基于强化学习的异构多智能体系统最优输出调节

熊春萍, 马 倩†

(南京理工大学 自动化学院, 江苏 南京 210094)

摘要: 本文研究了异构多智能体系统的最优输出调节问题. 通信网络拓扑含有向生成树. 首先, 设计了外部系统状态补偿器和状态反馈控制器, 应用图论和Lyapunov稳定性理论证明了所设计的补偿器和控制器可以解决一般输出调节问题. 然后, 通过最小化预定义的成本方程, 解决最优输出调节问题. 结合最优控制理论和强化学习技术, 提出了两种求解最优控制器的算法, 即基于模型的策略迭代算法和无模型off-policy算法. 利用无模型算法获取最优控制器的过程既不需要求解输出调节方程也不需要使用系统动态信息. 最后, 通过数值仿真验证了本文所提出的算法的有效性.

关键词: 异构多智能体系统; 最优输出调节; 策略迭代; 无模型算法; 强化学习

引用格式: 熊春萍, 马倩. 基于强化学习的异构多智能体系统最优输出调节, 2025, 42(3): 491 – 498

DOI: 10.7641/CTA.2023.30067

Optimal output regulation of heterogeneous multi-agent systems via reinforcement learning

XIONG Chun-ping, MA Qian†

(School of Automation, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, China)

Abstract: The optimal output regulation of heterogeneous multi-agent systems is investigated in this paper. A directed spanning tree is contained in the communication network. First of all, the exo-system state compensator and the state feedback controller are designed. Based on the graph theory and the Lyapunov stability theory, it is proved that the designed compensator and controller can achieve the general output regulation. Then, the optimal output regulation problem is worked out via minimizing a predefined cost function. Combining optimal control theory with reinforcement learning technology, two algorithms are proposed to deal with the optimal controller, which are model-based policy iteration algorithm and model-free off-policy algorithm. The process of obtaining the optimal controller by model-free algorithm does not need to solve the output regulation equation or use the information of system dynamics. Last but not least, a numerical example is proposed to verify the effectiveness of the proposed algorithms.

Key words: heterogeneous multi-agent systems; optimal output regulation; policy iteration; model-free algorithm; reinforcement learning

Citation: XIONG Chunping, MA Qian. Optimal output regulation of heterogeneous multi-agent systems via reinforcement learning. *Control Theory & Applications*, 2025, 42(3): 491 – 498

1 引言

多智能体系统的同步问题近年来一直是人们研究的热点, 被广泛应用于微电网、传感器网络、智能交通等领域, 其研究成果可参见文献[1–4]等. 多智能体系统可以分类为同构系统和异构系统. 对于同构多智能体系统, 研究者们通常考虑状态同步问题, 而对异构多智能体系统则通常研究输出同步问题. 在实际应用中, 异构多智能体系统相较于同构多智能体系统更适合描述物理模型, 特别是在任务复杂的环境中, 允许每个智能体可以拥有独立特性的异构多智能体系统更能发挥作用. 因此, 异构多智能体系统的输出同步问题被广泛关注, 目前已获得了相对成熟的研究成果, 比如文献[5–7]为线性系统的输出同步问题提供了有效的方法, 文献[8–10]则对非线性系统的输出同步问题的理论研究做出了贡献.

随着科技的蓬勃发展和资源的日渐短缺, 人们对

收稿日期: 2023−02−17; 录用日期: 2023−11−30.

†通信作者. E-mail: qma@njust.edu.cn; Tel.: +86 13951769356.

本文责任编委: 吴立刚.

国家自然科学基金项目(62173183), 湖北省科技计划项目(2022BBA026), 咸宁市科技计划项目(2021JBZXM02)资助.

Supported by the National Natural Science Foundation of China (62173183), the Science and Technology Program of Hubei Province (2022BBA026) and the Science and Technology Program of Xianning City (2021JBZXM02).

成本和能耗的要求越来越严格. 因此最优输出同步成为控制领域的热点话题, 相关成果可见文献[11–13]. 解决线性系统最优输出同步问题的经典方法是利用最优控制理论获得可以使成本方程取得最小值的最优控制器, 而最小化成本方程的过程通常被转化为求解相应的代数黎卡提方程(algebraic riccati equation, ARE)的过程, 该方程可以利用系统模型信息直接求解. 然而, 在实际应用中精确的系统模型信息是很难获得的, 这对ARE的求解带来了很大的困难. 因此, 人们提出了一类无模型算法, 即通过强化学习(reinforcement learning, RL)[14–15]思想或自适应动态规划[16–17]方法设计迭代算法. RL的基本思想是在系统中应用试验控制策略, 对性能结果进行评估, 并在此基础上更新控制策略以提高性能, 它可以通过测量系统的输入和输出数据来在线实时解决线性二次调节器问题, 因此, 基于数据设计的RL算法不需要知道系统动力学信息. 文献[18–19]为解决最优输出同步问题, 基于两个分布式观测器设计了状态反馈控制器, 并提出了无模型RL算法对其进行求解. 文献[20]则利用相对系统输出信息和off-policy RL算法获得最优控制器. 文献[21]针对一类离散系统的输出同步问题, 提出了无模型Q学习优化方法. 文献[22]利用内模信息和off-policy策略迭代思想提出了使异构多智能体系统在输出数据的驱动下实现输出同步的学习算法.

多智能体系统的输出调节是指设计控制器使得该系统对既能生成参考信号又会带来扰动的外部系统在抗干扰的同时实现追踪, 并能使不含扰动的闭环系统稳定, 相关研究成果参见文献[23–25]. 进一步, 最优输出调节是指该系统不仅能实现输出调节还能保证能量损耗最低, 此处能量损耗的标准通常由追踪误差构成的成本方程表示[26–29]. 文献[26]针对线性异构多智能体系统的最优输出调节问题利用输出调节方程的解和值迭代思想提出了基于RL的无模型算法. 文献[27]将自适应动态规划方法与内模相结合解决了多智能体系统的最优控制问题. 文献[28]针对模型信息已知的非线性系统的鲁棒最优输出调节问题, 利用内模设计了actor-critic算法. 文献[29]基于自适应动态规划思想提出了双补偿器设计方法, 在系统动态信息部分未知的情况下解决了异质多智能体系统的最优输出调节问题. 然而, 系统动力学信息在实际工程中几乎是不可能被精确获得的. 因此, 笔者思考是否可以在不使用内模又不实际求解调节方程的情况下解决系统动态信息完全未知的异构多智能体系统的最优输出调节问题.

鉴于以上思考, 在系统动力学信息完全未知的情况下, 本文基于RL方法提出了解决线性异构多智能体系统实现最优输出调节的方案. 首先, 为解决部分智能体无法直接获取外部系统信息的问题设计了外部系统状态补偿器, 并利用补偿器状态信息设计了状态反馈控制器, 利用图论和Lyapunov稳定性理论证明了所设计的补偿器和控制器可以解决一般输出调节问题. 然后, 应用最优控制理论得到可以使多智能体系统实现最优输出调节的最优控制器. 最后基于RL方法设计了基于模型的策略迭代算法和不基于模型的off-policy算法. 本文的主要贡献如下: 一方面, 最优输出调节控制器的设计不要求多智能体系统的动态信息已知. 文献[28]要求系统动力学信息完全已知, 文献[29]则允许部分信息未知, 而本文可以容许系统动态信息完全未知; 另一方面, 本文不需要求解输出调节方程, 因此简化了计算流程. 文献[26]利用调节方程的解分别计算最优前馈增益和最优反馈增益, 而本文可以在不求解调节方程的情况下一步获得最优控制增益.

符号说明: $\mathbb{C}^-$表示复平面的开左半平面; $\mathcal{R}^m$代表m维欧几里得向量空间; $\mathcal{R}^{m\times m}$代表$m\times m$维矩阵; $\lambda(X_0)$表示矩阵X_0的特征值; $X_1\otimes X_2$表示X_1和X_2的克罗内克积; $\Re(x)$表示复数x的实部; 对于向量文献$g\in\mathcal{R}^m$, 矩阵$L\in\mathcal{R}^{n\times n}$, 分别有$\mathrm{vecv}(g)=[g_1^2\ \ g_1g_2\ \ \cdots\ \ g_1g_m\ \ g_2^2\ \ g_2g_3\ \ \cdots\ \ g_{m-1}g_m\ \ g_m^2]^{\mathrm{T}}$, $\mathrm{vecs}(L)=[l_{11}\ \ 2l_{12}\ \ \cdots\ \ 2l_{1m}\ \ l_{22}\ \ 2l_{23}\ \ \cdots\ \ 2l_{(n-1),n}\ \ l_{nn}]^{\mathrm{T}}$.

2 预备知识和问题描述

2.1 图论

多智能体之间的通信拓扑可以由加权有向图$\mathcal{G}=\{\mathcal{A},\mathcal{E},\mathcal{N}\}$来表示, 其中: $\mathcal{A}=(a_{ij})_{N\times N}$, $\mathcal{N}=\{1,2,\cdots,N\}$和$\mathcal{E}\subseteq\mathcal{N}\times\mathcal{N}$分别表示图$\mathcal{G}$对应的加权邻接矩阵、节点(智能体)的顶点集合和边集合. 如果$(j,i)\in\mathcal{E}$, 则表示节点j可以直接接收来自节点i的信号, 此时定义邻接矩阵的元素$a_{ij}>0$, $a_{ii}=0$; 如果$(j,i)\notin\mathcal{E}$, 则定义$a_{ij}=0$. 节点i的邻居集合用$\mathcal{N}_i=\{j\in\mathcal{N}:(j,i)\in\mathcal{E}, j\neq i\}$表示. 图$\mathcal{G}$对应的拉普拉斯矩阵由$\mathcal{L}=(l_{ij})_{N\times N}$表示, 其中: $l_{ii}=\sum\limits_{j=1,j\neq i}^{N}a_{ij}$, $l_{ij}=-a_{ij}$, $i=1,2,\cdots,N$. 假如存在一个节点与其他任意节点之间均存在有向路径, 则称该节点为根节点. 假如一个图中至少存在一个根节点, 那么称该图含有向生成树.

2.2 问题描述

考虑以下由N个节点(智能体)组成的线性异构多智能体系统:

$$\begin{cases}\dot{x}_i=A_ix_i+B_iu_i+E_iv,\\ y_i=C_ix_i+D_iv,\\ e_i=C_ix_i+F_iv,\ i\in\mathcal{N},\end{cases}\tag{1}$$

其中: $x_i\in\mathcal{R}^{n_i}$, $y_i\in\mathcal{R}^{p_i}$, $e_i\in\mathcal{R}^{p_i}$和$u_i\in\mathcal{R}^{m_i}$分别表示智能体i的状态信息、测量输出、追踪误差和控制

输入; A_i, B_i, C_i, D_i, E_i, F_i均为适维的常数矩阵, 其中A_i, B_i, E_i未知; $v \in \mathcal{R}^q$表示外部系统的状态信息, 外部系统的动态方程为

$$\dot{v} = Sv. \tag{2}$$

注 1　外部系统(2)既会给系统(1)带来干扰又能生成被追踪的参考信号. 在有向图中, 外部系统由智能体标号0表示, 可以被看作被追随的领导者, 其他N个智能体可以被看作跟随者. 假设只有部分跟随者可以直接获得领导者信息, 此时定义$a_{i0} > 0$, 否则$a_{i0} = 0$.

不失一般性, 给出以下假设和定义:

假设 1　(A_i, B_i)是可稳定的, (A_i, C_i)是可检测的.

假设 2　图$\mathcal{G}$含有向生成树, 其根节点可以直接获取领导者信息.

假设 3　以下输出调节方程有解组(Π_i, Γ_i):

$$\begin{cases} \Pi_i S = A_i\Pi_i + B_i\Gamma_i + E_i, \\ 0 = C_i\Pi_i + F_i,\ i \in \mathcal{N}. \end{cases} \tag{3}$$

假设 4　S的特征值不重复且实部为零.

定义 1　如果以下条件均被满足: 1) 当$v=0$时, 整个闭环系统渐近稳定; 2) 在任意初始条件下, 随着时间的推移追踪误差收敛到零, 即$\lim\limits_{t\to\infty} e_i = 0$. 那么说明输出调节问题已被解决.

定义 2　如果最优控制输入u^*不仅可以满足定义1的所有条件, 还能使将被定义的成本方程取得最小值, 则表明最优输出调节问题可以被解决.

注 2　以上所有假设是解决由定义1描述的输出调节问题的标准假设.

3　基于状态反馈的输出调节

为使所有跟随者都能获得参考信息, 设计以下补偿器:

$$\dot{z}_i = Sz_i + cH\Big(\sum_{j\in\mathcal{N}_i} a_{ij}(z_j - z_i) + a_{i0}(v - z_i)\Big), \tag{4}$$

其中: H是待求解的补偿器控制增益矩阵, z_i是补偿器的状态信息, c是耦合强度.

利用以上补偿器设计分布式状态反馈控制器

$$u_i = K_{1i}x_i + K_{2i}z_i, \tag{5}$$

其中K_{1i}和K_{2i}是待求解的控制增益矩阵.

定义

$$A = \text{diag}\{A_1, \cdots, A_N\},\ B = \text{diag}\{B_1, \cdots, B_N\},$$
$$C = \text{diag}\{C_1, \cdots, C_N\},\ D = \text{diag}\{D_1, \cdots, D_N\},$$
$$E = \text{diag}\{E_1, \cdots, E_N\},\ F = \text{diag}\{F_1, \cdots, F_N\},$$
$$\Pi = \text{diag}\{\Pi_1, \cdots, \Pi_N\},\ \Gamma = \text{diag}\{\Gamma_1, \cdots, \Gamma_N\},$$
$$K_1 = \text{diag}\{K_{11}, \cdots, K_{1N}\},\ x = \text{col}(x_1, \cdots, x_N),$$
$$K_2 = \text{diag}\{K_{21}, \cdots, K_{2N}\},\ \bar{v} = \text{col}(v, \cdots, v),$$
$$z = \text{col}(z_1, \cdots, z_N),\ x_{\text{c}} = \text{col}(x, z),$$
$$e = \text{col}(e_1, \cdots, e_N).$$

对x_{c}求导, 得

$$\dot{x}_{\text{c}} = A_{\text{c}}x_{\text{c}} + B_{\text{c}}\bar{v}, \tag{6}$$

其中:

$$A_{\text{c}} = \begin{bmatrix} A + BK_1 & BK_2 \\ 0 & I_N \otimes S - c(\mathcal{L} + G) \otimes H \end{bmatrix},$$
$$B_{\text{c}} = \begin{bmatrix} E \\ c(\mathcal{L} + G) \otimes H \end{bmatrix},$$
$$G = \text{diag}\{a_{10}, \cdots, a_{N0}\}.$$

参考文献[30]得以下引理:

引理 1　如果存在$\bar{Q} > 0$和$\bar{R} > 0$, 那么可以根据$H = \bar{R}^{-1}P$选择补偿器控制增益矩阵, 其中P是以下ARE的唯一正定解:

$$0 = S^{\text{T}}P + PS + \bar{Q} - P\bar{R}^{-1}P. \tag{7}$$

在假设1–3均成立的情况下, 如果K_{1i}可以使$A_i + B_iK_{1i}$是赫尔维兹(Hurwitz)的, 并且耦合强度c满足以下不等式:

$$c \geqslant \frac{1}{2\min\limits_{i\in\mathcal{N}} \Re(\lambda_i)}, \tag{8}$$

其中λ_i是矩阵$\mathcal{L} + G$的第i个特征值, 那么A_{c}是赫尔维兹的.

定理 1　在假设1–4均成立的情况下, 补偿器(4)和分布式状态反馈控制器(5)在由引理1得到的c, H, K_{1i}和选择$K_{2i} = \Gamma_i - K_{1i}\Pi_i$的联合作用下可以使系统(1)解决输出调节问题.

证　令$\tilde{x}_i = x_i - \Pi_i v$, $K_{2i} = \Gamma_i - K_{1i}\Pi_i$, $\tilde{z}_i = z_i - v$, $X = \text{col}(\tilde{x}, \tilde{z})$, 对$\tilde{x}_i$求导

$$\dot{\tilde{x}}_i = \dot{x}_i - \Pi_i\dot{v} =$$
$$A_ix_i + B_iK_{1i}x_i + B_iK_{2i}z_i + E_iv - \Pi_iSv =$$
$$A_ix_i + B_iK_{1i}x_i + B_iK_{2i}z_i + E_iv - A_i\Pi_iv -$$
$$B_i\Gamma_iv - E_iv =$$
$$A_ix_i + B_iK_{1i}x_i + B_iK_{2i}z_i - A_i\Pi_iv -$$
$$B_iK_{2i}v - B_iK_{1i}\Pi_iv =$$
$$(A_i + B_iK_{1i})\,\tilde{x}_i + B_iK_{2i}\tilde{z}_i.$$

令$\tilde{x} = \text{col}(\tilde{x}_1, \cdots, \tilde{x}_N)$, $\tilde{z} = \text{col}(\tilde{z}_1, \cdots, \tilde{z}_N)$, 将上式写成矩阵形式得

$$\begin{bmatrix} \dot{\tilde{x}} \\ \dot{\tilde{z}} \end{bmatrix} = \begin{bmatrix} A + BK_1 & BK_2 \\ 0 & I_N \otimes S - c(\mathcal{L} + G) \otimes H \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix},$$

即

$$\dot{X} = A_{\rm c}X. \tag{9}$$

根据引理1可知$A_{\rm c}$是赫尔维兹的, 由此可以说明闭环系统(9)是稳定的, 即$\tilde{x} \to 0$, $\tilde{z} \to 0$. 近一步由下式可得追踪误差e_i最终收敛到零, 即$e_i \to 0$:

$$e = Cx + F\bar{v} = C(\tilde{x} + \Pi\bar{v}) + F\bar{v} = C\tilde{x} + (C\Pi + F)\bar{v} = C\tilde{x}.$$

根据定义1得系统(1)在补偿器(4)和分布式状态反馈控制器(5)的作用下可以实现输出调节. 证毕.

注 3 定理2中提到的K_{i2}要求实际求解调节方程(3)的解, 下一节将在不要求求解调节方程的情况下设计最优输出调节控制器.

4 基于强化学习的最优输出调节

本节将设计两个算法来解决最优输出调节问题.

首先, 定义$\tilde{u}_i = u_i - \Gamma_i v$, $\tilde{u} = {\rm col}(\tilde{u}_1, \cdots, \tilde{u}_N)$, 利用$\tilde{u}$, 对$\tilde{x}$, X求导得

$$\begin{aligned}&\dot{\tilde{x}} = \dot{x} - \Pi\dot{v} = \\ &Ax + Bu + E\bar{v} - \Pi(I_N \otimes S)\bar{v} = \\ &Ax + Bu + E\bar{v} - (A\Pi + B\Gamma + E)\bar{v} = \\ &A\tilde{x} + B\tilde{u}, \end{aligned} \tag{10}$$

$$\dot{X} = \begin{bmatrix} A & 0 \\ 0 & I_N \otimes S - c(\mathcal{L}+G)\otimes H \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} \tilde{u} = \bar{T}X + \bar{B}\tilde{u}, \tag{11}$$

其中:

$$\bar{T} = \begin{bmatrix} A & 0 \\ 0 & I_N \otimes S - c(L+G)\otimes H \end{bmatrix},\ \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}.$$

另外有

$$\begin{aligned}&\tilde{u} = u - \Gamma\bar{v} = \\ &K_1 x + K_2 z - \Gamma\bar{v} = \\ &K_1(\tilde{x} + \Pi\bar{v}) + (\Gamma - K_1\Pi)\, z - \Gamma\bar{v} = \\ &K_1\tilde{x} + K_2\tilde{z} = -KX, \end{aligned} \tag{12}$$

其中$K = [-K_1 \ \ -K_2]$.

然后, 定义成本方程

$$V = \int_t^\infty \left(\tilde{x}^{\rm T}Q\tilde{x} + \tilde{u}^{\rm T}R\tilde{u}\right){\rm d}\tau, \tag{13}$$

其中: $Q > 0$, $R > 0$.

根据式(12), 以上方程可以写成二次型

$$V = \int_t^\infty (X^{\rm T}\begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} X + \tilde{u}^{\rm T}R\tilde{u}){\rm d}\tau = \int_t^\infty (X^{\rm T}\tilde{Q}X + \tilde{u}^{\rm T}R\tilde{u}){\rm d}\tau = X^{\rm T}MX, \tag{14}$$

其中: $M > 0$, 并且$\tilde{Q} = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}$.

基于式(14), 定义哈密顿方程为

$$H_{\rm a} \equiv (\bar{T}X + \bar{B}\tilde{u})^{\rm T}MX + X^{\rm T}M(\bar{T}X + \bar{B}\tilde{u}) + X^{\rm T}\tilde{Q}X + \tilde{u}^{\rm T}R\tilde{u}. \tag{15}$$

以上方程对$\tilde{u}$求偏导得

$$\frac{\partial H_{\rm a}}{\partial \tilde{u}} = 2\bar{B}^{\rm T}MX + 2R\tilde{u}. \tag{16}$$

根据最优控制原理, 令$\dfrac{\partial H_{\rm a}}{\partial \tilde{u}} = 0$, 可以得到最优控制器$\tilde{u}^* = -R^{-1}\bar{B}^{\rm T}MX$, 由此获得最优控制增益$K^* = R^{-1}\bar{B}^{\rm T}M$. 将$\tilde{u}^* = -R^{-1}\bar{B}^{\rm T}MX$代入$H_{\rm a} = 0$, 得

$$\begin{aligned}0 =\ &(\bar{T}X + \bar{B}\tilde{u}^*)^{\rm T}MX + X^{\rm T}M(\bar{T}X + \bar{B}\tilde{u}^*) + X^{\rm T}\tilde{Q}X + (\tilde{u}^*)^{\rm T}R\tilde{u}^* = \\ &(\bar{T}X - \bar{B}R^{-1}\bar{B}^{\rm T}MX)^{\rm T}MX + X^{\rm T}M(\bar{T}X - \bar{B}R^{-1}\bar{B}^{\rm T}MX) + X^{\rm T}\tilde{Q}X + X^{\rm T}M\bar{B}R^{-1}\bar{B}^{\rm T}MX = \\ &X^{\rm T}(\bar{T}^{\rm T}M + M\bar{T} + \tilde{Q} - M\bar{B}R^{-1}\bar{B}^{\rm T}M)X. \end{aligned} \tag{17}$$

由此得以下ARE:

$$0 = \bar{T}^{\rm T}M + M\bar{T} + \tilde{Q} - M\bar{B}R^{-1}\bar{B}^{\rm T}M, \tag{18}$$

其中M是其唯一正定解.

由文献[31]可知, 基于假设1, 式(18)有唯一对称正定解M^*(可称为M的最优值), 由此最优控制增益也可以写作$K^* = R^{-1}\bar{B}^{\rm T}M^*$.

接下来提出算法1. 基于模型的策略迭代算法, 由以下步骤构成:

步骤 1 初始化: 令ρ表示学习过程的迭代次数, 且初始值设为$\rho = 0$. 给定控制输入$\tilde{u}^\rho = -K^0X$, 其中K^0是稳定的初始增益, 即满足$\lambda(\bar{T} - \bar{B}K^0) \subset \mathbb{C}^-$.

步骤 2 策略评估: 利用K^ρ, 对以下方程进行求解获得M^ρ:

$$(\bar{T} - \bar{B}K^\rho)^{\rm T}M^\rho + M^\rho(\bar{T} - \bar{B}K^\rho) + \tilde{Q} + (K^\rho)^{\rm T}RK^\rho = 0. \tag{19}$$

策略更新: 利用M^ρ并结合下式求解$K^{\rho+1}$:

$$K^{\rho+1} = R^{-1}\bar{B}^{\rm T}M^\rho. \tag{20}$$

步骤 3 如果不等式$\|K^{\rho+1} - K^\rho\| \leqslant o$($o$是接近于零的正数)可以被满足, 那么停止迭代, 否则设置$\rho = \rho + 1$并返回步骤2.

步骤 4 迭代过程结束后获得最优控制增益$K^* = K^{\rho+1}$.

注 4 文献[32]对该算法的收敛性进行了详细的解释, 本文不再赘述. 且由文献[22]可知如果初始化时选择稳定的

增益K_0, 那么在之后每次迭代过程中的增益将保持稳定, 即满足$\lambda(\bar{T}-\bar{B}K^\rho)\subset\mathbb{C}^-$. K_0的选择方法可参考文献[33]. 在已知系统模型信息的情况下可以直接求解式(19)和式(20), 但是在实际工程应用中, 精确的控制模型信息几乎是不可得的, 所以接下来将设计不基于模型的算法2.

首先, 将u代入$x_{\rm c}$的导数, 式(6)可以被写作

$$\begin{aligned}\dot{x}_{\rm c}=&\begin{bmatrix}A & 0\\ 0 & I_N\otimes S-c(\mathcal{L}+G)\otimes H\end{bmatrix}x_{\rm c}+\begin{bmatrix}B\\0\end{bmatrix}u+\\&\begin{bmatrix}E\\c(\mathcal{L}+G)\otimes H\end{bmatrix}\bar{v}=\\&\bar{T}x_{\rm c}+\bar{B}u+B_{\rm c}\bar{v}.\end{aligned}\tag{21}$$

定义$T^\rho=\bar{T}-\bar{B}K^\rho$, 得

$$\dot{x}_{\rm c}=T^\rho x_{\rm c}+\bar{B}\left(u+K^\rho x_{\rm c}\right)+B_{\rm c}\bar{v}.$$

联立式(19)和式(20), 得

$$\begin{aligned}&x_{\rm c}^{\rm T}(t+\Delta t)M^\rho x_{\rm c}(t+\Delta t)-x_{\rm c}^{\rm T}(t)M^\rho x_{\rm c}(t)=\\&\int_t^{t+\Delta t}x_{\rm c}^{\rm T}((T^\rho)^{\rm T}M+MT^\rho)x_{\rm c}{\rm d}\tau+\\&2\int_t^{t+\Delta t}(u+K^\rho x_{\rm c})^{\rm T}\bar{B}^{\rm T}M^\rho x_{\rm c}{\rm d}\tau+\\&2\int_t^{t+\Delta t}\bar{v}^{\rm T}B_{\rm c}^{\rm T}M^\rho x_{\rm c}{\rm d}\tau=\\&-\int_t^{t+\Delta t}x_{\rm c}^{\rm T}(\tilde{Q}+(K^\rho)^{\rm T}RK^\rho)x_{\rm c}{\rm d}\tau+\\&2\int_t^{t+\Delta t}(u+K^\rho x_{\rm c})^{\rm T}RK^{\rho+1}x_{\rm c}{\rm d}\tau+\\&2\int_t^{t+\Delta t}\bar{v}^{\rm T}B_{\rm c}^{\rm T}M^\rho x_{\rm c}{\rm d}\tau,\end{aligned}\tag{22}$$

其中:

$$\begin{aligned}&\bar{v}^{\rm T}B_{\rm c}^{\rm T}M^\rho x_{\rm c}=(x_{\rm c}^{\rm T}\otimes\bar{v}^{\rm T})\,{\rm vec}(B_{\rm c}^{\rm T}M^\rho),\\&x_{\rm c}^{\rm T}(\tilde{Q}+(K^\rho)^{\rm T}RK^\rho)x_{\rm c}=\\&(x_{\rm c}^{\rm T}\otimes x_{\rm c}^{\rm T})\,{\rm vec}(\tilde{Q}+(K^\rho)^{\rm T}RK^\rho),\\&(u+K^\rho x_{\rm c})^{\rm T}RK^{\rho+1}x_{\rm c}=\\&(x_{\rm c}^{\rm T}\otimes u^{\rm T})(I_N\otimes R)\,{\rm vec}(K^{\rho+1})+\\&(x_{\rm c}^{\rm T}\otimes x_{\rm c}^{\rm T})(I_N\otimes(K^\rho)^{\rm T}R)\,{\rm vec}(K^{\rho+1}).\end{aligned}$$

然后, 给定正整数h, 定义

$$\delta=\begin{bmatrix}{\rm vecv}\left(x_{\rm c}\left(t_1\right)\right)-{\rm vecv}\left(x_{\rm c}\left(t_0\right)\right)\\{\rm vecv}\left(x_{\rm c}\left(t_2\right)\right)-{\rm vecv}\left(x_{\rm c}\left(t_1\right)\right)\\\vdots\\{\rm vecv}\left(x_{\rm c}\left(t_h\right)\right)-{\rm vecv}\left(x_{\rm c}\left(t_{h-1}\right)\right)\end{bmatrix},$$

$$\Upsilon_{\rm XX}=\begin{bmatrix}\int_{t_0}^{t_1}x_{\rm c}^{\rm T}\otimes x_{\rm c}^{\rm T}{\rm d}\tau\\\int_{t_1}^{t_2}x_{\rm c}^{\rm T}\otimes x_{\rm c}^{\rm T}{\rm d}\tau\\\vdots\\\int_{t_{h-1}}^{t_h}x_{\rm c}^{\rm T}\otimes x_{\rm c}^{\rm T}{\rm d}\tau\end{bmatrix},$$

$$\Upsilon_{\rm XV}=\begin{bmatrix}\int_{t_0}^{t_1}x_{\rm c}^{\rm T}\otimes \bar{v}^{\rm T}{\rm d}\tau\\\int_{t_1}^{t_2}x_{\rm c}^{\rm T}\otimes \bar{v}^{\rm T}{\rm d}\tau\\\vdots\\\int_{t_{h-1}}^{t_h}x_{\rm c}^{\rm T}\otimes \bar{v}^{\rm T}{\rm d}\tau\end{bmatrix},$$

$$\Upsilon_{\rm XU}=\begin{bmatrix}\int_{t_0}^{t_1}x_{\rm c}^{\rm T}\otimes u^{\rm T}{\rm d}\tau\\\int_{t_1}^{t_2}x_{\rm c}^{\rm T}\otimes u^{\rm T}{\rm d}\tau\\\vdots\\\int_{t_{h-1}}^{t_h}x_{\rm c}^{\rm T}\otimes u^{\rm T}{\rm d}\tau\end{bmatrix},$$

其中$t_0<t_1<\cdots<t_h$.

令

$$\Phi^\rho=\begin{bmatrix}\delta\\-2(\Upsilon_{\rm XU}(I_N\otimes R)-2\Upsilon_{\rm XX}(I_N\otimes(K^\rho)^{\rm T}R))\\-2\Upsilon_{\rm XV}\end{bmatrix}^{\rm T},$$

$$\Theta^\rho=-\Upsilon_{\rm XX}\,{\rm vec}(\bar{Q}+(K^\rho)^{\rm T}RK^\rho).$$

根据式(22)得

$$\Phi^\rho\begin{bmatrix}{\rm vecs}(M^\rho)\\{\rm vec}(K^{\rho+1})\\{\rm vec}(B_{\rm c}^{\rm T}M^\rho)\end{bmatrix}=\Theta^\rho,\tag{23}$$

如果Φ^ρ列满秩, 可以用下式求解M^ρ和$K^{\rho+1}$:

$$\begin{bmatrix}{\rm vecs}(M^\rho)\\{\rm vec}(K^{\rho+1})\\{\rm vec}(B_{\rm c}^{\rm T}M^\rho)\end{bmatrix}=((\Phi^\rho)^{\rm T}\Phi^\rho)^{-1}(\Phi^\rho)^{\rm T}\Theta^\rho.\tag{24}$$

最后, 得到算法2. 不基于模型的off-policy算法, 该算法分为以下步骤:

步骤 1　初始化: 令ρ表示学习过程的迭代次数, 且初始化为$\rho=0$. 给定控制输入$u=-K^0X+\upsilon$, 其中υ是探测噪声, K^0是稳定的初始增益, 即满足$\lambda(\bar{T}-\bar{B}K^0)\subset\mathbb{C}^-$.

步骤 2　计算$\Upsilon_{\rm XX},\Upsilon_{\rm XU},\Upsilon_{\rm XV}$.

步骤 3　判断Φ^ρ是否满足列满秩条件, 如果满足则根据式(24)计算M^ρ和$K^{\rho+1}$, 否则继续进行数据收集.

步骤 4　如果不等式$\|K^{\rho+1}-K^\rho\|\leqslant o$可以被满足, 那么则停止迭代, 否则设置$\rho=\rho+1$并返回步骤3.

注 5　给定的控制输入u在off-policy算法中被称为行为策略, 用于生成系统数据$(x_{\rm c},u,\bar{v})$, 经过h次数据收集后, Φ^ρ可以满足列满秩的条件, 即满足Φ^ρ的列数等于$[\Upsilon_{\rm XX},\Upsilon_{\rm XU},\Upsilon_{XV}]$的秩, ${\rm rank}[\Upsilon_{\rm XX},\Upsilon_{\rm XU},\Upsilon_{\rm XV}]=(Nn_i+Nq)\times(\frac{Nn_i+Nq+1}{2}+Nm_i+Nq)$, 所以$h$需要满足不等式$h\geqslant$

$(Nn_i+Nq)(\frac{Nn_i+Nq+1}{2}+Nm_i+Nq)$. 在$\Phi^\rho$列满秩的情况下式(24)有唯一解, 因此$\Upsilon_{\rm XX}$, $\Upsilon_{\rm XU}$, $\Upsilon_{\rm XV}$可以由持续激励下的最小二乘法唯一确定[22]. 持续激励条件可以通过给控制输入添加探测噪声的方式满足, 而探测噪声的添加并不会影响off-policy RL算法的收敛性和系统的稳定性[34]. 在实际工程中, 随机噪声或正弦信号通常被用来表示探测噪声.

定理 2 算法2中, 随着迭代次数的增加K^ρ最终收敛到K^*, M^ρ最终收敛到其最优值M^*, 且最优控制输入为$u^*=-K^*x_{\rm c}$.

证 M^ρ和$K^{\rho+1}$在Φ^ρ列满秩的情况下可以通过式(24)唯一求解. 另一方面, M^ρ是ARE(19)的唯一解, 且$K^{\rho+1}$通过式(20)唯一确定. 所以求解式(24)等同于求解式(19)和式(20), 即M^ρ最终趋向于M^*, K^ρ最终收敛到K^*. 另外根据式(12), 显然得$u^*=-K^*x_{\rm c}$.

证毕.

注 6 由以上定理得最优输出调节控制器u^*只与K^*和$x_{\rm c}$有关, 即最优输出调节控制器的设计不需要求解输出调节方程. 且利用算法2求解K^*和采集$x_{\rm c}$数据过程中并未使用A, B和E的信息, 所以算法2可以解决多智能体系统动态信息完全未知情况下的最优输出调节问题.

注 7 基于模型的算法1使用系统模型来预测下一个状态和奖励(本文考虑的是能量损耗), 并使用这些信息来更新策略, 可以精确地预测多智能体系统在未来的状态和成本, 从而更加高效地更新策略, 适用于状态空间较小、可以使用可靠的环境模型预测的情况. 不基于模型的算法2不使用系统模型, 只依靠与环境之间的交互来更新策略, 可以在任意环境中使用, 无需事先对系统建模, 因此具有更广泛的适用性, 适用于状态空间较大或不确定的情况.

5 数例仿真

本节将通过数例仿真来验证算法1和算法2的有效性. 考虑如图1所示的网络拓扑结构, 其中标号为0的节点表示领导者, 其余节点代表跟随者. 给出如下系统矩阵参数:

$$A_1=\begin{bmatrix}-1 & -1\\ 1 & 0\end{bmatrix},\ B_1=\begin{bmatrix}1\\1\end{bmatrix},\ C_1=[1\ \ 1],$$

$$A_2=\begin{bmatrix}-2 & 0\\ 1 & -1\end{bmatrix},\ B_2=\begin{bmatrix}1\\0\end{bmatrix},\ C_2=[1\ \ 1],$$

$$A_3=\begin{bmatrix}-3 & 2\\ 2 & -2\end{bmatrix},\ B_3=\begin{bmatrix}0\\1\end{bmatrix},\ C_3=[1\ \ 2],$$

$$E_1=E_2=E_3=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},$$

$$D_1=D_2=D_3=[1\ \ 1],$$

$$F_1=F_2=F_3=[0.5\ \ 0.5],\ S=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix},$$

$$\bar{Q}=\mathrm{diag}\{1,2\},\ \bar{R}=\mathrm{diag}\{1,1\},$$

$$Q=\mathrm{diag}\{10,10,0.1,0.1,1,1\},\ R=\mathrm{diag}\{1,2,3\}.$$

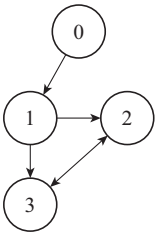

图 1 拓扑结构

Fig. 1 The topology structure

根据引理1, 选择耦合强度$c=0.5$, 那么可以计算得补偿器的控制增益为

$$H=\begin{bmatrix}1.0917 & -0.1010\\ -0.1010 & 1.3371\end{bmatrix}.$$

选择正弦信号作为探测噪声, 并选择$o=0.01$. 图2–6为仿真结果.

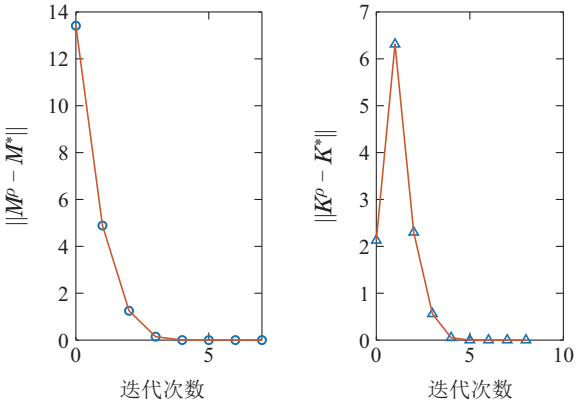

图 2 算法1中M^ρ与M^*的差和K^ρ与K^*的差

Fig. 2 The comparison of M^ρ and M^* and the comparison of K^ρ and K^* of Algorithm 1

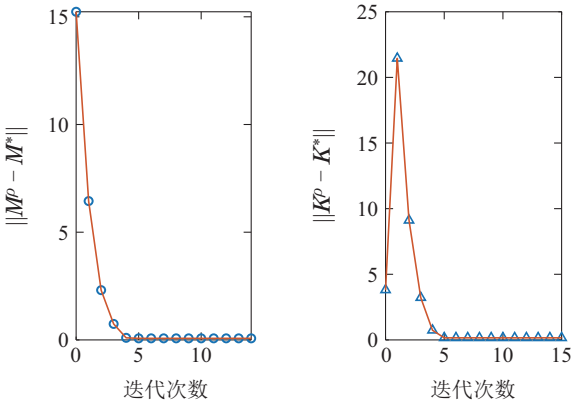

图 3 算法2中M^ρ与M^*的差和K^ρ与K^*的差

Fig. 3 The comparison of M^ρ and M^* and the comparison of K^ρ and K^* of Algorithm 2

图2描述了算法1仿真结果M^ρ与M^*的差的曲线和K^ρ与K^*的差的曲线, 表明M^ρ和K^ρ都最终收敛到

了各自的最优值. 图3描述了算法2仿真结果M^{ρ}与M^{*}的差的曲线和K^{ρ}与K^{*}的差的曲线, 表明M^{ρ}和K^{ρ}都最终收敛到了各自的最优值. 比较图2和图3可知, 利用模型信息和李雅普诺夫方程求解方法的算法1与利用在线采集的数据和最小二乘法的算法2相比能更快的得到收敛结果, 即需要的迭代次数较少. 图4描述了多智能体状态轨迹, 表明闭环系统最终是稳定的. 图5表示领导者与跟随者之间的追踪误差在最优控制器作用下的仿真曲线, 由此可以发现追踪误差随着时间的推移收敛到0, 即表示所有智能体已经追踪上了外部系统提供的参考信号.

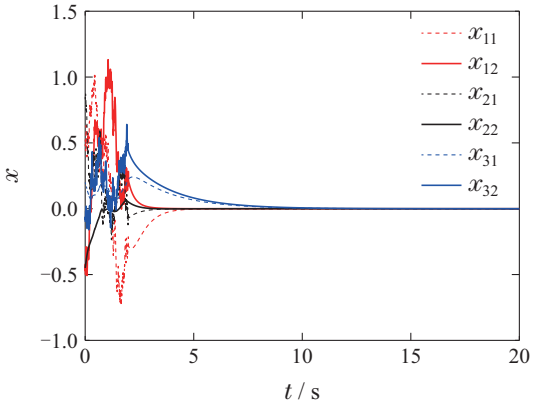

图 4　智能体状态曲线x

Fig. 4　The curves of agent's state x

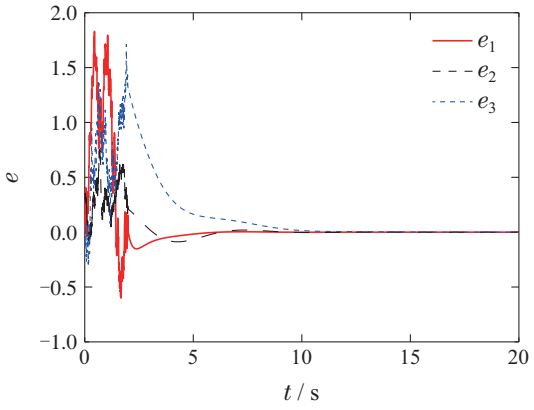

图 5　跟踪误差曲线e

Fig. 5　The curves of tracking error e

根据文献[26]的思想, 利用A_i, B_i, Q, R可以得到如图6所示的M和K_1的收敛结果, 该结果表示在通过有限次迭代后可以获得最优反馈控制增益K_{i1}^{*}如下:

$$K_{11}^{*}=[-1.1323 \quad -3.6369],$$
$$K_{21}^{*}=[-0.0166 \quad -0.0083],$$
$$K_{31}^{*}=[-0.1337 \quad -0.2064],$$

然后利用Sylvester映射计算得到输出调解方程的解

$$\Pi_1=\begin{bmatrix} 0 & -1 \\ \dfrac{1}{2} & \dfrac{1}{2} \end{bmatrix},\ \Gamma_1=[-\frac{1}{2} \quad -\frac{1}{2}],$$

$$\Pi_2=\begin{bmatrix} -\dfrac{2}{5} & -\dfrac{4}{5} \\ -\dfrac{1}{10} & \dfrac{3}{10} \end{bmatrix},\ \Gamma_2=[-1 \quad -2],$$

$$\Pi_3=\begin{bmatrix} \dfrac{3}{34} & -\dfrac{5}{34} \\ \dfrac{5}{17} & -\dfrac{3}{17} \end{bmatrix},\ \Gamma_3=[-\frac{10}{17} \quad -\frac{23}{17}].$$

根据$K_{2i}=\Gamma_i-K_{1i}\Pi_i$得

$$K_{12}^{*}=[-2.3185 \quad 0.1862],$$
$$K_{22}^{*}=[-0.5042 \quad -0.5125],$$
$$K_{32}^{*}=[-0.6023 \quad -0.5314].$$

比较图2和图6可知, 本文的算法和文献[26]的算法均可以在有限次迭代后获得最优控制增益, 而本文获得的最优控制增益是$[K_{1i}^{*} \ \ K_{2i}^{*}]$, 文献[26]则需要进一步计算才能获得$K_{2i}^{*}$. 所以本文的算法在不需要求解调节方程的情况下可以减少计算量节约计算时间.

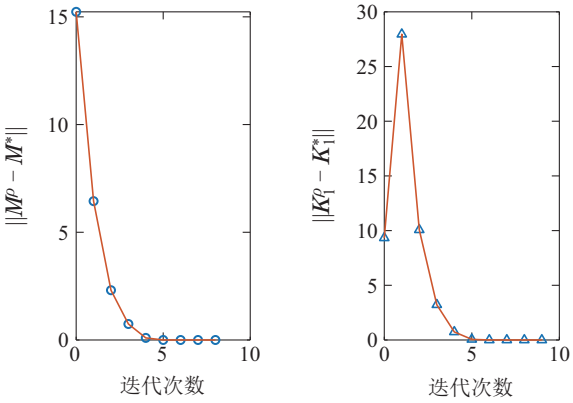

图 6　文献[26]算法1中M^{ρ}与M^{*}的差和K_1^{ρ}与K_1^{*}的差

Fig. 6　The comparison of M^{ρ} and M^{*} and the comparison of K_1^{ρ} and K_1^{*} of Algorithm 1 in [26]

6　结论

本文通过设计可行的控制协议和两个有效的算法解决了由系统(1)描述的异构多智能体系统的最优输出调节问题. 首先, 设计了补偿器(4)和状态反馈控制器(5)来解决一般输出调节问题. 然后, 在此基础上利用最优控制理论和RL技术设计了两个算法来解决最优输出调节问题. 当系统动态信息完全未知时基于模型的算法不再适用, 因此使用不基于模型的算法来避免对系统动态信息的要求, 且避免了对输出调节方程的求解. 最后, 利用数例仿真验证了两个算法的有效性. 本文的无模型算法在既不求解输出调节方程又不要求系统动态信息完全已知的情况下解决了异构多智能体系统的最优输出调节问题. 但是, 本文的研究结果要求理想的通信环境. 然而, 不含时滞或噪声的通信几乎是不存在的, 因此通信约束下的多智能体系统输出调节问题将作为笔者下一步的研究内容.

参考文献:

[1] LIU Z, GUO L. Synchronization of multi-agent systems without connectivity assumptions. *Automatica*, 2009, 45(12): 2744 – 2753.

[2] TRENTELMAN H L, TAKABA K, MONSHIZADEH N. Robust synchronization of uncertain linear multi-agent systems. *IEEE Transactions on Automtic Control*, 2013, 58(6): 1511 – 1523.

[3] SU S Z, LIN Z L, GARCIA A. Distributed synchronization control of multiagent systems with unknown nonlinearities. *IEEE Transactions on Cybernetics*, 2016, 46(1): 325 – 338.

[4] MA Q, XU S Y. Intentional delay can benefit consensus of second-order multi-agent systems. *Automatica*, 2023, 147: 110750.

[5] WU Y Q, SU H Y, SHI P, et al. Output synchronization of nonidentical linear multiagent systems. *IEEE Transactions on Cybernetics*, 2017, 47(1): 130 – 141.

[6] XUE M Q, TANG Y, REN W, et al. Practical output synchronization for asynchronously switched multi-agent systems with adaption to fast-switching perturbations. *Automatica*, 2020, 116: 108917.

[7] MA Q, MIAO G Y. Output consensus for heterogeneous multi-agent systems with linear dynamics. *Applied Mathematics and Computation*, 2015, 271: 548 – 555.

[8] RAN M P, XIE L H. Practical output consensus of nonlinear heterogeneous multi-agent systems with limited data rate. *Automatica*, 2021, 129: 109624.

[9] YANG Yang, LIU Qidong, CHEN Didi, et al. Predictor-based neural dynamic surface output consensus control of a class of nonlinear multi-agent systems. *Control Theory & Applications*, 2021, 38(8): 1197 – 1212.
(杨杨, 刘奇东, 陈笛笛, 等. 基于预估器的一类多智能体系统神经动态面输出一致控制. 控制理论与应用, 2021, 38(8): 1197 – 1212.)

[10] KHAN G D, CHEN Z Y, MENG H F. Output synchronization of nonlinear heterogeneous multi-agent systems with switching networks. *Systems & Control Letters*, 2019, 125: 45 – 50.

[11] LIU Y Y, WANG Z S. Optimal output synchronization of heterogeneous multi-agent systems using measured input-output data. *Information Sciences*, 2022, 582: 462 – 479.

[12] XIA L N, LI Q, SONG R Z, et al. Optimal synchronization control of heterogeneous asymmetric input-constrained unknown nonlinear MASs via reinforcement learning. *IEEE/CAA Journal of Automatica Sinica*, 2022, 9(3): 520 – 532.

[13] JIAO J, TRENTELMAN H L, CAMLIBEL M K. H_2 suboptimal output synchronization of heterogeneous multi-agent systems. *Systems & Control Letters*, 2021, 149: 104872.

[14] QIN J H, LI M, SHI Y, et al. Optimal synchronization control of multiagent systems with input saturation via off-policy reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(1): 85 – 96.

[15] LI Q, XIA L N, SONG R Z, et al. Output event-triggered tracking synchronization of heterogeneous systems on directed digraph via model-free reinforcement learning. *Information Sciences*, 2021, 559: 171 – 190.

[16] WANG B J, XU L, YI X L, et al. Semiglobal suboptimal output regulation for heterogeneous multi-agent systems with input saturation via adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 35(3): 3242 – 3250.

[17] GAO W N, JIANG Z P. Adaptive dynamic programming and adaptive output regulation of linear systems. *IEEE Transactions on Automatic Control*, 2016, 61(12): 4164 – 4169.

[18] MODARES H, NAGESHRAO S P, LOPES G A D, et al. Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning. *Automatica*, 2016, 71: 334 – 341.

[19] YANG Y L, MODARES H, WUNSCH D C, et al. Leader-follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(6): 2139 – 2153.

[20] CHEN C, LEWIS F L, XIE K, et al. Off-policy learning for adaptive optimal output synchronization of heterogeneous multi-agent systems. *Automatica*, 2020, 119: 109081.

[21] KIUMARSI B, LEWIS F L. Output synchronization of heterogeneous discrete-time systems: A model-free optimal approach. *Automatica*, 2017, 84: 86 – 94.

[22] CHEN C, LEWIS F L, XIE K, et al. Distributed output data-driven optimal robust synchronization of heterogeneous multi-agent systems. *Automatica*, 2023, 153: 111030.

[23] QIAN Y Y, LIU L, FENG G. Cooperative output regulation of linear multiagent systems: An event-triggered adaptive distributed observer approach. *IEEE Transactions on Automatic Control*, 2021, 66(2): 833 – 840.

[24] ZHANG D, DENG C, FENG G. Resilient cooperative output regulation for nonlinear multiagent systems under DoS attacks. *IEEE Transactions on Automatic Control*, 2023, 68(4): 2521 – 2528.

[25] WEI Wenjun, MA Yangqin, LI Zonggang. Cooperative output regulation of heterogeneous multi-agent system in finite time. *Control Theory & Applications*, 2019, 36(6): 885 – 892.
(魏文军, 马羊琴, 李宗刚. 有限时间内异构多智能体系统的协同输出调节. 控制理论与应用, 2019, 36(6): 885 – 892.)

[26] JIANG Y, KIUMARSI B, FAN J L, et al. Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning. *IEEE Transactions on Cybernetics*, 2020, 50(7): 3147 – 3156.

[27] GAO W N, JIANG Z P, LEWIS F L, et al. Leader-to-formation stability of multiagent systems: An adaptive optimal control approach. *IEEE Transactions on Automatic Control*, 2018, 63(10): 3581 – 3587.

[28] JIN P, MA Q, ZHOU G P, et al. Reinforcement learning-based robust optimal output regulation for constrained nonlinear systems with static and dynamic uncertainties. *International Journal of Robust and Nonlinear Control*, 2022, 33(3): 2022 – 2040.

[29] ZHANG H G, LIANG H J, WANG Z S, et al. Optimal output regulation for heterogeneous multiagentsystems via adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28(1): 18 – 29.

[30] MA Q, XU S Y, LEWIS F L, et al. Cooperative output regulation of singular heterogeneous multiagent systems. *IEEE Transactions on Cybernetics*, 2016, 46(6): 1471 – 1475.

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

[32] KLEINMAN D. On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, 1968, 13(1): 114 – 115.

[33] CHEN C, LEWIS F L, LI B. Homotopic policy iteration-based learning design for unknown linear continuous-time systems. *Automatica*, 2022, 138: 110153.

[34] KIUMARSI B, LEWIS F L, JIANG Z P. H_∞ control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, 2017, 78: 144 – 152.

作者简介:

熊春萍 博士研究生, 目前研究方向为多智能体系统分布式控制、强化学习、博弈论, E-mail: chunp_bear@163.com;

马 倩 教授, 博士生导师, 目前研究方向为非线性系统的镇定与控制、时滞系统的稳定性分析及鲁棒控制、多智能体系统协调控制、强化学习, E-mail: qma@njust.edu.cn.