SCIENCE CHINA
Information Sciences

July 2023, Vol. 66 170202:1–170202:16
https://doi.org/10.1007/s11432-022-3687-1

• RESEARCH PAPER •

Special Topic: Entanglement of Control and Optimization

Data-driven cooperative optimal output regulation for linear discrete-time multi-agent systems by online distributed adaptive internal model approach

Kedi XIE[1], Yi JIANG[2], Xiao YU[1,3*] & Weiyao LAN[1,3]

[1] *Department of Automation, Xiamen University, Xiamen* 361005*, China;*
[2] *Department of Electrical Engineering, City University of Hong Kong, Hong Kong* 999077*, China;*
[3] *Key Laboratory of Control and Navigation (Xiamen University), Fujian Province University, Xiamen* 361005*, China*

Received 24 April 2022/Revised 21 October 2022/Accepted 16 December 2022/Published online 26 June 2023

Abstract In this study, a data-driven learning algorithm was developed to estimate the optimal distributed cooperative control policy, which solves the cooperative optimal output regulation problem for linear discrete-time multi-agent systems. Notably, the dynamics of all the agent systems and exo-system is completely unknown. By combining adaptive dynamic programming with an internal model, a model-free off-policy learning method is proposed to estimate the optimal control gain and the distributed adaptive internal model by only accessing the measurable data of multi-agent systems. Moreover, different from the traditional cooperative adaptive controller design method, a distributed internal model is approximated online. Convergence and stability analyses show that the estimate controller generated by the proposed data-driven learning algorithm converges to the optimal distributed controller. Finally, simulation results verify the effectiveness of the proposed method.

Keywords adaptive dynamic programming, cooperative control, distributed adaptive internal model, multi-agent systems, optimal output regulation

Citation Xie K D, Jiang Y, Yu X, et al. Data-driven cooperative optimal output regulation for linear discrete-time multi-agent systems by online distributed adaptive internal model approach. Sci China Inf Sci, 2023, 66(7): 170202, https://doi.org/10.1007/s11432-022-3687-1

1 Introduction

The cooperative output regulation problem (CORP) [1–3], which simultaneously addresses the problem of asymptotic tracking and disturbance rejection for multi-agent systems, has attracted widespread interest in the last decades due to its wide applications in many areas, such as multiple unmanned aerial vehicles [4], dynamic traffics [5], distributed robotics [6], and distributed sensor networks [6, 7]. Necessary and sufficient conditions for solving the output regulation problem were first established by [8] in terms of the well-known regulator equations. By viewing the leader and disturbance as an exo-system, the leader-following output consensus problem with disturbance rejection can be viewed as a CORP, where the exo-system lumps reference inputs and disturbances together. This scenario has been extensively investigated by many researchers, and some recent interesting results are provided in [9–14].

In general, two classes of control schemes are used to solve the output regulation problem, i.e., the feedback-feedforward controller [8, 15] and dynamic controller embedding with the internal model [15, 16]. The CORP was addressed in [1] and solved by designing a feedback-feedforward controller with a distributed observer. The distributed observer proposed in [1] is used to estimate the exo-system state, which has been developed for solving the CORP for switching networks [2, 11] and nonlinear systems [3]. However, as stated in [15, 17], the feedforward control gain is dependent on the whole system parameters, so with varying system parameters, the desired feedforward control gain would also vary. Thus, the feedback-feedforward controller can only deal with the deterministic system dynamics without

[*] Corresponding author (email: xiaoyu@xmu.edu.cn)

uncertainty parameters. To further solve the cooperative robust output regulation problem, in [18], an internal model [16, 19] was employed to estimate the steady state of the system without solving the regulator equations. This method ensures the robustness, with respect to parameter uncertainties, of a controller embedded with the internal model. Later on, the cooperative robust output regulation for nonlinear systems was investigated (see [11, 20, 21]). Due to the communication limitation, a distributed adaptive internal model scheme was proposed in [9] for discrete-time multi-agent systems, and a similar result for the continuous-time system is given in [10]. In [9], using the prior knowledge of the exo-system dynamics, not only the estimation of the exo-system state but also the estimation of the internal model for each agent can be achieved by the proposed adaptive distributed observer. Recently, the solution in [9] was extended to solve the CORP for time-delay multi-agent systems [22].

Considering the output regulation and optimal control problems, an optimal controller design methodology was proposed in [23] by minimizing a predefined cost function that involves transient responses with the constraint of the steady-state output regulation requirement, which is the so-called optimal output regulation. Under the framework presented in [23], several studies on investigating the optimal output regulation problem were presented for discrete-time linear systems [24], sample-data systems [25], and nonlinear systems [26]. However, system matrices should be known prior to solving the related algebraic Riccati equation (ARE) for linear systems [23, 24] or the Hamilton-Jacobi-Bellman (HJB) equation for nonlinear systems [26]. In addition, accurate knowledge of system dynamics is required to construct the corresponding controllers in most existing studies on the CORP, e.g., [1–3, 9–14, 18, 20–22, 27]. This limits the applications of these solutions due to the existence of unknown parameters in practice.

Recently, combining the adaptive dynamic programming (ADP) algorithm and reinforcement learning (RL) method, several online learning algorithms have been proposed to solve optimal control problems, e.g., optimal regulation problems [28–34], H_∞ tracking problems [35,36], and zero-sum games [37]. In [28], a data-driven cooperative optimal cruise control was proposed for heterogeneous vehicle platoons using the ADP algorithm with the policy-iteration (PI) scheme. A general containment control of the discrete-time multi-agent system was presented in [32] by employing the value-iteration (VI)-based learning scheme. Compared to the PI-based learning method, the VI-based learning algorithm relaxes the requirement on the initial stabilizing control policy at the expense of the convergence rate. Taking the disturbance rejection into account, a developed ADP-based learning algorithm with distributed observers was proposed in [38] to solve the leader-following control problem. Later on, with the distributed adaptive internal model designed in [10], the cooperative optimal output regulation problem (COORP) was addressed in [39], and it was solved by the ADP-based learning method with the PI and VI schemes. However, in most of these existing studies, e.g., [29, 31, 34, 38–41], the exo-system dynamics should still be known, such that the internal model and distributed observer can be efficiently designed for each agent.

To solve the COORP for unknown multi-agent system dynamics, a data-driven ADP-based learning algorithm, which aims at estimating the optimal dynamic feedback controller embedding with the adaptive distributed internal model, was developed in this study.

The contributions of this study are threefold: First, compared to [9] where the system dynamics should be known, this study aims at developing a model-free learning algorithm to solve the COORP with unknown multi-agent systems, including the unknown exo-system. Two different finite exciting (FE)-based estimation update laws are proposed in this paper to approximate the exo-system dynamics, which is used to establish the online distributed internal model. Second, different from [14, 27, 30], where the internal model can be directly obtained for all agents, in this study, the distributed internal model of each agent can be estimated by introducing an online adaptive distributed observer. Third, in most existing methods for solving the COORP, for instance, [9, 10, 30, 38, 39, 41], not only the exo-system dynamics are assumed to be known, but also the exo-system state should be estimated by an observer network among the agents. While in this study, by only accessing the data of the input, output, and state of each agent, a data-driven learning algorithm is proposed, which does not require the approximation of the exo-system state or the use of any prior knowledge of the multi-agent system and the exo-system. This could lead to a reduction in the data computational load and data communication load. Moreover, estimation laws for estimating the exo-system dynamics and a data-driven learning algorithm for approximating the optimal control policy are designed without the requirement on the persistence of excitation condition.

The remainder of this paper is organized as follows: In Section 2, we present the basic assumptions, preliminaries, and control objectives. In Section 3, we propose two online estimation methods to approximate the exo-system dynamics, which is used to design the distributed internal model, and then, we developed an ADP-based data-driven learning method to estimate the optimal control policy without using any prior knowledge of the multi-agent system. We also present the convergence analysis of the proposed algorithm in Section 3. In Section 4, we provide a numerical example to illustrate the proposed method. Finally, in Section 5, we draw the conclusion.

Notations. The following notations are used throughout the paper. $I_n \in \mathbb{R}^{n\times n}$ is a unit matrix. $\sigma(A)$ and $\rho(A)$ are the complex spectrum and spectral radius of matrix A, respectively. For a matrix $B \in \mathbb{R}^{m\times n}$, $\mathrm{vec}(B) = [{b_1}^{\mathrm{T}}, {b_2}^{\mathrm{T}}, \ldots, {b_n}^{\mathrm{T}}]^{\mathrm{T}}$, where $b_i \in \mathbb{R}^m$ is the ith column of B. For a symmetric matrix $C \in \mathbb{R}^{n\times n}$, $\mathrm{vech}(C) = [c_{11}, c_{12}, \ldots, c_{1n}, c_{22}, c_{23}, \ldots, c_{n-1,n}, c_{nn}]^{\mathrm{T}} \in \mathbb{R}^{\frac{n(n+1)}{2}}$, where c_{ij} is the element of C. For two column vectors $v_1 \in \mathbb{R}^{n_1}$ and $v_2 \in \mathbb{R}^{n_2}$, $\mathrm{col}(v_1, v_2) = [v_1^{\mathrm{T}}, v_2^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n_1+n_2}$, and $\mathrm{dia}(v_1)$ denotes a diagonal matrix $V \in \mathbb{R}^{n_1\times n_1}$, where V_{ii} is the i-th element of v_1. For any matrix G_i, $\mathcal{G} = \text{block diag}\{G_i, \ldots, G_n\}$ is an augmented block diagonal matrix with block matrix $\mathcal{G}_{ii} = G_i$. $\emptyset$ denotes the null set. $\mathbb{P}^n$ denotes the normed space of all n-by-n real symmetric matrices and $\mathbb{P}^n_+ := \{P \in \mathbb{P}^n : P \geqslant 0\}$. $\|\cdot\|$ is the Euclidean norm of vectors and the Frobenius norm of matrices. $\otimes$ denotes the Kronecker product.

2 Plant and control objective

Consider a class of discrete-time multi-agent systems described by

$$x_i(t+1) = A_i x_i(t) + B_i u_i(t) + E_i v(t), \tag{1}$$

$$y_i(t) = C_i x_i(t), \tag{2}$$

$$e_i(t) = y_i(t) - Fv(t), \tag{3}$$

where $x_i \in \mathbb{R}^{n_i}$, $u_i \in \mathbb{R}^{m_i}$, $y_i \in \mathbb{R}^p$, and $e_i \in \mathbb{R}^p$ denote the state, control input, output, and tracking error of the i subsystem with $i = 1, 2, \ldots, N$, respectively. $v(t)$ is the exo-system state generated by the following autonomous exo-system:

$$v(t+1) = Sv(t), \tag{4}$$

and $A_i \in \mathbb{R}^{n_i\times n_i}$, $B_i \in \mathbb{R}^{n_i\times m_i}$, $E_i \in \mathbb{R}^{n_i\times q}$, $C_i \in \mathbb{R}^{p\times n_i}$, $F \in \mathbb{R}^{p\times q}$, and $S \in \mathbb{R}^{q\times q}$ are constant system matrices.

Define a directed graph $\mathcal{D} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{0, 1, \ldots, N\}$ is the set of nodes with node 0 denoting the exo-system and the other N nodes associated with the N agents, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes the edge set. Let $\mathcal{N}_i = \{j \in \mathcal{V} | (j, i) \in \mathcal{E}\}$ denote the set of incoming neighbors of agent i. Let $\mathcal{V}_0^+ := \mathcal{V}/\{0\}$ and $\mathcal{N}_i^+ := \mathcal{N}_i/\{0\}$. In addition, the virtual tracking error is described as

$$\hat{e}_i \triangleq \sum_{j\in\mathcal{N}_i} \frac{a_{ij}(y_i - y_j)}{\sum_{j=0}^N a_{ij}}, \quad i = 1, \ldots, N, \tag{5}$$

where $y_0 = Fv$, a_{ij} is the element of a weighted adjacency matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{(N+1)\times(N+1)}$, which has positive elements $a_{ij} > 0$ if $(j, i) \in \mathcal{E}$ and otherwise $a_{ij} = 0$. Define the graph Laplacian matrix $W = [w_{ij}] \in \mathbb{R}^{N\times N}$ with $w_{ii} = \sum_{j=1}^N a_{ij}$ and $w_{ij} = -a_{ij}$, for any $i \neq j$. Let $\Delta = \mathrm{dia}\left([a_{10}, a_{20}, \ldots, a_{N0}]\right)$ and $\mathcal{H} = W + \Delta$.

The following standard assumptions are made on the multi-agent system.

Assumption 1. The pair (A_i, B_i) is stabilizable, $\forall i \in \mathcal{V}_0^+$.

Assumption 2. $\mathrm{rank}[\begin{smallmatrix} A_i - \lambda I & B_i \\ C_i & 0 \end{smallmatrix}] = n_i + p$, $\forall \lambda \in \sigma(S)$, $i \in \mathcal{V}_0^+$.

Assumption 3. The communication graph $\mathcal{D}$ is acyclic which contains a directed spanning tree with node 0 as the root.

Assumption 4. The eigenvalues of S are with modulus as 1.

Assumption 5. All of the system matrices A_i, B_i, C_i, E_i, F, and S are completely unknown, and the exo-system state with a known dimension q is only available to the agents i, $0 \in \mathcal{N}_i$.

Now, let the performance index

$$J(t_0) = \sum_{i=1}^N \sum_{t=t_0}^{\infty} \mathcal{C}_i(x_i(t), u_i(t)), \tag{6}$$

where $\mathcal{C}_i(x_i(t), u_i(t)) = \|x_i - x_d^i\|_{Q_i} + \|u_i - u_d^i\|_{R_i}$ is the cost function with weight matrices $Q_i = Q_i^{\mathrm{T}} \geqslant 0 \in \mathbb{R}^{n_i\times n_i}$, $R_i = R_i^{\mathrm{T}} > 0 \in \mathbb{R}^{m_i\times m_i}$ and the pair $(A_i, \sqrt{Q_i})$ being detectable. x_d^i and u_d^i are the steady state and stabilizing control input for each agent $i \in \mathcal{V}_0^+$, respectively, when the closed-loop system achieves the reference tracking and disturbance rejection simultaneously.

The objective of this study is to determine the cooperative optimal control input u_i^* for each agent $i \in \mathcal{V}_0^+$, such that the following conditions are satisfied:

(i) The closed-loop system of each subsystem described in (1) is stable.

(ii) The tracking error of each agent converges to zero as $t \to \infty$, i.e., $\lim_{t\to\infty} e_i(t) = 0$.

(iii) The performance index $J(t_0)$ is minimized.

Remark 1. Assumptions 1 and 2 are the necessary and sufficient conditions used to solve the traditional output regulation problem, which aims at simultaneously achieving the asymptotically tracking and the disturbance rejection as shown in [9, 15, 38, 41]. In Assumption 3, the graph containing a directed spanning tree is a necessary condition that ensures the consensus of multi-agent systems [39, 40]. Under Assumption 3, one can always label all the agents such that $\mathcal{A}$ is a lower triangular matrix, which is also required in [39]. Assumption 4 indicates that the exo-system is marginally stable and the exo-system state is bounded, which is made without loss of any generality. Assumption 5 is used as the main setup of this study. In addition, in Assumption 5, only the agents i, $0 \in \mathcal{N}_i$ have access to the exo-system state, but all agents, including the agents i, $0 \in \mathcal{N}_i$, cannot obtain the system matrix S of the exo-system.

3 Main results

In this section, an ADP-based data-driven learning algorithm is proposed to estimate the optimal cooperative control policy with a distributed adaptive internal model for solving the COORP. First, two FE-based estimation methods are presented to establish the online distributed internal model design scheme. Then, a model-free VI-based learning algorithm is developed to estimate the optimal control policy. Finally, the stability and convergence analyses are presented.

3.1 Online distributed adaptive internal model design method

To solve the CORP for multi-agent systems, the internal model framework is usually introduced for converting the CORP into a tractable cooperative stabilizing problem. In general, the distributed adaptive internal model can be designed by the prior knowledge of the exo-system dynamics. The challenge arises when the exo-system matrix S is unknown. Thus, for agents i, $0 \in \mathcal{N}_i$, we first propose two FE-based methods to estimate the exo-system matrix S online without the requirement on the persistence of excitation condition, where the estimation error converges to zero.

To begin with, the FE condition, which is similar to the FE condition shown in [42], is given as follows.

Definition 1 (Finite exciting condition)**.** A bounded signal $\rho(t) \in \mathbb{R}^{m\times n}$ is finite exciting over a time series $t = t_0, t_0+1, \ldots, t_0+T_s$ with $T_s \in \mathbb{Z}^+$, for a given $0 < \gamma < 1$, if $\exists\ \beta > 0$ such that

$$\sum_{t=t_0}^{t_0+T_s} \gamma^t \rho(t)^{\mathrm{T}} \rho(t) \geqslant \beta I_n \tag{7}$$

holds, where $I_n \in \mathbb{R}^{n\times n}$ is a unit matrix and β denotes the degree of excitation.

The exo-system dynamics described in (4) can be linearly parameterized as

$$v(t+1) = \chi_v(t)\theta, \tag{8}$$

where $\theta = \mathrm{vec}(S^{\mathrm{T}}) \in \mathbb{R}^{q^2}$ is the vectorization of the matrix S, and $\chi_v(t) = I_q \otimes v(t)^{\mathrm{T}} \in \mathbb{R}^{q\times q^2}$ is the regressor matrix with the unit matrix $I_q \in \mathbb{R}^{q\times q}$.

Consider the following filter equations:

$$F(t+1) = \gamma F(t) + \chi_v(t)^{\mathrm{T}} \chi_v(t), \qquad F(t_0) = 0, \tag{9}$$

$$H(t+1) = \gamma H(t) + \chi_v(t)^{\mathrm{T}} v(t+1), \qquad H(t_0) = 0, \tag{10}$$

where $\gamma \in (0,1)$ is a tunable gain for adjusting the convergence rate and $F(t) \in \mathbb{R}^{q^2\times q^2}$ and $H(t) \in \mathbb{R}^{q^2}$ are the integrated-filtered regressors.

We introduce two different parameter estimation update laws for θ as follows:

(1) The one-step estimation update law of $\hat{\theta}$ is given as

$$\hat{\theta} = F(t_{T_s})^{-1}H(t_{T_s}), \tag{11}$$

where t_{T_s} is the instant when $\chi_v(t)$ satisfies the FE condition (7).

(2) The dynamic estimation update law of $\hat{\theta}(t)$ is given as

$$\hat{\theta}(t+1) = \hat{\theta}(t) + k_{\theta_1}\chi_v(t)^{\mathrm{T}}\left(v(t+1) - \chi_v(t)\hat{\theta}(t)\right) + k_{\theta_2}\left(H(t) - F(t)\hat{\theta}(t)\right), \tag{12}$$

where $k_{\theta_1} > 0$ and $k_{\theta_2} > 0$ are the tuning gains for adjusting the estimation convergence rate.

Now, we present the following Theorem 1 to show the convergence of the FE-based estimation methods for approximating the exo-system matrix S.

Theorem 1. For agents i, $0 \in \mathcal{N}_i$, if there exists an instant $t_{T_s} = t_0 + T_s$, such that $\chi_v(t)$ satisfies the FE condition (7), then the exo-system matrix S can be estimated by either the one-step estimation update law (11) or the dynamic estimation update law (12), where both the estimation errors converge to zero.

Proof. Note that the estimation process for approximating the exo-system dynamics only exists in the agents i, $0 \in \mathcal{N}_i$, since the exo-system state is only available for the agents i, $0 \in \mathcal{N}_i$. To begin with, it follows from (8)–(10) that

$$H(t+1) - F(t+1)\theta = \gamma\left[H(t) - F(t)\theta\right]. \tag{13}$$

Let an error $\varepsilon(t) = H(t) - F(t)\theta$. Then, it follows from (13) that the error dynamics is

$$\varepsilon(t+1) = \gamma\varepsilon(t). \tag{14}$$

Clearly, $\varepsilon(t) = 0$ for $t = t_0, t_0+1, \ldots$, since $F(t_0) = 0$ and $G(t_0) = 0$. In addition, even though $\varepsilon(t) \neq 0$ suddenly suffers from the measurement error, the error dynamics of $\varepsilon(t)$ can still be stabilized by setting the tunable gain $\gamma \in (0,1)$. Thus, we have

$$H(t) = F(t)\theta, \quad t = t_0, t_0+1, \ldots. \tag{15}$$

Since $\chi_v(t)^{\mathrm{T}}\chi_v(t) \geqslant 0$, $F(t)$ defined in (9) is a nondecreasing positive semidefinite time-varying square matrix, i.e., $F(t_2) \geqslant F(t_1) \geqslant 0$ for $t_2 > t_1$. Considering that $\gamma \in (0,1)$, if $\chi_v(t)$ satisfies the FE condition at t_{T_s}, then

$$F(t_{T_s}) = \sum_{t=t_0}^{t_0+T_s}\gamma^{t-t_0-1}\chi_v(t)^{\mathrm{T}}\chi_v(t) \geqslant \beta I_{q^2}.$$

Thus, we have $F(t) > 0$, $\forall t \geqslant t_{T_s}$, which indicates that the square matrix $F(t)$ is of full rank for all $t \geqslant t_{T_s}$. The full rank of $F(t)$ ensures the existence of a unique solution to the estimation equation (15). Therefore, the parameters of S, i.e., θ, can be estimated by calculating the following equation:

$$\hat{\theta} = F(t_{T_s})^{-1}H(t_{T_S}), \tag{16}$$

where t_{T_s} is the instant when the FE condition is met. Accordingly, the proof of the one-step estimation update law for θ is completed.

Let $\tilde{\theta}(t) = \hat{\theta}(t) - \theta$ denote the estimation error. It follows from (8), (12), and (15) that the dynamics of the estimation error $\tilde{\theta}(t)$ is

$$\begin{aligned}\tilde{\theta}(t+1) &= \hat{\theta}(t) + k_{\theta_1}\chi_v(t)^{\mathrm{T}}\left(v(t+1) - \chi_v(t)\hat{\theta}(t)\right) + k_{\theta_2}\left(H(t) - F(t)\hat{\theta}(t)\right) - \theta\\ &= \tilde{\theta}(t) + k_{\theta_1}\chi_v(t)^{\mathrm{T}}\left(\chi_v(t)\theta - \chi_v(t)\hat{\theta}(t)\right) + k_{\theta_2}\left(F(t)\theta(t) - F(t)\hat{\theta}(t)\right)\\ &= \tilde{\theta}(t) - k_{\theta_1}\chi_v(t)^{\mathrm{T}}\chi_v(t)\tilde{\theta}(t) - k_{\theta_2}F(t)\tilde{\theta}(t).\end{aligned} \tag{17}$$

Consider the following Lyapunov function:

$$V(t) = \frac{1}{2}\tilde{\theta}^{\mathrm{T}}(t)\tilde{\theta}(t).$$

If $\chi_v(t)$ satisfies the FE condition at t_{T_s}, it follows from (17) that

$$\begin{aligned} V(t+1)-V(t) &= -\tilde{\theta}^{\mathrm{T}}(t)\left(k_{\theta_1}\chi_v(t)^{\mathrm{T}}\chi_v(t)+k_{\theta_2}F(t)\right)\tilde{\theta}(t) \\ &\leqslant -\tilde{\theta}^{\mathrm{T}}(t)k_{\theta_2}F(t)\tilde{\theta}(t) \\ &\leqslant -k_{\theta_2}\beta V(t), \quad t \geqslant t_{T_s}. \end{aligned} \tag{18}$$

Therefore, since $k_{\theta_2} > 0$ and $\beta > 0$, the estimation error $\tilde{\theta}$ converges to zero under the dynamic update law (12). This completes the proof.

Remark 2. Compared to the estimation update law (12) where the start instant for estimating S is t_0 and the estimation error converges to zero as $t \to \infty$, the estimation update law (11) starts from t_{T_s} when the matrix $F(t)$ is of full rank. In addition, if there exists no measurement error for all $t > t_0$, the estimation error of the estimation update law (11) is zero for $t \geqslant t_{T_s}$.

For the agents i, $0 \in \mathcal{N}_i$, which have access to the exo-system state, using the estimation solution $\hat{\theta}$ to (11) or (12), the minimal polynomial of the estimated exo-system matrix $\hat{S}$ can be obtained as

$$\Lambda_{\hat{S}}(\lambda) = \lambda^{q_m} + \alpha_{q_m-1}\lambda^{q_m-1} + \cdots + \alpha_1\lambda + \alpha_0, \tag{19}$$

with $q_m \leqslant q$. Then, combining (19) with the internal model principle given in [15, Chapter 1], design the dynamics of the internal state as

$$z_i(t+1) = \mathcal{G}_1^0 z_i(t) + \mathcal{G}_2\hat{e}_i(t), \tag{20}$$

where $z_i \in \mathbb{R}^{pq_m}$ is the internal state, $\hat{e}_i$ is defined in (5), and the pair $(\mathcal{G}_1^0, \mathcal{G}_2)$ is the designed internal model parameters which can be explicitly given as

$$\mathcal{G}_1^0 = \text{block diag}\{\underbrace{G_1^0,\ldots,G_1^0}_{p\text{-tuple}}\} \in \mathbb{R}^{pq_m\times pq_m}, \quad \mathcal{G}_2 = \text{block diag}\{\underbrace{G_2,\ldots,G_2}_{p\text{-tuple}}\} \in \mathbb{R}^{pq_m\times p}, \tag{21}$$

with

$$G_1^0 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \cdots & -\alpha_{q_m-1} \end{bmatrix}, \quad G_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b \end{bmatrix}, \tag{22}$$

where α_j is the coefficients of λ in (19) with $j = 0, 1, \ldots, q_m - 1$, and b is a nonzero constant that ensures the controllability of the pair (G_1^0, G_2).

For the agents i, $0 \notin \mathcal{N}_i$, due to the inaccessibility of the state and dynamics of the exo-system, the matrix $\mathcal{G}_1^0$ cannot be directly obtained. Inspired by [9], although $\mathcal{G}_1^0$ is only obtainable for the agents i, $0 \in \mathcal{N}_i$, we can still design the online distributed adaptive internal model $\mathcal{G}_1^i(t)$ for all agents $i \in \mathcal{V}_0^+$ as

$$z_i(t+1) = \mathcal{G}_1^i(t)z_i(t) + \mathcal{G}_2\hat{e}_i(t), \tag{23}$$

with $\mathcal{G}_2$ defined in (21) and

$$\mathcal{G}_1^i(t+1) = \mathcal{G}_1^i(t) + \mu_1\sum_{j\in\mathcal{N}_i} a_{ij}\left(\mathcal{G}_1^j(t) - \mathcal{G}_1^i(t)\right), \quad \mathcal{G}_1^i(t_0) = 0, \tag{24}$$

where $0 < \mu_1 < 2/\rho(\mathcal{H})$ is a positive tuning constant and $\mathcal{G}_1^i$ has the same form as $\mathcal{G}_1^0$, i.e.,

$$\mathcal{G}_1^i = \text{block diag}\{\underbrace{G_1^i,\ldots,G_1^i}_{p\text{-tuple}}\} \in \mathbb{R}^{pq_m\times pq_m}. \tag{25}$$

Now, we show the stability of the closed-loop multi-agent system under the cooperative control policy embedding with the distributed adaptive internal model designed in (23) and (24), which is summarized in Theorem 2.

Remark 3. According to the internal model principle given in [15, Chapter 1], $\mathcal{G}_1^0$ must have exactly p invariant factors, each of which is divisible by the minimal polynomial of S. Hence, $\mathcal{G}_1^0$ is determined by (19). Once the estimation process for approximating S is completed, $\mathcal{G}_1^0$ can be designed for the agents i, $0 \in \mathcal{N}_i$. Then, for the agents i, $0 \notin \mathcal{N}_i$, using (24) where the information of $\mathcal{G}_1^j$ of the agent j, $j \in \mathcal{N}_i$ is required, the internal model parameter $\mathcal{G}_1^i$ of the agents i, $0 \notin \mathcal{N}_i$ can be directly approximated without estimating the exo-system matrix S.

Theorem 2. Under Assumptions 1–5, if there exists an instant t_{T_s}, such that $\chi_v(t)$ satisfies the FE condition, then the CORP of the multi-agent system (1)–(4) is solved by the dynamic feedback controller with the online distributed adaptive internal model given as

$$\begin{cases} u_i(t) = -K_x^i x_i(t) - K_z^i z_i(t), \\ z_i(t+1) = \mathcal{G}_1^i(t) z_i(t) + \mathcal{G}_2 \hat{e}_i(t), \\ \mathcal{G}_1^i(t+1) = \mathcal{G}_1^i(t) + \mu_1 \sum\limits_{j \in \mathcal{N}_i} a_{ij}(\mathcal{G}_1^j(t) - \mathcal{G}_1^i(t)), \end{cases} \tag{26}$$

with $i \in \mathcal{V}_0^+$, $\mathcal{G}_1^i(t_0) = 0 \in \mathbb{R}^{pq_m \times pq_m}$, and the pair $(K_x^i,\ K_z^i)$ satisfying that the matrix

$$A_c^i = \begin{bmatrix} A_i - B_i K_x^i & -B_i K_z^i \\ \mathcal{G}_2 C_i & \mathcal{G}_1^0 \end{bmatrix} \tag{27}$$

is Schur.

Proof. Under Assumptions 1 and 2, there exists a Schur matrix A_c^i for each agent $i \in \mathcal{V}_0^+$, such that the following augmented regulator equations [15]:

$$X_i S = (A_i - B_i K_x^i) X_i - B_i K_z^i Z_i + E_i, \tag{28}$$

$$Z_i S = \mathcal{G}_1^0 Z_i, \tag{29}$$

$$0 = C_i X_i - F \tag{30}$$

have a unique solution (X_i^*, Z_i^*), where $\mathcal{G}_1^0$ defined in (21) is determined by $\hat{\theta}$. As proven in Theorem 1, $\hat{\theta}$ is the estimation of the exo-system dynamics where the estimation error converges to zeros once $\chi_v(t)$ satisfies the FE condition.

Let

$$\bar{x}_i = x_i - X_i^* v, \tag{31}$$

$$\bar{z}_i = z_i - Z_i^* v. \tag{32}$$

Using the above transformation equations (31) and (32), the closed-loop system (1)–(4) of each agent $i \in \mathcal{V}_0^+$ with the dynamic feedback controller (26) can be transformed into the following compact system:

$$\begin{aligned} \bar{x}_z^i(t+1) &= A_c^i \bar{x}_z^i(t) - \sum_{j \in \mathcal{N}_i} \frac{a_{ij}}{\sum_{j=0}^N a_{ij}} \begin{bmatrix} 0 \\ \mathcal{G}_2 \bar{C}_j \end{bmatrix} \bar{x}_z^j(t) + \begin{bmatrix} 0 & 0 \\ 0 & \tilde{\mathcal{G}}_1^i(t) \end{bmatrix} \bar{x}_z^i(t) + \begin{bmatrix} 0 \\ \tilde{\mathcal{G}}_1^i(t) Z_i^* v(t) \end{bmatrix} \\ &:= A_c^i \bar{x}_z^i(t) - L_0^i(t) + L_1^i(t) \bar{x}_z^i(t) + L_2^i(t), \end{aligned} \tag{33}$$

$$e_i(t) = \bar{C}_i \bar{x}_z^i(t), \tag{34}$$

where $\bar{x}_z^i = \mathrm{col}(\bar{x}_i, \bar{z}_i)$, $\bar{x}_z^0 = 0$, $\bar{C}_i = [C_i,\ 0]$, $\tilde{\mathcal{G}}_1^i(t) = \mathcal{G}_1^i(t) - \mathcal{G}_1^0$ is the estimation error of internal model, and

$$L_0^i = \sum_{j \in \mathcal{N}_i} \frac{a_{ij}}{\sum_{j=0}^N a_{ij}} \begin{bmatrix} 0 \\ \mathcal{G}_2 \bar{C}_j \end{bmatrix} \bar{x}_z^j(t), \quad L_1^i(t) = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{\mathcal{G}}_1^i(t) \end{bmatrix}, \quad L_2^i(t) = \begin{bmatrix} 0 \\ \tilde{\mathcal{G}}_1^i(t) Z_i^* v(t) \end{bmatrix}. \tag{35}$$

Following the same derivation from [39], one can always label all the agents such that $i < j$ if $(i, j) \in \mathcal{E}$ under Assumption 3. Thus, the representation that contains the dynamic of all the agents is given as

$$\bar{x}_z(t+1) = A_c \bar{x}_z(t) + L_1(t) \bar{x}_z(t) + L_2(t), \tag{36}$$

where

$$\bar{x}_z = \operatorname{col}\left(\bar{x}_z^1, \ldots, \bar{x}_z^N\right), \tag{37}$$

$$A_c = \text{block diag}\left\{A_c^1, \ldots, A_c^N\right\}, \tag{38}$$

$$L_1 = \text{block diag}\left\{L_1^1, \ldots, L_1^N\right\}, \tag{39}$$

$$L_2 = \operatorname{col}\left(L_2^1, \ldots, L_2^N\right). \tag{40}$$

According to [9, Lemma 2], the estimation error of the adaptive internal model $\tilde{\mathcal{G}}_1^i(t)$ converges to zero, which implies that $\lim_{t\to\infty} L_1(t) = 0$. In addition, since the exo-system state is bounded, we have $\lim_{t\to\infty} L_2(t) = 0$. As stated in [15, Chapter 1.6], for each agent $i \in \mathcal{V}_0^+$, there always exists a pair $(K_x^i,\ K_z^i)$, such that the matrix A_c^i is Schur if Assumptions 1 and 2 are satisfied. This implies that there exist pairs $(K_x^i,\ K_z^i)$, such that A_c is Schur, which indicates that $\lim_{t\to\infty} \bar{x}_z(t) = 0$. Then, it follows from (34) that the tracking error $\lim_{t\to\infty} e_i(t) = 0$, which means that the CORP of the multi-agent system is solved by the proposed dynamic feedback controller (26). Thus, the proof is completed.

3.2 Data-driven learning algorithm for solving the COORP

In this subsection, we first define a class of cost functions for the performance index described in (6). Then, combining the existence of the solution to regulator equations, the COORP is transformed into a cooperative optimal control problem. A model-free ADP-based learning algorithm with employing the VI scheme is developed to estimate the optimal control gain by approximating the solution to the corresponding ARE. Finally, the convergence analysis is given to show that the estimated distributed controllers converge to the optimal distributed control policies.

3.2.1 *COORP transformation*

Since $\mathcal{G}_1^i$ converges to $\mathcal{G}_1^0$ as shown in [9, Lemma 2], using the transformation equations (31) and (32), the augmented multi-agent system of the representation (36) with the online desired internal model (20) can be rewritten as

$$\bar{x}_z(t+1) = \bar{A}\bar{x}_z(t) + \bar{B}\bar{u}(t), \tag{41}$$

where

$$\bar{u} = [\bar{u}_i^{\mathrm{T}}, \ldots, \bar{u}_N^{\mathrm{T}}]^{\mathrm{T}} \tag{42}$$

is the control input with $\bar{u}_i = -K_i\bar{x}_z^i$ and $K_i = [K_x^i,\ K_z^i]$, $\bar{A} = \text{block diag}\left\{\bar{A}_1, \ldots, \bar{A}_N\right\}$ with $\bar{A}_i = \left[\begin{smallmatrix} A_i & 0 \\ \mathcal{G}_2 C_i & \mathcal{G}_1^0 \end{smallmatrix}\right]$, and $\bar{B} = \text{block diag}\left\{\bar{B}_1, \ldots, \bar{B}_N\right\}$ with $\bar{B}_i = \left[\begin{smallmatrix} B_i \\ 0 \end{smallmatrix}\right]$.

It follows from (31) and (32) that

$$\bar{u}_i = -K_i\bar{x}_z^i = u_i - (-K_x^i X_i^* v - K_z^i Z_i^* v),$$

where $u_i = -K_x^i x_i - K_z^i z_i$.

Now, modify the cost function in (6) for each agent $i \in \mathcal{V}_0^+$ as

$$\mathcal{C}_i(x_i(t), u_i(t)) = \|x_i - x_d^i\|_{Q_i^x} + \|z_i - z_d^i\|_{Q_i^z} + \|u_i - u_d^i\|_{R_i}, \tag{43}$$

where $Q_i^x = (Q_i^x)^{\mathrm{T}} > 0$, $Q_i^z = (Q_i^z)^{\mathrm{T}} > 0$, and $R_i = (R_i)^{\mathrm{T}} > 0$ are the weight matrices with appropriate dimensions.

Let $x_d^i = X_i^* v$, $z_d^i = Z_i^* v$, and $u_d^i = -K_x^i X_i^* v - K_z^i Z_i^* v$. It follows from (26) and (28)–(32) that the cost function (43) can be rewritten as

$$\mathcal{C}_i(x_i(t), u_i(t)) = \left(\bar{x}_z^i\right)^{\mathrm{T}} Q_i \left(\bar{x}_z^i\right) + \left(\bar{u}_i\right)^{\mathrm{T}} R_i \left(\bar{u}_i\right), \tag{44}$$

where $Q_i = \text{block diag}\{Q_i^x,\ Q_i^z\}$.

That is, the COORP for the multi-agent system can be transformed as a tractable cooperative optimal control problem subject to the augmented compact system (41) as formulated in Problem 1.

Problem 1.

$$\min_{\bar{u}} \sum_{i=0}^{N} \sum_{t=t_0}^{\infty} \left(\bar{x}_z^i\right)^{\mathrm{T}} Q_i \left(\bar{x}_z^i\right) + \left(\bar{u}_i\right)^{\mathrm{T}} R_i \left(\bar{u}_i\right) \tag{45}$$
$$\text{subject to } \ \bar{x}_z(t+1) = \bar{A}\bar{x}_z(t) + \bar{B}\bar{u}(t).$$

Minimizing the cost function with respect to the policy $\bar{u}$ gives the following optimal control policy:

$$\bar{u}_i^* = -K_i^* \bar{x}_z^i, \tag{46}$$

where $K_i^* = \left(R_i + \bar{B}_i^{\mathrm{T}} P_i^* \bar{B}_i\right)^{-1} \bar{B}_i^{\mathrm{T}} P_i^* \bar{A}_i$ is the optimal control gain and P_i^* is the symmetric positive definite solution to the following ARE:

$$\bar{A}_i^{\mathrm{T}} P_i \bar{A}_i - P_i + Q_i - \bar{A}_i^{\mathrm{T}} P_i \bar{B}_i \left(R_i + \bar{B}_i^{\mathrm{T}} P_i \bar{B}_i\right)^{-1} \bar{B}_i^{\mathrm{T}} P_i \bar{A}_i = 0. \tag{47}$$

Then, we introduce a one-step recursion of the value update in the model-based VI scheme [43, Chapter 17.5] for solving the ARE (47), which is given as

$$P_i^{(k+1)} = \bar{A}_i^{\mathrm{T}} P^{(k)} \bar{A}_i - \bar{A}_i^{\mathrm{T}} P_i^{(k)} \bar{B}_i \left(R_i + \bar{B}_i^{\mathrm{T}} P_i^{(k)} \bar{B}_i\right)^{-1} \bar{B}_i^{\mathrm{T}} P_i^{(k)} \bar{A}_i + Q_i. \tag{48}$$

The policy evaluation is

$$K_i^{(k+1)} = \left(R_i + \bar{B}_i^{\mathrm{T}} P_i^{(k+1)} \bar{B}_i\right)^{-1} \bar{B}_i^{\mathrm{T}} P_i^{(k+1)} \bar{A}_i. \tag{49}$$

According to [43, Lemma 17.5.4], for $P_i^{(k+1)}$, $k = 0, 1, 2, \ldots$, evaluated in (48) with any initial symmetric positive matrix $P_i^{(0)} \in \mathbb{P}_+^{n_i}$, we have $\lim_{k\to\infty} P_i^{(k)} = P_i^*$ and $\lim_{k\to\infty} K_i^{(k)} = K_i^*$ with $(\bar{A}_i - \bar{B}_i K_i^*)$ being a Schur matrix.

Notably, if the system matrices $\bar{A}_i$ and $\bar{B}_i$ are unknown, the solutions to the regulator equations (28)–(30) and ARE (47) are unavailable. Hence, the steady state x_d^i, steady input u_d^i, and the optimal control gain K_i^* are unknown. Moreover, the data $(\bar{x}_z, \bar{u})$ of the system (41) are unavailable due to the absence of knowledge of x_d^i, z_d^i, and u_d^i, which makes the design of the online learning algorithm for solving the COORP more difficult.

3.2.2 *Model-free online VI-based learning algorithm*

To begin with, we combine the original multi-agent system (1) with the distributed adaptive internal model (23) to obtain

$$x_z^i(t+1) = \bar{A}_i x_z^i(t) + \bar{B}_i u_i + \begin{bmatrix} E_i \\ -\varrho_{i0}\mathcal{G}_2 F \end{bmatrix} v(t) - \sum_{j\in\mathcal{N}_i} \varrho_{ij} \begin{bmatrix} 0 \\ \mathcal{G}_2 C_j \end{bmatrix} x_j(t) + \begin{bmatrix} 0 \\ \tilde{\mathcal{G}}_1^i(t) z_i(t) \end{bmatrix}, \tag{50}$$

where $x_z^i = \mathrm{col}(x_i, z_i)$, $x_0 = v$, $C_0 = F$, and $\varrho_{ij} = a_{ij}/\sum_{j=0}^{N} a_{ij}$.

The main idea of the online ADP learning algorithm is that, by replacing the system matrices in the ARE with the online data, a corresponding iterative learning equation is established for estimating the solution to ARE without using any prior knowledge of system dynamics. Note that the solution to the ARE (47) only relies on the system matrices $\bar{A}_i$ and $\bar{B}_i$. If there exists a related system that has the same system matrices $(\bar{A}_i, \bar{B}_i)$ as those of the system (41), then the ADP-based learning algorithm can be established by the online data from the related system without using any knowledge of $(\bar{A}_i, \bar{B}_i)$.

Moreover, using the online data x_z^i and u_i from (50) to replace the data $\bar{x}_z^i$ and $\bar{u}_i$ in (41), since it has the same system matrices $(\bar{A}_i, \bar{B}_i)$ as those of system (41), the online ADP-based learning algorithm can be established for estimating the solution to the ARE, i.e., K_i^*, without using any knowledge of $(\bar{A}_i, \bar{B}_i)$.

To deal with the issue that the exo-system state $v(t)$ is unavailable to the agents i, $0 \notin \mathcal{N}_i$, in most of the existing studies for solving the CORP, for instance, [9,10,30,38,39,41], the adaptive observer network is introduced in the agents i, $0 \notin \mathcal{N}_i$ to estimate the exo-system state, which could demand a large data communication load to transmit the estimated state data among the agents. Instead of estimating the state $v(t)$, we consider a transformation method to construct a signal $\hat{v}(t)$ to replace $v(t)$ in (50), as shown below.

As $G_1^i(t)$ converges to G_1^0, the minimal polynomial of S is obtained for each agent $i \in \mathcal{V}_0^+$. Thus, we can always find a signal $\hat{v}(t)$ and a known matrix $\hat{G}_1^i(t) \in \mathbb{R}^{q_m \times q_m}$ satisfying

$$\hat{v}_i(t+1) = \hat{G}_1^i(t)\hat{v}_i(t), \tag{51}$$

$$v(t) = S_t^i \hat{v}_i(t), \tag{52}$$

where $S_t^i \in \mathbb{R}^{q \times q_m}$ is an unknown matrix.

Thus, combining (51) and (52), the system (50) can be rewritten as

$$\begin{aligned} x_z^i(t+1) &= \bar{A}_i x_z^i(t) + \bar{B}_i u_i + \begin{bmatrix} E_i S_t^i \\ -\varrho_{i0}\mathcal{G}_2 F S_t^i \end{bmatrix} \hat{v}_i(t) - \sum_{j\in\mathcal{N}_i} \varrho_{ij} \begin{bmatrix} 0 \\ \mathcal{G}_2 C_j \end{bmatrix} x_j(t) + \begin{bmatrix} 0 \\ \tilde{\mathcal{G}}_1^i(t) z_i(t) \end{bmatrix} \\ &:= \bar{A}_i x_z^i(t) + \bar{B}_i u_i + \bar{E}_i \hat{v}_i(t) + \bar{D}_i x_j^+(t) + \varepsilon_i(t), \end{aligned} \tag{53}$$

where $x_j^+(t)$ is the vector lumping all $x_j(t)$ for $j \in \mathcal{N}_i$ and $\bar{D}_i$ is the matrix lumping all $\bar{D}_j$ for $j \in \mathcal{N}_i$ with

$$\bar{D}_j = -\varrho_{ij} \begin{bmatrix} 0 \\ \mathcal{G}_2 C_j \end{bmatrix}, \quad \bar{E}_i = \begin{bmatrix} E_i S_t^i \\ -\varrho_{i0}\mathcal{G}_2 F S_t^i \end{bmatrix}, \quad \varepsilon_i(t) = \begin{bmatrix} 0 \\ \tilde{\mathcal{G}}_1^i(t) z_i(t) \end{bmatrix}. \tag{54}$$

The above statements indicate that $v(t)$ is replaced by a signal $\hat{v}_i(t)$ generated by the agent i itself without interacting with the estimated state from other agents, which could lead to a reduction in the data communication load and computational load, compared to [9, 10, 30, 38, 39, 41] where an adaptive observer network is needed to interact the estimated state among the agents.

Combining (48) with the data from system (53), we have

$$\begin{aligned} x_z^i(t+1)^{\mathrm{T}} P_i^{(k)} x_z^i(t+1) &= \begin{bmatrix} x_z^i(t) \\ u_i(t) \\ \hat{v}_i(t) \\ x_j^+(t) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \bar{A}_i^{\mathrm{T}} P_i^{(k)} \bar{A}_i & \bar{A}_i^{\mathrm{T}} P_i^{(k)} \bar{B}_i & \bar{A}_i^{\mathrm{T}} P_i^{(k)} \bar{E}_i & \bar{A}_i^{\mathrm{T}} P_i^{(k)} \bar{D}_i \\ * & \bar{B}_i^{\mathrm{T}} P_i^{(k)} \bar{B}_i & \bar{B}_i^{\mathrm{T}} P_i^{(k)} \bar{E}_i & \bar{B}_i^{\mathrm{T}} P_i^{(k)} \bar{D}_i \\ * & * & \bar{E}_i^{\mathrm{T}} P_i^{(k)} \bar{E}_i & \bar{E}_i^{\mathrm{T}} P_i^{(k)} \bar{D}_i \\ * & * & * & \bar{D}_i^{\mathrm{T}} P_i^{(k)} \bar{D}_i \end{bmatrix} \begin{bmatrix} x_z^i(t) \\ u_i(t) \\ \hat{v}_i(t) \\ x_j^+(t) \end{bmatrix} + \tilde{\varepsilon}_i(t) \\ &:= \begin{bmatrix} x_z^i(t) \\ u_i(t) \\ \hat{v}_i(t) \\ x_j^+(t) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathcal{P}_{11}^{(i,k)} & \mathcal{P}_{12}^{(i,k)} & \mathcal{P}_{13}^{(i,k)} & \mathcal{P}_{14}^{(i,k)} \\ * & \mathcal{P}_{22}^{(i,k)} & \mathcal{P}_{23}^{(i,k)} & \mathcal{P}_{24}^{(i,k)} \\ * & * & \mathcal{P}_{33}^{(i,k)} & \mathcal{P}_{34}^{(i,k)} \\ * & * & * & \mathcal{P}_{44}^{(i,k)} \end{bmatrix} \begin{bmatrix} x_z^i(t) \\ u_i(t) \\ \hat{v}_i(t) \\ x_j^+(t) \end{bmatrix} + \tilde{\varepsilon}_i(t), \quad t > t_{T_d}, \end{aligned} \tag{55}$$

where t_{T_d} is the instant when $\|\mathcal{G}_1^i(t) - \mathcal{G}_1^j(t)\| < \epsilon_1$ is achieved for all $(j,i) \in \mathcal{E}$ with a small threshold $\epsilon_1 > 0$, and $\tilde{\varepsilon}_i(t) = (2x_z^i(t{+}1) - \varepsilon_i(t))^{\mathrm{T}} P_i^{(k)} \varepsilon_i(t)$. As shown in [9, Lemma 2], the estimation error of $\tilde{\mathcal{G}}_1^i(t)$ converges to zero by tuning $0 < \mu_1 < 2/\rho(\mathcal{H})$. Since $P_i^{(k)}$ and $x_z^i(t)$ are usually bounded, the term $\tilde{\varepsilon}_i(t)$ could be ignored by setting a small enough ϵ_1.

Next, we define the functions Γ_{v_1,v_2} and Δ_{v_1,v_1} associated with the column vectors v_1 and v_2 as

$$\Gamma_{v_1,v_2}(t) = [v_1(t) \otimes v_2(t)]^{\mathrm{T}}, \quad \Delta_{v_1,v_1}(t) = \mathrm{vech}\left(2v_1(t)v_1(t)^{\mathrm{T}} - \mathrm{dia}\left(v_1(t)\right)\mathrm{dia}\left(v_1(t)\right)\right)^{\mathrm{T}}. \tag{56}$$

Using the functions defined in (56), Eq. (55) can be rewritten as

$$\bar{\Psi}_i(T) \begin{bmatrix} \mathrm{vech}(\mathcal{P}_{11}^{(i,k)}) \\ \mathrm{vec}(\mathcal{P}_{12}^{(i,k)}) \\ \mathrm{vech}(\mathcal{P}_{22}^{(i,k)}) \\ \Pi \end{bmatrix} = \bar{\Phi}_i^{(k)}(T+1), \tag{57}$$

where

$$\bar{\Psi}_i(T) = [\Psi_i(t_{T_d})^{\mathrm{T}}, \Psi_i(t_{T_d}+1)^{\mathrm{T}}, \ldots, \Psi_i(t_{T_d}+T)^{\mathrm{T}}]^{\mathrm{T}}, \tag{58}$$

$$\bar{\Phi}_i^{(k)}(T+1) = [\Phi_i^{(k)}(t_{T_d}+1)^{\mathrm{T}}, \Phi_i^{(k)}(t_{T_d}+2)^{\mathrm{T}}, \ldots, \Psi_i^{(k)}(t_{T_d}+T+1)^{\mathrm{T}}]^{\mathrm{T}}, \tag{59}$$

$$\Pi = \left[\mathrm{vec}(\mathcal{P}_{13}^{(i,k)})^{\mathrm{T}}, \mathrm{vec}(\mathcal{P}_{14}^{(i,k)})^{\mathrm{T}}, \mathrm{vec}(\mathcal{P}_{23}^{(i,k)})^{\mathrm{T}}, \mathrm{vec}(\mathcal{P}_{24}^{(i,k)})^{\mathrm{T}}, \mathrm{vech}(\mathcal{P}_{33}^{(i,k)})^{\mathrm{T}}, \mathrm{vec}(\mathcal{P}_{34}^{(i,k)})^{\mathrm{T}}, \mathrm{vech}(\mathcal{P}_{44}^{(i,k)})^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{60}$$

with

$$\Phi_i^{(k)}(t+1) = x_z^i(t+1)^{\mathrm{T}} P_i^{(k)} x_z^i(t+1), \tag{61}$$

$$\Psi_i(t) = \left[\Delta_{x_z^i x_z^i}, 2\Gamma_{x_z^i u_i}, \Delta_{u_i u_i}, 2\Gamma_{x_z^i \hat{v}_i}, 2\Gamma_{x_z^i x_j^+}, 2\Gamma_{u_i \hat{v}_i}, 2\Gamma_{u_i x_j^+}, \Delta_{\hat{v}_i \hat{v}_i}, 2\Gamma_{\hat{v}_i x_j^+}, \Delta_{x_j^+ x_j^+}\right]. \tag{62}$$

Note that Eq. (57) is the so-called online VI-based iterative learning equation, which requires the data of x_z^i, u_i, $\hat{v}_i$, $P_i^{(k)}$, and x_j, $j \in \mathcal{N}_i$ to construct the related matrices $\bar{\Psi}_i$ and $\bar{\Phi}_i^{(k)}$ for each agent $i \in \mathcal{V}_0^+$ at each iteration k. It follows from (48) and (49) that, given an initial $P_i^{(0)} \in \mathbb{P}_+^{n_i}$, for $k = 0, 1, 2, \ldots$, the online VI-based value update and control policy evaluation are

$$P_i^{(k+1)} = \mathcal{P}_{11}^{(i,k)} - \mathcal{P}_{12}^{(i,k)} \left(R_i + \mathcal{P}_{22}^{(i,k)}\right)^{-1} \left(\mathcal{P}_{12}^{(i,k)}\right)^{\mathrm{T}} + Q_i, \tag{63}$$

$$K_i^{(k+1)} = \left(R_i + \mathcal{P}_{22}^{(i,k)}\right)^{-1} \left(\mathcal{P}_{12}^{(i,k)}\right)^{\mathrm{T}}, \tag{64}$$

where $\mathcal{P}_{11}^{(i,k)}$, $\mathcal{P}_{12}^{(i,k)}$, and $\mathcal{P}_{22}^{(i,k)}$ are obtained by solving (57), and Q_i and R_i are user-defined weight matrices. As shown in most existing ADP-based learning algorithms, for instance, [30,39,41], the iterative learning equation (57) can be solved by the pseudo-inverse method, when $\bar{\Psi}_i(T)$ is of full-column rank but not a square matrix. However, it should be pointed out that there may exist calculation errors when the pseudo-inverse method is used.

Finally, the data-driven ADP-based learning algorithm for solving the COORP is described in Algorithm 1. The convergence and stability of Algorithm 1 are shown in Theorem 3.

3.3 Convergence and stability analyses

Theorem 3. For each agent $i \in \mathcal{V}_0^+$, if there exists an instant $t_{s^*} = t_{T_d} + s^*$ with $s^* \in \mathbb{Z}^+$, such that the rank condition

$$\mathrm{rank}\left(\bar{\Psi}_i(T)\right) = \frac{(n_i + pq_m + m_i + q_m + \sum_{j\in\mathcal{N}_i} n_j)(1 + n_i + pq_m + m_i + q_m + \sum_{j\in\mathcal{N}_i} n_j)}{2} \tag{65}$$

is satisfied for all $t_s > t_{s^*}$, then the COORP for the multi-agent system (1)–(4) is solved by employing the control policy estimated by Algorithm 1.

Proof. First, since $\bar{\Psi}_i(T)$ only relies on the collected data and is independent of either $P_i^{(k)}$ or $K_i^{(k)}$, $\bar{\Psi}_i(T)$ is of full column rank for each iteration k at any instant $t_s \geqslant t_{s^*}$, once the rank condition (65) is satisfied at t_{s^*}. If the above rank condition (65) holds for each agent $i \in \mathcal{V}_0^+$, then there always exists a unique solution to the iterative learning equation (57). Note that Eq. (57) is transformed from the value evaluation equation (48). Therefore, it follows from the properties of the VI scheme shown in [43, Chapter 17.5], i.e., $\lim_{k\to\infty} K_i^{(k)} = K_i^*$, that the estimated control gain $K_i^{(k)}$ converges to its optimal value.

Then, according to [43], the optimal solution K_i^* for minimizing the cost function (44) makes $(\bar{A}_i - \bar{B}_i K_i^*)$ be Schur matrix. Combining the stability analysis proven in Theorem 2, the controller (26) with the estimated optimal control gain $\hat{K}_i^*$ and the online distributed adaptive internal model by employing Algorithm 1 can achieve the aforementioned three objectives (i)–(iii) given in Section 2, which indicates that the COORP is solved by Algorithm 1. Thus, the proof is completed.

The initial control policy used in the proposed Algorithm 1 has no demand on a stabilizing control gain, yet it only needs two requirements: (1) The state response of a closed-loop system with the initial control policy is bounded; (2) The rank condition (65) can be met. The rank condition can be easily satisfied by only adding the exploration noise in the data collection process, and the exploration noise would be cut off once the rank condition is satisfied, which was utilized in most existing ADP-based learning algorithms, for instance, [28,29,32,38,39]. The judgment on the rank condition for each agent i would not affect each other if the initial control policies with exploration noise are chosen appropriately.

Algorithm 1 ADP-based data-driven learning algorithm for solving the COORP

```
1: Initialize: Give an arbitrary initial control policy sequences $\{u_0^i(t)\}_{t=0}^{\infty}$ with exploration noise series $\{\xi_0^i(t)\}_{t=0}^{\infty}$, initial $\hat{v}_i(0)$, and the weight matrices $Q_i$ and $R_i$ for each agent $i = 1, \dots, N$. Select $\gamma \in (0, 1)$, $k_{\theta_1} > 0$, $k_{\theta_2} > 0$, $0 < \mu_1 < 2/\rho(\mathcal{H})$, two small thresholds $\epsilon_1$ and $\epsilon_2$, and a maximum iteration number $k_N$. Define $\mathcal{V}_d = \mathcal{V}_0^+$.
2: Compute $F(t)$ and $H(t)$ by (9) and (10), respectively;
3: if the FE condition (7) of $F(t)$ holds
4:     Solve (11) or update (12) to obtain $\hat{S}$;
5:     Find the pair $(G_1^0, b)$ by (22);
6:     Construct the distributed adaptive internal model $(\mathcal{G}_1^i(t), \mathcal{G}_2)$ obtained by (24) for the agent $i \in \mathcal{V}_d$;
7: else
8:     Let $t = t + 1$, and then go to Step 2;
9: end if
10: if $\|\mathcal{G}_1^i(t) - \mathcal{G}_1^j(t)\| < \epsilon_1$ holds for all $(j, i) \in \mathcal{E}$
11:     Collect data and compute $\Psi_i(t)$ by (62) for agent $i \in \mathcal{V}_d$; Let $T = t$, $i = 1$;
12:     for $i \in \mathcal{V}_d$ do
13:         if the rank condition (65) holds for $\bar{\Psi}_i(T)$
14:             Let $k = 0$, and initialize $P_i^{(0)} \in \mathbb{P}_+^n$;
15:             Solve (57) to obtain $\mathcal{P}_{11}^{(i,k)}$, $\mathcal{P}_{12}^{(i,k)}$, and $\mathcal{P}_{22}^{(i,k)}$;
16:             Update $P_i^{k+1} = \mathcal{P}_{11}^{(i,k)} - \mathcal{P}_{12}^{(i,k)}(R_i + \mathcal{P}_{22}^{(i,k)})^{-1}(\mathcal{P}_{12}^{(i,k)})^{\mathrm{T}} + Q_i$;
17:             if $\|P_i^{k+1} - P_i^k\| < \epsilon_2$
18:                 Update control law $u_i = -\hat{K}_i^* x_z^i$ with $\hat{K}_i^* = (R_i + \mathcal{P}_{22}^{(i,k)})^{-1}(\mathcal{P}_{12}^{(i,k)})^{\mathrm{T}}$ for agent $i$;
19:                 Update the set $\mathcal{V}_d$ with deleting the element $i$ from $\mathcal{V}_d$;
20:             else if $k < k_N$
21:                 Let $k = k + 1$, and then go to Step 15;
22:             end if
23:         else
24:             Let $i = i + 1$;
25:             if $i \notin \mathcal{V}_d$ and $i < N$
26:                 Go to Step 24;
27:             end if
28:         end if
29:     end for
30:     if $\mathcal{V}_d \subset \emptyset$
31:         break
32:     else
33:         Let $t = t + 1$, and then go to Step 11.
34:     end if
35: end if
```

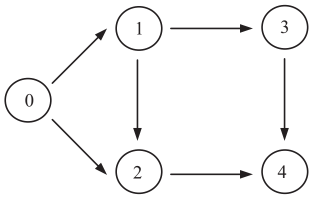

Figure 1 Communication graph of the multi-agent system.

4 Simulation

In this section, a numerical example is given to validate the effectiveness of the proposed data-driven distributed cooperative optimal output regulation control scheme. Consider a multi-agent system with four agents whose communication graph is shown in Figure 1, and the system dynamics for agent $i = 1, 2, 3, 4$ is described as

$$\begin{aligned} x_i(t+1) &= \begin{bmatrix} 0 & 1 \\ -i \times 0.1 & 1.1 \end{bmatrix} x_i(t) + \begin{bmatrix} 0 \\ 1 - i \times 0.1 \end{bmatrix} u_i(t) + \begin{bmatrix} 0 & 0 \\ 0 & -0.05 \times i \end{bmatrix} v(t), \\ v(t+1) &= \begin{bmatrix} 0.99500 & -0.09983 \\ 0.09983 & 0.99500 \end{bmatrix} v(t), \\ e_i(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x_i(t) + \begin{bmatrix} 2 & 0 \end{bmatrix} v(t), \end{aligned}$$

with the state initial values $x_i(0) = [i,\ i]^{\mathrm{T}}$ and $v(0) = [2,\ 0]^{\mathrm{T}}$. The weight matrices in the performance function for each agent i are set to $Q_i = i \times I_4 \in \mathbb{R}^{4\times4}$ and $R_i = i$ where I_4 is the unit matrix. In

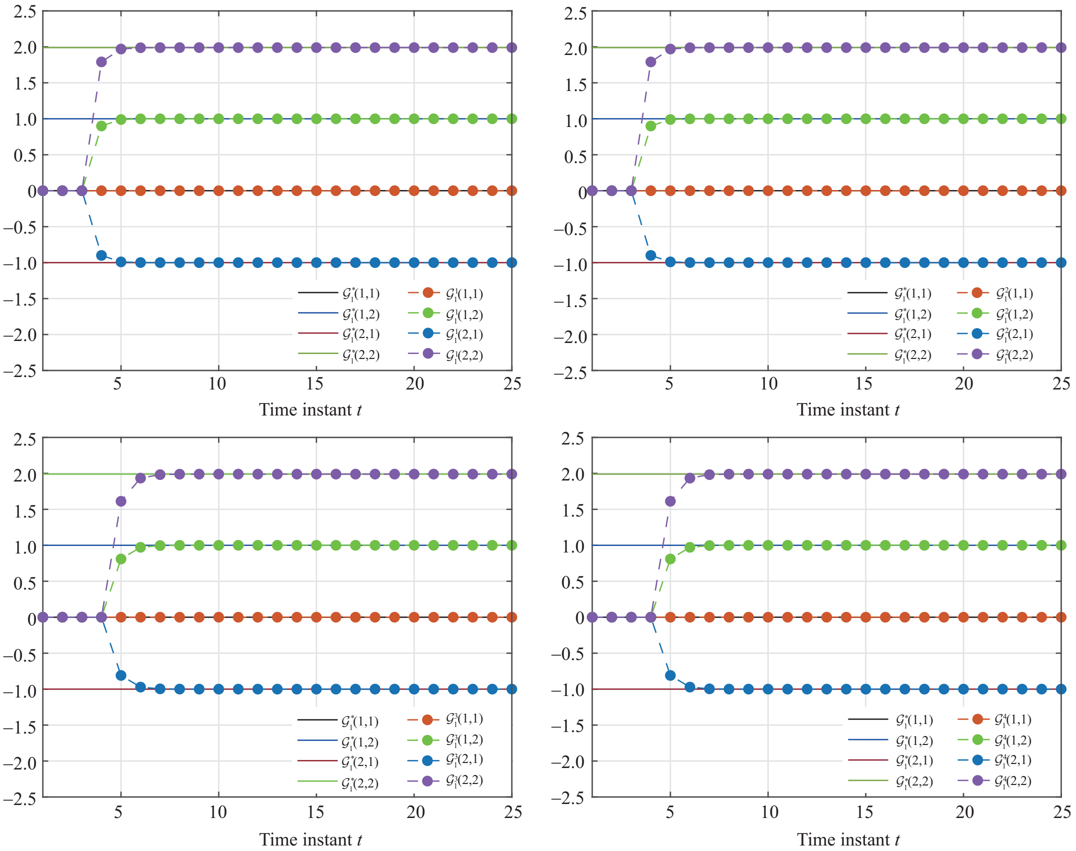

Figure 2 (Color online) Evolution of the distribution adaptive internal model parameters $\mathcal{G}_1^i(t)$.

addition, for each agent i, $\hat{v}_i(0) = [1,\ 1]^{\mathrm{T}}$. The parameters in the proposed algorithm are chosen as $\gamma = 0.9$, $\mu_1 = 0.9$, $\epsilon_1 = \epsilon_2 = 10^{-8}$, and the maximum iteration number is $k_N = 1000$. The initial control gain is chosen as $K_i^{(0)} = 0$, and the probing noise is selected as $\sum_{j=0}^{10} c_j \sin(2\pi\omega_j \Delta T \times t)$ with ΔT being the sampling time, which is a combination of several sinusoidal signals with different amplitudes c_j and frequencies ω_j. Then, we collect the online data to estimate the exo-system dynamics, distributed adaptive internal model, and optimal control gain for solving the COORP.

To illustrate the effectiveness of the proposed Algorithm 1, according to the given multi-agent system dynamics with known weight matrices Q_i and R_i, we first give the optimal control policy with the desired distributed internal model for each agent $i = 1, 2, 3, 4$ as

$$\begin{cases} u_i(t) = -K_i^* x_z^i(t), \\ z_i(t+1) = \mathcal{G}_1^* z_i(t) + \mathcal{G}_2^* \hat{e}_i(t), \end{cases} \tag{66}$$

with

$$\begin{aligned} K_1^* &= [-0.0098,\ 0.9423,\ \ -7.1773,\ 7.2764], \\ K_2^* &= [-0.0973,\ 0.9732,\ \ -7.6751,\ 7.7789], \\ K_3^* &= [-0.1964,\ 0.9981,\ \ -8.2404,\ 8.3476], \\ K_4^* &= [-0.3076,\ 1.0129,\ \ -8.8885,\ 8.9960], \end{aligned}$$

and

$$\mathcal{G}_1^* = \begin{bmatrix} 0 & 1 \\ -1.0000 & 1.9900 \end{bmatrix}, \quad \mathcal{G}_2^* = \begin{bmatrix} 0 \\ 0.01 \end{bmatrix}.$$

The dynamic state feedback controller with the distributed adaptive internal model used in the pro-

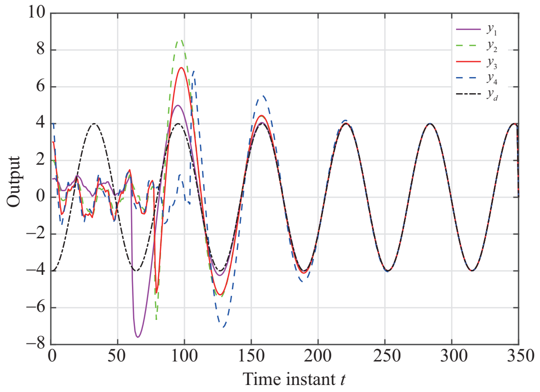

Figure 3 (Color online) Reference and outputs of each agent.

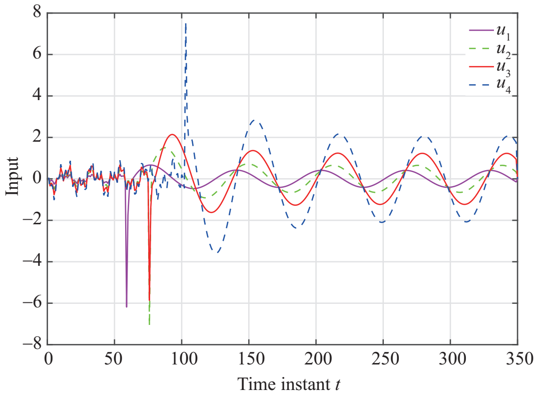

Figure 4 (Color online) Control inputs of each agent.

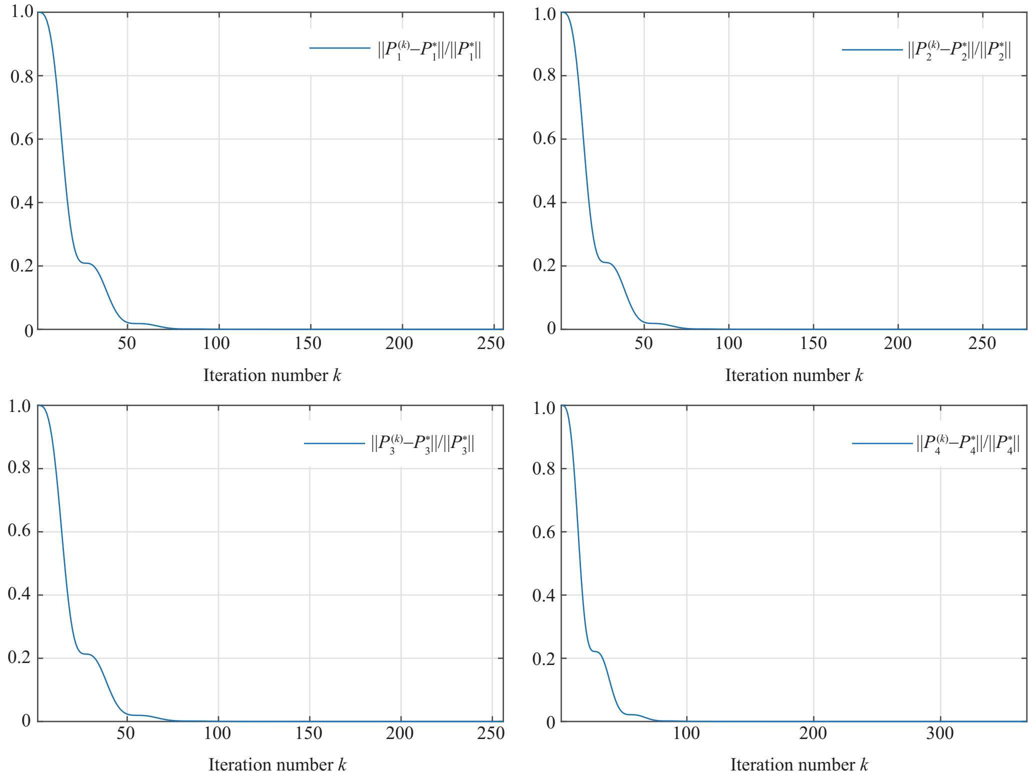

Figure 5 (Color online) Convergence performances of the online VI-based learning algorithm.

posed Algorithm 1 is given as

$$\begin{cases} u_i(t) = -\hat{K}_i^* x_z^i(t), \\ z_i(t+1) = \mathcal{G}_1^i(t) z_i(t) + \mathcal{G}_2 \hat{e}_i(t), \\ \mathcal{G}_1^i(t+1) = \mathcal{G}_1^i(t) + \mu_1 \sum_{j\in\mathcal{N}_i} a_{ij}(\mathcal{G}_1^j(t) - \mathcal{G}_1^i(t)). \end{cases} \tag{67}$$

Figure 2 shows that all the $\mathcal{G}_1^i(t)$ for each agent $i = 1, 2, 3, 4$ converges to $\mathcal{G}_1^*$ and the error accuracy ϵ_1 is achieved around the instant $t = 25$, which illustrates the convergence of the distributed adaptive internal model parameters. The trajectories of outputs and inputs are drawn in Figures 3 and 4, respectively, and Figure 5 shows the convergence performances of the online VI-based learning algorithm. In particular, Figure 3 shows that the control policies of each agent $i = 1, 2, 3, 4$ estimated by Algorithm 1 can achieve the reference tracking with disturbance rejection. In addition, Figures 4 and 5 indicate that the estimated control policies u_1, u_2, u_3, and u_4 are updated after 255, 276, 256, and 368 iterations, respectively, when the rank condition for agent $i = 1, 2, 3, 4$ is satisfied at $t = 58$, $t = 75$, $t = 75$, and $t = 102$, respectively, and the estimate values are

$$\begin{aligned}
\hat{K}_1^* &= [-0.0097,\ 0.9414,\ -7.1698,\ 7.2684],\\
\hat{K}_2^* &= [-0.0965,\ 0.9728,\ -7.7171,\ 7.8235],\\
\hat{K}_3^* &= [-0.1963,\ 0.9983,\ -8.2546,\ 8.3625],\\
\hat{K}_4^* &= [-0.3075,\ 1.0134,\ -8.8989,\ 9.0034].
\end{aligned}$$

The simulation results illustrate that, by employing Algorithm 1 without using any prior knowledge of the exo-system and multi-agent system or estimating the exo-system state, the estimated optimal control policies for each agent with a distributed adaptive internal model can solve the COORP.

5 Conclusion

This paper addresses the distributed cooperative optimal output regulation problem with completely unknown multi-agent systems. Two parameter approximation update laws are designed to estimate the exo-system dynamics, which ensures that the online distributed adaptive internal model is established without any prior knowledge of the exo-system dynamics. By only accessing the input, state, and output data, the data-driven learning algorithm is proposed by using the ADP method with the VI scheme to estimate the optimal control policy. The FE condition and rank condition guarantee the convergence and stability of the proposed model-free ADP-based learning algorithm. Our future work will focus on data-driven algorithms for adaptive optimal measurement/output feedback control with unmeasurable external disturbance for unknown systems.

Acknowledgements This work was supported in part by National Key R&D Program of China (Grant No. 2021ZD0112600), National Natural Science Foundation of China (Grant Nos. 61873219, 62173283), and Natural Science Foundation of Fujian Province of China (Grant No. 2021J01051).

References

1 Su Y F, Huang J. Cooperative output regulation of linear multi-agent systems. IEEE Trans Automat Contr, 2012, 57: 1062–1066

2 Su Y F, Huang J. Cooperative output regulation with application to multi-agent consensus under switching network. IEEE Trans Syst Man Cybern B, 2012, 42: 864–875

3 Dong Y, Huang J. Cooperative global output regulation for a class of nonlinear multi-agent systems. IEEE Trans Automat Contr, 2014, 59: 1348–1354

4 Tran A T, Sakamoto N, Sato M, et al. Control augmentation system design for quad-tilt-wing unmanned aerial vehicle via robust output regulation method. IEEE Trans Aerosp Electron Syst, 2017, 53: 357–369

5 Zhuan X, Xia X. Speed regulation with measured output feedback in the control of heavy haul trains. Automatica, 2008, 44: 242–247

6 Wang J, Guo Y. Leaderless cooperative control of robotic sensor networks for monitoring dynamic pollutant plumes. IET Control Theor Appl, 2019, 13: 2670–2680

7 Roberto C. Blockchain-based distributed cooperative control algorithm for WSN monitoring. In: Proceedings of International Symposium on Distributed Computing and Artificial Intelligence, 2019. 414–417

8 Francis B A. The linear multivariable regulator problem. SIAM J Control Optim, 1977, 15: 486–505

9 Huang J. The cooperative output regulation problem of discrete-time linear multi-agent systems by the adaptive distributed observer. IEEE Trans Automat Contr, 2017, 62: 1979–1984

10 Cai H, Lewis F L, Hu G, et al. The adaptive distributed observer approach to the cooperative output regulation of linear multi-agent systems. Automatica, 2017, 75: 299–305

11 Liu T, Huang J. Cooperative output regulation for a class of nonlinear multi-agent systems with unknown control directions subject to switching networks. IEEE Trans Automat Contr, 2018, 63: 783–790

12 Dong S, Liu L, Feng G, et al. Cooperative output regulation quadratic control for discrete-time heterogeneous multiagent Markov jump systems. IEEE Trans Cybern, 2021, 52: 9882–9892

13 Zhang Y, Su Y F. Cooperative output regulation for linear uncertain MIMO multi-agent systems by output feedback. Sci China Inf Sci, 2018, 61: 092206

14 Yan Y M, Huang J. Cooperative robust output regulation problem for discrete-time linear time-delay multi-agent systems. Int J Robust Nonlinear Control, 2018, 28: 1035–1048

15 Huang J. Nonlinear Output Regulation: Theory and Applications. Philadelphia: SIAM, 2004

16 Wieland P, Sepulchre R, Allgöwer F. An internal model principle is necessary and sufficient for linear output synchronization. Automatica, 2011, 47: 1068–1074

17 Yu W W, Wang H, Hong H F, et al. Distributed cooperative anti-disturbance control of multi-agent systems: an overview. Sci China Inf Sci, 2017, 60: 110202

18 Su Y, Hong Y, Huang J. A general result on the robust cooperative output regulation for linear uncertain multi-agent systems. IEEE Trans Automat Contr, 2013, 58: 1275–1279

19 Francis B A, Wonham W M. The internal model principle of control theory. Automatica, 1976, 12: 457–465

20 Su Y, Huang J. Cooperative global output regulation of heterogeneous second-order nonlinear uncertain multi-agent systems. Automatica, 2013, 49: 3345–3350

21 Su Y, Huang J. Cooperative semi-global robust output regulation for a class of nonlinear uncertain multi-agent systems. Automatica, 2014, 50: 1053–1065

22 Yan Y, Chen Z. Cooperative output regulation of linear discrete-time time-delay multi-agent systems by adaptive distributed observers. Neurocomputing, 2019, 331: 33–39

23 Saberi A, Stoorvogel A A, Sannuti P, et al. On optimal output regulation for linear systems. Int J Control, 2003, 76: 319–333

24 Lee J W, Khargonekar P P. Optimal output regulation for discrete-time switched and markovian jump linear systems. SIAM J Control Optim, 2008, 47: 40–72

25 Ullah S, Liquat M. Optimal output regulation on sample-data systems. In: Proceedings of International Conference on Control, Electronics, Renewable Energy and Communications, Bandung, 2015

26 Tran A T, Sakamoto N, Kikuchi Y, et al. Pilot induced oscillation suppression controller design via nonlinear optimal output regulation method. Aerospace Sci Tech, 2017, 68: 278–286

27 Yan Y, Huang J. Cooperative output regulation of discrete-time linear time-delay multi-agent systems under switching network. Neurocomputing, 2017, 241: 108–114

28 Song X L, Ding F, Xiao F, et al. Data-driven optimal cooperative adaptive cruise control of heterogeneous vehicle platoons with unknown dynamics. Sci China Inf Sci, 2020, 63: 190204

29 Gao W, Jiang Z P, Lewis F L, et al. Cooperative optimal output regulation of multi-agent systems using adaptive dynamic programming. In: Proceedings of the 2017 American Control Conference (ACC), Seattle, 2017

30 Gao W, Liu Y, Odekunle A, et al. Adaptive dynamic programming and cooperative output regulation of discrete-time multi-agent systems. Int J Control Autom Syst, 2018, 16: 2273–2281

31 Zhang H, Liang H, Wang Z, et al. Optimal output regulation for heterogeneous multiagent systems via adaptive dynamic programming. IEEE Trans Neural Netw Learn Syst, 2017, 28: 18–29

32 Peng Z N, Hu J P, Ghosh B K. Data-driven containment control of discrete-time multi-agent systems via value iteration. Sci China Inf Sci, 2020, 63: 189205

33 Peng Z N, Zhang J F, Hu J P, et al. Optimal containment control of continuous-time multi-agent systems with unknown disturbances using data-driven approach. Sci China Inf Sci, 2020, 63: 209205

34 Jiang Y, Gao W, Wu J, et al. Reinforcement learning and cooperative H_∞ output regulation of linear continuous-time multi-agent systems. Automatica, 2023, 148: 110768

35 Peng Y, Chen Q, Sun W. Reinforcement Q-learning algorithm for H_∞ tracking control of unknown discrete-time linear systems. IEEE Trans Syst Man Cybern Syst, 2020, 50: 4109–4122

36 Liu Y, Zhang H, Yu R, et al. H_∞ tracking control of discrete-time system with delays via data-based adaptive dynamic programming. IEEE Trans Syst Man Cybern Syst, 2020, 50: 4078–4085

37 Fu Y, Fu J, Chai T. Robust adaptive dynamic programming of two-player zero-sum games for continuous-time linear systems. IEEE Trans Neural Netw Learn Syst, 2015, 26: 3314–3319

38 Gao W, Jiang Z P, Lewis F L, et al. Leader-to-formation stability of multiagent systems: an adaptive optimal control approach. IEEE Trans Automat Contr, 2018, 63: 3581–3587

39 Gao W, Mynuddin M, Wunsch D C, et al. Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model. IEEE Trans Neural Netw Learn Syst, 2021, 33: 5229–5240

40 Jiang Y, Fan J, Gao W, et al. Cooperative adaptive optimal output regulation of nonlinear discrete-time multi-agent systems. Automatica, 2020, 121: 109149

41 Gao W, Jiang Y, Davari M. Data-driven cooperative output regulation of multi-agent systems via robust adaptive dynamic programming. IEEE Trans Circuits Syst II, 2019, 66: 447–451

42 Roy S B, Bhasin S, Kar I N. Combined MRAC for unknown MIMO LTI systems with parameter convergence. IEEE Trans Automat Contr, 2018, 63: 283–290

43 Lancaster P, Rodman L. Algebraic Riccati Equations. New York: Oxford University Press Inc., 1995