

基于强化学习的数据驱动多智能体系统最优一致性综述

李金娜, 程薇燃

(辽宁石油化工大学信息与控制工程学院, 辽宁 抚顺 113000)

摘要: 多智能体系统因其在工程、社会科学和自然科学等多学科领域具有潜在、广泛的应用性, 在过去的 20 年里引起了研究者的广泛关注。实现多智能体系统的一致性通常要求解相关矩阵方程离线设计控制协议, 这要求系统模型精确已知。然而, 实际上多智能体系统具有大规模尺度、非线性耦合性特征, 并且环境动态变化, 使得系统精确建模非常困难, 这给模型依赖的多智能体一致性控制协议设计带来了挑战。强化学习技术因其可以利用沿系统轨迹的测量数据实时学习控制问题的最优解, 被广泛用于解决复杂系统最优控制和决策问题。综述了利用强化学习技术, 采用数据驱动方式实时在线求解多智能体系统最优一致性控制问题的现有理论和方法, 分别从连续和离散、同构和异构、抗干扰的鲁棒性等多个方面介绍了数据驱动的强化学习技术在多智能体系统最优一致性控制问题中的应用。最后讨论了基于数据驱动的多智能体系统最优一致性问题的未来研究方向。

关键词: 强化学习; 多智能体系统; 最优一致性; 数据驱动

中图分类号: O232

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202035

An overview of optimal consensus for data driven multi-agent system based on reinforcement learning

LI Jinna, CHENG Weiran

School of Information and Control Engineering, Liaoning Shihua University, Fushun 113000, China

Abstract: Multi-agent system has attracted extensive attention in the past two decades because of its potential applications in engineering, social science and natural science, etc. To achieving the consensus of multi-agent system, it is usually necessary to solve the correlation matrix equation to design the control protocol offline, which requires system model to be known accurately. However, the actual multi-agent system has the characteristics of large-scale, nonlinear coupling, and dynamic change of environment, which makes it very difficult to accurately model the system. This brings challenges to the design of model dependent multi-agent consensus protocol. Reinforcement learning is widely used to solve the optimal control and decision-making problems of complex systems because it can learn the optimal solution of control problems in real time by using the measurement data along the trajectory of the system. The existing theories and methods of online solving the optimal consensus of multi-agent system in real-time by using reinforcement learning technology were summarized. The application of data-driven reinforcement learning technology in multi-agent system optimal consensus was introduced from the aspects of continuous and discrete, homogenous and heterogeneous, anti-interference robustness and so on. Finally, the future research direction of the optimal consensus problem of multi-agent system based on data-driven technology was discussed.

Key words: reinforcement learning, multi-agent system, optimal consensus, data driven

收稿日期: 2020-11-18; 修回日期: 2020-12-03

通信作者: 李金娜, lijinna_721@126.com

基金项目: 国家自然科学基金资助项目 (No.61673280, No.62073158); 辽宁省重点领域联合开放基金资助项目 (No.2019-KF-03-06); 辽宁石油化工大学研究基金资助项目 (No.2018XJJ-005)

Foundation Items: The National Natural Science Foundation of China (No.61673280, No.62073158), The Open Project of Key Field Alliance of Liaoning Province (No.2019-KF-03-06), The Project of Liaoning Shihua University (No.2018XJJ-005)

1 引言

多智能体系统由于具有更好的鲁棒性、灵活性和可扩展性,在工程、社会科学和自然科学等多学科领域被广泛地应用^[1-2]。多智能体一致性问题的目标是根据每个智能体及其相邻智能体的局部信息制定控制协议,使所有智能体在一定数量的收益量上达成一致或跟踪一个参考轨迹^[3-6]。近年来,对多智能体的分布式一致性问题的研究因得到各界的广泛关注而迅速发展,在群集^[7]、智能电网^[8]、卫星群^[9]和无人机^[10]等各个领域应用广泛。大量文献提出了各种分布式一致性控制协议来保证系统的一致性^[3-6,11]。随着对一致性问题的深入研究,在获得各个智能体一致性的同时,人们开始研究最小化由智能体之间的局部邻居误差和控制输入组成的预设的性能指标的技术方法,该性能指标代表了多智能体系统的能耗、效率和精度。换言之,每个智能体通过计算局部信息(即自身和邻居的信息)来寻找分布式一致性问题的最优解,这被称作最优一致性控制问题。近年来,大量学者沿着不同的思路 and 方向对多智能体系统最优一致性问题进行了研究,从连续系统^[12-14]和离散系统^[15-16]、有领导者^[17-18]和无领导者^[19]、同构^[20-21]和异构^[22-23]等多个方面进行了研究。研究多智能体的分布式最优一致性问题是发挥多智能体系统可以用更灵活更智能的方式处理复杂的问题、解决单智能体不能解决的大规模复杂问题等优势的关键,具有重要的现实意义和理论价值。

目前许多国内外研究者致力于对多智能体系统的分布式控制的研究,对于离散的多智能体系统的分布式一致性问题的研究^[24-27]存在有限时间^[24,28]、参数调节^[27]、时延^[25]等不同的研究问题,并且对异构多智能体系统的研究更广泛^[28-33]。随着自适应动态规划方法的不断发展,这种控制方法也被深入应用到了多智能体的最优一致性问题上^[29-30]。但是随着多智能体系统规模的不断扩大以及数据的不断增多,需要发展更智能的控制技术来解决这些困难。

强化学习(reinforcement learning, RL)是机器学习的一个子领域,其受哺乳动物学习机制的启发,研究如何根据观察到的来自环境的响应系统地修改智能体的行为。强化学习算法指与环境相互作用的智能体利用环境的响应来学习最优控制策略,并从未知环境中找出最优行为的算法^[34]。1991年, Werbos P J^[35]较早提出了基于强化学习的自适应动

态规划(adaptive dynamic programming, ADP)控制,首创采用强化学习技术求解离散系统的最优调节器。这种控制算法采用策略迭代(policy iteration, PI)技术,仅要求对系统动力学有部分了解。2009年, Doya K^[36]提出了将强化学习技术应用到求解连续时间系统的控制器中。参考文献[37-38]提出了利用积分强化学习(integral reinforcement learning, IRL)的在线学习算法,用于解决系统模型部分未知的线性或非线性系统的最优跟踪控制问题。强化学习算法也被用来求解 H_∞ 控制问题,例如利用强化学习算法解决 H_∞ 控制问题的Q-学习方法^[39]以及神经动态规划方法^[40]。近年来,将强化学习技术应用于多智能体系统的最优一致性控制的研究越来越受到人们的关注。参考文献[41-42]引入基于模型的强化学习算法来求解多人微分对策问题。参考文献[43-44]考虑了最优协同控制问题,采用基于模型的自适应动态规划方法来学习耦合哈密顿-雅可比-贝尔曼(Hamilton-Jacobi-Bellman, HJB)方程的解。

需要注意的是,参考文献[41-44]提出的利用强化学习算法解决多智能体系统最优一致性的方法要求系统模型已知或者部分已知,并使用策略学习方法求解最优一致性协议。随着数字传感器技术的迅速发展和广泛应用,人们可以采集到大量承载系统信息的数据。人们希望利用这些数据开发数据驱动的最优控制协议。数据驱动技术已经被广泛应用于解决单智能体系统的最优控制问题^[45-47],与策略强化学习(on-policy RL)相比,非策略强化学习(off-policy RL)中引进了行为策略和目标策略,通过行为策略来生成系统的数据,在丰富数据挖掘的同时更新目标策略,以寻找最优策略。非策略强化学习算法克服了策略强化学习算法在应用中产生的两个缺点:一是数据只能由一种特定的方法生成,导致数据挖掘能力非常有限;二是为了充分激励系统,在目标策略中加入探测噪声,使贝尔曼方程的解产生偏差。非策略强化学习是一种更实用、更有效的处理最优控制问题的技术。随着数据处理技术和人工智能技术的深入发展,数据驱动的强化学习技术也越来越多地被应用到多智能体的最优一致性问题上。

基于数据驱动的强化学习技术利用实时测量的数据,在系统动态完全未知的情况下,解决多智能体系统的最优一致性控制协议设计或者决策问题,这对于研究更具复杂性的多智能体系统具有重

大的推动作用,是十分重要的方法。智能体之间的数据交换导致了智能体动力学的耦合性,这给数据驱动的多智能体最优一致性控制问题带来了挑战。本文综述了基于数据驱动方式解决分布式多智能体系统最优一致性问题的主要成果和进展,将该领域的最新研究成果分为连续、离散、同构、异构、含有扰动等最优一致性问题。最后讨论了基于强化学习的数据驱动多智能体最优一致性控制的未来的研究方向和挑战性问题。

2 多智能体系统最优一致性基本理论

本节首先介绍研究多智能体系统最优一致性需要用到的图论理论,然后从有领导者和无领导者两个方面介绍多智能体系统最优一致性的定义,最后介绍解决多智能体系统最优一致性所需要的博弈理论。

2.1 图论理论

图论是一致性问题研究分析中非常重要的工具,通常采用图来表示多智能体之间传递信息的关系。考虑由 N 个多智能体组成的系统,其网络拓扑图为 $G = (V, E)$ 。其中, $V = \{v_1, v_2, v_3, \dots, v_N\}$, 表示网络拓扑图的顶点集; $E \subseteq V \times V$ 表示网络拓扑图的边; $E = [e_{ij}]$ 被称为连通矩阵,如果 $(v_j, v_i) \in E$, 则 $e_{ij} > 0$, 否则 $e_{ij} = 0$ 。 $(v_j, v_i) \in E$ 表示从顶点 v_j 到顶点 v_i 的一条边。定义 $N_i = \{v_j : (v_j, v_i) \in E\}$ 为节点 v_i 的邻居集。 $D = \text{diag}(d_1, \dots, d_N)$ 被称为度矩阵,其中 $d_i = \sum_j e_{ij}$ 为顶点 v_i 的加权重。定义网络拓扑图的拉普拉斯矩阵为 $L = D - E$ [48]。

2.2 多智能体的最优一致性问题

在解决多智能体的最优一致性问题时,通常将其分为有领导者和无领导者两类进行研究。

2.2.1 有领导者的多智能体系统最优一致性问题

考虑由 N 个智能体组成的多智能体系统,式(1)为智能体的动力学模型:

$$\dot{\mathbf{x}}_i(t) = \mathbf{A}\mathbf{x}_i(t) + \mathbf{B}\mathbf{u}_i(t) \quad (1)$$

其中, $\dot{\mathbf{x}}_i(t) = [\dot{x}_{i1}(t), \dots, \dot{x}_{in}(t)]^\top$, $\mathbf{x}_i(t) = [x_{i1}(t), \dots, x_{in}(t)]^\top$ 和 $\mathbf{u}_i(t) = [u_{i1}(t), \dots, u_{im}(t)]^\top$ (m, n 为正整数) 分别为第 i 个智能体的系统状态导数、状态量和控制输入量。 $\mathbf{A}$ 、 $\mathbf{B}$ 为适当维数的多智能体系统的动态矩阵。

领导者的动态为:

$$\dot{\mathbf{x}}_0 = \mathbf{A}\mathbf{x}_0 \quad (2)$$

其中, $\mathbf{x}_0$ 、 $\dot{\mathbf{x}}_0$ 为领导者系统状态量和状态导数, $\mathbf{A}$ 为领导者系统动态矩阵。

定义 1 [49-55] 多智能体系统的一致性问题是指设计局部控制协议 $\mathbf{u}_i$ 使所有智能体的状态与领导者达成一致,即对于 $\forall i, i = 1, \dots, N$, $\lim_{t \rightarrow \infty} \mathbf{x}_i(t) = \mathbf{x}_0(t)$ 。

根据多智能体系统最优一致性的概念,为了使式(1)的系统达到最优一致性,可以为每个智能体定义一个局部二次性能指标:

$$\int_{t_0}^{\infty} [\mathbf{Q}(\mathbf{x}_i(t)) + \mathbf{u}_i^\top \mathbf{R}_i \mathbf{u}_i] dt \quad (3)$$

其中, $\mathbf{R}_i$ 是正定矩阵。通过最小化预定义的式(3)的性能指标,达到最优控制的目的。

2.2.2 无领导者的多智能体系统最优一致性问题

在多智能体系统中,如果各智能体的地位和作用是平等的,则称这样的系统是无领导者的多智能体系统。

定义 2 [50,56] 设计局部控制协议 $\mathbf{u}_i$, 满足: (a) 所有的状态达到一致,即对于 $\forall i, i = 1, \dots, N$, $\lim_{t \rightarrow \infty} \mathbf{x}_i(t) = \mathbf{x}_j(t)$; (b) 最小化式(3)的性能指标。

2.3 博弈理论

博弈理论是解决分布式问题的有力工具 [57-60], 为研究多人决策和控制问题提供了一个理想的环境。参考文献[54]介绍了多智能体系统的合作控制与博弈理论之间的联系。参考文献[61]将合作控制、强化学习和博弈理论结合起来,提出了一种多智能体团队博弈的在线求解方法。在动态系统中,每个节点的动态性和性能指标只依赖于局部邻居信息,参考文献[61]提出了图形博弈的概念,并给出了“交互式纳什均衡”的新定义。

假设图 G 是强连通的,根据 N 个智能体组成的多智能体系统(式(1))和领导者(式(2))定义局部邻居一致性误差变量:

$$\delta_i = \sum_{j \in N_i} a_{ij} (\mathbf{x}_i - \mathbf{x}_j) + \mathbf{g}_i (\mathbf{x}_i - \mathbf{x}_0) \quad (4)$$

其中, a_{ij} 为连通矩阵, $\mathbf{g}_i$ 为智能体的顶点增益, δ_i 为一致性误差变量。其动态模型为:

$$\dot{\delta} = \mathbf{A}\delta + (\mathbf{d}_i + \mathbf{g}_i)\mathbf{B}\mathbf{u}_i - \sum_{j \in N_i} a_{ij}\mathbf{B}\mathbf{u}_j \quad (5)$$

其中, $\mathbf{d}_i$ 为智能体顶点的加权重。定义一个局部二次性能指标:

$$\int_{t_0}^{\infty} [\delta_i^\top \mathbf{Q}_i \delta_i + \mathbf{u}_i^\top \mathbf{R}_i \mathbf{u}_i] dt \quad (6)$$

其中, Q_i 和 R_i 为适当维数的矩阵。全局纳什均衡的定义如下。

定义 3^[62-64] N 个智能体博弈的全局纳什均衡解是一个 N 元组 $\{u_1^*, u_2^*, \dots, u_N^*\}$, 并满足:

$$J_i^* = J(\delta(t_0), u_i^*, u_{G-i}^*) \leq J(\delta(t_0), u_i, u_{G-i}^*) \quad (7)$$

其中, 全部 i 满足 $i \in N$, 对于 $\forall u_i, u_{G-i} = \{u_j : j \in V, j \neq i\}$, u_{G-i}^* 为最优值。博弈 N 元组的值 $\{J_1^*, J_2^*, \dots, J_N^*\}$ 被称为 N 个智能体博弈的纳什均衡结果。达到一个纳什均衡, 每个参与者的价值函数都能得到优化。在这种纳什均衡下, 没有一个博弈者能够通过改变其输入策略来提高其绩效指标。

3 基于强化学习的数据驱动多智能体最优一致性

在研究多智能体的各种分布式一致性方法时, 遇到的共同难题是缺乏整个系统的全局知识, 每个智能体只能与其邻居进行交互, 以实现某些全局行为。同时, 为了寻找最优策略, 需要求解耦合的哈密顿-雅可比 (Hamilton-Jacobi, HJ) 方程。随着数字传感器技术的迅速发展和广泛应用, 大量的数据可以通过数据驱动的方式进行采集, 这些数据承载了系统信息。研究者希望利用这些数据来开发基于模型数据的最优控制协议。强化学习技术的发展使得研究者可以利用沿系统轨迹的测量数据实时学习优化控制问题的最优解。近几年, 很多研究致力于利用强化学习技术, 通过基于数据驱动的方式来学习多智能体系统一致性问题的最优的控制策略。本节从同构、异构和具有扰动的多智能体系统 3 个方面, 总结近年来利用基于强化学习的数据驱动技术解决多智能体最优一致性的研究成果和结论。

3.1 基于强化学习的同构多智能体系统最优一致性

基于数据驱动的同构多智能体系统的分布式一致性问题涉及许多关于连续时间的系统。参考文献[49]针对多智能体系统的最优同步问题, 提出了一种非策略强化学习算法, 并利用产生的数据, 采用 actor-critic 神经网络和最小二乘法来逼近目标控制策略和值函数, 得到每个智能体近似最优控制策略。参考文献[50]提出了一种非策略强化学习算法, 用可测状态数据代替状态时滞系统动力学知识来学习耦合时变 HJB 方程的两阶段最优一致解。然后, 对于每一个智能体, 利用单临界神经网络来逼近时变值函数, 并帮助计算出最优一致性控制策

略。在加权残差法的基础上, 参考文献[50]提出了基于加权残差的自适应权值更新律。

对于离散时间的同构多智能体系统, 近几十年来, 强化学习算法的策略迭代和值迭代技术被广泛应用于求解离散时间的多智能体系统的最优一致性控制。然而, 大多数物理系统是非线性的, 具有高阶动力学特性, 很难对其进行精确建模。作为基于数据的最优控制方法, QD 学习 (distributed Q-learning)^[51]和 Q 学习^[52]已经被用来学习离散的多智能体系统一致性问题的最优解。对于离散系统, 参考文献[60]给出了线性系统的动态图形对策, 开发了一种算法来求解博弈问题。参考文献[54]提出了一种实时求解离散线性系统动态图形对策的无模型策略迭代算法, 并证明了算法的收敛性。参考文献[55]将这种无模型的策略迭代算法用到了非线性系统优化控制问题中。参考文献[62]研究离散时间多人非零和博弈问题, 提出了一种基于数据驱动的基于动作的启发式动态规划方法。下面分别从连续多智能体系统和离散多智能体系统两个方面介绍基于强化学习的数据驱动多智能体最优一致性控制研究成果。

3.1.1 连续多智能体系统

连续多智能体系统^[49]以式 (1) 的多智能体系统的动态特性为系统模型, 将控制输入解释为依赖于局部邻居跟踪误差的策略, 对应于性能指标 (式 (3)) 的值函数为:

$$V_i(\delta(t_i)) = \int_t^\infty [\delta_i^\top Q \delta_i + u_i^\top R u_i] d\tau \quad (8)$$

其中, δ_i 为局部误差变量, u_i 为智能体控制输入量, Q 和 R 为固定的适当维数矩阵。

得到如下耦合的博弈 HJB 方程:

$$\begin{aligned} H(\delta_i, \nabla V_i^*, u_i^*, u_{-i}^*) = & \delta_i^\top Q \delta_i + \frac{1}{4} (d_i + g_i)^2 (\nabla V_i^*)^\top B R_i^{-1} B^\top \nabla V_i^* + \\ & (\nabla V_i^*)^\top (A \delta_i - \frac{1}{2} (d_i + g_i)^2 B R_i^{-1} B^\top \nabla V_i^*) + \\ & \frac{1}{2} \sum_{j \in N_i} e_{ij} (d_i + g_i) B R_j^{-1} B^\top \nabla V_j^* \end{aligned} \quad (9)$$

其中, $\Delta V_i^* = \frac{\partial V}{\partial \delta_i}$ 为梯度算子, e_{ij} 为连通矩阵, g_i 为智能体的顶点增益, d_i 为智能体顶点的加权重, A 、 B 为适当维数的多智能体系统的动态矩阵。最优控

制策略为:

$$\mathbf{u}_i^*(t) = -\frac{1}{2}(\mathbf{d}_i + \mathbf{g}_i) \mathbf{R}^{-1} \mathbf{B}_i^\top \nabla V_i^* \quad (10)$$

假设值函数是二次函数:

$$V_i = \delta_i^\top \mathbf{P}_i \delta_i \quad (11)$$

其中, $\mathbf{P}_i$ 是正定矩阵。耦合 HJB 方程 (式 (9)) 等价于耦合代数 Riccati 方程:

$$2\delta_i^\top \mathbf{P}_i^\top (\mathbf{A}\delta_i + \sum_{j \in N_i} \mathbf{e}_{ij}(\mathbf{d}_i + \mathbf{g}_i) \mathbf{B} \mathbf{R}_j^{-1} \mathbf{B}^\top \mathbf{P}_i \delta_i) + \delta_i^\top \mathbf{Q} \delta_i - (\mathbf{d}_i + \mathbf{g}_i)^\top \delta_i^\top \mathbf{P}_i^\top \mathbf{B} \mathbf{R}_i^{-1} \mathbf{B}^\top \mathbf{P}_i \delta_i = 0 \quad (12)$$

其中, $\mathbf{R}_i$ 为智能体 i 的输入量动态矩阵, $\mathbf{R}_j$ 为智能体 j 的输入量动态矩阵, 将最优控制 (式 (10)) 改写为:

$$\mathbf{u}_i^*(t) = -(\mathbf{d}_i + \mathbf{g}_i) \mathbf{R}_i^{-1} \mathbf{B}_i^\top \mathbf{P}_i \delta_i \quad (13)$$

为了达到全局纳什均衡, 需要通过求解 N 个智能体博弈问题的 N 个耦合偏微分 HJB 方程 (式 (9)) 来计算每个智能体的最优响应。如果所有的智能体选择自己的最优响应, 并且网络拓扑图是强连通的, 那么系统是渐近稳定的。因此, 所有智能体都是同步的。同时, 这 N 个智能体都处于全局纳什均衡状态。

对每个智能体的动态, 引入辅助变量 $\mathbf{u}_i^{(s)}$ 和 $\mathbf{u}_{-i}^{(s)}$, 有:

$$\begin{aligned} \dot{\delta}_i &= \mathbf{A}\delta_i^\top + (\mathbf{d}_i + \mathbf{g}_i) \mathbf{B} \mathbf{u}_i^{(s)} - \\ &\sum_{j \in N_i} \mathbf{e}_{ij} \mathbf{B} \mathbf{u}_j^{(s)} + (\mathbf{d}_i + \mathbf{g}_i) \mathbf{B} (\mathbf{u}_i - \mathbf{u}_i^{(s)}) + \\ &\sum_{j \in N_i} \mathbf{e}_{ij} \mathbf{B} (\mathbf{u}_j - \mathbf{u}_j^{(s)}) \end{aligned} \quad (14)$$

其中, $\mathbf{u}_i$ 为实际应用于系统的行为策略。 $\mathbf{u}_i^{(s)}$ 为目标策略。

对 $V_i^{(s+1)}$ 进行微分, 通过式 (14) 得到:

$$\begin{aligned} \frac{dV_i^{(s+1)}(\delta_i)}{dt} &= \nabla V_i^{(s+1)} \dot{\delta}_i = \\ &(\nabla V_i^{(s+1)})(\mathbf{A}\delta_i^\top + (\mathbf{d}_i + \mathbf{g}_i) \mathbf{B} \mathbf{u}_i^{(s)} - \sum_{j \in N_i} \mathbf{e}_{ij} \mathbf{B} \mathbf{u}_j^{(s)}) + \\ &(\nabla V_i^{(s+1)})((\mathbf{d}_i + \mathbf{g}_i) \mathbf{B} (\mathbf{u}_i - \mathbf{u}_i^{(s)}) - \sum_{j \in N_i} \mathbf{e}_{ij} \mathbf{B} (\mathbf{u}_j - \mathbf{u}_j^{(s)})) \end{aligned} \quad (15)$$

进一步, 有:

$$\begin{aligned} V_i^{(s+1)}(\delta_i) - V_i^{(s+1)}(\delta_i') &= \\ &\int_t^{t'} [\delta_i(\tau)^\top \mathbf{Q} \delta_i(\tau) + (\mathbf{u}_i^{(s)}(\delta_i(\tau)))^\top \mathbf{R}_i \mathbf{u}_i^{(s)}(\delta_i(\tau))] d\tau - \\ &2 \int_t^{t'} [(\mathbf{u}_i^{(s+1)}(\delta_i(\tau)))^\top \mathbf{R}_i (\mathbf{u}_i^{(s)}(\delta_i(\tau)) - \mathbf{u}_i(\delta_i(\tau)))] d\tau + \\ &2(\mathbf{d}_i + \mathbf{g}_i)^{-1} \int_t^{t'} \sum_{j \in N_i} [\mathbf{e}_{ij} (\mathbf{u}_i^{(s+1)}(\delta_i(\tau)))^\top \mathbf{R}_i \cdot \\ &(\mathbf{u}_j^{(s)}(\delta_i(\tau))) (-\mathbf{u}_j(\delta_i(\tau)))] d\tau \end{aligned} \quad (16)$$

算法 1 给出了非策略积分强化学习算法。

算法 1^[49] 多智能体博弈的非策略强化学习

步骤 1: 对于 $\forall i$ 从一个允许的初始控制策略 $\mathbf{u}_i^0$ 开始计算, 令 s 代表控制策略迭代系数, 并且 $s=0$;

步骤 2: 根据非策略贝尔曼方程 (式 (16)) 求解 V_{i+1} 和 $\mathbf{u}_{i+1}$;

步骤 3: 当 $\|V_i^{(s)} - V_i^{(s+1)}\| \leq \varepsilon$ 时, 停止。

需要注意的是, 算法 1 中不包含系统模型的参数信息, 完全利用系统产生的数据来学习得到最优控制策略, 并且不受模型不精确或辨识系统模型简化的影响。

目前值函数和控制策略估计的方法往往采用最小二乘法或者神经网络估计方法^[49-66]。算法 1 这种非策略强化学习算法具有显著的优点, 其可以在不需要智能体系统的动力学知识、无须系统辨识的情况下学习连续多智能体系统的近似最优控制协议, 在不影响最优控制协议精度的前提下, 消除因为模型辨识不准确所带来的负面影响。不同于现有的基于模型的多智能体系统最优控制, 算法 1 体现了非策略强化学习算法的优势。

3.1.2 离散多智能体系统

考虑一个由通信图 G 描述的具有 N 个智能体的离散时间的多智能体系统。每个智能体的动态如下:

$$\mathbf{x}_i(k+1) = \mathbf{A} \mathbf{x}_i(k) + \mathbf{B} \mathbf{u}_i(k) \quad (17)$$

其中, $\mathbf{x}_i(k) = [\mathbf{x}_{i1}(k), \dots, \mathbf{x}_{im}(k)]^\top$ 和 $\mathbf{u}_i(k) = [\mathbf{u}_{i1}(k), \dots, \mathbf{u}_{im}(k)]^\top$ 分别是第 i 个智能体的系统状态量和控制输入量。

参考文献[63]考虑如下领导者动态:

$$\mathbf{x}_0(k+1) = \mathbf{A} \mathbf{x}_0 \quad (18)$$

定义局部邻居跟踪误差为:

$$\varepsilon_i(k) = \sum_{j \in N} \mathbf{a}_{ij} (\mathbf{x}_i(k) - \mathbf{x}_j(k)) + \mathbf{g}_i (\mathbf{x}_i(k) - \mathbf{x}_0(k)) \quad (19)$$

选择将智能体 i 的输入 u_i 作为加权分布式控制协议:

$$u_i(k) = -c(1 + d_i + g_i)^{-1} K \varepsilon_i(k) \quad (20)$$

其中, c 是一个标量耦合增益, K 是反馈控制增益。

求解离散时间的多智能体系统的最优一致性控制, 得到分布式控制协议 (式 (20)) 需要以下两个部分的计算^[63]。

利用离散时间系统的黎卡提方程:

$$\begin{aligned} P - A^T P A + A^T P B (R + B^T P B)^{-1} B^T P A &= Q \\ K &= (R + B^T P B)^{-1} B^T P A \end{aligned} \quad (21)$$

得到反馈增益 K 。

根据耦合增益 c 需要满足的关系:

$$\max_{k \in N} |1 - c \lambda_k| < \mathfrak{R} \quad (22)$$

$$\mathfrak{R} = \sigma_{\max}^{-\frac{1}{2}} (Q_a^{-T} A^T P B (R + B^T P B) B^T P A Q_a^{-1}) \quad (23)$$

得到耦合增益 c 。

参考文献[63]在系统未知的情况下利用数据驱动的方式解决最优一致性控制问题。首先通过定义 Q 函数, 得到关于 Q 函数的贝尔曼方程:

$$\begin{aligned} Z^T(k) H Z(k) &= x^T(k) Q x(k) + [u^i(k)]^T R [u^i(k)] + \\ &Z^T(k+1) H Z(k+1) \end{aligned} \quad (24)$$

并利用算法, 得到反馈控制器增益 K 。

算法 2^[63] Q 学习无模型强化学习算法

步骤 1: 给定一个控制策略 $u(k) = -K^0 x(k)$, K^0 是使 (A, B) 稳定的控制器增益;

步骤 2: 通过下式求解 H^{i+1} ;

$$\begin{aligned} Z^T(k) H^{i+1} Z(k) &= \\ x^T(k) Q x(k) &+ [u^i(k)]^T R [u^i(k)] + \\ Z^T(k+1) H^{i+1} Z(k+1) \end{aligned}$$

步骤 3: 进行如下计算;

$$u^{i+1}(k) = -(H_{uu}^{i+1})^{-1} H_{ux}^{i+1} x(k)$$

步骤 4: 重复步骤 2、步骤 3, 直到 $\|H_{i+1} - H_i\| < \varepsilon$, ε 为任意的极小值, 停止迭代;

步骤 5: $u(k) = -K_i x(k)$ 为近似最优控制策略, K_i 是最优反馈增益。

其次, 根据算法 2, 若利用数据驱动方式求解多智能体一致性, 还需要计算可行的耦合增益 c , 为得到 $\mathfrak{R}_i$, 需要计算式 (25):

$$\Gamma := A^T P B (R + B^T P B)^{-1} B^T P A \quad (25)$$

由 Q 函数, 得到矩阵 $\mathfrak{R}_i$ 的最大奇异值 Γ_i 。

$$\Gamma_i = H_{xu}^i H_{uu}^{i-1} H_{ux}^i \quad (26)$$

因此可以利用矩阵 H 来学习 $\mathfrak{R}_i$, 而不需要系统状态 (A, B) 的信息。

$$\mathfrak{R}_i = \sigma_{\max}^{-\frac{1}{2}} (Q_a^{-T} \hat{\Gamma}_i Q_a^{-1}) \quad (27)$$

最终耦合增益可以通过式 (28) 得到:

$$\frac{r_0}{c_0} < \mathfrak{R}_i, \quad c = \frac{1}{c_0} \quad (28)$$

式 (28) 可以避免使用系统矩阵 (A, B) 。利用求解的反馈增益 K 和耦合增益 c , 得到离散时间多智能体系统的分布式控制协议。

算法 2 在研究基于 LQR 的协同控制设计的基础上, 成功地引入了一种近似动态规划技术来解决线性多智能体系统 (部分或完全) 无模型协同控制问题。由于 A 是不确定的, 因此在使用强化学习算法时, 可以依靠设计者的经验来获得对智能体的初始稳定控制。并且 HDP 算法可以从任意的初值函数开始, 避免了对稳定控制的要求, 但仍然存在困难是输入矩阵 B 必须已知, 因此算法 2 不能实现完全的无模型控制。

3.2 基于强化学习的异构多智能体系统最优一致性

在广泛的实际应用中, 单个系统可能具有不同的动力学特性, 事实上它们的状态可能具有不同的维数。因此, 许多研究致力于解决异构多智能体系统网络中的一致性, 其中智能体的动态性可能不同。对于离散的异构多智能体一致性, 参考文献[64]研究了离散时间异构多智能体系统最优一致性控制的 Q 学习方法。在 Bellman-Q 函数的基础上, 得到了 Bellman-Q-Q 函数。采用策略迭代法迭代求解最优控制, 并采用最小二乘法对实现过程进行激励。参考文献[66]提出了一种自适应最优分布式算法, 用于求解存在外部干扰的未知非线性约束输入系统的多智能体离散时间图形对策。该算法基于数值迭代启发式动态规划, 在不需要系统动力学知识的情况下求解耦合的哈密顿-雅可比-埃萨克斯 (Hamilton-Jacobi-Isaacs, HJI) 方程组。参考文献[67]建立了一个折扣性能指标, 介绍了用于激励的策略迭代算法。为了实现所提出的在线行为相关启发式动态规划方法, 分别利用 critic 神经网络和 actor 神经网络实时逼近迭代性能指标函数和控制策略。因

为异构多智能体系统可以有不同的状态维数, 所以对于异构的多智能体系统问题一般不考虑状态一致性, 大多考虑输出一致性, 并假设输出具有相同的维数。关于多智能体系统的输出一致性, 参考文献[68]研究了网络连接多智能体系统的一致性输出调节问题。每个智能体或子系统都是一个相对度为 1 的输出反馈形式的非线性系统, 但子系统在不同的非线性函数甚至不同的系统阶数下可能具有不同的动力学特性。参考文献[68]提出了一种一致性控制设计方法, 利用内模设计策略, 保证各子系统的输出收敛到同一期望输出轨迹上。参考文献[69-70]研究了具有多个未知领导者的一般线性异构多智能体系统的自适应输出包容控制问题。输出控制的目标是保证每个跟随器的输出收敛到由领导者输出的动态凸壳上。参考文献[69]利用状态反馈和动态输出反馈两种分布式控制协议, 提出了一种自适应调节律来求解相关的输出调节器方程组。参考文献[70]设计了一个分布式自适应观测器来估计每个智能体的领导者状态, 然后将输出同步问题转化为最优控制问题, 并提出了一种新的无模型脱离策略强化学习算法, 实时在线求解最优输出同步问题。参考文献[71]将多智能体系统最优输出同步问题考虑到非线性系统中, 设计了一个分布式观测器来估计每个智能体的领导状态; 推导了一个增广的 HJB 方程, 利用 HJB 的解隐式得到最优解, 并通过数据驱动的方式学习每个智能体的最优控制协议。下面分别列举几类利用数据驱动求解异构多智能体系统一致性问题的算法。

3.2.1 连续异构多智能体系统

对于具有完全未知系统动力学和不确定参数变化的异构多智能体系统的最优鲁棒输出控制问题, 参考文献[72]提出了一种基于模型的算法, 利用离线策略迭代来解决标称系统模型可用时的鲁棒输出抑制问题。利用 p-copy 内模原理处理参数变化。为了解决系统模型不可用的问题, 从误差动力学和智能体动力学出发, 构造了一个新的增广系统, 然后为每个智能体引入一个折扣性能函数, 从而建立了一个具有有界 L_2 增益的最优输出反馈设计问题。只使用在线计算的状态输出数据, 利用贝尔曼方程计算出最优控制策略, 并同时找到更新的控制策略。最后, 利用这个贝尔曼方程, 在不需要任何系统动力学知识的情况下, 利用非策略积分强化学习开发了一个无模型算法来实时求解异构多

智能体系统的最优鲁棒输出包容问题。

考虑分布在网络拓扑图 G 上的 $N+1$ 个节点, 其中 N 个不同的节点可描述为:

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \mathbf{A}_i \mathbf{x}_i(t) + \mathbf{B}_i \mathbf{u}_i(t) \\ \mathbf{y}_i(t) = \mathbf{C}_i \mathbf{x}_i(t) \end{cases} \quad (29)$$

首先为每个跟随器设计以下分布式控制器, 该控制器由一个局部分布式观测器和采用状态反馈设计的控制协议组成:

$$\begin{cases} \dot{\mathbf{z}}_i = \mathbf{F} \mathbf{z}_i + \mathbf{G} \mathbf{e}_{y_i} \\ \mathbf{u}_i = \mathbf{K}_i \mathbf{x}_i + \mathbf{H}_i \mathbf{z}_i \end{cases} \quad (30)$$

局部相对输出信息如下:

$$\mathbf{e}_{y_i} = -\mathbf{y}_i + \sum_{j=1}^n \mathbf{a}_{ij} \mathbf{y}_j + \sum_{v=n+1}^{n+m} \mathbf{g}_i^v \mathbf{y}_v \quad (31)$$

跟随者的动态具有以下闭环形式:

$$\begin{aligned} \begin{bmatrix} \dot{\mathbf{x}}_i \\ \dot{\mathbf{z}}_i \end{bmatrix} &= \begin{bmatrix} \mathbf{A}_i + \mathbf{B}_i \mathbf{K}_i & \mathbf{B}_i \mathbf{H}_i \\ -\mathbf{G}_i \mathbf{C}_i & \mathbf{F}_i \end{bmatrix} \begin{bmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{bmatrix} + \\ &\begin{bmatrix} \mathbf{0} \\ \mathbf{G}_i \end{bmatrix} \times \sum_{j=1}^n \mathbf{a}_{ij} \begin{bmatrix} \mathbf{C}_i & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_j \\ \mathbf{z}_j \end{bmatrix} + \\ &\begin{bmatrix} \mathbf{0} \\ \mathbf{G}_i \end{bmatrix} \sum_{v=n+1}^{n+m} \mathbf{g}_i^v \mathbf{R} \xi_v \end{aligned} \quad (32)$$

参考文献[72]得到如下闭环系统:

$$\begin{aligned} \dot{\boldsymbol{\varepsilon}}_i^c &= \mathbf{A} \boldsymbol{\varepsilon}_i^c + \bar{\mathbf{G}}_i \sum_{j=1}^n \mathbf{a}_{ij} \mathbf{C}_j^c \boldsymbol{\varepsilon}_j^c = \\ &\begin{bmatrix} \mathbf{A}_i & \mathbf{0} \\ -\mathbf{G}_i \mathbf{C}_i & \mathbf{F}_i \end{bmatrix} \boldsymbol{\varepsilon}_i^c + \begin{bmatrix} \mathbf{B}_i \\ \mathbf{0} \end{bmatrix} [\mathbf{K}_i \quad \mathbf{H}_i] \boldsymbol{\varepsilon}_i^c + \bar{\mathbf{G}}_i \bar{\mathbf{z}}_i \end{aligned} \quad (33)$$

其中,

$$\dot{\boldsymbol{\varepsilon}}_i^c = \hat{\mathbf{A}}_i \boldsymbol{\varepsilon}_i^c + \hat{\mathbf{B}}_i \bar{\mathbf{u}}_i + \bar{\mathbf{G}}_i \bar{\mathbf{z}}_i \quad (34)$$

根据状态反馈 L_2 增益定理, 令 $(\mathbf{F}_i, \mathbf{G}_i)$ 为 S 的最小 p-copy 内部模型, 选择 $\gamma_i < 1/\rho(\mathbf{A})$ 。利用控制器 (式 (30)) 解决鲁棒输出抑制问题, 存在矩阵 $\bar{\mathbf{K}}_i$, 且:

$$\bar{\mathbf{K}}_i = -\frac{1}{\alpha_i} \mathbf{R}_i^{-1} \hat{\mathbf{B}}_i^\top \mathbf{P}_i \quad (35)$$

参考文献[56,72]提出了一种非策略积分强化学习算法, 通过构造新的增广系统, 利用固定的控制策略得到完整的状态信息。因为这个增广系统的解依赖于未知的系统动力学信息, 所以定义的状态信息无法使用。为了解决这个问题, 笔者给出了以下方法来重建一个新的增广系统, 其中状态和输出信息都是可用的。

为了实现非策略积分强化学习算法, 需要利用固定的控制策略找到的完整状态的信息。在解决鲁棒输出抑制问题时, 不考虑增广系统, 而是求解增广系统的有界 L_2 增益输出反馈可镇定控制问题。

定义全局变量 $\mathbf{y}_d = [\mathbf{y}_{d1}^\top, \dots, \mathbf{y}_{dn}^\top]^\top$ 及 $\xi_d = [\xi_{d1}^\top, \dots, \xi_{dn}^\top]^\top$, 构造如下的全局增广系统动力学方程:

$$\begin{cases} \bar{\mathbf{X}}_i = \bar{\mathbf{A}}_i \mathbf{X}_i + \bar{\mathbf{B}}_i \mathbf{u}_i + \bar{\mathbf{G}}_i \bar{\mathbf{z}}_i \\ \boldsymbol{\eta}_i = \bar{\mathbf{C}}_i \bar{\mathbf{X}}_i \\ \bar{\mathbf{u}}_i = \bar{\mathbf{K}}_i \mathbf{w}_i \end{cases} \quad (36)$$

改写系统动态为:

$$\begin{aligned} \bar{\mathbf{X}}_i &= \bar{\mathbf{A}}_i \bar{\mathbf{X}}_i + \bar{\mathbf{B}}_i \bar{\mathbf{K}}_i^k \mathbf{w}_i + \bar{\mathbf{G}}_i \bar{\mathbf{z}}_i^k + \\ &\quad \bar{\mathbf{B}}_i (\bar{\mathbf{u}}_i - \bar{\mathbf{K}}_i^k \mathbf{w}_i) + \bar{\mathbf{G}}_i (\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_i^k) \end{aligned} \quad (37)$$

得到输出反馈设计的非策略积分强化学习贝尔曼方程:

$$\begin{aligned} e^{-\delta_i \theta t} \bar{\mathbf{X}}_i(t + \theta t)^\top \mathbf{P}_i^k \bar{\mathbf{X}}_i(t + \theta t) - \bar{\mathbf{X}}_i(t)^\top \mathbf{P}_i^k \bar{\mathbf{X}}_i(t) = \\ \int_t^{t+\theta t} e^{-\delta_i(\tau-t)} (-\boldsymbol{\eta}_i^\top \mathbf{Q}_i \boldsymbol{\eta}_i - \mathbf{a}_i \mathbf{w}_i^\top (\mathbf{K}_i^k)^\top \mathbf{R}_i \mathbf{K}_i^k \mathbf{w}_i) d\tau \cdot \\ \int_t^{t+\theta t} e^{-\delta_i(\tau-t)} \gamma_i^2(\mathbf{z}_i^k) d\tau - \\ \int_t^{t+\theta t} e^{-\delta_i(\tau-t)} \mathbf{a}_i \mathbf{w}_i^\top (\bar{\mathbf{K}}_i^{k+1})^\top \mathbf{R}_i (\bar{\mathbf{u}}_i - \bar{\mathbf{K}}_i^k \mathbf{w}_i) d\tau - \\ \int_t^{t+\theta t} e^{-\delta_i(\tau-t)} 2 \mathbf{X}_i^\top (\mathbf{L}_i^{k+1})^\top (\bar{\mathbf{u}}_i - \bar{\mathbf{K}}_i^k \mathbf{w}_i) d\tau + \\ \int_t^{t+\theta t} e^{-\delta_i(\tau-t)} 2 \gamma_i^2(\mathbf{z}_i^{k+1}) (\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_i^k) d\tau \end{aligned} \quad (38)$$

算法 3^[72] 基于非策略强化学习的无模型最优控制

步骤 1: 对于每个跟随者 i , 令 $(\mathbf{F}_i, \mathbf{G}_i)$ 为 S 的最小 p-copy 内部模型。定义系统的混合控制策略 $\bar{\mathbf{u}}_i = \bar{\mathbf{K}}_i^0 \mathbf{w}_i + \mathbf{e}_i$, 其中 $\bar{\mathbf{K}}_i^0$ 是稳定的, $\mathbf{e}_i$ 是一个扩展干扰。收集系统状态, 输出控制输入的反馈信号。得到 $\gamma_i < 1/\rho(\mathbf{A})$ 、 $\mathbf{a}_i > 0$ 和 $\mathbf{R}_i$;

步骤 2: 利用收集到的系统信号求解积分强化学习的贝尔曼方程, 并得到 $\mathbf{P}_i^k$ 、 $\bar{\mathbf{K}}_i^{k+1}$ 、 $\mathbf{L}_i^{k+1}$ 和 $\bar{\mathbf{z}}_i^{k+1}$ 。由于积分强化学习的贝尔曼方程在一定的激励持续性条件下, 这些解可以用最小二乘法唯一确定;

步骤 3: $\|\bar{\mathbf{K}}_i^k - \bar{\mathbf{K}}_i^{k-1}\| \leq \varepsilon$ 则停止迭代, 令 $\bar{\mathbf{K}}_i = \bar{\mathbf{K}}_i^k$ 。

算法 3 解决了未知系统参数不确定时的鲁棒控制问题, 创新点是构造了新的增广系统, 这表明, 利用从实际系统收集的状态、输出和控制输入信息来稳定增广系统的方法可以解决多智能体鲁棒控制问题。

3.2.2 非线性异构多智能体系统

参考文献[71]研究了具有未知非线性动力学多智能体主从系统的最优输出同步问题, 设计了一个分布式观测器来估计每个智能体的领导状态; 为每个智能体定义了一个折扣性能函数, 并导出了一个增广 HJB 方程, 求其最小值。HJB 解依赖于局部状态和分布观测状态的轨迹。基于 HJB 解决方案的控制协议保证了所有代理的同步误差局部渐近快速地归零。与标准输出同步方法相比, 该方法有两个主要优点: 第一, 它不仅使稳态同步误差为零, 而且使瞬态误差最小; 第二, 它不需要输出调节器方程的显式解, 因为 HJB 解隐式地提供了最优解。参考文献[71]通过利用强化学习技术来学习每个智能体的最优控制协议, 而不需要任何代理或领导者动态的知识; 开发了一种非策略强化学习算法, 以学习 HJB 方程的解, 从而在线找到每个代理的最优控制协议, 而不需要任何关于代理或领导者动态的知识。非策略强化学习算法通过使用两个神经网络, 即 actor 和 critic 神经网络, 近似得到每个智能体的值函数以及更新控制策略。

$$\hat{\mathbf{V}}_i^j(\mathbf{X}) = \hat{\mathbf{W}}_{li}^\top \boldsymbol{\sigma}_i(\mathbf{X}) \quad (39)$$

$$\hat{\mathbf{u}}_i^{j+1}(\mathbf{X}) = \hat{\mathbf{W}}_{2i}^\top \boldsymbol{\phi}(\mathbf{X}) \quad (40)$$

得到贝尔曼近似误差:

$$\begin{aligned} \mathbf{e}_i(t) &= \hat{\mathbf{W}}_{li}^\top \left(e^{-\mathbf{a}_i^\top \tau} \boldsymbol{\sigma}_i(\mathbf{X}_i(t+T)) - \boldsymbol{\sigma}_i(\mathbf{X}_i(t)) \right) - \\ &\quad \int_t^{t+T} e^{-\mathbf{a}_i(\tau-t)} \left(-\mathbf{X}_i^\top \mathbf{Q}_i \mathbf{X}_i - \mathbf{u}_i^\top \mathbf{R} \mathbf{u}_i \right) d\tau + \\ &\quad 2 \sum_{l=1}^m r_l \int_t^{t+T} e^{-\mathbf{a}_i(\tau-t)} \hat{\mathbf{W}}_{2i,l}^\top \boldsymbol{\phi}(\mathbf{X}_i(t)) \mathbf{v}_l^1 d\tau - \\ &\quad \int_t^{t+T} e^{-\mathbf{a}_i(\tau-t)} \hat{\mathbf{W}}_{li}^\top \Delta \boldsymbol{\sigma}_i(\mathbf{X}_i(t)) \mathbf{D} \mathbf{z}_i d\tau \end{aligned} \quad (41)$$

利用最小二乘法使贝尔曼方程误差最小化。贝尔曼近似误差可以改写成:

$$\mathbf{y} \mathbf{o}_i(t) + \mathbf{e}_i(t) = \hat{\mathbf{W}}_i^\top \mathbf{t} \mathbf{o}_i(t) \quad (42)$$

其中,

$$\hat{\mathbf{W}}_i = [\hat{\mathbf{W}}_{li}^\top, \hat{\mathbf{W}}_{2i,1}^\top, \dots, \hat{\mathbf{W}}_{2i,m}^\top]^\top \quad (43)$$

$$\mathbf{t} \mathbf{o}_i(t) = \begin{bmatrix} e^{-\mathbf{a}_i^\top \tau} \boldsymbol{\sigma}_i(\mathbf{X}_i(t+T)) - \boldsymbol{\sigma}_i(\mathbf{X}_i(t)) + \\ \int_t^{t+T} e^{-\mathbf{a}_i(\tau-t)} \hat{\mathbf{W}}_{li}^\top \Delta \boldsymbol{\sigma}_i(\mathbf{X}_i(t)) \mathbf{D} \mathbf{z}_i d\tau \\ 2r_1 \int_t^{t+T} e^{-\mathbf{a}_i(\tau-t)} \boldsymbol{\phi}(\mathbf{X}_i(t)) \mathbf{v}_1^1 d\tau \\ \vdots \\ 2r_m \int_t^{t+T} e^{-\mathbf{a}_i(\tau-t)} \boldsymbol{\phi}(\mathbf{X}_i(t)) \mathbf{v}_m^m d\tau \end{bmatrix} \quad (44)$$

$$\mathbf{y}_i(t) = \int_t^{t+T} e^{-\alpha_i(\tau-t)} (-\mathbf{X}_i^\top \mathbf{Q}_T \mathbf{X}_i - \mathbf{u}_i^{j\top} \mathbf{R} \mathbf{u}_i^j) d\tau \quad (45)$$

在状态空间中通过从点 t_1 到 t_N 得到 $\mathbf{y}_i(t)$ 和 $\mathbf{t}_i(t)$ 。采用最小二乘法得到:

$$\hat{\mathbf{W}}_i = (\mathbf{T}_i \mathbf{T}_i^\top)^{-1} \mathbf{T}_i \cdot \mathbf{Y}_i \quad (46)$$

其中, $\mathbf{T}_i = [\mathbf{t}_i(t_1), \dots, \mathbf{t}_i(t_N)]$, $\mathbf{Y}_i = [\mathbf{y}_i(t_1), \dots, \mathbf{y}_i(t_N)]$, 所提出的求解 HJB 的方法不需要了解代理或领导动态。每个智能体需要通过知道领导者的状态维数和它自己的状态矩阵来执行非策略强化学习算法。

参考文献[71]设计了一个自适应的分布式观测器来估计领导者的状态, 并无须任何智能体的动态知识, 利用强化学习方式求解黎卡提方程。其优点是这种方法不要求输出调节器方程, 不需要通过在每个智能体的控制器中加入领导者动态方程的方式来求解异构系统的输出同步问题。

3.3 具有扰动的多智能体系统最优一致性

在多智能体博弈中, 为了寻找最优策略, 需要求解耦合的 HJ 方程。在现实世界中, 外部干扰是不可避免的, 其会影响系统的性能和稳定性。因此, 在存在未知外部干扰的情况下, 求解多智能体博弈问题也是研究多智能体系统最优一致性的一个关键问题。为了求解具有外部扰动的多智能体对策, 需要求解耦合的 HJI 方程。然而, 由于系统状态的不确定性和求解这些非线性方程组的困难性, 需要通过基于强化学习的数据驱动技术来近似求解这些方程组。

研究含有扰动的多智能体一致性问题通常采用输出调节^[72-73]、H ∞ 控制^[74]、内模原理^[72]、增广矩阵等方式。在求解具有未知动力学的离散 HJI 方程的研究中, 参考文献[72]研究了二阶离散多智能体系统的合作鲁棒输出调节问题, 基于数据驱动设计了一种分布式动态状态反馈控制律, 并采用内模方法求解鲁棒输出调节问题。对于存在外部干扰的未知非线性约束输入系统的多智能体离散时间图形对策问题, 参考文献[66]基于数值迭代启发式动态规划, 使用一个 actor-critic 结构来近似得出值函数、控制策略和最坏情况下的干扰策略, 并验证了闭环系统的稳定性和对纳什均衡的收敛性。参考文献[74]提出的非线性多智能体系统的分布式协同 H ∞ 最优跟踪控制算法对每个智能体只使用一个神经

网络, 并保证闭环系统中的神经网络权值逼近误差和协作跟踪误差等所有信号最终一致有界。对于连续系统, 参考文献[75]研究了一类多智能体系统的最优一致问题, 这类问题的跟随者动力学是未知的, 领导者由一个扰动的外系统来模拟产生, 文中提出了一种数据驱动方法, 该方法对不可测量的先导扰动具有一定的鲁棒性。参考文献[76]提出使用基于数据驱动的多智能体系统在线策略迭代算法来求解耦合的 HJI 方程。参考文献[73]利用鲁棒自适应动态规划方法, 通过在线输入和状态数据, 设计了无模型的分布式控制器来解决具有参数和动态不确定性的多智能体系统的协同输出调节问题, 使得每个跟随器都能实现渐近跟踪和非奇异干扰抑制。

3.3.1 离散含扰动多智能体系统

考虑一个具有 N 个智能体的离散时间含干扰的异构多智能体系统:

$$\mathbf{x}_i(k+1) = \mathbf{f}_i(\mathbf{x}_i(k)) + \mathbf{g}_i(\mathbf{x}_i(k))\mathbf{u}_i(k) + \mathbf{h}_i(\mathbf{x}_i(k))\mathbf{w}_i(k) \quad (47)$$

其中, $\mathbf{x}_i(k)$ 、 $\mathbf{u}_i(k)$ 和 $\mathbf{w}_i(k)$ 分别为智能体 i 的状态、输入和外部干扰。 $\mathbf{f}_i(\mathbf{x}_i)$ 、 $\mathbf{g}_i(\mathbf{x}_i)$ 和 $\mathbf{h}_i(\mathbf{x}_i)$ 分别为状态、输入和外部干扰的未知动态。

考虑系统的领导者动态为:

$$\mathbf{x}_0(k+1) = \mathbf{f}_0(\mathbf{x}_0(k)) \quad (48)$$

对于分布式一致性问题, 定义每个智能体的局部跟踪误差, 通过系统状态得到了智能体 i 的局部邻居跟踪误差的动力学表达式:

$$\begin{aligned} \varepsilon_i(k+1) = & \sum_{j \in N_i} \mathbf{e}_{ij} (\mathbf{f}_j(\mathbf{x}_{jk}) - \mathbf{f}_i(\mathbf{x}_{ik})) + \\ & \mathbf{b}_i (\mathbf{f}_0(\mathbf{x}_0) - \mathbf{f}_i(\mathbf{x}_{ik})) - (\mathbf{d}_i + \mathbf{b}_i) \times \\ & (\mathbf{g}_i(\mathbf{x}_{ik})\mathbf{u}_{ik} + \mathbf{h}_i(\mathbf{x}_{ik})\mathbf{w}_{ik}) + \\ & \sum_{j \in N} \mathbf{e}_{ij} (\mathbf{g}_j(\mathbf{x}_{jk})\mathbf{u}_{jk} + \mathbf{h}_j(\mathbf{x}_{jk})\mathbf{w}_{jk}) \end{aligned} \quad (49)$$

基于纳什均衡理论, 得到了以下耦合 HJI 方程^[66]:

$$\begin{aligned} H(\varepsilon_i, \nabla V_i(\varepsilon_{ik+1}), \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{w}_i, \mathbf{w}_{-i}) = & \\ \nabla V_i(\varepsilon_{ik+1})^\top [& \sum_{j \in N_i} \mathbf{e}_{ij} (\mathbf{f}_j(\mathbf{x}_{jk}) - \mathbf{f}_i(\mathbf{x}_{ik})) + \\ & \mathbf{b}_i (\mathbf{f}_0(\mathbf{x}_0) - \mathbf{f}_i(\mathbf{x}_{ik})) - (\mathbf{d}_i + \mathbf{b}_i) (\bar{\mathbf{U}} \mathbf{g}_i(\mathbf{x}_{ik})) \times \\ & \varphi((\bar{\mathbf{U}} \mathbf{R})^{-1} (\mathbf{d}_i + \mathbf{b}_i) \mathbf{g}_i(\mathbf{x}_{ik})) \nabla V_i(\varepsilon_{ik+1}) \mathbf{u}_{ik} - \\ & \mathbf{h}_i(\mathbf{x}_{ik}) \frac{1}{\gamma^2} (\mathbf{d}_i + \mathbf{b}_i) \mathbf{T}_{ii}^{-1} \mathbf{h}_i^\top(\mathbf{x}_{ik}) \nabla V_i(\varepsilon_{ik+1})) + \end{aligned}$$

$$\begin{aligned}
& \sum_{j \in N_i} e_{ij} (\bar{U} g_j(x_{jk})) \times \\
& \phi((\bar{U} R)^{-1} (d_j + b_j) g_j^T(x_{j(k)}) \nabla V_j(\varepsilon_{jk+1}) - \\
& h_j(x_{jk}) \frac{1}{\gamma^2} (d_j + b_j) T_{jj}^{-1} h_j^T(x_{j(k)}) \nabla V_j(\varepsilon_{jk+1})) + \\
& \frac{1}{2} [\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + W(u_{ik}^*) + \sum_{j \in N_i} W(u_{jk}^*) - \frac{1}{\gamma^2} (d_i + b_i)^2 \times \\
& \nabla V_i^T(\varepsilon_{i(k+1)}) h_i(x_{j(k)}) T_{ii}^{-1} h_i^T(x_{i(k)}) \nabla V_i(\varepsilon_{i(k+1)}) - \\
& \frac{1}{\gamma^2} \sum_{j \in N_i} (d_i + b_i)^2 \nabla V_j(\varepsilon_{j(k+1)}) h_j(x_{j(k)}) T_{jj}^{-1} T_{ij} T_{jj}^{-1} \times \\
& h_j^T(x_{j(k)}) \nabla V_j(\varepsilon_{j(k+1)})] = 0 \quad (50)
\end{aligned}$$

利用神经网络辨识器辨识多智能体系统的未知动力学。在系统辨识过程中, 假设输入层和隐藏层之间的权重矩阵是常数, 只需调整隐藏层和输出层之间的权重矩阵。因此, 神经网络输出表示为:

$$\hat{x}_i(k+1) = \hat{W}_{is}^T \phi_i(\hat{z}_{is}(k)) \quad (51)$$

为实现 HDP 算法, 参考文献[66]中设计了一个 actor-critic 框架。对于每个智能体, 构造 critic 神经网络来执行策略评估, 并逼近最优值函数; 构造 actor 逼近器来执行策略改进, 这些改进估计了最优控制和最坏情况下的扰动策略。如下所示:

$$\hat{u}_{ik}(\hat{W}_{ia}) = \hat{W}_{ia}^T Z_{ik} \quad (52)$$

$$\hat{w}_{ik}(\hat{W}_{id}) = \hat{W}_{id}^T Z_{ik} \quad (53)$$

$$\hat{V}_{ik}(\hat{W}_{ic}) = Z_{ik}^T \hat{W}_{ic}^T Z_{ik} \quad (54)$$

actor 神经网络对控制输入的逼近误差可以定义为:

$$e_{ai} = \hat{W}_{ia}^T Z_{ik} - u_{ik} \quad (55)$$

控制输入可以改写为:

$$u_{ik} = \bar{U} \phi((\bar{U} R)^{-1} (d_i + b_i) g_i^T Q_i \hat{W}_{ic}^T Z_{ik}) \quad (56)$$

目标值函数 $\hat{V}_{ik}^l$ 由式 (57) 给出:

$$\begin{aligned}
\hat{V}_{i(k)}^l &= \frac{1}{2} (\varepsilon_i^T Q_{ii} \varepsilon_i + W(\hat{u}_{ik}^l) + \sum_{j \in N_i} W(\hat{u}_{jk}^l) - \\
& \gamma^2 \hat{w}_{ik}^{lT} T_{ii} \hat{w}_{ik}^l - \gamma^2 \hat{w}_{jk}^{lT} T_{ij} \hat{w}_{jk}^l) + \hat{V}_{i(k+1)}^l \quad (57)
\end{aligned}$$

算法 4^[66] 未知动态图形游戏的 actor-critic 神经网络权重在线调整

步骤 1: 将 actor 和辨识器的权重值随机地初始化, 将 critic 权重初始化为 0;

步骤 2: 初始化所有智能体的初始状态 $x_i(0)$,

$x_0(0)$ 为任意领导者;

步骤 3: 通过以下计算式更新:

在系统轨迹上, 计算局部跟踪误差 ε_{i0} ;

通过式 (52) 计算控制策略 $\hat{u}_{ik}^l$;

通过式 (53) 计算干扰策略 $\hat{w}_{ik}^l$;

通过式 (51) 计算估计状态 $\hat{x}_{i(k+1)}^l$;

使用估计状态计算局部跟踪误差 $\hat{\varepsilon}_{i(k+1)}^l$;

通过式 (54) 计算值函数 $\hat{V}_{i(k+1)}^l$;

更新 critic 权重:

$$\hat{W}_{ic}^{(1+l)T} = \hat{W}_{ic}^{lT} - \hat{\mu}_{ic} (V_{ik} - Z_{ik}^T \hat{W}_{ic}^{lT} Z_{ik}) Z_{ik} Z_{ik}^T$$

其中通过式 (57) 得到 V_{ik} ;

更新 actor 权重:

$$\hat{W}_{ia}^{(1+l)T} = \hat{W}_{ia}^{lT} - \hat{\mu}_{ia} (\hat{W}_{ia}^{lT} Z_{ik} - u_{ik}^l) Z_{ik}^T$$

$$\hat{W}_{id}^{(1+l)T} = \hat{W}_{id}^{lT} - \hat{\mu}_{id} (\hat{W}_{id}^{lT} Z_{ik} - u_{ik}^l) Z_{ik}^T$$

更新 identifier 权重:

$$\hat{W}_{is}^{(1+l)T} = \hat{W}_{is}^{lT} - \hat{\mu}_{is} \phi_i(\hat{W}_{is}^{lT} \phi_i - x_{ik})^T$$

步骤 4: 对于所有 i , 当 $\|V_i^{l+1}(\varepsilon_{ik}) - V_i^l(\varepsilon_{ik})\| \leq \varepsilon$

时, 停止迭代。

算法 4 的优点是将算法以 actor-critic 结构的形式实现, 以逼近 actor 的最优值函数、最优控制和最坏情况下的干扰策略, 解决了存在外部干扰的未知非线性约束输入系统的多智能体零和动态图形对策问题。该算法利用数据驱动的方式, 不需要精确的系统模型。

3.3.2 连续含扰动多智能体系统

对于连续的多智能体系统, 研究具有参数和动态不确定性的多智能体系统的协同输出调节问题。参考文献[73]利用鲁棒自适应动态规划方法, 通过在线输入和状态数据, 设计了无模型的分布式控制器, 利用循环增益定理保证闭环系统的渐近稳定性, 从而解决了协同输出调节问题。每个跟随器都能实现渐近跟踪和非奇异干扰抑制。参考文献[73]提出了一种新的数据驱动控制方案, 通过鲁棒 ADP 实现一类部分线性多智能体系统的协同输出调节。不同于现有的多智能体系统无模型控制^[68-72], 参考文献[73]首次尝试将鲁棒 ADP、输出调节理论和循环小增益技术相结合。其次, 与现有的鲁棒控制策略通常处理有界静态不确定性的方式不同, 参考文献[73]提出了一种新的数据驱动分布式控制方法, 该方法能够处理表示为主体间耦合的动态不确定

性。它严格保证了每个智能体在抑制非奇异干扰的同时能够实现渐近跟踪。

首先, 考虑一类异构的多智能体系统:

$$\begin{aligned}\dot{\mathbf{x}}_i &= \mathbf{A}_i \mathbf{x}_i + \mathbf{B}_i (\mathbf{u}_i + \boldsymbol{\psi}_i(\mathbf{e}, \mathbf{v})) + \mathbf{D}_i \mathbf{v} \\ \mathbf{e}_i &= \mathbf{C}_i \mathbf{x}_i + \mathbf{F}_i \mathbf{v}\end{aligned}\quad (58)$$

考虑有外部干扰:

$$\dot{\mathbf{v}} = \mathbf{E} \mathbf{v} \quad (59)$$

利用线性最优控制理论和输出调节理论, 设计了分散最优控制器, 之后建立鲁棒分布式控制器:

$$\begin{aligned}\dot{\mathbf{u}}_i &= -\mathbf{K}_i^* \mathbf{x}_i + \mathbf{L}_i^* \boldsymbol{\xi}_i \\ \dot{\boldsymbol{\xi}}_i &= \mathbf{E} \boldsymbol{\xi}_i + \sum_{j \in N_i} \mathbf{a}_{ij} (\boldsymbol{\xi}_j - \boldsymbol{\xi}_i) + \mathbf{a}_{i0} (\mathbf{v} - \boldsymbol{\xi}_i)\end{aligned}\quad (60)$$

考虑到耦合多智能体系统与外部系统相关联, 利用输出调节理论和循环小增益定理, 设计了一种新型的分布式控制器。然后, 在不依赖对象动力学知识的情况下, 提出了一种基于鲁棒 ADP 的数据驱动控制方法。

定义:

$$\bar{\mathbf{x}}_{ij} = \mathbf{x}_i - \mathbf{X}_{ij} \mathbf{v} \quad (61)$$

其中, $\mathbf{x}_{i0} = \mathbf{0}_{n_i \times q}$, 且:

$$\mathbf{C}_i \mathbf{x}_{i1} + \mathbf{F}_i = \mathbf{0} \quad (62)$$

得到:

$$\begin{aligned}\dot{\bar{\mathbf{x}}}_{ij} &= \mathbf{A}_i \mathbf{x}_i + \mathbf{B}_i \hat{\mathbf{u}}_i + (\mathbf{D}_i - \mathbf{X}_{ij} \mathbf{E}) \mathbf{v} = \\ &\mathbf{A}_{ik} \bar{\mathbf{x}}_{ij} + \mathbf{B}_i (\mathbf{K}_{ik} \bar{\mathbf{x}}_{ij} + \hat{\mathbf{u}}_i) + (\mathbf{D}_i - \mathbf{S}_i(\mathbf{X}_{ij})) \mathbf{v}\end{aligned}\quad (63)$$

算法 5^[73] 协同输出调节问题的鲁棒 ADP 算法

步骤 1: 令 $i=1$, 选择一个足够小的常数 $s > 0$;

步骤 2: 计算矩阵 $\mathbf{X}_{i0}, \mathbf{X}_{i1}, \dots, \mathbf{X}_{i, h_i+1}$, 在区间 $[t_0, t_s]$ 上设置一个初始策略 $\mathbf{u}_i = -\mathbf{K}_{i0x} + \boldsymbol{\xi}_i$, 其中勘探噪声为 $\boldsymbol{\xi}_i$;

步骤 3: 令 $k=0$;

步骤 4: 求解 $\mathbf{P}_{ik}$ 和 $\mathbf{P}_{ik+1}$;

步骤 5: 令 $k=k+1$, 重复步骤 4、步骤 5, 直到 $\|\mathbf{P}_{ik} - \mathbf{P}_{i, k+1}\| < s$. 令 $k^* = k$, $j=1$;

步骤 6: 求解 $\mathbf{S}_i(\mathbf{X}_{ij})$;

步骤 7: 令 $j=j+1$, 重复步骤 6、步骤 7, 直到 $j=h_i+2$;

步骤 8: 求解 $(\mathbf{X}_i^*, \mathbf{U}_i^*)$, 得到 $\mathbf{L}_{i, k} = \mathbf{U}_i^* + \mathbf{K}_{i, k} \cdot \mathbf{X}_i^*$;

步骤 9: 令 $i=i+1$, 重复以上步骤, 直到 $i=N+1$ 。

算法 5 的优点在于利用鲁棒自适应动态规划方法, 通过在线输入和状态数据, 设计了无模型的分

布式控制器, 利用循环小增益定理保证闭环系统的渐近稳定性, 从而解决了协同输出调节问题。

4 未来研究方向

4.1 动态拓扑系统

在多智能体系统中, 若智能体之间的信息交互是始终保持不变的, 则称为固定拓扑。在实际应用中, 还存在智能体间的信息交互是时变的情况, 这种情况则被称为动态拓扑。具有动态拓扑的多智能体系统是多智能体系统一致性中需要考虑和研究的问题。与固定交互拓扑情况相比, 由于每个智能体的邻居集是时变的, 具有动态拓扑结构的多智能体系统的最优一致性问题会更加复杂。基于数据驱动的强化学习技术已经在具有固定拓扑的多智能体系统协同一致性问题上取得了众多研究成果以及有了深入的发展, 如参考文献[77]研究了有向交换拓扑下离散时间异构多智能体系统的输出同步问题。然而, 具有动态拓扑的多智能体系统的协同问题还需要进行深入的研究。因此, 将基于数据驱动的强化学习算法扩展到研究动态拓扑结构的多智能体系统中是一个具有潜力和价值的研究方向。

4.2 提高算法的精度和速度

在数据驱动技术中利用 actor 和 critic 神经网络进行估计, 需要考虑估计精度的问题, 如何更好地提高神经网络的估计精度也是一个值得深入探索的研究方向。在多智能体协同问题的研究中, 通过强化学习算法求解最优的控制策略的方法大多可以保证结果是严格收敛的, 但在保证估计精度准确的同时无法保证收敛速度, 在保证收敛速度的同时无法确保估计的准确性。如何同时提高收敛速度和收敛精度仍然是研究所面临的难题, 也是未来的一个研究方向。

4.3 探索与开发权衡

强化学习面临的挑战之一是探索与开发 (exploration and exploitation) 之间的权衡。为了获得奖励, 除了利用已知的信息, 智能体也需要进行探索, 以便在未来做出更好的行动选择。由于强化学习已经扩展到多智能体环境中, 静态环境的假设通常不再成立。一个智能体必须适应其他智能体的不断变化。在这种情况下, 出现了一种新的探索与开发权衡: 一个智能体因为环境的变化而不断地探索, 但是这种探索不应过度。在这样一个非平稳的环境中, 智能体应该学习选择合适的时机进行探

索：智能体学习可以更及时地适应不断变化的环境，同时在环境不变的情况下将探索保持在最小限度。以往的研究中，已提出了许多基于模型的强化学习算法用于克服探索和开发的权衡问题。这些算法都需要通过转移概率来构造模型，导致无法在线学习。因此如何利用强化学习算法在不需要环境模型的情况下解决探索和开发的权衡，是数据驱动技术未来可以发展研究的一个方向。

5 结束语

本文阐述了基于强化学习的数据驱动多智能体系统最优一致性问题的研究现状，分别从同构、异构和含扰动3类多智能体系统总结了通过数据驱动技术解决一致性问题的几种算法，表明了数据驱动技术对多智能体系统分布式最优一致性问题的作用，并分析了目前还需要优化的问题，总结了未来基于强化学习的数据驱动技术需要深入研究的方向。

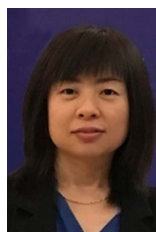
参考文献：

- [1] XU Y, LIU W. Novel multiagent based load restoration algorithm for microgrids[J]. IEEE Transactions on Smart Grid, 2011, 2(1): 152-161.
- [2] YU C, REN F. Collective learning for the emergence of social norms in networked multiagent systems[J]. IEEE Transactions on Cybernetics, 2014, 44(12): 2342-2355.
- [3] WEN G, HU G, YU W, et al. Distributed H_∞ consensus of higher order multiagent systems with switching topologies[J]. IEEE Transactions on Circuits and Systems-II: Express Briefs, 2014, 61(5): 359-363.
- [4] SU S, LIN Z. Distributed consensus control of multi-agent systems with higher order agent dynamics and dynamically changing directed interaction topologies[J]. IEEE Transactions on Automatic Control, 2016, 61(2): 515-519.
- [5] ZHANG H, YUE D, YIN X, et al. Finite-time distributed event-triggered consensus control for multi-agent systems[J]. Journal of Information Science, 2016, 339: 132-142.
- [6] REHAN M, JAMEEL A, AHN C K. Distributed consensus control of one-sided Lipschitz nonlinear multiagent systems[J]. IEEE Transactions on Systems, Man and Cybernetics, 2018, 48(8): 1297-1308.
- [7] OLFATI-SABER R. Flocking for multi-agent dynamic systems: algorithms and theory[J]. IEEE Transactions on Automatic Control, 2006, 51(3): 401-420.
- [8] DOU C, YUE D, GUERRERO J M, et al. Multi-agent system-based distributed coordinated control for radial DC microgrid considering transmission time delays[J]. IEEE Transactions on Smart Grid, 2017, 8(5): 2370-2381.
- [9] BANERJEE B. A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(4): 741-745.
- [10] ZHAO W, LI R, ZHANG H. Finite-time distributed formation tracking control of multi-UAVs with a time-varying reference trajectory[J]. IMA Journal of Mathematical Control and Information, 2018, 35(4): 1297-1318.
- [11] GE X, HAN Q L. Distributed formation control of networked multi-agent systems using a dynamic event-triggered communication mechanism[J]. IEEE Transactions on Industrial Electronics, 2017, 64(10): 8118-8127.
- [12] LEWIS F L, ZHANG H, HENGSTER-MOVRIC K, et al. Cooperative control of multi-agent systems: optimal and adaptive design approaches[M]. London: Springer-Verlag, 2014.
- [13] WANG C, WANG X, JI H. A continuous leader-following consensus control strategy for a class of uncertain multi-agent systems[J]. IEEE/CAA Journal of Automatica Sinica, 2014, 1(2): 187-192.
- [14] DONG W. Distributed optimal control of multiple systems[J]. International Journal of Control, 2010, 83(10): 2067-2079.
- [15] ABOUHEAF M, LEWIS F L, VAMVOUDAKIS K, et al. Multi-agent discrete-time graphical games and reinforcement learning solutions[J]. Automatica, 2014, 50(12): 3038-3053.
- [16] AL-TAMIMI A, LEWIS F L, ABU-KHALAF M. Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 38(4): 943-949.
- [17] TATARI F, NAGHIBI-SISTANI M B, VAMVOUDAKIS K G. Distributed learning algorithm for non-linear differential graphical games[J]. Transactions of the Institute of Measurement and Control, 2017, 39(2): 173-182.
- [18] WEI Q, LIU D, LEWIS F L. Optimal distributed synchronization control for continuous-time heterogeneous multi-agent differential graphical games[J]. Information Sciences, 2015, 317: 96-113.
- [19] REN W, MOORE K L, CHEN Y. High-order and model reference consensus algorithms in cooperative control of multivehicle systems[J]. Journal of Dynamic Systems, 2007, 129(5): 678-688.
- [20] LI Z, REN W, LIU X, et al. Distributed consensus of linear multi-agent systems with adaptive dynamic protocols[J]. Automatica, 2013, 49(7): 1986-1995.
- [21] ZHAO Y, DUAN Z, WEN G, et al. Distributed finite-time tracking of multiple non-identical second-order nonlinear systems with settling time estimation[J]. Automatica, 2016, 64(C): 86-93.
- [22] KIM H, SHIM H, SEO J H. Output consensus of heterogeneous uncertain linear multi-agent systems[J]. IEEE Transactions on Automatic Control, 2011, 56(1): 200-206.
- [23] YAGHMAIE F A, LEWIS F L, SU R. Output regulation of linear heterogeneous multi-agent systems via output and state feedback[J]. Automatica, 2016, 67: 157-164.
- [24] 曹伟, 孙明. 离散时变多智能体系统有限时间一致性迭代学习控制[J]. 控制与决策, 2019, 34(4): 891-896.
- CAO W, SUN M. Finite-time consensus iterative learning control of discrete time-varying multi-agent systems[J]. Control and Decision, 2019, 34(4): 891-896.
- [25] 徐君, 张国良, 曾静, 等. 具有时延和切换拓扑的高阶离散时间多智能体系统鲁棒保性能一致性[J]. 自动化学报, 2019, 45(2): 360-373.
- XU J, ZHANG G L, ZENG J, et al. Robust guaranteed cost consensus for high-order discrete-time multi-agent systems with switching topologies and time delays[J]. Acta Automatica Sinica, 2019, 45(2): 360-373.
- [26] 何吕龙, 柏鹏, 梁晓龙, 等. 多智能体系统离散时间一致性问题中

- 的参数设计[J]. 控制与决策, 2018, 33(8): 1455-1460.
- HE L L, BAI P, LIANG X L, et al. Parameters design for consensus in multi-agent systems with second-order discrete-time dynamics[J]. Control and Decision, 2018, 33(8): 1455-1460.
- [27] 王静蓉, 李宗刚, 杜亚江. 基于 LQR 的异构多智能体系统的最优一致性[J]. 信息与控制, 2018, 47(4): 468-472.
- WANG J R, LI Z G, DU Y J. LQR-based optimal leader-follower consensus in heterogeneous multi-agent systems[J]. Information and Control, 2018, 47(4): 468-472.
- [28] 董新民, 丁超, 陈勇, 等. 完全分布式异构多智能体系统有限时间跟踪[J]. 控制与决策, 2020, 35(4): 105-111.
- DONG X M, DING C, CHEN Y, et al. Fully distributed finite-time tracking of heterogeneous multi-agent systems[J]. Control and Decision Making, 2020, 35(4): 105-111.
- [29] 李耿, 秦雯, 王婷, 等. 异质多智能体系统滞后一致性跟踪控制[J]. 计算机应用, 2018, 38(12): 37-42.
- LI G, QIN W, WANG T, et al. Lag consensus tracking control for heterogeneous multi-agent systems[J]. Journal of Computer Applications, 2018, 38(12): 37-42.
- [30] 张霓, 杜伟, 何熊熊, 等. 基于中间状态值的多智能体系统安全一致性控制[J]. 控制与决策, 2019, 34(3): 567-571.
- ZHANG N, DU W, HE X X, et al. Secure consensus control of multi-agent systems based on median state strategy[J]. Control and Decision, 2019, 34(3): 567-571.
- [31] 王巍. 基于自适应动态规划的多智能体系统一致性方法[D]. 武汉: 中国地质大学, 2019.
- WANG W. Unified method of multi-agent system based on adaptive dynamic programming[D]. Wuhan: China University of Geosciences, 2019.
- [32] 李健, 沈艳军, 刘允刚. 线性多智能体系统一致性的自适应动态规划求解方法[J]. 系统科学与数学, 2016, 36(7): 1016-1030.
- LI J, SHEN Y J, LIU Y G. Adaptive dynamic programming technique for the consensus of linear multi-agent system[J]. Journal of Systems Science and Mathematical Sciences, 2016, 36(7): 1016-1030.
- [33] 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[J]. 控制理论与应用, 2019, 36(6): 885-892.
- WEI W J, MA Y Q, LI Z G. Cooperative output regulation of heterogeneous multi-agent system in finite time[J]. Control Theory and Applications, 2019, 36(6): 885-892.
- [34] SI J, WANG Y T. Online learning control by association and reinforcement[J]. IEEE Transactions on Smart Grid, 2001, 12(2): 264-276.
- [35] WERBOS P J. A menu of designs for reinforcement learning over-time[M]//Neural Networks for Control. Cambridge: MIT Press, 1991: 67-95.
- [36] DOYA K. Reinforcement learning in continuous-time and space[J]. Neural Computation, 2000, 12: 219-245.
- [37] MODARES H, LEWIS F L. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning[J]. IEEE Transactions on Automatic Control, 2014, 59(11): 3051-3056.
- [38] MODARES H, LEWIS F L. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning[J]. Automatica, 2014, 50(7): 1780-1792.
- [39] GADEWADIKAR J, LEWIS F L. Necessary and sufficient conditions for Hstatic output-feedback control[J]. Journal of Guidance Control and Dynamics, 2006, 29(4): 915-920.
- [40] BITTANTI S, COLANERI P. Lyapunov and Riccati equations: periodic inertia theorems[J]. IEEE Transactions on Automatic Control, 1986, 31(7): 659-661.
- [41] VAMVOUDAKIS K G, LEWIS F L, HUDAS G R. Multi-agent differential graphical games: online adaptive learning solution for synchronization with optimality[J]. Automatica, 2012, 48(8): 1598-1611.
- [42] ABOUHEAF M I, LEWIS F L, VAMVOUDAKIS K G, et al. Multi-agent discrete-time graphical games and reinforcement learning solutions[J]. IEEE Transactions on Smart Grid, 2014, 50(12): 152-161.
- [43] ZHANG H, ZHANG J, YANG G H, et al. Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(1): 152-163.
- [44] ZHANG H, YUE D, ZHAO W, et al. Distributed optimal consensus control for multiagent systems with input delay[J]. IEEE Transactions on Cybernetics, 2018, 48(6): 1747-1759.
- [45] LUO B, WU H N, HUANG T, et al. Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design[J]. Automatica, 2014, 50(12): 3281-3290.
- [46] HOU Z S, WANG Z. From model-based control to data-driven control: survey, classification and perspective[J]. Information Sciences, 2013, 235: 3-35.
- [47] BIAN T, JIANG Y, JIANG Z P. Decentralized adaptive optimal control of large-scale systems with application to power systems[J]. IEEE Transactions on Industrial Electronics, 2015, 62(4): 2439-2447.
- [48] MOVRIC K H, LEWIS F L. Cooperative optimal control for multi-agent systems on directed graph topologies[J]. IEEE Transactions on Smart Grid, 2014, 59(3): 769-774.
- [49] LI J, MODARES H, CHAI T, et al. Off-policy reinforcement learning for synchronization in multiagent graphical games[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2434-2445.
- [50] ZHANG H, PARK J H, YUE D, et al. Finite-horizon optimal consensus control for unknown multiagent state-delay systems[J]. IEEE Transactions on Cybernetics, 2020, 50(2): 402-413.
- [51] KAR S, MOURA J M, POOR H V. QD-learning: a collaborative distributed strategy for multi-agent reinforcement learning through consensus innovations[J]. IEEE Transaction on Smart Grid, 2013, 61(7): 1848-1862.
- [52] KAYA M, ALHAJJ R. A novel approach to multiagent reinforcement learning: utilizing OLAP mining in the learning process[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2005, 35(4): 152-161.
- [53] ABOUHEAF M I, LEWIS F L, VAMVOUDAKIS K G, et al. Multi-agent discrete-time graphical games and reinforcement learning solutions[J]. Automatica, 2014, 50(12): 3038-3053.
- [54] ABOUHEAF M I, LEWIS F L, MAHMOUD M S, et al. Discrete-time dynamic graphical games: model-free reinforcement learning solution[J]. IEEE Transactions on Smart Grid, 2015, 13(1): 55-69.
- [55] WANG W, CHEN X, FU H, et al. Model-free distributed consensus control based on actor-critic framework for discrete-time nonlinear multiagent systems[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(11): 4123-4134.
- [56] ZHANG H, ZHANG J, YANG G H, et al. Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(1): 152-163.

- [57] ZHANG H, YUE D, ZHAO W, et al. Distributed optimal consensus control for multiagent systems with input delay[J]. IEEE Transactions on Cybernetics, 2018, 48(6): 1747-1759.
- [58] XUE L, SUN C, YU F. A game theoretical approach for distributed resource allocation with uncertainty[J]. International Journal of Intelligent Computing and Cybernetics, 2017, 10(1): 52-67.
- [59] XUE L, SUN C, WUNSCH D, et al. An adaptive strategy via reinforcement learning for the prisoner's dilemma game[J]. IEEE-CAA Journal of Automatica Sinica, 2018, 5(1): 301-310.
- [60] LI N, MARDEN J R. Designing games for distributed optimization[J]. IEEE Journal on Selected Topics in Signal Processing, 2013, 7(2): 230-242.
- [61] VAMVOUDAKIS K G, LEWIS F L, HUDAS G R. Multi-agent differential graphical games: online adaptive learning solution for synchronization with optimality[J]. Automatica, 2012, 48(8): 1598-1611.
- [62] SU L C, ZHU F. Design of a novel omnidirectional stereo vision system[J]. Acta Automatica Sinica, 2017, 275: 649-658.
- [63] ZHANG H G, FENG T, LIANG H, et al. LQR-based optimal distributed cooperative design for linear discrete-time multi agent systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 28(3): 599-611.
- [64] MU C X, ZHAO Q, SUN C Y. Q-learning solution for optimal consensus control of discrete-time multi-agent systems using reinforcement learning[J]. IEEE Transactions on Smart Grid, 2019, 356(13): 6946-6967.
- [65] MARDEN J R, ARSLAN G, SHAMMA J S. Cooperative control and potential games[J]. IEEE Transactions on Systems, Man and Cybernetics, 2009, 39(6): 1393-1407.
- [66] JAHAN Z, DIDEBAN A, TATARI F. Heuristic dynamic programming for nonlinear zero-sum dynamic graphical games with unknown dynamics and input constraint[J]. IET Research Journals, 2015: 1-10.
- [67] ZHANG H, JIANG H, LUO Y, 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.
- [68] DING Z. Consensus output regulation of a class of heterogeneous nonlinear systems[J]. IEEE Transactions on Automatic Control, 2013, 58(10): 2648-2653.
- [69] ZUO S, SONG Y, LEWIS F L, et al. Adaptive output containment control of heterogeneous multi-agent systems with unknown leaders[J]. Automatica, 2018, 92: 235-239.
- [70] 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.
- [71] 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.
- [72] ZUO S, SONG Y D, LEWIS F L, et al. Optimal robust output containment of unknown heterogeneous multiagent system using off-policy reinforcement learning[J]. IEEE Transactions on Cybernetics, 2018, 48(11): 3197-3207.
- [73] GAO W N, JIANG Y, DAVARI M. Data-driven cooperative output regulation of multi-agent systems via robust adaptive dynamic programming[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2019, 66(3): 447-451.
- [74] LUY T N. Distributed cooperative H_∞ optimal tracking control of MIMO nonlinear multi-agent systems in strict-feedback form via adaptive dynamic programming[J]. IEEE Transactions on Smart Grid, 2017: 952-968.
- [75] GAO W N, JIANG Z P, LEWIS F L, et al. Leader-to-formation stability of multi-agent systems: an adaptive optimal control approach[J]. IEEE Transactions on Automatic Control, 2018, 68(10): 3581-3587.
- [76] JIAO Q, MODARES H, XU S, et al. Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control[J]. Automatica, 2016, 69(1): 24-34.
- [77] MU C X, ZHAO Q, SUN C Y. Optimal model-free output synchronization of heterogeneous multi-agent systems under switching topologies[J]. IEEE Transactions on Industrial Electronics, 2020, 67(12): 10951-10964.

[作者简介]



李金娜 (1977-), 女, 博士, 辽宁石油化工大学信息与控制工程学院教授, 博士生导师, 主要研究方向为数据驱动控制、运行优化控制、强化学习、网络控制等。



程薇燃 (1996-), 女, 辽宁石油化工大学信息与控制工程学院硕士生, 主要研究方向为强化学习、多智能体控制、最优控制、数据驱动控制。