

异构多智能体系统的输出同步：一个基于数据的强化学习方法

刘莹莹, 王占山

(东北大学信息科学与工程学院, 辽宁 沈阳 110819)

摘要: 通过强化学习研究了异构多智能体系统的输出同步问题。根据多智能体系统的拓扑结构, 定义一个具有邻居控制输入的性能指标和价值函数。为克服已有控制方法需要系统模型的弊端, 提出一个基于系统数据的强化学习算法, 使输出同步控制器也可以被应用于模型未知的情况。此外, 通过调节价值函数中的权重矩阵, 可以减少每个智能体的控制成本。最后, 通过一个仿真示例验证了该方法的有效性和定义的价值函数的优越性。

关键词: 多智能体系统; 强化学习; 输出同步; 基于数据

中图分类号: TP273, TP301

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202042

Output synchronization of heterogeneous multi-agent system: a reinforcement learning approach based on data

LIU Yingying, WANG Zhanshan

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

Abstract: The output synchronization of heterogeneous multi-agent system was studied by reinforcement learning. According to the topology of multi-agent system, the performance index and value function with neighbor control input were defined. To overcome the disadvantage of existing control methods that require system model, a reinforcement learning algorithm based on system data was proposed. Hence, the output synchronization controller can also be applied when the system model was unknown. In addition, by adjusting the weight matrix in value function, the control cost of each agent can be reduced. Finally, a simulation example was given to illustrate the effectiveness of proposed method and the superiority of defined value function.

Key words: multi-agent system, reinforcement learning, output synchronization, based on data

1 引言

多智能体系统 (multi-agent system, MAS) 可以描述一组系统的集体行为, 并且有很多重要的应用^[1-3]。近些年涌现了大量关于 MAS 问题的研究^[4-8]。其中, 输出同步问题是 MAS 的一个基本问题。在很多实际应用中, 每个智能体的动态和维度可能是不同的, 因此, 研究异构 MAS 的输出同步是十分

重要的^[9]。

目前, 已经有许多关于异构 MAS 输出同步问题的分布式控制器设计方法的研究^[10-12]。大部分控制器设计方法需要求解 MAS 的输出调节方程, 并且不可避免地需要系统模型信息。参考文献[11]基于内模原理提出了一种统一的方法来解决异构 MAS 的输出群同步问题。参考文献[12]通过设计分布式观测器研究了离散 MAS 的协同输出调节问题。

收稿日期: 2020-11-16; 修回日期: 2020-12-01

通信作者: 王占山, zhanshan_wang@163.com

基金项目: 国家自然科学基金资助项目 (No.61973070); 辽宁省“兴辽英才计划”资助项目 (No.XLYC1802010); 流程工业综合自动化基础研究基金资助项目 (No.2018ZCX22)

Foundation Items: The National Natural Science Foundation of China (No.61973070), Liaoning Revitalization Talents Program (No.XLYC1802010), SAPI Fundamental Research Funds (No.2018ZCX22)

参考文献[11-12]都是基于系统模型得到的结果。然而，在许多实际应用中是无法获取 MAS 的精确模型的。因此，已有的分布式控制方法不能应用于这些系统模型未知的情况。

强化学习起源于“试错学习”，强调使用基于环境的反馈修改行为，它由英国生物和心理学家 Conway Morgan 正式提出^[13]。随后，Werbos P J 将强化学习用于求解单系统的最优调节问题^[14-15]。由于强化学习可以利用系统数据迭代求解最优控制问题，它也被广泛应用于求解 MAS 问题^[7,16-17]。例如，参考文献[16]利用离策略强化学习算法求解 MAS 的最优输出同步问题。

作为强化学习的一种，Q 学习可以利用系统状态和控制信息求解最优控制问题^[18-19]。因此，可以借助系统数据，利用 Q 学习求解异构 MAS 的输出同步问题。参考文献[19]通过 Q 学习算法求取离散时间异构 MAS 的最优分布式控制策略。

对于大多数求解异构 MAS 输出同步问题的强化学习算法，其价值函数由追踪误差和智能体本身的控制输入组成^[16-17,19]。这样的价值函数虽然能在一定程度上减少控制成本，但是没有考虑 MAS 拓扑结构对价值函数的影响，即没有考虑邻居智能体对智能体本身的影响。因此需要定义一个包含邻居控制输入信息的性能指标和价值函数，从而减少每个智能体的控制成本。

基于以上讨论，在价值函数中考虑异构 MAS 的拓扑信息，并借助系统数据，使用强化学习方法求解异构 MAS 的输出同步问题是十分迫切的。因此，本文定义了一个含有邻居控制输入信息的性能指标，然后基于强化学习中的 Q 学习提出一个求解 MAS 控制器的强化学习算法。

2 基本知识和问题描述

本节首先描述图理论和 MAS 的连接拓扑，然后阐述 MAS 的输出同步问题。

2.1 图理论和 MAS 的连接拓扑

一般来说，MAS 的连接拓扑可以用图理论描述。构成图 $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{A})$ 的 3 个要素分别是节点集 $\mathcal{V}=\{1,\cdots,N\}$ 、边集 $\mathcal{E}\in\mathcal{V}\times\mathcal{V}$ 和邻接矩阵 $\mathcal{A}=[a_{ij}], i,j=1,\cdots,N$ 。图 $\mathcal{G}$ 的边由 $v_{ij}=(j,i)$ 表示，意味着信息由节点 j 流向节点 i 。如果 $v_{ij}\in\mathcal{E}$ ，则 $a_{ij}>0$ ，否则 $a_{ij}=0$ 。除此之外，节点 i 的邻居集定

义为 $\mathcal{N}_i=\{j|v_{ij}\in\mathcal{E}\}$ 。图 $\mathcal{G}$ 的度矩阵 $\mathcal{D}$ 定义为

$$\mathcal{D}=\text{diag}\{d_i\}, d_i=\sum_{j=1}^N a_{ij}。由此可得图 \mathcal{G} 的拉普拉斯$$

(Laplacian) 矩阵 $\mathcal{L}=\mathcal{D}-\mathcal{A}$ 。

如果 MAS 存在领导者，并且每个智能体都可以直接或间接获得领导者的信息，那么图 $\mathcal{G}$ 包含了一个生成树。如果领导者直接连接到节点 i 上，那么牵引增益 $g_i=1$ ，否则 $g_i=0$ 。图 $\mathcal{G}$ 的牵引矩阵为 $\mathcal{G}=\text{diag}\{g_i\}, i=1,\cdots,N$ 。

2.2 MAS 的输出同步

考虑如下线性离散时间异构 MAS:

$$\begin{aligned} \mathbf{x}_i(k+1) &= \mathbf{A}_i \mathbf{x}_i(k) + \mathbf{B}_i \mathbf{u}_i(k) \\ \mathbf{y}_i(k) &= \mathbf{C}_i \mathbf{x}_i(k), i=1,\cdots,N \end{aligned} \quad (1)$$

其中, N 代表智能体的个数, $\mathbf{x}_i(k)\in\mathbb{R}^{n_i}$ 、 $\mathbf{u}_i(k)\in\mathbb{R}^{m_i}$ 和 $\mathbf{y}_i(k)\in\mathbb{R}^q$ 分别代表智能体 i 的系统状态、控制输入(即分布式控制器)和输出, n_i 、 m_i 、 q 分别表示 $\mathbf{x}_i$ 、 $\mathbf{u}_i$ 、 $\mathbf{y}_i$ 的维度。系统矩阵 $\mathbf{A}_i\in\mathbb{R}^{n_i\times n_i}$ 、 $\mathbf{B}_i\in\mathbb{R}^{n_i\times m_i}$ 、 $\mathbf{C}_i\in\mathbb{R}^{q\times n_i}$ 是时不变的。同时, 假设 $(\mathbf{A}_i, \mathbf{B}_i)$ 是可控的。

领导者的动态表示如下:

$$\begin{aligned} \mathbf{x}_0(k+1) &= \mathbf{A}_0 \mathbf{x}_0(k) \\ \mathbf{y}_0(k) &= \mathbf{C}_0 \mathbf{x}_0(k) \end{aligned} \quad (2)$$

其中, $\mathbf{x}_0(k)\in\mathbb{R}^p$ 和 $\mathbf{y}_0(k)\in\mathbb{R}^q$ 表示领导者的系统状态和输出, p 表示 $\mathbf{x}_0$ 的维度。系统矩阵 $\mathbf{A}_0\in\mathbb{R}^{p\times p}$ 、 $\mathbf{C}_0\in\mathbb{R}^{q\times p}$ 。

根据式(2)和图 $\mathcal{G}$ ，给出 MAS 输出同步问题的通用假设^[20]，具体如下。

假设 1: 领导者的所有特征值都在单位圆上，并且不重复。

假设 2: 图 $\mathcal{G}$ 存在生成树。

对于线性离散时间异构 MAS, 输出同步问题就是设计一个分布式控制器 $\mathbf{u}_i(k)$, 使所有智能体的输出 $\mathbf{y}_i(k)$ 和领导者的输出 $\mathbf{y}_0(k)$ 达到同步。追踪误差 $\mathbf{e}_i(k), i=1,\cdots,N$ 定义为:

$$\mathbf{e}_i(k) = \mathbf{y}_i(k) - \mathbf{y}_0(k) \quad (3)$$

并且追踪误差需要满足:

$$\lim_{k\rightarrow\infty} \mathbf{e}_i(k) = 0, \forall i=1,\cdots,N \quad (4)$$

3 利用强化学习解决 MAS 的输出同步问题

在第 3.1 节定义了一个具有邻居控制输入的性能指标和价值函数，根据定义的价值函数得到 MAS

的贝尔曼 (Bellman) 方程。基于 Bellman 最优原则, 得到一个基于系统模型的最优控制表达式。为了在系统模型未知的情况下也可以控制 MAS 达到同步, 在第3.2节基于强化学习中的Q学习提出了一个求解 MAS 控制器的强化学习算法。

3.1 基于模型的最优控制器

根据式 (3) 和式 (4), 定义需要最小化的性能指标:

$$\begin{aligned} J_i(\mathbf{e}_i(k), \mathbf{u}_i(k), \mathbf{u}_j(k)) = & \\ & \sum_{l=k}^{\infty} \lambda^{l-k} L_i[\mathbf{e}_i(l), \mathbf{u}_i(l), \mathbf{u}_j(l)] = \\ & \sum_{l=k}^{\infty} \lambda^{l-k} [(\mathbf{y}_i(l) - \mathbf{y}_0(l))^{\top} \mathbf{W}_i (\mathbf{y}_i(l) - \mathbf{y}_0(l)) + \\ & \mathbf{u}_i^{\top}(l) \mathbf{R}_i \mathbf{u}_i(l) + \sum_{j \in \mathcal{N}_i} \mathbf{u}_j^{\top}(l) \mathbf{R}_{ij} \mathbf{u}_j(l)] \end{aligned} \quad (5)$$

其中, λ 为折扣因子, $\mathbf{W}_i \in \mathbb{R}^{q \times q}$ 、 $\mathbf{R}_i \in \mathbb{R}^{m_i \times m_i} > \mathbf{0}$ 、 $\mathbf{R}_{ij} \in \mathbb{R}^{m_j \times m_j} \geq \mathbf{0}$ 为对称矩阵。由式 (5) 可知, 智能体 i 的追踪性能不仅取决于其追踪误差和控制输入信息, 还取决于与其相连的邻居智能体 j 的输入信息, 这与 MAS 的拓扑结构相关。因此, 最小化性能指标可以减少每个智能体的控制成本, 并完成 MAS 的共同任务——输出同步。

基于式 (5), 如果控制策略 $\mathbf{u}_i$ 是可允许的 (即控制策略 $\mathbf{u}_i$ 是可以得到并输入系统的), 则每个智能体的价值函数可以定义为:

$$V_i(k) = J_i(\mathbf{e}_i(k), \mathbf{u}_i(k), \mathbf{u}_j(k)) \quad (6)$$

定义增广状态 $\mathbf{X}_i(k) = [\mathbf{x}_i(k)^{\top}, \mathbf{x}_0(k)^{\top}]^{\top} \in \mathbb{R}^{n_i+p}$ 、 $\mathbf{Y}_i(k) = [\mathbf{y}_i(k)^{\top}, \mathbf{y}_0(k)^{\top}]^{\top} \in \mathbb{R}^{2q}$ 。根据式 (1) 和式 (2), 相应的增广系统为:

$$\begin{aligned} \mathbf{X}_i(k+1) &= \bar{\mathbf{A}}_i \mathbf{X}_i(k) + \bar{\mathbf{B}}_i \mathbf{u}_i(k) \\ \mathbf{Y}_i(k) &= \bar{\mathbf{C}}_i \mathbf{X}_i(k) \end{aligned} \quad (7)$$

其中, 系统矩阵 $\bar{\mathbf{A}}_i = \begin{bmatrix} \mathbf{A}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_0 \end{bmatrix} \in \mathbb{R}^{(n_i+p) \times (n_i+p)}$ 、

$\bar{\mathbf{B}}_i = \begin{bmatrix} \mathbf{B}_i \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n_i+p) \times m_i}$ 、 $\bar{\mathbf{C}}_i = \begin{bmatrix} \mathbf{C}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_0 \end{bmatrix} \in \mathbb{R}^{2q \times (n_i+p)}$ 。

根据增广状态 $\mathbf{X}_i(k)$, 价值函数可写为:

$$\begin{aligned} V_i(k) = & \sum_{l=k}^{\infty} \lambda^{l-k} (\mathbf{X}_i^{\top}(l) \bar{\mathbf{W}}_i \mathbf{X}_i(l) + \mathbf{u}_i^{\top}(l) \mathbf{R}_i \mathbf{u}_i(l) + \\ & \sum_{j \in \mathcal{N}_i} \mathbf{u}_j^{\top}(l) \mathbf{R}_{ij} \mathbf{u}_j(l)) \end{aligned} \quad (8)$$

其中, $\bar{\mathbf{W}}_i = \begin{bmatrix} \mathbf{C}_i^{\top} \mathbf{W}_i \mathbf{C}_i & -\mathbf{C}_i^{\top} \mathbf{W}_i \mathbf{C}_0 \\ -\mathbf{C}_0^{\top} \mathbf{W}_i \mathbf{C}_i & \mathbf{C}_0^{\top} \mathbf{W}_i \mathbf{C}_0 \end{bmatrix} \in \mathbb{R}^{(n_i+p) \times (n_i+p)}$ 。

注释 1: 不同于已有的研究异构 MAS 输出同步问题的文章 (如参考文献[16-17,19]), 本文的价值函数包含邻居智能体 $j(j \in \mathcal{N}_i)$ 的控制输入 $\mathbf{u}_j$ 和邻居控制权重矩阵 $\mathbf{R}_{ij}$ 。可以通过调节 $\mathbf{R}_{ij}$ 来调节 $\mathbf{u}_j$ 在价值函数中的占比。因此, 本文可以在减少自身控制成本的同时, 减少邻居智能体的控制成本。

对于由线性系统 (式 (1)) 和二次价值函数 (式 (8)) 构成的线性二次追踪问题, 其价值函数在状态上是二次的。此外, 因为价值函数 (式 (8)) 考虑了邻居智能体的控制输入, 所以智能体 i 的价值函数与由自身状态和邻居状态构成的二次型函数近似。因此, 价值函数 (式 (8)) 可以近似表示为如下二次型形式:

$$V_i(k) = \begin{bmatrix} \mathbf{X}_i(k)^{\top} & \sum_{j \in \mathcal{N}_i} \mathbf{X}_j(k)^{\top} \end{bmatrix} \mathbf{P}_i \begin{bmatrix} \mathbf{X}_i(k) \\ \sum_{j \in \mathcal{N}_i} \mathbf{X}_j(k) \end{bmatrix} \quad (9)$$

其中, $\mathbf{P}_i = \begin{bmatrix} \mathbf{P}_i^{11} & \mathbf{P}_i^{12} \\ \mathbf{P}_i^{21} & \mathbf{P}_i^{22} \end{bmatrix} = \mathbf{P}_i^{\top} > \mathbf{0}$ 且 $\mathbf{P}_i^{11} \in \mathbb{R}^{(n_i+p) \times (n_i+p)}$ 、 $\mathbf{P}_i^{12} \in \mathbb{R}^{(n_i+p) \times \sum_{j \in \mathcal{N}_i} (n_j+p)}$ 、 $\mathbf{P}_i^{21} \in \mathbb{R}^{\sum_{j \in \mathcal{N}_i} (n_j+p) \times (n_i+p)}$ 、 $\mathbf{P}_i^{22} \in \mathbb{R}^{\sum_{j \in \mathcal{N}_i} (n_j+p) \times \sum_{j \in \mathcal{N}_i} (n_j+p)}$ 。注意矩阵 $\mathbf{P}_i$ 是构成价值函数 $V_i(k)$ 二次型形式的内核矩阵, 下一节将用矩阵 $\mathbf{P}_i$ 推导出 MAS 的 Q 学习算法。

结合 Bellman 原则, 可以得到 MAS 的 Bellman 方程:

$$\begin{aligned} V_i(k) = & \mathbf{X}_i^{\top}(k) \bar{\mathbf{W}}_i \mathbf{X}_i(k) + \mathbf{u}_i^{\top}(k) \mathbf{R}_i \mathbf{u}_i(k) + \\ & \sum_{j \in \mathcal{N}_i} \mathbf{u}_j^{\top}(k) \mathbf{R}_{ij} \mathbf{u}_j(k) + \lambda V_i(k+1) \end{aligned} \quad (10)$$

根据 Bellman 最优原则, 可以得到 MAS 的 Bellman 最优方程:

$$\begin{aligned} V_i^*(k) = & \min_{\mathbf{u}_i(k)} \{ \mathbf{X}_i^{\top}(k) \bar{\mathbf{W}}_i \mathbf{X}_i(k) + \mathbf{u}_i^{\top}(k) \mathbf{R}_i \mathbf{u}_i(k) + \\ & \sum_{j \in \mathcal{N}_i} \mathbf{u}_j^{\top}(k) \mathbf{R}_{ij} \mathbf{u}_j(k) + \lambda V_i^*(k+1) \} \end{aligned} \quad (11)$$

相应的最优控制为:

$$\begin{aligned} \mathbf{u}_i^*(k) = & \arg \min_{\mathbf{u}_i(k)} \{ \mathbf{X}_i^{\top}(k) \bar{\mathbf{W}}_i \mathbf{X}_i(k) + \mathbf{u}_i^{\top}(k) \mathbf{R}_i \mathbf{u}_i(k) + \\ & \sum_{j \in \mathcal{N}_i} \mathbf{u}_j^{\top}(k) \mathbf{R}_{ij} \mathbf{u}_j(k) + \lambda V_i^*(k+1) \} \end{aligned} \quad (12)$$

为了求得最优控制的具体表达式, 将价值函数的

二次型形式(式(9))代入式(10),得到:

$$\begin{aligned} & \left[X_i(k)^\top \sum_{j \in N_i} X_j(k)^\top \right] P_i \left[\sum_{j \in N_i} X_j(k) \right] = \\ & X_i^\top(k) \bar{W}_i X_i(k) + u_i^\top(k) R_i u_i(k) + \sum_{j \in N_i} u_j^\top(k) R_{ij} u_j(k) + \\ & \lambda \left[X_i(k+1)^\top \sum_{j \in N_i} X_j(k+1)^\top \right] P_i \left[\sum_{j \in N_i} X_j(k+1) \right] \end{aligned} \quad (13)$$

由此定义相应的哈密顿(Hamiltonian)函数 $H_i(k)$ 为:

$$\begin{aligned} H_i(k) = & X_i^\top(k) \bar{W}_i X_i(k) + u_i^\top(k) R_i u_i(k) + \sum_{j \in N_i} u_j^\top(k) R_{ij} u_j(k) + \\ & \lambda \left[X_i(k+1)^\top \sum_{j \in N_i} X_j(k+1)^\top \right] P_i \left[\sum_{j \in N_i} X_j(k+1) \right] - \\ & \left[X_i(k)^\top \sum_{j \in N_i} X_j(k)^\top \right] P_i \left[\sum_{j \in N_i} X_j(k) \right] \end{aligned} \quad (14)$$

将式(7)代入式(14),通过

$$\frac{\partial H_i(k)}{\partial u_i(k)} = 0 \quad (15)$$

可以求得最优控制的具体表达式为:

$$\begin{aligned} u_i^*(k) = & \bar{K}_i \eta_i(k) = \\ & \bar{K}_i^1 X_i(k) + \bar{K}_i^2 \sum_{j \in N_i} X_j(k) + \bar{K}_i^3 \sum_{j \in N_i} u_j(k) \end{aligned} \quad (16)$$

其中, $\bar{K}_i$ 、 $\bar{K}_i^1$ 、 $\bar{K}_i^2$ 、 $\bar{K}_i^3$ 分别为控制器的控制增益:

$$\begin{aligned} \bar{K}_i^1 = & -\lambda(R_i + \lambda \bar{B}_i^\top P_i^{11} \bar{B}_i)^{-1} \bar{B}_i^\top P_i^{11} \bar{A}_i \\ \bar{K}_i^2 = & -\lambda(R_i + \lambda \bar{B}_i^\top P_i^{11} \bar{B}_i)^{-1} \bar{B}_i^\top P_i^{12} \sum_{j \in N_i} \bar{A}_j \\ \bar{K}_i^3 = & -\lambda(R_i + \lambda \bar{B}_i^\top P_i^{11} \bar{B}_i)^{-1} \bar{B}_i^\top P_i^{12} \sum_{j \in N_i} \bar{B}_j \end{aligned} \quad (17)$$

并且 P_i 满足式(13)。

由式(17)可知,所获得的控制器是基于系统模型(A_i, B_i, C_i)的,它无法在系统模型未知的情况下完成 MAS 的输出同步控制。因此,下一节将给出一个基于数据的强化学习算法,通过该算法进一步求解 MAS 的输出同步问题。

3.2 基于 Q 学习的 MAS 强化学习算法

在第 3.1 节针对模型已知的情况给出了一个需要系统模型的控制器。但是,在一些实际系统中,被控 MAS 的精确模型是无法获取的。为了提高系统的适应性和学习性,本节提出了一个基于系统数据的 Q 学

习算法,通过该算法求解 MAS 的同步控制器。

根据 Bellman 式(式(10)),定义 MAS 的 Q 函数为:

$$Q_i(k) = X_i^\top(k) \bar{W}_i X_i(k) + u_i^\top(k) R_i u_i(k) + \sum_{j \in N_i} u_j^\top(k) R_{ij} u_j(k) + \lambda V_i(k+1) \quad (18)$$

结合增广系统(式(7))和价值函数(式(9)),式(18)的 Q 函数变为:

$$\begin{aligned} Q_i(k) = & X_i^\top(k) \bar{W}_i X_i(k) + u_i^\top(k) R_i u_i(k) + \\ & \sum_{j \in N_i} u_j^\top(k) R_{ij} u_j(k) + \lambda s_i^\top(k) U^\top P_i U_i s_i(k) = \\ & s_i^\top(k) E_i s_i(k) \end{aligned} \quad (19)$$

其中, $s_i(k) = \begin{bmatrix} u_i^\top(k) & X_i^\top(k) & \sum_{j \in N_i} X_j^\top(k) & \sum_{j \in N_i} u_j^\top(k) \end{bmatrix}^\top$,

$$U_i = \begin{bmatrix} \bar{B}_i & \bar{A}_i & 0 & 0 \\ 0 & 0 & \sum_{j \in N_i} \bar{A}_j & \sum_{j \in N_i} \bar{B}_j \end{bmatrix}, E_i = \begin{bmatrix} R_i & 0 & 0 & 0 \\ * & \bar{W}_i & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & \sum_{j \in N_i} R_{ij} \end{bmatrix} + \lambda U_i^\top P_i U_i = \begin{bmatrix} E_{uu}^i & E_{uX}^i & E_{u\setminus X}^i & E_{u\setminus u}^i \\ * & E_{XX}^i & E_{X\setminus X}^i & E_{X\setminus u}^i \\ * & * & E_{\setminus X\setminus X}^i & E_{\setminus X\setminus u}^i \\ * & * & * & E_{\setminus u\setminus u}^i \end{bmatrix}。$$

然后根据

$$\frac{\partial Q_i(k)}{\partial u_i(k)} = 0 \quad (20)$$

得到控制器:

$$u_i(k) = K_i \eta_i(k) = -\left(E_{uu}^i\right)^{-1} \begin{bmatrix} E_{uX}^i & E_{u\setminus X}^i & E_{u\setminus u}^i \end{bmatrix} \eta_i(k) \quad (21)$$

根据式(21),要想得到控制器 $u_i(k)$,就要先求矩阵 E_i 。因为

$$V_i^*(k) = \min_{u_i(k)} Q_i(k) = Q_i^*(k) \quad (22)$$

所以 Q 函数的 Bellman 方程为:

$$Q_i(k) = X_i^\top(k) \bar{W}_i X_i(k) + u_i^\top(k) R_i u_i(k) + \sum_{j \in N_i} u_j^\top(k) R_{ij} u_j(k) + \lambda Q_i(k+1) \quad (23)$$

根据式(19),式(23)变为:

$$\begin{aligned} s_i^\top(k) E_i s_i(k) = & X_i^\top(k) \bar{W}_i X_i(k) + u_i^\top(k) R_i u_i(k) + \\ & \sum_{j \in N_i} u_j^\top(k) R_{ij} u_j(k) + \lambda s_i^\top(k+1) E_i s_i(k+1) \end{aligned} \quad (24)$$

由此, 可以通过式 (24) 求得矩阵 E_i , 进而求得控制器 (式 (21))。然而, 此时求得的控制要求每个智能体都能知道领导者的信息 $\mathbf{x}_0(k)$, 这在不同的拓扑结构下是无法保证的, 因此需要引入一个分布式观测器 $\mathbf{x}_0(k)$:

$$\mathbf{x}_0^i(k+1) = \mathbf{A}_0 \mathbf{x}_0^i(k) + \mathbf{c}(1 + d_i + g_i)^{-1} \mathbf{M} \times \left[\sum_{j=1}^N a_{ij} (\mathbf{x}_0^j(k) - \mathbf{x}_0^i(k)) + g_i (\mathbf{x}_0(k) - \mathbf{x}_0^i(k)) \right] \quad (25)$$

其中, $\mathbf{c}$ 是一个正常数, $\mathbf{M}$ 是一个待设计的矩阵。通过观测器, 未被牵引到领导者的智能体 i 也可以获得领导者的估计信息。

定义观测误差 $\tilde{\mathbf{x}}_0^i(k) = \mathbf{x}_0^i(k) - \mathbf{x}_0(k)$, 所以

$$\mathbf{X}_i(k) = \hat{\mathbf{X}}_i(k) - \tilde{\mathbf{X}}_i(k) = \begin{bmatrix} \mathbf{x}_i(k) \\ \mathbf{x}_0^i(k) \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{x}}_0^i(k) \end{bmatrix} \quad (26)$$

基于式 (26), $\mathbf{s}_i(k)$ 和 $\boldsymbol{\eta}_i(k)$ 可以改写为:

$$\mathbf{s}_i(k) = \begin{bmatrix} \mathbf{u}_i^\top(k) \\ \hat{\mathbf{X}}_i^\top(k) \\ \sum_{j \in N_i} \hat{\mathbf{X}}_j^\top(k) \\ \sum_{j \in N_i} \mathbf{u}_j^\top(k) \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{X}}_i^\top(k) \\ \sum_{j \in N_i} \tilde{\mathbf{X}}_j^\top(k) \\ \mathbf{0} \end{bmatrix} = \hat{\mathbf{s}}_i(k) - \tilde{\mathbf{s}}_i(k) \quad (27)$$

$$\boldsymbol{\eta}_i(k) = \begin{bmatrix} \hat{\mathbf{X}}_i(k) \\ \sum_{j \in N_i} \hat{\mathbf{X}}_j(k) \\ \sum_{j \in N_i} \mathbf{u}_j(k) \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{X}}_i(k) \\ \sum_{j \in N_i} \tilde{\mathbf{X}}_j(k) \\ \mathbf{0} \end{bmatrix} = \hat{\boldsymbol{\eta}}_i(k) - \tilde{\boldsymbol{\eta}}_i(k) \quad (28)$$

需要说明的是, 本文引入的观测器 (式 (25)) 在参考文献[19]中已有研究。考虑到篇幅的限制, 笔者进行了简要的描述。根据参考文献[19], 观测误差 $\tilde{\mathbf{x}}_0^i(k)$ 是稳定的, 因此当 $k \rightarrow \infty$ 时, $\tilde{\mathbf{X}}_i(k) \rightarrow \mathbf{0}$ 。

结合式 (21)、式 (24) 及式 (26) ~ (28), 可用如下基于 Q 学习的强化学习算法求解 MAS 同步控制器。

算法 1 基于 Q 学习的强化学习算法

步骤 1: 初始化。设置迭代数 $l=0$, 选择任意的控制策略 $\mathbf{u}_i^0(k), i=1, \dots, N$ 。

步骤 2: 价值更新。

$$\hat{\mathbf{s}}_i^\top(k) \mathbf{E}^{l+1} \hat{\mathbf{s}}_i(k) = \hat{\mathbf{X}}_i^\top(k) \bar{\mathbf{W}}_i \hat{\mathbf{X}}_i(k) + \mathbf{u}_i^l(k)^\top \mathbf{R}_i \mathbf{u}_i^l(k) + \sum_{j \in N_i} \mathbf{u}_j^l(k)^\top \mathbf{R}_j \mathbf{u}_j^l(k) + \lambda \hat{\mathbf{s}}_i^\top(k+1) \mathbf{E}^l \hat{\mathbf{s}}_i(k+1) \quad (29)$$

步骤 3: 策略提升。

$$\mathbf{u}_i^{l+1}(k) = \mathbf{K}_i^{l+1} \hat{\boldsymbol{\eta}}_i(k) = - \left(\left(\mathbf{E}_{uu}^i \right)^{l+1} \right)^{-1} \left[\mathbf{E}_{uX}^i \quad \mathbf{E}_{u \setminus X}^i \quad \mathbf{E}_{u \setminus u}^i \right]^{l+1} \hat{\boldsymbol{\eta}}_i(k) \quad (30)$$

步骤 4: 判断终止条件。设 ε 是一个小的正值, 如果 $|(K_i^{l+1} - (K_i^l)^l)| \leq \varepsilon$, 停止迭代, 否则 $l=l+1$, 并跳转到步骤 2。

注释 2: 从算法 1 可以看出, 式 (30) 基于系统数据求解异构 MAS 的输出同步问题。因此, 本文通过强化学习中的 Q 学习克服了已有研究需要系统模型的弊端^[12-13]。

4 仿真示例

本节给出一个仿真示例来验证所提方法的有效性。对于线性离散时间异构 MAS (式 (1)), $N=3$, 其中,

$$\begin{aligned} \mathbf{A}_1 &= [2], \quad \mathbf{B}_1 = [3], \quad \mathbf{C}_1 = [1], \\ \mathbf{A}_2 &= \begin{bmatrix} -1 & 2 \\ 2.2 & 1.7 \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} 2 \\ 1.6 \end{bmatrix}, \quad \mathbf{C}_2 = [1 \quad 2], \\ \mathbf{A}_3 &= \begin{bmatrix} 1 & 1 \\ 10 & 3 \end{bmatrix}, \quad \mathbf{B}_3 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \quad \mathbf{C}_3 = [1 \quad 1] \end{aligned} \quad (32)$$

对于领导者动态 (式 (2)), 其中,

$$\mathbf{A}_0 = \begin{bmatrix} \cos(0.1) & \sin(0.1) \\ -\sin(0.1) & \cos(0.1) \end{bmatrix}, \quad \mathbf{C}_0 = [1 \quad 1] \quad (33)$$

从式 (31) 和式 (32) 可以看出, 系统动态矩阵 $\mathbf{A}_i, i=1, \dots, N$ 和领导者动态矩阵 $\mathbf{A}_0$ 的维度和数值是不同的, 因此可以通过上述示例验证异构 MAS 的输出同步问题。MAS 的有向拓扑如图 1 所示。

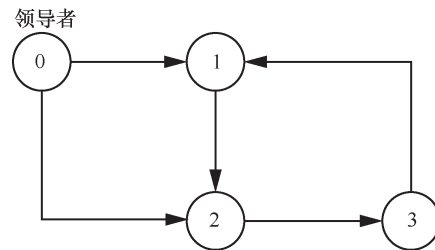


图 1 MAS 的有向拓扑

从以上描述可知, $n_1=1, n_2=n_3=2, q=1, m_1=m_2=m_3=1, p=2$, 则由图 1 可知,

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

式 (15) 中参数如下: $\lambda=0.9$ 、 $R_i=10$ 、 $W_i=40$, $i=1, \dots, N$ 、 $R_{13}=10$ 、 $R_{21}=10$ 、 $R_{32}=10$ 。

基于以上信息, 通过算法 1 可以获得系统的控制律。图 2 和图 3 分别是控制律作用于系统上的输出和跟踪误差。最终, 智能体的输出 $y_i(k)$, $i=1, \dots, N$ 追踪到了领导者的输出 $y_0(k)$, 跟踪误差趋于 0, 从而解决了异构 MAS 的输出同步问题, 验证了本文方法的有效性。

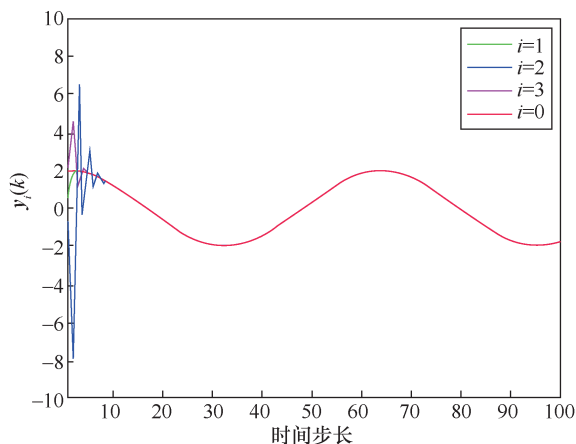


图2 智能体的输出 $y_i(k)$ 和领导者的输出 $y_0(k)$

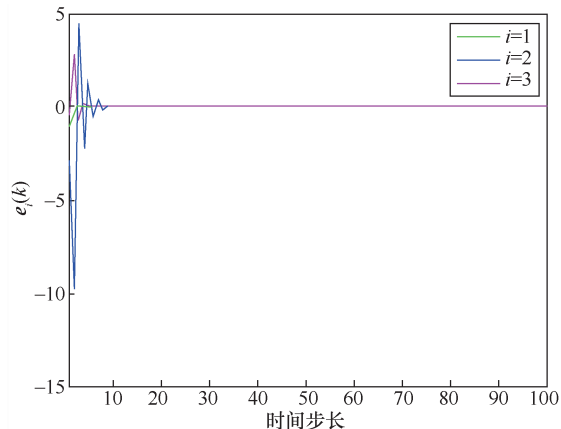


图3 跟踪误差 $e_i(k)$

为了进一步验证注释 1, 选取 $R_{13}=0.01$ 、 $R_{21}=0.01$ 、 $R_{32}=0.01$, 即智能体 i 的价值函数中几乎没有邻居智能体的输入。图 4~图 6 展示了 MAS 控制过程中不同 R_{ij} , $j \in \mathcal{N}_i$ 下控制输入的变化。

由图 4~图 6 可知, 小权重 $R_{ij}=0.01$ 对应大的控制输入幅值, 大权重 $R_{ij}=10$ 对应小的控制输入幅值。因此, 可以通过调节 R_{ij} 减少每个智能体的控制成本, 从而验证了本文定义的具有邻居控制输入的性能指标和价值函数的优越性。

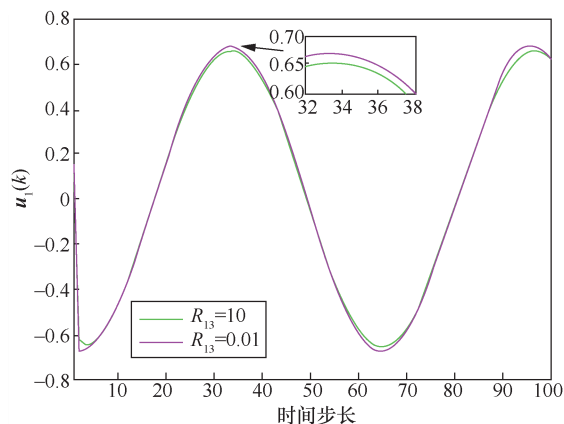


图4 不同权重矩阵下 $u_1(k)$ 的变化

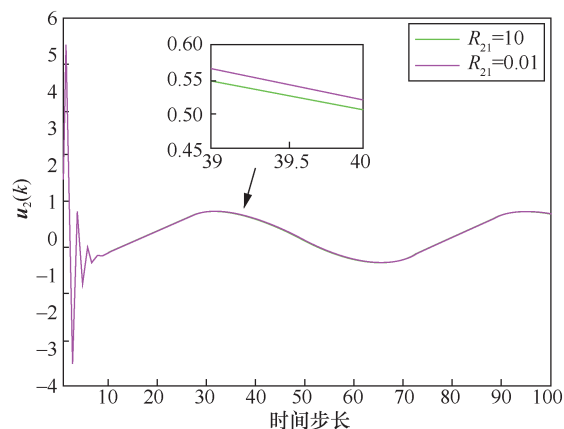


图5 不同权重矩阵下 $u_2(k)$ 的变化

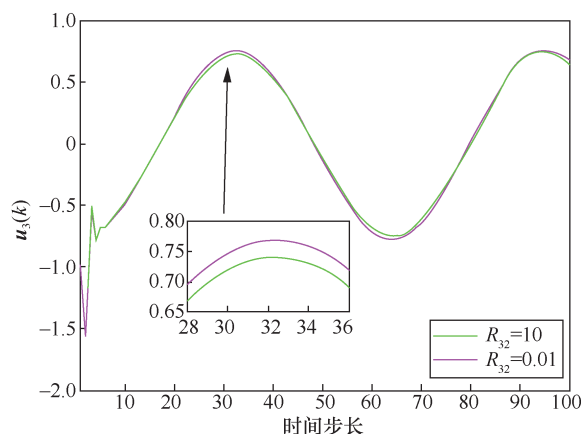


图6 不同权重矩阵下 $u_3(k)$ 的变化

5 结束语

本文研究了异构 MAS 的输出同步问题。首先, 通过定义一个具有邻居控制输入的性能指标和价值函数, 得到 MAS 的 Bellman 方程; 然后, 根据最优原则得到基于模型的 MAS 输出同步问题的控制器; 为了使控制器也可以应用于模型未知的情况, 提出一种基于 Q 学习的强化学习算法; 最后,

给出一个仿真示例,并验证了本文方法的有效性。通过调节权重矩阵,体现了本文定义的价值函数的优越性。

参考文献:

- [1] LI X W, SOH Y C, XIE L H, et al. Cooperative output regulation of heterogeneous linear multi-agent networks via H_∞ performance allocation[J]. IEEE Transactions on Automatic Control, 2019, 64(2): 683-696.
- [2] 孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报, 2020, 7(46): 1301-1312.
SUN C Y, MU C X. Important scientific problems of multi-agent deep reinforcement learning[J]. Acta Automatica Sinica, 2020, 7(46): 1301-1312.
- [3] 裴国旭, 杜晓明, 薛昭, 等. 多智能体系统在军事仿真领域的应用现状[J]. 飞航导弹, 2017(2): 46-49, 73.
PEI G X, DU X M, XUE Z, et al. Application status of multi-agent system in military simulation field[J]. Aerodynamic Missile Journal, 2017(2): 46-49, 73.
- [4] ABDELKADER A, ABDELHAMID T. Distributed output regulation of heterogeneous linear multi-agent systems with communication constraints[J]. Automatica, 2018, 91: 152-158.
- [5] 史乐, 李辉, 原江波. 基于消息通信的多智能体系统的应用[J]. 计算机应用, 2008, 28(2): 531-534.
SHI L, LI H, YUAN J B. Multi-agent system based on the message communication[J]. Journal of Computer Applications, 2008, 28(2): 531-534.
- [6] QIN J H, MA Q C, SHI Y, et al. Recent advances in consensus of multi-agent systems: a brief survey[J]. IEEE Transactions on Industrial Electronics, 2017, 64(6): 4972-4983.
- [7] ZHANG H G, JIANG H, LUO Y H, et al. Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method[J]. IEEE Transactions on Industrial Electronics, 2017, 64(5): 4091-4100.
- [8] WANG X, YANG G H. Fault-tolerant consensus tracking control for linear multi-agent systems under switching directed network[J]. IEEE Transactions on Cybernetics, 2020, 50(5): 1921-1930.
- [9] ZHANG J C, ZHU F L. Observer-based output consensus of a class of heterogeneous multi-agent systems with unmatched disturbances[J]. Communications in Nonlinear Science and Numerical Simulation, 2018, 56: 240-251.
- [10] LIU L. Adaptive cooperative output regulation for a class of nonlinear multi-agent systems[J]. IEEE Transactions on Automatic Control, 2015, 60(6): 1677-1682.
- [11] MA Q, QIN J, ZHENG W X, et al. Output group synchronization for networks of heterogeneous linear systems under internal model principle[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2018, 65(5): 1684-1695.
- [12] YAN Y M, HUANG J. Cooperative output regulation of discrete-time linear time-delay multi-agent systems[J]. IET Control Theory & Applications, 2016, 10(16): 2019-2026.
- [13] 王飞跃, 曹东璞, 魏庆来. 强化学习: 迈向知行合一的智能机制与算法[J]. 智能科学与技术学报, 2020, 2(2): 101-106.
WANG F Y, CAO D P, WEI Q L. Reinforcement learning: toward action-knowledge merged intelligent mechanisms and algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(2): 101-106.
- [14] LEWIS F L, VRABIE D. Reinforcement learning and adaptive dynamic programming for feedback control[J]. IEEE Circuits and Systems Magazine, 2009, 9(3): 32-50.
- [15] LIU Y Y, WANG Z S, SHI Z. H_∞ tracking control for linear discrete-time systems via reinforcement learning[J]. International Journal of Robust and Nonlinear Control, 2020, 30(1): 282-301.
- [16] MODARES H, NAGESHRAO S P, LOPES G A D, et al. Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning[J]. Automatica, 2016, 71: 334-341.
- [17] MODARES H, LEWIS F L, KANG W, et al. Optimal synchronization of heterogeneous nonlinear systems with unknown dynamics[J]. IEEE Transactions on Automatic Control, 2018, 63(1): 117-131.
- [18] KIUMARSI B, LEWIS F L, MODARES H, et al. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics[J]. Automatica, 2014, 50: 1167-1175.
- [19] KIUMARSI B, LEWIS F L. Output synchronization of heterogeneous discrete-time systems: A model-free optimal approach[J]. Automatica, 2017, 84: 86-94.
- [20] YANG Y L, MODARES H, WUNSCH D C, et al. Leader-follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2139-2153.

[作者简介]



刘莹莹(1995—),女,东北大学信息科学与工程学院博士生,主要研究方向为数据驱动的系统无模型控制、基于强化学习的系统最优控制、多智能体系统最优控制等。



王占山(1971—),男,博士,东北大学信息科学与工程学院教授、博士生导师,主要研究方向为神经动力系统稳定性理论、复杂网络同步性与一致性以及数据驱动的智能故障诊断与自适应容错控制等。