目标是: 对于离散时间系统 (1) ~ (6), 设计鲁棒最优控制器为 $u(k) = -K^*(x(k) - X^*v(k)) + U^*v(k) = -K^*x(k) - L^*v(k)$, 其中 $L^* = U^* + K^*X^*$, 使得 $\lim_{k\to\infty} e(k) = \lim_{k\to\infty} Cx(k) + Fv(k) = 0$, 那么系统在满足下述假设 1 ~ 5 条件下可完成干扰抑制和渐近跟踪. 其中 $X \in \mathbf{R}^{n\times q}$ 和 $U \in \mathbf{R}^{m\times q}$ 满足下面的线性调节器方程

$$\begin{aligned} XE &= AX + BU + D \\ 0 &= CX + F \end{aligned} \tag{7}$$

假设 1. E 的特征值在单位圆上且不重复[20].

假设 2. 存在一个充分光滑的函数 $\zeta(v)$, $\zeta(0) = 0$, 对于任意 $v \in \mathbf{R}^q$ 满足下面的方程[5, 7].

$$\begin{aligned} \zeta(Ev(k)) &= g(\zeta(v(k)), r(k), v(k)) \\ 0 &= \Delta(\zeta(v(k)), r(k), v(k)) \end{aligned} \tag{8}$$

假设 3.

$$\text{rank} \begin{bmatrix} A - \lambda I & B \\ C & 0 \end{bmatrix} = n + r, \quad \forall \lambda \in \sigma(E) \tag{9}$$

注 1. 假设 3 可保证对于任意的 D 和 F, 式 (7) 为调节器方程是可解的[5], 且解是唯一的[17].

根据式 (7) 和式 (8), 并令 $\bar{x}(k) = x(k) - Xv(k)$, $\bar{u}(k) = u(k) - Uv(k)$, $\bar{\zeta}(k) = \zeta(k) - \zeta(v(k))$, 可将原系统 (1) ~ (6) 写成如下的误差系统:

$$\bar{\zeta}(k+1) = \bar{g}(\bar{\zeta}(k), e(k), v(k)) \tag{10}$$

$$\bar{x}(k+1) = A\bar{x}(k) + B(\bar{u}(k)+ \bar{\Delta}(\bar{\zeta}(k), e(k), v(k))) \tag{11}$$

$$e(k) = C\bar{x}(k) \tag{12}$$

其中

$$\bar{g}(\bar{\zeta}(k), e(k), v(k)) = g(\zeta(k), y(k), v(k)) - g(\zeta(v(k)), r(k), v(k))$$

$$\bar{\Delta}(\bar{\zeta}(k), e(k), v(k)) = \Delta(\zeta(k), y(k), v(k)) - \Delta(\zeta(v(k)), r(k), v(k))$$

对于变换后的误差系统 (10) ~ (11) 中的 $\bar{\zeta}$ 子系统做如下假设, 相似的假设见文献 [7]:

假设 4. 存在 $\beta_s \in \mathcal{KL}$ 和 $\gamma_{s1}, \gamma_{s2} \in \mathcal{K}$ 使得对于任意的可测量且局部本质有界的输入 e, 任何初始条件 $\bar{\zeta}(0) = \bar{\zeta}_0$ 和任意的 v, $\bar{\zeta}(k)$ 满足:

$$|\bar{\zeta}(k)| \le \max\{\beta_s(|\bar{\zeta}(0)|, k), \gamma_{s1}(\|e_{[k-1]}\|), \gamma_{s2}(\|\bar{\Delta}_{[k-1]}\|)\}, \quad \forall k \in \mathbf{Z}_+$$

假设 5. 存在 $\beta_{\bar{\Delta}} \in \mathcal{KL}$ 和 $\gamma_e^{\bar{\Delta}} \in \mathcal{K}$, 使得对于任意的初始状态 $\bar{\zeta}(0) = \bar{\zeta}_0$ 和任意的可测量且局部本质有界的输入 e 和任意的 v, 使得下式成立:

$$|\bar{\Delta}(k)| \le \max\{\beta_{\bar{\Delta}}(|\bar{\zeta}(0)|, k), \gamma_e^{\bar{\Delta}}(\|e_{[k-1]}\|)\}, \quad \forall k \in \mathbf{Z}_+$$

注 2. 假设 4 使得 $\bar{\zeta}$ 子系统具有以 e 为输入, $\bar{\Delta}(\bar{\zeta}, e, v)$ 为输出的零偏差的强无界能观 (Strong unboundedness observability, SUO) 性质, 假设 5 使得 $\bar{\zeta}$ 子系统具有以 e 为输入, $\bar{\Delta}(\bar{\zeta}, e, v)$ 为输出的输入输出稳定 (Input-to-output stability, IOS) 性质.

下面将给出最优输出调节问题当中的两个规划问题.

1.2 输出调节问题中的两个规划问题

受文献 [7–8, 18] 启示, 对于最优输出调节问题的求解, 可拆分成两个规划问题, 分别为受约束的静态规划问题和动态规划问题. 通过解静态规划问题 1 可以确定输出调节器方程的解 X^*, U^*, 解动态规划问题 2 可以确定最优反馈控制增益 K^*, 则可得到最优控制器 $u^*(k) = -K^*(x(k) - X^*v(k)) + U^*v(k)$.

问题 1. 静态规划问题

通过解下面的静态规划问题确定线性调节器方

程的唯一解 (X, U)

$$\begin{cases} \min\limits_{(X,U)} \operatorname{tr}(X^{\mathrm{T}}QX + U^{\mathrm{T}}RU) \\ \text{s.t.} \\ \quad XE = AX + BU + D \\ \quad 0 = CX + F \end{cases} \tag{13}$$

式中, $Q = Q^{\mathrm{T}} > 0$, $R = R^{\mathrm{T}} > 0$. 式 (13) 有约束的规划问题等价于下面的形式:

$$\begin{cases} \min \left(\begin{bmatrix} \operatorname{vec}(X) \\ \operatorname{vec}(\mathrm{U}) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I_q \otimes Q & 0 \\ 0 & I_q \otimes R \end{bmatrix} \begin{bmatrix} \operatorname{vec}(X) \\ \operatorname{vec}(\mathrm{U}) \end{bmatrix} \right) \\ \text{s.t.} \\ \quad XE = AX + BU + D \\ \quad 0 = CX + F \end{cases} \tag{14}$$

下面先介绍当系统模型参数已知的情况下, 静态规划问题的解, 即是线性调节器方程的解, 并将静态规划问题 1 重新改写形式. 此部分为第二部分数据驱动求解静态规划问题做铺垫.

定义一个 Sylvester 映射, $\underline{A} : \mathbf{R}^{n\times q} \to \mathbf{R}_{n\times q}$:

$$\underline{A}(X) = XE - AX, \quad X \in \mathbf{R}_{n\times q} \tag{15}$$

选一个常数矩阵 $X_0 = 0_{n\times q}$ 和 $X_1 \in \mathbf{R}_{n\times q}$ 使得 $CX_1 + F = 0$. 选 $X_i \in \mathbf{R}_{n\times q}$, 其中 $i = 2, \cdots, h+1$, 其中 $h = q(n-r)$, 使得所有的 $\operatorname{vec}(X_i)$ 构成 $I_q \otimes C$ 的核, 其中 h 是 $I_q \otimes C$ 的零空间的维数, 即 $CX_i = 0$. 那么 X 可由 $X_1, \cdots, X_{h+1}$ 进行线性表示, 因此式 (8) 的通解[8] 可以进行如下描述:

$$X = X_1 + \sum_{i=2}^{h+1} \alpha_i X_i \tag{16}$$

通过式 (8)、式 (15) 和式 (16) 可得:

$$\underline{A}(X) = \underline{A}(X_1) + \sum_{i=2}^{h+1} \alpha_i \underline{A}(X_i) = BU + D \tag{17}$$

将式 (16) 和式 (17) 进行联立, 移项并展开, 将 Λ, ς 作为已知项, 并对其进行分块划分, χ 作为待求项, 可写为:

$$\Lambda\chi = \varsigma \tag{18}$$

其中

$$\Lambda = \begin{bmatrix} \operatorname{vec}(\underline{A}(X_2)) & \cdots & \operatorname{vec}(\underline{A}(X_2)) & 0 & -I_q \otimes B \\ \operatorname{vec}(X_2) & \cdots & \operatorname{vec}(X_{h+1}) & -I_{n\times q} & 0 \end{bmatrix} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

$$\chi = [\alpha_2 \quad \cdots \quad \alpha_{h+1} \quad \operatorname{vec}(X)^{\mathrm{T}} \quad \operatorname{vec}(U)^{\mathrm{T}}]^{\mathrm{T}}$$

$$\varsigma = \begin{bmatrix} \operatorname{vec}(-\underline{A}(X_1) + D) \\ \operatorname{vec}(X_1) \end{bmatrix} = \begin{bmatrix} \varsigma_1 \\ \varsigma_2 \end{bmatrix}$$

且 Λ_{21} 是非奇异矩阵.

将式 (18) 进行展开计算,并把 χ 中的调节器方程的解 $\operatorname{vec}(X)$ 和 $\operatorname{vec}(U)$ 分离出来, 可以得到式 (19).

定理 1. 通过解式 (19), 可得线性调节器方程的解 (X, U) :

$$\Pi \begin{bmatrix} \operatorname{vec}(X) \\ \operatorname{vec}(U) \end{bmatrix} = \Psi \tag{19}$$

式中, $\Pi = -\Lambda_{11}\Lambda_{21}^{-1}\Lambda_{22} + \Lambda_{21}$, $\Psi = -\Lambda_{11}\Lambda_{21}^{-1}\varsigma_2 + \varsigma_1$.

那么, 问题 1 可以重写为:

$$\begin{cases} \min \left(\begin{bmatrix} \operatorname{vec}(X) \\ \operatorname{vec}(U) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I_q \otimes Q & 0 \\ 0 & I_q \otimes R \end{bmatrix} \begin{bmatrix} \operatorname{vec}(X) \\ \operatorname{vec}(U) \end{bmatrix} \right) \\ \text{s.t.} \\ \quad \Pi \begin{bmatrix} \operatorname{vec}(X) \\ \operatorname{vec}(U) \end{bmatrix} = \Psi \end{cases} \tag{20}$$

问题 2. 动态规划问题

解决如下问题来确定最优反馈增益 K^*:

$$\begin{cases} \min V(k) = \sum\limits_{i=k}^{\infty} \bar{x}^{\mathrm{T}}(i)\bar{Q}\bar{x}(i) + \bar{u}^{T}(i)\bar{R}\bar{u}(i) \\ \text{s.t.} \\ \quad \bar{x}(k+1) = A\bar{x}(k) + B\bar{u}(k) \\ \quad e(k) = C\bar{x}(k) \end{cases} \tag{21}$$

式中, $\bar{Q} = \bar{Q}^{\mathrm{T}} > 0$, $\bar{R} = \bar{R}^{\mathrm{T}} > 0$, 且 $(A, \sqrt{\bar{Q}})$ 是可观测的.

当不考虑非线性不确定性时, 问题 2 是一个线性二次型最优调节器问题, 目标是设计一个状态反馈控制器 $\bar{u}(k) = -K\bar{x}(k)$ 使得 (11) 中系统的状态趋于 0, 同时使得规定的值函数 $V(k) = \bar{x}^{T}(k)P\bar{x}(k)$ 最小.

那么由线性最优控制理论, 对哈密顿函数求控制输入 $\bar{u}(k)$ 的导数, 得到最优反馈增益为 $K = (\bar{R} + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA$, 其中 P 是下面黎卡提方程的解:

$$\bar{Q} - P + A^{\mathrm{T}}PA - A^{\mathrm{T}}PB(\bar{R} + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA = 0 \tag{22}$$

式中, $P = P^{\mathrm{T}} > 0$. 求解黎卡提方程中的正定矩阵 P, 可以采用策略迭代 (Policy iteration, PI) 方法. 算法 1 的收敛性见文献 [21−22]. 算法 1 为算法 2 的推导做一个简单的铺垫.

算法 1. 策略迭代算法

1) 初始化: 选一个可镇定系统的初始控制策略 K^0, 迭代下面两个步骤, 直到第 j 步, P 收敛.

2) 策略评估: 用下式求解矩阵 P.

$$P^j = \bar{Q} + {K^j}^{\mathrm{T}}\bar{R}K^j + (A - BK^j)^{\mathrm{T}}P^j(A - BK^j) \tag{23}$$

3) 策略改进:

$$K^{j+1} = (\bar{R} + B^{\mathrm{T}}P^jB)^{-1}B^{\mathrm{T}}P^jA \tag{24}$$

4) 当

$$\|P^j - P^{j-1}\|_2 \le \varepsilon \tag{25}$$

时停止, 否则 $j \leftarrow j+1$ 返回 2). ε 是一个数值很小的正数.

注 3. 动态规划问题的求解是针对于线性系统, 即不考虑系统存在非线性不确定性时, 求得的最优反馈增益. 第 1.3 节对该最优反馈控制器对非线性不确定性是否存在鲁棒性, 即是否可以全局渐近镇定误差系统 (10) ~ (12) 进行说明.

1.3 系统最优输出调节问题的可解性

本节将原系统最优输出调节问题的可解性转化为误差系统的全局渐近稳定性, 通过提出两个定理进行说明. 定理 1 说明了最优输出调节控制器使得闭环误差系统是全局渐近稳定的, 定理 2 说明了原系统的最优输出调节问题是可解的.

定理 1. 在假设 1 ~ 5 下, 令 $\bar{Q} > (\gamma_x - 1)I_n$, $\bar{R} = I_m$, $0 < \gamma_x < \lambda_{\max}(P^*)$, 若满足:

$$\gamma_e^{\bar{\Delta}}(s) < {\gamma_{\bar{\Delta}}^e}^{-1}(s) \tag{26}$$

那么最优反馈控制器 $\bar{u}^*(k) = -K^*\bar{x}^*(k) = (\bar{R}+B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*A\bar{x}^*(k)$ 可以全局渐近镇定误差系统 (10) ~ (12).

证明. 取值函数 $V(\bar{x}(k)) = \bar{x}^{\mathrm{T}}(k)P^*\bar{x}(k)$, 值函数满足 $\alpha_1(|\bar{x}(k)|) < V(\bar{x}(k)) < \alpha_2(|\bar{x}(k)|)$, 其中 $\alpha_1(s) = \lambda_{\min}(P^*)(s)$, $\alpha_2(s) = \lambda_{\max}(P^*)(s)$.

对李雅普诺夫函数 $V(\bar{x}(k)) = \bar{x}^{\mathrm{T}}(k)P^*\bar{x}(k)$ 进行差分, 通过不等式进行缩放, 可得:

$$\begin{aligned}
&V(\bar{x}(k+1)) - V(\bar{x}(k)) = \\
&\quad \bar{x}^{\mathrm{T}}(A-BK)^{\mathrm{T}}P^*(A-BK)\bar{x}+ \\
&\quad 2\bar{x}^{\mathrm{T}}(A-BK)^{\mathrm{T}}P^*B\bar{\Delta} + \bar{\Delta}^{\mathrm{T}}B^{\mathrm{T}}P^*B\bar{\Delta} - \bar{x}^{\mathrm{T}}P^*\bar{x} = \\
&\quad -\bar{x}^{\mathrm{T}}(\bar{Q}+K^{\mathrm{T}}\bar{R}K)\bar{x} + 2\bar{x}^{\mathrm{T}}(A-BK)^{\mathrm{T}}P^*B\bar{\Delta}+ \\
&\quad \bar{\Delta}^{\mathrm{T}}B^{\mathrm{T}}P^*B\bar{\Delta} = \\
&\quad -\bar{x}^{\mathrm{T}}(\bar{Q}+A^{\mathrm{T}}P^*B((\bar{R}+B^{\mathrm{T}}P^*B)^{-1})^{\mathrm{T}} \\
&\quad (\bar{R}+B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*A)^{\mathrm{T}}\bar{x}+ \\
&\quad 2\bar{x}^{\mathrm{T}}(A-B(\bar{R}+B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*A)^{\mathrm{T}}P^*B\bar{\Delta}+ \\
&\quad \bar{\Delta}^{\mathrm{T}}B^{\mathrm{T}}P^*B\bar{\Delta} = -\bar{x}^{\mathrm{T}}\bar{Q}\bar{x} - \bar{x}^{\mathrm{T}}(\bar{A}\bar{\bar{R}}(\bar{\bar{R}})^{\mathrm{T}}\bar{A}^{\mathrm{T}})\bar{x}+ \\
&\quad 2\bar{x}^{\mathrm{T}}\bar{A}\bar{\Delta} - 2\bar{x}^{\mathrm{T}}\bar{A}\bar{\bar{R}}\bar{B}\bar{\Delta} + \bar{\Delta}^{\mathrm{T}}\bar{B}\bar{\Delta} = -\bar{x}^{\mathrm{T}}\bar{Q}\bar{x}- \\
&\quad |\bar{B}\bar{\Delta} + \bar{\bar{R}}^{\mathrm{T}}\bar{A}^{\mathrm{T}}\bar{x}|^2 + \bar{\Delta}^{\mathrm{T}}B^{\mathrm{T}}B\bar{\Delta} + 2\bar{x}^{\mathrm{T}}\bar{A}\bar{\Delta}+ \\
&\quad \bar{\Delta}^{\mathrm{T}}\bar{B}\bar{\Delta} \le -\bar{x}^{\mathrm{T}}\bar{Q}\bar{x} - |\bar{A}\bar{\Delta} - \bar{x}|^2 + \bar{\Delta}^{\mathrm{T}}\bar{A}^{\mathrm{T}}\bar{A}\bar{\Delta}+ \\
&\quad \bar{\Delta}^{\mathrm{T}}\bar{B}^{\mathrm{T}}\bar{B}\bar{\Delta} + \bar{\Delta}^{\mathrm{T}}\bar{B}\bar{\Delta} + \bar{x}^{\mathrm{T}}\bar{x} \le \\
&\quad -\bar{x}^{\mathrm{T}}(\bar{Q}-I)\bar{x} + \bar{\Delta}^{\mathrm{T}}(\bar{A}^{\mathrm{T}}\bar{A} + \bar{B}^{\mathrm{T}}\bar{B} + \bar{B})\bar{\Delta} \le \\
&\quad -\gamma_x|\bar{x}|^2 + \lambda_{\max}(\bar{\bar{A}})\|\bar{\Delta}\|^2 \le \\
&\quad -\frac{\gamma_x}{\lambda_{\max}(P^*)}V(k) + \lambda_{\max}(\bar{\bar{A}})\|\bar{\Delta}\|^2
\end{aligned} \tag{27}$$

式中, $\alpha = \gamma_x/\lambda_{\max}(P^*)$, $\sigma = \lambda_{\max}(\bar{\bar{A}})$, $\bar{A} = A^{\mathrm{T}}P^*B$, $\bar{B} = B^{\mathrm{T}}P^*B$, $\bar{\bar{R}} = ((\bar{R} + B^{\mathrm{T}}P^*B)^{-1})^{\mathrm{T}}$, $\bar{\bar{A}} = \bar{A}^{\mathrm{T}}\bar{A} + \bar{B}^{\mathrm{T}}\bar{B} + \bar{B}$. $\alpha \in \mathcal{K}_\infty$, $\sigma \in \mathcal{K}$. 为不丢失一般性[18], 需 $\mathrm{Id} - \alpha \in \mathcal{K}$, 故 $0 < \gamma_x < \lambda_{\max}(P^*)$ 由此而来.

由文献 [23] 可知, 如果有不等式 $V(\bar{x}(k+1)) - V(\bar{x}(k)) \le -\alpha(V(\bar{x}(k))) + \sigma(\|\bar{\Delta}\|)$, 那么就一个 $\rho \in \mathcal{K}_\infty$, $\mathrm{Id} - \rho \in \mathcal{K}$, 使得函数 $\alpha_1^{-1} \circ \alpha^{-1} \circ (\mathrm{Id} + \rho) \circ \sigma(\mathrm{s})$ 可以作为一个系统的输入状态稳定–增益函数 $\gamma_{\bar{x}}(s)$, 且存在一个 $\mathcal{KL}$ 类函数 β 使得下式成立:

$$|\bar{x}(k)| \le \beta_{\bar{x}}(\bar{x}(0), k) + \gamma_{\bar{x}}(\|\bar{\Delta}_{[k-1]}\|), \quad \forall k \in \mathbf{Z}_+ \tag{28}$$

注 4. $\gamma_{\bar{x}}(s)$ 是 $\bar{x}$ 子系统中, 以 $\bar{\Delta}$ 为输入, $\bar{x}$ 为状态的输入–状态增益函数.

并通过利用广义三角不等式[24]:

$$\begin{aligned}
&\max\{a, b\} \le a + b \le \\
&\max\{(\mathrm{Id}+\delta^{-1})(a), (\mathrm{Id}+\delta)(b)\}
\end{aligned}$$

对任意 $a, b > 0$ 和任意 $\delta \in \mathcal{K}_\infty$ 都成立. 那么可以将加型的不等式 (28) 写成如下的 max 型不等式:

$$|\bar{x}(k)| \le \{\beta_{\hat{\bar{x}}}(\bar{x}(0), k), \gamma_{\hat{\bar{x}}}(\|\bar{\Delta}_{[k-1]}\|)\}, \ \forall k \in \mathbf{Z}_+ \tag{29}$$

注 5. 为得到一个与 $\gamma_{\bar{x}}$ 十分接近的新的 $\gamma_{\hat{\bar{x}}}$, 可以找一个非常小的 δ, 这样有可能会使得新的 β 很大[25].

那么通过式 (12) 和式 (29), 自然可得:

$$\begin{aligned}
|e(k)| \le \{&\max\{|C|\beta_{\hat{\bar{x}}}(\bar{x}(0), k), |C|\gamma_{\hat{\bar{x}}}(\|\bar{\Delta}_{[k-1]}\|)\}, \\
&\forall k \in \mathbf{Z}_+
\end{aligned} \tag{30}$$

注 6. 式 (29) 说明 $\bar{x}$ 子系统具有以 $\bar{\Delta}(\bar{\zeta}, e, v)$ 为输入, e 为输出的输入状态稳定性质 (Input-to-state stability, ISS). 式 (30) 说明 $\bar{x}$ 子系统具有以 $\bar{\Delta}(\bar{\zeta}, e, v)$ 为输入, e 为输出的输入输出稳定性质. $\gamma_{\bar{\Delta}}^e = |C|\gamma_{\hat{\bar{x}}}$.

那么现在具有输入输出稳定和强无界能观性质的 $\bar{\zeta}$ 子系统是和具有输入输出稳定和强无界能观性质的 $\bar{x}$ 子系统, 在下面的小增益条件

$$\gamma_e^{\bar{\Delta}} \circ \gamma_{\bar{\Delta}}^e < \mathrm{Id} \tag{31}$$

成立时, 关联的误差系统在原点处全局渐近稳定.□

注 7. $\gamma_e^{\bar{\Delta}}$ 是 $\bar{\zeta}$ 子系统中的输入–输出增益, $\gamma_{\bar{\Delta}}^e$ 是 $\bar{x}$ 子系统中输入–输出增益. 当两个子系统都是强无界能观和输入输出稳定的, 且在输入输出稳定小增益条件成立下, 两个子系统的输出都趋于零, 那么由 $\bar{x}$ 子系统的输入状态稳定性质和 $\bar{\zeta}$ 子系统的零偏差强无界能观性质, 可以知道两个关联系统的状态也是趋于零的.

定理 2. 在定理 1 的条件下, 那么鲁棒最优控制器 $u^*(k) = -K^*(x(k) - X^*v(k)) + U^*v(k)$ 对于系统 (1) ~ (6) 的输出调节问题可解. □

证明. 通过定理 1 可知存在控制器使得误差系统在原点处全局渐近稳定, 所以 $\lim_{k\to\infty}\bar{\zeta}(k) = 0$, $\lim_{k\to\infty}\bar{x}^*(k)$=0, 那么 $\lim_{k\to\infty}e(k)$=$\lim_{k\to\infty}C\bar{x}^*(k) + (CX^* + F)v(k) = 0$, 即该系统的输出调节问题可解.

注 8. 最优控制器 $\bar{u}^*(k) = -K^*\bar{x}^*(k)$ 与最优控制器 $u^*(k) = -K^*(x(k) - X^*v(k)) + U^*v(k)$ 等价.

原系统最优输出调节问题的可解性得以证明后, 下部分将对该最优控制器进行学习. 第 2 节针对于具有未知系统模型参数的离散时间的部分线性系统, 用基于强化学习的数据驱动方法, 利用测量数据在线求解其最优输出调节问题.

2　数据驱动在线求解最优输出调节问题

强化学习中学习的方式分为离线策略学习算法和在线策略学习算法两种. 离线策略更新算法中的行为策略和目标策略不是同一策略, 行为策略用于产生数据, 目标策略则是被评估和提高的策略. 而在线策略算法则是行为与目标策略一致. 本文提出一个仅利用在线数据基于强化学习的离线策略的数据驱动方法, 用于求解离散时间部分线性系统的最优输出调节问题. 由于本文系统的模型参数是未知的, 首先求解动态规划问题求得最优反馈增益, 然后基于动态规划问题的解, 本文提出一种数据驱动方法, 在无法获取系统模型参数的情况下在线求解静态规划问题的解.

2.1　数据驱动求解动态优化问题

假设 $\Delta(k)$ 和 $v(k)$ 是可测的, X 可由 $X_1, \cdots, X_{h+1}$ 表示, 又 $\bar{x}(k) = x(k) - Xv(k)$, 现定义一个新的状态 $\bar{x}_i(k) = x(k) - X_iv(k)$, 其中 $i = 0, 1, 2, \cdots, h+1$, $X_0 = 0_{n\times q}$. 那么有:

$$\begin{aligned}\bar{x}_i(k+1) &= x(k+1) - X_iv(k+1) = \\ &Ax(k) + B(u(k) + \Delta(k)) + (D - X_iE)v(k) = \\ &A^j\bar{x}_i(k) + B(K^j\bar{x}_i(k) + w(k)) - \\ &(\underline{A}(X_i) - D)v(k)\end{aligned} \tag{32}$$

式中, $A^j = A - BK^j$, $w(k) = u(k) + \Delta(k)$.

写出 $k+1$ 时刻的值函数减去 k 时刻的值函数, 将式 (32) 代入, 可得:

$$\begin{aligned}&\bar{x}_i^{\mathrm{T}}(k+1)P^{j+1}\bar{x}_i(k+1) - \bar{x}_i^{\mathrm{T}}(k)P^{j+1}\bar{x}_i(k) = \\ &\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}A^j\bar{x}_i(k) + \\ &2\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}B(K^j\bar{x}_i(k) + w(k)) + \\ &(K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}B^{\mathrm{T}}P^{j+1}B(K^j\bar{x}_i(k) + w(k)) + \\ &2\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)v(k) + \\ &2(K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}B^{\mathrm{T}}P^{j+1}((\underline{A}(X_i) - D)v(k)) + \\ &v^{\mathrm{T}}(k)(\underline{A}(X_i) - D)^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)v(k) - \\ &\bar{x}_i^{\mathrm{T}}(k)P^{j+1}\bar{x}_i(k)\end{aligned} \tag{33}$$

用 $A^j = A - BK^j$ 代替 A^j, 将式 (23) 代入上式整理得到:

$$\begin{aligned}&\bar{x}_i^{\mathrm{T}}(k+1)P^{j+1}\bar{x}_i(k+1) - \bar{x}_i^{\mathrm{T}}(k)P^{j+1}\bar{x}_i(k) = \\ &\bar{x}_i^{\mathrm{T}}(k)(-\bar{Q} - {K^j}^{\mathrm{T}}\bar{R}K^j)\bar{x}_i(k) - \\ &2w^{\mathrm{T}}(k)B^{\mathrm{T}}P^{j+1}((\underline{A}(X_i) - D)v(k)) + \\ &(-K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}B^{\mathrm{T}}P^{j+1}B(K^j\bar{x}_i(k) + w(k)) + \\ &2\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}B(K^j\bar{x}_i(k) + w(k)) - \\ &2\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)v(k) + \\ &v^{\mathrm{T}}(k)(\underline{A}(X_i) - D)^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)v(k)\end{aligned} \tag{34}$$

为将上式的数据与矩阵参数进行分离, 将式 (34) 各项用克罗内克积和矩阵的拉直运算进行表示, 即根据 $a^{\mathrm{T}}Wb = (a^{\mathrm{T}}\otimes b^{\mathrm{T}})\mathrm{vec}(W)$, 可得上式对应的各式可以等价的表示如下:

$$\begin{aligned}&\bar{x}_i^{\mathrm{T}}(k)P^{j+1}\bar{x}_i(k) = (\bar{x}_i^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}(P^{j+1}) \\ &\bar{x}_i^{\mathrm{T}}(k)(-\bar{Q} - {K^j}^{\mathrm{T}}\bar{R}K^j)\bar{x}_i(k) = \\ &(\bar{x}_i^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}(-\bar{Q} - {K^j}^{\mathrm{T}}\bar{R}K^j) \\ &w^{\mathrm{T}}(k)B^{\mathrm{T}}P^{j+1}((\underline{A}(X_i) - D)v(k)) = \\ &(v^{\mathrm{T}}(k)\otimes w^{\mathrm{T}}(k))\mathrm{vec}(B^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)) \\ &(-K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}B^{\mathrm{T}}P^{j+1}B(K^j\bar{x}_i(k) + w(k)) = \\ &((-K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}\otimes(-K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}) \\ &\mathrm{vec}(B^{\mathrm{T}}P^{j+1}B) \\ &\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}B(K^j\bar{x}_i(k) + w(k)) = \\ &((K^j\bar{x}_i(k) + w(k))^{\mathrm{T}}\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}({A^j}^{\mathrm{T}}P^{j+1}B) \\ &\bar{x}_i^{\mathrm{T}}(k){A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)v(k) = \\ &(v^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}({A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)) \\ &v^{\mathrm{T}}(k)(\underline{A}(X_i) - D)^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D)v(k) = \\ &(v^{\mathrm{T}}(k)\otimes v^{\mathrm{T}}(k)) \\ &\mathrm{vec}((\underline{A}(X_i) - D)^{\mathrm{T}}P^{j+1}(\underline{A}(X_i) - D))\end{aligned} \tag{35}$$

因此, 式 (34) 可以用式 (35) 的形式表示为:

$$\begin{aligned}&((\bar{x}_i^{\mathrm{T}}(k+1)\otimes\bar{x}_i^{\mathrm{T}}(k+1)-(\bar{x}_i^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k)))\\&\quad\mathrm{vec}(P^{j+1})-2((K^j\bar{x}_i(k)+w(k))^{\mathrm{T}}\otimes\bar{x}_i^{\mathrm{T}}(k))\\&\quad\mathrm{vec}({A^j}^{\mathrm{T}}P^{j+1}B)-((K^j\bar{x}_i(k)+w(k))^{\mathrm{T}}\otimes\\&\quad(-K^j\bar{x}_i(k)+w(k))^{\mathrm{T}})\mathrm{vec}(B^{\mathrm{T}}P^{j+1}B)-\\&\quad(v^{\mathrm{T}}(k)\otimes v^{\mathrm{T}}(k))\\&\quad\mathrm{vec}((\underline{A}(X_i)-D)^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D))+\\&\quad2(v^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}({A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D))+\\&\quad2(v^{\mathrm{T}}(k)\otimes w^{\mathrm{T}}(k))\mathrm{vec}(B^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D))=\\&\quad(\bar{x}_i^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}(-\bar{Q}-{K^j}^{\mathrm{T}}\bar{R}K^j)\end{aligned}\tag{36}$$

为了对参数矩阵进行学习, 将式 (36) 写成式 (41) 的形式, 则需定义待求的参数矩阵如式 (37) 和数据组 (38) 和 (39) 如下, 式 (38) 收集的是式 (36) 中等式右边的 t 组数据组成数据向量 $\varphi_i^j(k)$, 式 (39) 收集的是式 (36) 中等式左边的 t 组数据组成数据矩阵 $\Psi_i^j(k)$.

$$\begin{aligned}L_1^{j+1}&={A^j}^{\mathrm{T}}P^{j+1}B\\L_2^{j+1}&=B^{\mathrm{T}}P^{j+1}B\\L_{3i}^{j+1}&=(\underline{A}(X_i)-D)^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D)\\L_{4i}^{j+1}&={A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D)\\L_{5i}^{j+1}&=B^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D)\end{aligned}\tag{37}$$

$$\varphi_i^j(k)=\begin{bmatrix}(\bar{x}_i^{\mathrm{T}}(k)\otimes\bar{x}_i^{\mathrm{T}}(k))\mathrm{vec}(-\bar{Q}-{K^j}^{\mathrm{T}}\bar{R}K^j)\\(\bar{x}_i^{\mathrm{T}}(k+1)\otimes\bar{x}_i^{\mathrm{T}}(k+1))\mathrm{vec}(-\bar{Q}-{K^j}^{\mathrm{T}}\bar{R}K^j)\\\vdots\\(\bar{x}_i^{\mathrm{T}}(k+t)\otimes\bar{x}_i^{\mathrm{T}}(k+t))\mathrm{vec}(-\bar{Q}-{K^j}^{\mathrm{T}}\bar{R}K^j)\end{bmatrix}\tag{38}$$

$$\Psi_i^j(k)=\begin{bmatrix}\Phi_{11}&\Phi_{12}&\cdots&\Phi_{16}\\\Phi_{21}&\Phi_{22}&\cdots&\Phi_{26}\\\vdots&\vdots&\ddots&\vdots\\\Phi_{t1}&\Phi_{t2}&\cdots&\Phi_{t6}\end{bmatrix}\tag{39}$$

其中

$$\begin{aligned}\Phi_{l1}&=(\bar{x}_i^{\mathrm{T}}(k+l+1)\otimes\bar{x}_i^{\mathrm{T}}(k+l+1))-\\&\quad(\bar{x}_i^{\mathrm{T}}(k+l)\otimes\bar{x}_i^{\mathrm{T}}(k+l))\\\Phi_{l2}&=-2((K^j\bar{x}_i(k+l)+w(k+l))^{\mathrm{T}}\otimes\bar{x}_i^{\mathrm{T}}(k+l))\\\Phi_{l3}&=-((K^j\bar{x}_i(k+l)+w(k+l))^{\mathrm{T}}\otimes\\&\quad(-K^j\bar{x}_i(k+l)+w(k+l))^{\mathrm{T}})\\\Phi_{l4}&=-(v^{\mathrm{T}}(k+l)\otimes v^{\mathrm{T}}(k+l))\end{aligned}$$

$$\begin{cases}\Phi_{l5}=2(v^{\mathrm{T}}(k+l)\otimes\bar{x}_i^{\mathrm{T}}(k+l))\\\Phi_{l6}=2(v^{\mathrm{T}}(k+l)\otimes w^{\mathrm{T}}(k+l))\end{cases}\tag{40}$$

并且应满足 $t\geq t_0$, $t_0=((n\times(n+1)/2))+((m\times(m+1)/2))+((q\times(q+1)/2))+n\times m+n\times q+m\times q-1$.

那么式 (33) 可以由式 (36) ~ (39) 表示为:

$$\begin{aligned}&\Psi_i^j(k)[\mathrm{vec}(P^{j+1})^{\mathrm{T}},\mathrm{vec}(L_1^{j+1})^{\mathrm{T}},\mathrm{vec}(L_2^{j+1})^{\mathrm{T}},\\&\quad\mathrm{vec}(L_{3i}^{j+1})^{\mathrm{T}},\mathrm{vec}(L_{4i}^{j+1})^{\mathrm{T}},\mathrm{vec}(L_{5i}^{j+1})^{\mathrm{T}}]^{\mathrm{T}}=\varphi_i^j(k)\end{aligned}\tag{41}$$

式 (41) 可以用最小二乘法进行求解:

$$\begin{aligned}&[\mathrm{vec}(P^{j+1})^{\mathrm{T}},\mathrm{vec}(L_1^{j+1})^{\mathrm{T}},\mathrm{vec}(L_2^{j+1})^{\mathrm{T}},\\&\quad\mathrm{vec}(L_{3i}^{j+1})^{\mathrm{T}},\mathrm{vec}(L_{4i}^{j+1})^{\mathrm{T}},\mathrm{vec}(L_{5i}^{j+1})^{\mathrm{T}}]^{\mathrm{T}}=\\&\quad(\Psi_i^{j\mathrm{T}}(k)\Psi_i^j(k))^{-1}\Psi_i^j(k)\varphi_i^j(k)\end{aligned}\tag{42}$$

由此迭代的反馈增益矩阵可以表示为:

$$K^{j+1}=(R+L_2^{j+1})^{-1}(L_1^{j+1})^{\mathrm{T}}\tag{43}$$

通过多次的迭代学习, 可得到近似的最优反馈增益矩阵 K^*.

注 9. 式 (41) 中有 t_0 个未知数, 因此至少需要 t_0 组数据对方程进行求解, 且如果 $\Psi_i^j(k)$ 列满秩时, 式 (41) 的解是唯一的.

2.2 数据驱动求解静态优化问题

前面已经介绍了当模型参数已知时, 受约束的静态规划问题应如何求解, 并将原静态规划问题 1 的形式重新改写. 在此基础上, 下面提出数据驱动的拉格朗日乘子法来求解式 (20) 这个受约束的静态规划问题. 该方法无需知道系统的模型参数, 仅使用测量的数据.

$$\begin{aligned}\min J=&\left(\begin{bmatrix}\mathrm{vec}(X)\\\mathrm{vec}(U)\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}I_q\otimes Q&0\\0&I_q\otimes R\end{bmatrix}\right.\\&\left.\begin{bmatrix}\mathrm{vec}(X)\\\mathrm{vec}(U)\end{bmatrix}\right)+\lambda^{\mathrm{T}}\mathrm{vec}(\Pi\begin{bmatrix}\mathrm{vec}(X)\\\mathrm{vec}(U)\end{bmatrix}-\Psi)\end{aligned}\tag{44}$$

为避免需要知道系统准确的模型参数, 根据动态规划问题的解来求得静态规划问题的解. 通过解动态规划问题可以求得 L_{4i}^{j+1}即 ${A^j}^{\mathrm{T}}P^{j+1}(\underline{A}(X_i)-D)$, 定义如下:

$$S(X_i)=\underline{A}(X_i)-D\tag{45}$$

$$\bar{S}(X_i)=A^{\mathrm{T}}P^{j+1}S(X_i)\tag{46}$$

$$\bar{\underline{A}}(X_i)=A^{\mathrm{T}}P^{j+1}\underline{A}(X_i)\tag{47}$$

其中

$$\bar{S}(X_i) = A^{\mathrm{T}}P^{j+1}S(X_i) = L_{4i}^{j+1}$$

$$\bar{S}(X_0) = A^{\mathrm{T}}P^{j+1}D = L_{40}^{j+1}$$

那么则有:

$$\underline{\bar{A}}(X_i) = A^{\mathrm{T}}P^{j+1}(S(X_i) - S(X_0)) = L_{4i}^{j+1} - L_{40}^{j+1} \tag{48}$$

由于无法直接求得 $BU + D$, 而通过解动态规划问题可得到 L_{4i}^{j+1}, L_{40}^{j+1}, 因此定义式 (48) 即 $A^{\mathrm{T}}P^{j+1}(BU + D)$, 那么式 (17) 则变形如下:

$$\underline{\bar{A}}(X) = \underline{\bar{A}}(X_1) + \sum_{i=2}^{h+1}\alpha_i\underline{\bar{A}}(X_i) = A^{\mathrm{T}}P^{j+1}(BU + D) \tag{49}$$

因此, 式 (19) 应重写如下:

$$\bar{\Pi}\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix} = \bar{\Psi} \tag{50}$$

式中, $\bar{\Pi} = -\bar{\Lambda}_{11}\bar{\Lambda}_{21}^{-1}\bar{\Lambda}_{22} + \bar{\Lambda}_{21}$, $\bar{\Psi} = -\bar{\Lambda}_{11}\bar{\Lambda}_{21}^{-1}\bar{\varsigma}_2 + \bar{\varsigma}_1$. 并且

$$\bar{\Lambda} = \begin{bmatrix}\mathrm{vec}(\underline{\bar{A}}(X_2)) & \cdots & \mathrm{vec}(\underline{\bar{A}}(X_{h+1}))\\ \mathrm{vec}(X_2) & \cdots & \mathrm{vec}(X_{h+1})\end{bmatrix}\begin{bmatrix}0 & -I_q\otimes(A^TP^{j+1}B)\\ -I_{n\times q} & 0\end{bmatrix} = \begin{bmatrix}\mathrm{vec}(L_{42}^{j+1} - L_{40}^{j+1}) & \cdots & \mathrm{vec}(L_{4(m+1)}^{j+1} - L_{40}^{j+1})\\ \mathrm{vec}(X_2) & \cdots & \mathrm{vec}(X_{m+1})\end{bmatrix}\begin{bmatrix}0 & -I_q\otimes L_1^{j+1}\\ -I_{n\times q} & 0\end{bmatrix} = \begin{bmatrix}\bar{\Lambda}_{11} & \bar{\Lambda}_{12}\\ \bar{\Lambda}_{21} & \bar{\Lambda}_{22}\end{bmatrix}$$

$$\bar{\varsigma} = \begin{bmatrix}\mathrm{vec}(-\underline{\bar{A}}(X_1) - \bar{S}(X_0))\\ \mathrm{vec}(X_1)\end{bmatrix} = \begin{bmatrix}\mathrm{vec}(-L_{41}^{j+1})\\ \mathrm{vec}(X_1)\end{bmatrix} = \begin{bmatrix}\bar{\varsigma}_1\\ \bar{\varsigma}_2\end{bmatrix}$$

那么受约束的静态规划问题 (20) 可重写为:

$$\min\ J = \left(\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}I_q\otimes Q & 0\\ 0 & I_q\otimes R\end{bmatrix}\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}\right) + \lambda^{\mathrm{T}}((I_q\otimes\bar{\Pi})\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix} - \mathrm{vec}(\bar{\Psi})) \tag{51}$$

对 J 求 $\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}$ 的偏导数, 即可求得静态规划问题的解 (X^*, U^*).

$$\frac{\partial J}{\partial\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix}} = 2\begin{bmatrix}I_q\otimes Q & 0\\ 0 & I_q\otimes R\end{bmatrix}\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix} + \lambda^{\mathrm{T}}(I_q\otimes\bar{\Pi}) = 0$$

$$\frac{\partial J}{\partial\lambda^{\mathrm{T}}} = (I_q\otimes\bar{\Pi})\begin{bmatrix}\mathrm{vec}(X)\\ \mathrm{vec}(U)\end{bmatrix} - \mathrm{vec}(\bar{\Psi}) = 0 \tag{52}$$

算法 2. 数据驱动离线策略更新算法

1) 迭代求解最优反馈增益: 选一个初始的稳定的反馈增益 K^0. 选择矩阵满足 $\bar{Q} > (\gamma_x - 1)I_n$, $\bar{R} = I_m$, $0 < \gamma_x < \lambda_{\max}(P^0)$, 并使得小增益定理条件成立. 并且计算矩阵 X_0, $X_1, \cdots, X_{h+1}$. 用 $u(k) = -K^0x(k) + \xi(k)$ 作为控制输入[18], 其中 $\xi(k)$ 为探测噪声. 令 $i = 0$, $j = 0$.

2) 策略评估: 解式 (42) 可得:

P^{j+1}, L_1^{j+1}, L_2^{j+1}, L_{3i}^{j+1}, L_{4i}^{j+1}, L_{5i}^{j+1}

3) 策略改进:

$K^{j+1} = (R + L_2^{j+1})^{-1}(L_1^{j+1})^{\mathrm{T}}$

4) 令 $j = j + 1$ 直到 $\|K^{j+1} - K^{j+}\|_2 \le \varepsilon$, ε 是一个数值很小的正数.

找输出调节器方程的最优解:

5) 令 $j = j^*$, $i \leftarrow i + 1$, 解 L_{4i}^{j+1} 直到 $i = h + 1$. 然后通过解式 (52) 找到解 (X^*, U^*).

6) 令 $u^*(k) = -K^*(x(k) - X^*v(k)) + U^*v(k)$.

注 10. 在算法 2 的控制输入中加入探测噪声不影响参数的学习效果.

定理 3. 给一个初始的可镇定系统的反馈增益 K^0, 若 $\Psi_i^j(k)$ 是列满秩的, 那么有 $\lim_{j\to\infty}P^j = P^*$, $\lim_{j\to\infty}K^j = K^*$.

证明. 给一个稳定的 K^j, 如果 $P^j = P^{j^{\mathrm{T}}}$ 是式 (23) 的解, K^{j+1} 是由式 (24) 决定的. 通过式 (33), 可知矩阵 P^{j+1}, L_1^{j+1}, L_2^{j+1} 满足式 (42). 当 $\Psi_i^j(k)$ 列满秩条件成立时, 矩阵 P^j、L_1^j、L_2^j、K^j 是唯一的, 并且又因为算法 1 具有收敛性, 即 $\lim_{j\to\infty}P^j = P^*$, $\lim_{j\to\infty}K^j = K^*$. 那么算法 2 中的 P^j、K^j 具有收敛性. □

3　仿真实验

本节首先建立一个仿真实验, 来说明本文方法的有效性; 然后进行对比实验, 用本文方法与对比方法进行仿真实验, 用评价指标结果说明本文方法的优越性.

3.1　仿真实验参数选择

考虑下面这个离散时间的部分线性系统:

$$x(k+1)=\begin{bmatrix}-1 & 2\\ 2.2 & 1.7\end{bmatrix}x(k)+\begin{bmatrix}-2\\ 1.6\end{bmatrix}\times$$
$$(u(k)+v_2(k)\zeta(k))+\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}v(k)$$
$$\zeta(k+1)=0.1e(k)\zeta(k)$$
$$v(k+1)=\begin{bmatrix}\cos(0.1) & \sin(0.1)\\ -\sin(0.1) & \cos(0.1)\end{bmatrix}v(k)$$
$$e(k)=[1\quad 0]\,x(k)+[0\quad 1]\,v(k)$$

此例中, $\zeta(v(k))=0$ 满足假设 2. 当增益函数 $\gamma_e^{\bar{\Delta}}(s)=0.4s^2$, 若 $\gamma_{\bar{\Delta}}^e(s)<\sqrt{2.5}s^{1/2}$, 那么关联的误差系统就可以认为在原点全局渐近稳定. 选择初始策略为 $K^0=[-0.3\quad 1.1]$ 和 $L^0=[0\quad 0]$. 在仿真中选择探测噪声为随机噪声, 并且对于 $i=0, 1, 2, 3$, 选择矩阵 X_i 为:

$$X_0=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix},\qquad X_1=\begin{bmatrix}0 & -1\\ 0 & 0\end{bmatrix}$$
$$X_2=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix},\qquad X_3=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}$$

对于静态规划问题 1 选择权重矩阵 $Q=5I_2$ 和 $R=1$, 对于动态规划问题 2 选择加权矩阵 $\bar{Q}=3I_2$ 和 $\bar{R}=1$. 通过计算得调节器方程解为 $X=[0\quad -1;-1.1389\quad -2.997]$, $U=[0.6888\quad 1.9995]$, 通过解黎卡提方程得最优的 P^*, $P^*=[35.8976\quad 0.7433;0.7433\quad 4.0401]$和最优策略 $K^*=[-0.3475\quad 0.9987]$, 那么就可计算最优 L^*, $L^*=U^*+K^*X^*=[-0.4486\quad -0.6462]$.

3.2 仿真结果

在仿真实验中, 算法 2 经过迭代学习 4 次收敛, 得到 $P^{j+1}=[35.8976\quad 0.7433;0.7433\quad 4.0401]$ 和增益 $K^{j+1}=[-0.3475\quad 0.9987]$. 学到最优增益后找调节器方程最优解为 $X=[4.281\times 10^{-17}\quad -1;-1.139\quad -2.997]$ 和 $U=[0.6888\quad 1.9995]$. 从而得到 $L=[-0.4486\quad -0.6461]$.

仿真结果见图 1 ~ 5. 图 1 给出了算法 2 的系统输出、参考输入和跟踪误差, 图 2 给出了控制输入. 由图 1 可知, 鲁棒最优输出调节控制器在由如图 3 系统干扰和存在非线性不确定的情况下, 仍可使得 $y(k)$ 跟踪参考输入 $r(k)$. 图 4 给出了在学习阶段 P 和 K 收敛到最优值的收敛情况, 由图 4 可知, 通过 4 次的迭代学习就可以求出最优的 P 和 K. 图 5 给出了误差系统的状态, 图 5 说明了误差系统在原点处是全局渐近稳定的, 同时也表明闭环系统的稳定性. 在仿真结果中, 跟踪误差从 100 步之后明显减小; 从第 120 步起, 跟踪误差的最大数量级为 10^{-9}, 控制输入中存在的动态非线性不确定性的大小从第 10 步起的最大数量级为 10^{-9}, 说明跟踪效果好, 且对于动态的非线性不确定性有良好的鲁棒性. 仿真结果表明, 本文算法在模型参数未知、存在干扰和输入中存在非线性不确定情况下, 只利用系统数据, 就可以实现具有鲁棒性的最优输出调节控制.

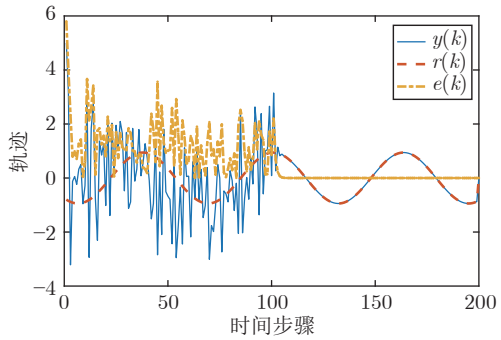


图 1 系统输出与参考轨迹及跟踪误差

Fig. 1 Trajectories of system output and reference and tracking error

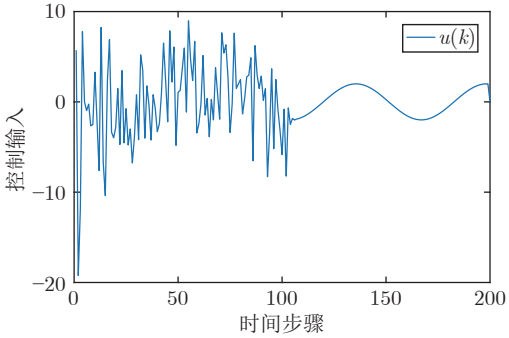


图 2 控制输入轨迹

Fig. 2 The control input trajectory

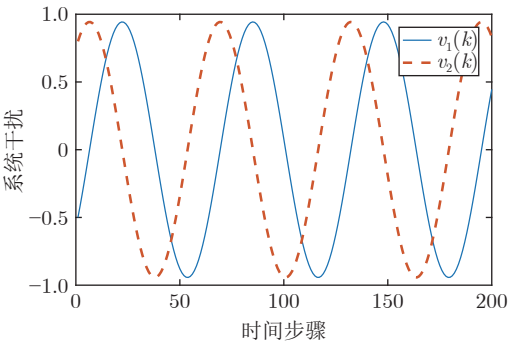


图 3 系统干扰

Fig. 3 The disturbance of system

3.3 对比实验

对比实验 1 采用本文提出的鲁棒最优输出调节

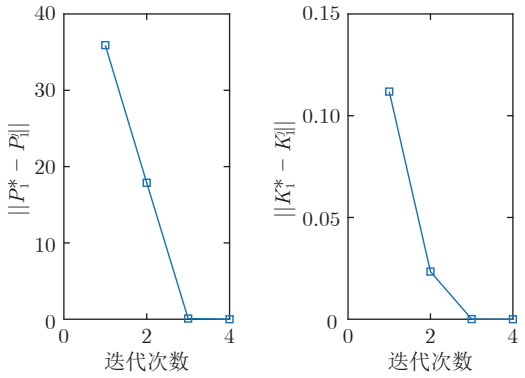


图 4　学习阶段 P 和 K 的收敛情况

Fig. 4　The convergence of P, K during learning phase

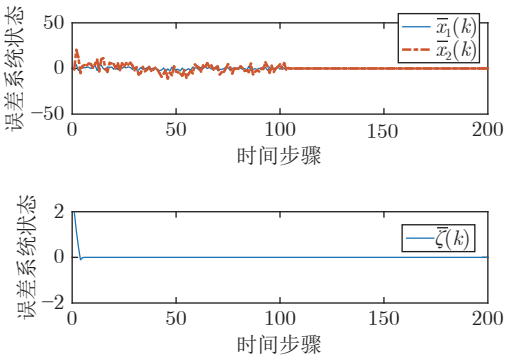


图 5　误差系统状态轨迹

Fig. 5　The error system state trajectory

的方法来跟踪参考信号, 且满足本文的假设条件. 对比实验 2 是文献 [12] 的方法, 在模型参数未知时采用 Q-学习的方法解决线性最优二次跟踪问题来跟踪参考信号. 2 个对比实验的未知模型参数和参考信号相同, 不同的是对比实验 1 还在控制输入中加入了非线性不确定性. 对比实验仿真结果见图 6 ~ 7.

对比实验 1 模型为:

$$x(k+1) = \begin{bmatrix} -1 & 2 \\ 2.2 & 1.7 \end{bmatrix} x(k) + \begin{bmatrix} -2 \\ 1.6 \end{bmatrix} \times (u(k) + v(k)\zeta(k))$$

$$\zeta(k+1) = 0.01e(k)\zeta(k)$$

$$v(k+1) = -v(k)$$

$$y(k) = [1 \quad 2]\, x(k)$$

对比实验 2 模型为:

$$x(k+1) = \begin{bmatrix} -1 & 2 \\ 2.2 & 1.7 \end{bmatrix} x(k) + \begin{bmatrix} -2 \\ 1.6 \end{bmatrix} u(k)$$

$$y(k) = [1 \quad 2]\, x(k)$$

本文用绝对误差积分 (Integral absolute error, IAE) 和均方根误差 (Root mean square error, RMSE) 两个指标[18, 26–29] 来评价本仿真实验的控制效果, 结果见表 1.

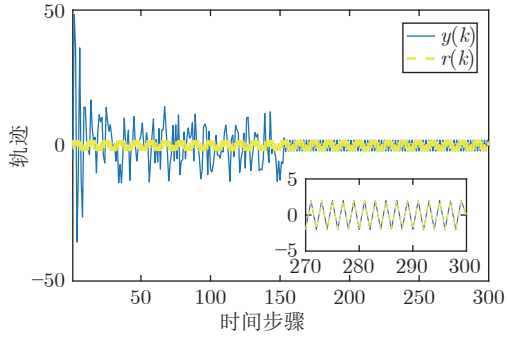


图 6　对比实验 1 仿真结果图

Fig. 6　The result of comparison experiment 1

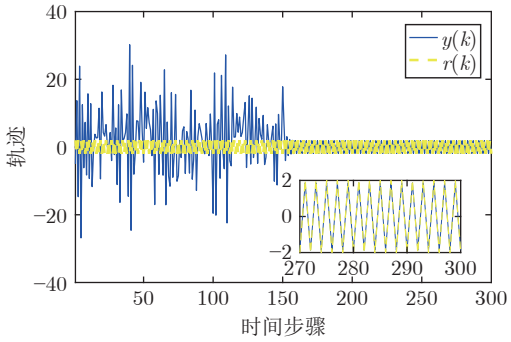


图 7　对比实验 2 仿真结果图

Fig. 7　The result of comparison experiment 2

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

$$RMSE_y = \sqrt{\frac{1}{k^*} \sum_{k=1}^{k^*} |w(k) - y(k)|^2}$$

表 1　对比实验评价指标

Table 1　Performance index of comparison experiment

$220 < k < 280$	IAE	RMSE
本文方法	1.8330×10^{-6}	3.6653×10^{-8}
对比方法	8.2293	0.1349

由图 6 ~ 7 可知, 对比实验 1 和 2 都能较好地跟踪设定值. 对比实验 1 相较于对比实验 2 还增加了非线性不确定性, 又从表 1 可知, 对比实验 1 的跟踪性能指标较对比实验 2 更好, 这也说明了本文提出算法的优越性.

4 结束语

本文提出一个基于强化学习的数据驱动算法, 用于解具有未知模型参数的离散时间部分线性系统的最优输出调节问题. 首先将原系统的输出调节问题的可解性转化为误差系统的全局渐近稳定问题, 给出了原问题的可解性说明; 然后在未知系统模型参数的条件下, 利用在线数据利用基于强化学习的数据驱动的离线策略算法求解最优反馈控制律, 并给出该算法的收敛性说明. 该控制律可以完成系统的干扰抑制和渐近跟踪且对于系统中存在的非线性不确定性存在鲁棒性. 仿真结果验证了本文方法的有效性, 通过对比实验和性能指标的比较, 说明了本文所提方法的优越性. 与跟踪问题相比, 本文方法不仅可以实现跟踪, 当系统本身存在干扰时, 同时可以抑制干扰达到闭环系统的稳定性. 本文方法与完全线性系统的输出调节问题相比, 对输入中存在的动态非线性不确定性存在鲁棒性. 本文将数据驱动的强化学习方法和小增益原理进行结合, 该方法可实现鲁棒强化学习, 从而也为更多控制问题的解决提供了思路.

References

1 Francis B A. The linear multivariable regulator problem. *SIAM Journal on Control Optimization*, 1977, **15**(3): 486−505

2 Davison E, Goldenberg A. Robust control of a general servomechanism problem: The servo compensator. *Automatica*, 1975, **11**(5): 461−471

3 Davison E. The robust control of a servomechanism problem for linear time-invariant multivariable systems. *IEEE Transactions on Automatic Control*, 1976, **1**(1): 25−34

4 Sontag E D. Adaptation and regulation with signal detection implies internal model. *System. & Control Letters*, 2003, **50**(2): 119−126

5 Huang J. *Nonlinear Output Regulation: Theory and Applications*. Philadelphia: Society for Industrial and Applied Mathematics, 2004.

6 Saberi A, Stoorvogel A A, Sannuti P, Shi G Y. On optimal output regulation for linear systems. *International Journal of Control*, 2003, **76**(4): 319−333

7 Gao W N, Jiang Z P. Global optimal output regulation of partially linear systems via robust adaptive dynamic programming. *IFAC-Papers OnLine*, 2015, **48**(11): 742−747

8 Gao W N, Jiang Z P. Adaptive dynamics programming and adptive optimal output regulation of linear systems. *IEEE Transactions on Automatic Control*, 2016, **61**(12): 4164−4169

9 Kiumarsi B, Vamvoudakis K G, Modares H, Lewis F L. Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(6): 2042−2062

10 Li Zhen, Fan Jia-Lu, Jiang Yi, Chai Tian-You. A model-free $H\infty$ method based on off-policy with output data feedback. *Acta Automatica Sinic*, 2021, **47**(9): 2182−2193
(李臻, 范家璐, 姜艺, 柴天佑. 一种基于Off-policy的无模型输出数据反馈$H\infty$控制方法. 自动化学报, 2021, **47**(9): 2182−2193)

11 Jiang Yi. Research on Data-driven Operational Optimization Control Approach for Complex Industrial Processes[Ph.D. dissertation], Northeastern University, China, 2020
(姜艺. 数据驱动的复杂工业过程运行优化控制方法研究[博士论文], 东北大学, 中国, 2020)

12 Kiumarsi B, Lewis F L, Modares H, Karimpour A, Naghibi M B. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica*, 2014, **50**(4): 1167−1175

13 Kiumarsi B, Lewis F L, Naghibi M B, Karimpour A. Optimal tracking control of unknown discrete-time linear systems using input-output measured data. *IEEE Transactions on Cybernetics*, 2015, **4**(12): 2770−2779

14 Kiumarsi B, Lewis F L. Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(1): 140−151

15 Kiumarsi B, Lewis F L, Jiang Z P. $H\infty$ control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, 2017, **78**: 144−152

16 Modares H, Lewis F L, Jiang Z P. $H\infty$ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(10): 2550−2562

17 Jiang Y, Fan J L, Chai T Y, Lewis F L, Li J N. Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(10): 4607−4620

18 Jiang Y, Kiumarsi B, Fan J L, Chai T Y, Li J N, Lewis F L. Optimal output regulation of linear discrete-time system with unknow dynamics using reinforcement learning. *IEEE Transactions on Cybernetics*, 2020, **50**(4): 3147−3156

19 Khalil H K, Grizzle J W. *Nonlinear Systems*. Upper Saddle River: Prentice hall, 2002.

20 Lan W Y, Huang J. Robust output regulation for discrete-time nonlinear systems. *International Journal of Robust and Nonlinear Control*, 2005, **15**(2): 63−81

21 Hewer G. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 1971, **16**(4): 382−384

22 Werbos P J. Neural network for control and system identification. In: Proceedings of the 28th IEEE Conference on Decision and Control. Tampa, USA: 1989, 260−265

23 Jiang Z P, Wang Y. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 2001, **37**: 857−869

24 Jiang Z P, Teel A R, Praly L. Small-gain theorem for ISS systems and applications. *Mathematics of Control Signals and Systems*, 1994, **7**(2): 95−120

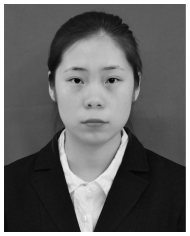
25 Liu Teng-Fei, Jiang Zhong-Ping. *Nonlinear Control Under Information Constraints*, Beijing: Science Press, 2018.
(刘腾飞, 姜钟平. 信息约束下的非线性控制, 北京: 科学出版社, 2018.)

26 Jiang Y, Fan J L, Chai T Y, Lewis F L. Dual-rate operational optimal control for flotation industrial process with unknown operational model. *IEEE Transactions on Industrial Electronics*, 2019, **66**(6): 4587−4599

27 Jiang Y, Fan J L, Chai T Y, Li J N, Lewis F L. Data driven flotation industrial process operational optimal control based on reinforcement learning. *IEEE Transactions on Industrial Informatics*, 2018, **66**(5): 1974−1989

28 Wu Qian, Fan Jia-Lu, Jiang Yi, Chai Tian-You. Data-driven dual-rate control for mixed separation thickening process in a wireless network environment. *Acta Automatica Sinica*, 2019, **45**(6): 1128−1141
(吴倩, 范家璐, 姜艺, 柴天佑. 无线网络环境下数据驱动混合选别浓密过程双率控制方法. 自动化学报, 2019, **45**(6): 1128−1141)

29 Jiang Yi, Fan Jia-Lu, Jia Yao, Chai Tian-You. Data-driven flotation process operational feedback decoupling control. *Acta Automatica Sinica*, 2019, **45**(4): 759−770
(姜艺, 范家璐, 贾瑶, 柴天佑. 数据驱动的浮选过程运行反馈解耦控制方法. 自动化学报, 2019, **45**(4): 759−770)



庞文砚 东北大学流程工业综合自动化国家重点实验室硕士研究生. 主要研究方向为工业过程运行控制和强化学习. E-mail: pangwy799@163.com
(**PANG Wen-Yan** Master student at the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. Her research interest covers industrial process operational control and reinforcement learning.)

范家璐 东北大学流程工业综合自动化国家重点实验室副教授. 2011 年获浙江大学博士学位. 主要研究方向为工业过程运行控制, 工业无线传感器网络与强化学习. 本文通信作者.
E-mail: jlfan@mail.neu.edu.cn
(**FAN Jia-Lu** Associate professor at the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. She received her Ph.D. degree from Zhejiang University in 2011. Her research interest covers industrial process operational control, industrial wireless sensor networks and reinforcement learning. Corresponding author of this paper.)

姜 艺 中国香港城市大学博士后. 2020 年获东北大学控制理论与控制工程专业博士学位. 主要研究方向为工业过程运行控制, 网络控制, 自适应动态规划和强化学习.
E-mail: yjian22@cityu.edu.hk
(**JIANG Yi** Postdoctor at City University of Hong Kong, China. He received his Ph.D. degree in control theory and engineering from Northeastern University in 2020. His research interest covers industrial process operational control, networked control, adaptive dynamic programming and reinforcement learning.)

LEWIS Frank Leroy 德克萨斯大学阿灵顿分校教授. 主要研究方向为反馈控制, 强化学习, 智能系统, 协同控制系统和非线性系统.
E-mail: lewis@uta.edu
(**LEWIS Frank Leroy** Professor at University of Texas at Arlington. His research interest covers feedback control, reinforcement learning, intelligent systems, cooperative control systems and nonlinear systems.)