Cooperative optimal output regulation of multi-agent systems using adaptive dynamic programming

Gao, W.; Jiang, Z.-P.; Lewis, F.; Wang, Y.

TR2017-064 May 2017

Abstract
This paper proposes a novel solution to the adaptive optimal output regulation problem of continuous-time linear multi-agent systems. A key strategy is to resort to reinforcement learning and approximate/adaptive dynamic programming. A data-driven, non-model-based algorithm is given to design a distributed adaptive suboptimal output regulator in the presence of unknown system dynamics. The effectiveness of the proposed computational control algorithm is demonstrated via cooperative adaptive cruise control of connected and autonomous vehicles.

American Control Conference (ACC)
Cooperative optimal output regulation of multi-agent systems using adaptive dynamic programming

Weinan Gao, Zhong-Ping Jiang, Frank L. Lewis, and Yebin Wang

Abstract—This paper proposes a novel solution to the adaptive optimal output regulation problem of continuous-time linear multi-agent systems. A key strategy is to resort to reinforcement learning and approximate/adaptive dynamic programming. A data-driven, non-model-based algorithm is given to design a distributed adaptive suboptimal output regulator in the presence of unknown system dynamics. The effectiveness of the proposed computational control algorithm is demonstrated via cooperative adaptive cruise control of connected and autonomous vehicles.

I. INTRODUCTION

Output regulation problems focus on designing feedback controllers for a plant to realize asymptotic tracking while rejecting external disturbance; see, e.g. [1]–[3]. It is a general mathematical framework that can describe numerous control problems in the real world. The consensus and coordinated control of multi-agent systems have been under extensive investigation in the last decade; see [4], [5] and references therein. At the same time, the cooperative output regulation problem has received considerable attention [6]–[8]. There are usually two groups of agents in the cooperative output regulation problem. The agents in the first group can directly access the leader information (modeled via an exo-system) for feedback control, while the other agents in the second group cannot. The leader-follower consensus problems can be treated as special cases.

Using traditional output regulation solutions, there are two major strategies for addressing cooperative output regulation problems: feedback-feedforward [6] and internal model principle [7], [9]. By means of the internal model principle, one can convert an output regulation problem to a stabilization problem of an augmented system composed of the plant and a dynamic compensator named as internal model. By taking unknown control direction and large parameter uncertainties into account, reference [8] proposes a distributed adaptive control design approach for the cooperative output regulation of a class of multi-agent dynamical systems. However, the issue of adaptive and optimal controller design for the cooperative output regulation of multi-agent systems with unknown dynamics remains open.

A recent trend in adaptive optimal control is to invoke reinforcement learning [10] and approximate/adaptive dynamic programming (ADP) [11] for feedback control of dynamical systems. Among all the different ADP approaches for both continuous-time and discrete-time systems, much attention has been paid to achieve the adaptive optimal stabilization of linear or nonlinear plants [12]–[18]. The generalization to adaptive optimal tracking control is studied by [19]–[21], and an integration of ADP and output regulation has been proposed for the first time in [22]. For non-model-based optimal stabilization of large-scale systems, some interesting results appear in [23], [24] by combining (robust) ADP with game theory or small-gain theory [25]. Nevertheless, an application of ADP to achieve cooperative output regulation of multi-agent systems remains an open problem.

The main purpose of this paper is to develop a novel ADP methodology to realize cooperative optimal output regulation of uncertain continuous-time multi-agent linear systems via online learning. As the first contribution of this paper, we design a distributed and suboptimal controller for cooperative optimal output regulation problem (COORP) whereby each agent can achieve disturbance rejection and asymptotic tracking. As the second contribution, we develop a non-model-based learning method that implements the policy iteration using only the real-time input/state data collected online along the trajectories of the multi-agent system, when the perfect system knowledge is not available. This method can be regarded as a generalization of our recent work on the centralized adaptive optimal output regulation of continuous-time linear systems [22].

The remainder of this paper is organized as follows. In Section II, we briefly review the cooperative output regulation problem and a corresponding model-based solution. Then, the COORP is formulated in Section III and a suboptimal controller is designed by solving several optimization problems. After that, a data-driven ADP based design for COORP is presented in Section IV which can produce an approximate solution of the corresponding algebraic Riccati equations and regulator equations. The convergence and stability of the proposed algorithm are also rigorously analyzed. An application to the cooperative adaptive cruise control (CACC) of connected vehicles is shown in Section V. Finally, some conclusions are drawn in Section VI.

Notations. Throughout this paper, $\| \cdot \|$ represents the Euclidean norm for vectors and the induced norm for matrices. $\otimes$ indicates the Kronecker product operator. $\text{vec}(A) = [a_1^T, a_2^T, \ldots, a_m^T]^T$, where $a_i \in \mathbb{R}^n$ are the columns of $A \in \mathbb{R}^{n \times m}$. 
For a symmetric matrix \( P \in \mathbb{R}^{m \times m} \), \( \text{vecs}(P) = [p_{11}, 2p_{12}, \ldots, 2p_{1m}, p_{22}, \ldots, 2p_{m-1,m}, p_{mm}]^T \in \mathbb{R}^{m(m+1)} \), \( \text{vecs}(v) = [v_1^T v_2, \ldots, v_1 v_m, v_2^T v_3, \ldots, v_{m-1} v_m, v_m^T]^T \in \mathbb{R}^{m(m+1)} \) for an arbitrary column vector \( v \in \mathbb{R}^m \). \( |v|^P \) denotes \( v^T P v \).

II. PROBLEM STATEMENT

Consider the following linear multi-agent system

\[
\dot{v} = Ev, \quad (1)
\]

\[
z_i = A_i x_i + B_i u_i + D_i v, \quad (2)
\]

where \( x_i \in \mathbb{R}^n, u_i \in \mathbb{R}^m \) and \( e_i \in \mathbb{R}^p \) are the state, control input and tracking error of the \( i \)th subsystem, and \( v \in \mathbb{R}^q \) is the state of the exosystem which generates the disturbance \( D_i v \) and the reference signal \( -F_i v \) of each subsystem. Given the exosystem (1) and the plant (2), define a digraph \( G = (V, E) \). \( V = \{0, 1, \ldots, N\} \) is the node set with node 0 denoting the leader modeled via the exosystem (1) and the remaining \( N \) nodes being identified as followers described by (2). \( E \subset V \times V \) refers to the edge set. Denote \( N_i \) the set of all the nodes \( j \) such that \((j, i) \in E \). The adjacency matrix \( A = [a_{ij}] \in \mathbb{R}^{(N+1) \times (N+1)} \) is defined such that \( a_{ij} > 0 \) is a positive weight if \((j, i) \in E \) and otherwise \( a_{ij} = 0 \). Then, the corresponding Laplacian \( \mathcal{L} \) of \( G \) is

\[
\mathcal{L} = \begin{bmatrix}
\sum_{j=1}^{N} a_{0j} & -a_{01} & \cdots & -a_{0N} \\
0 & \ddots & \vdots & \vdots \\
-\Delta a & \cdots & 0 & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
-\Delta a & \cdots & 0 & -a_{NN}
\end{bmatrix}
\]

where \( \Delta = \text{diag}\{a_{01}, a_{20}, \ldots, a_{N0}\} \), and \( \mathcal{H} = [h_{ij}] \in \mathbb{R}^{N \times N} \) is defined by \( h_{ii} = \left( \sum_{j=0}^{N} a_{ij} \right) - a_{ii} \) and \( h_{ij} = -a_{ij} \) for all \( i \neq j \).

Some standard assumptions for the solvability of traditional cooperative output regulation problem are made on the system (1)-(2).

**Assumption 1:** \((A_i, B_i)\) is stabilizable for \( i = 1, \ldots, N \).

**Assumption 2:** \( \text{rank} \begin{bmatrix} A_i - \lambda I & B_i \\ C_i & 0 \end{bmatrix} = n_i + p_i, \forall \lambda \in \sigma(E_i), i = 1, \ldots, N \).

**Assumption 3:** The digraph \( G \) contains a directed spanning tree with the node 0 as its root.

**Remark 1:** It has been shown in the Lemma 1 of [6] that all the eigenvalues of \( \mathcal{H} \) have positive real parts.

The cooperative output regulation is to design a distributed control policy such that the overall system is asymptotically stable (in the absence of \( v \)) and \( \lim_{t \to \infty} e_i(t) = 0 \), for \( i = 1, 2, \ldots, N \).

**Remark 2:** The cooperative output regulation problem studied in this paper includes leader-follower consensus problem as a special case if we let the \( v = x_0 \), and matrices \( C_i \) and \( F_i \) as identity matrices for all \( i = 1, 2, \ldots, N \). Moreover, if all the \( N \) subsystems in (2) share the same system dynamics, the cooperative output regulation problem will be reduced as a synchronized output regulation problem [26].

A technical solution to the cooperative output regulation problem is recalled as follows.

**Theorem 1 ( [6] ):** Under Assumptions 1-3, choose a large \( \mu > 0 \) such that for any \( i = 1, 2, \ldots, q \) and any \( j = 1, 2, \ldots, N \), \( \text{Re}(\lambda_i(E) - \mu \lambda_j(H)) < 0 \). Let \( K_i, i = 1, 2, \ldots, N \), be such that \( A_i - B_i K_i \) is Hurwitz and matrices \( L_i = U_i + K_i X_i \) where pairs \((X_i, U_i)\) solve the following regulator equations

\[
X_i E = A_i X_i + B_i U_i + D_i, \quad (4)
\]

\[
0 = C_i X_i + F_i, \quad (5)
\]

Then, the cooperative output regulation problem is solved by a distributed control policy

\[
u_i = -K_i x_i + L_i \zeta_i, i = 1, 2, \ldots, N, \quad (6)
\]

\[
\zeta_i = E \zeta_i + \mu \left[ \sum_{j \in N_i} a_{ij} (\zeta_j - \zeta_i) + a_{i0} (v - \zeta_i) \right]. \quad (7)
\]

III. FORMULATION AND MODEL-BASED SOLUTION OF COORP

The COORP studied in this paper considers both steady and transient performance of each subsystem. The formulation follows the traditional linear optimal output regulation problem [22], [27] that the optimal distributed control policy need not only solve the cooperative output regulation problem, but also address the following two problems.

**Problem 1:**

\[
\min_{(X_i, U_i)} \text{Tr}(X_i^T \tilde{Q}_i X_i + U_i^T \tilde{R}_i U_i) \quad (8)
\]

subject to (4)-(5),

where \( \tilde{Q}_i = \tilde{Q}_i^T > 0, \tilde{R}_i = \tilde{R}_i^T > 0 \).

**Remark 2:** The optimal feedback controller \( \hat{u}_i^* = -K_i \hat{x}_i \) is found by solving the following constrained minimization problem.

**Problem 2:**

\[
\min_{\hat{u}_i} \int_0^\infty (|e_i|^2 + |\hat{u}_i|^2 R_i) dt \quad (9)
\]

subject to (9),

where \( Q_i = Q_i^T \geq 0, R_i = R_i^T > 0 \), with \((A_i, \sqrt{Q_i} C_i)\) observable.

When the system parameters are known and \( v \) can be immediately used by all the other agents for feedback, the cooperative output regulation for system (1)-(2) is equivalent to achieve output regulation for each subsystem. Moreover, the COORP is solvable through designing the following decentralized controller

\[
\hat{u}_i^* = -K_i^* x + L_i^* v \quad (10)
\]

where, for \( i = 1, 2, \ldots, N \):
1) $K^*_i$ is computed by solving Problem 2, to arrive at
$$K^*_i = -R_i^{-1}B_i^TP_i^*$$
with $P_i^* = (P_i^*)^T > 0$ the solution to the following algebraic Riccati equation
$$A_i^TP_i^* + P_i^*A_i + C_i^TQ_iC_i - P_i^*BR_iR_i^{-1}B_i^TP_i^* = 0,$$
and in this case, the cost of Problem 2 for each $i$th subsystem achieves its minimum $J_i^* := \{x_i(0)\}_{P_i^*}$. 
2) $L_i^* = U_i^* + K_i^*X_i^*$, where $(X_i^*, U_i^*)$ is the minimizer of Problem 1.

Remark 3: Instead of solving (12) which is nonlinear in $P_i^*$, a policy iteration algorithm, Algorithm 1, is recalled which approximate $P_i^*$ by iteratively solving linear Lyapunov equations.

Algorithm 1 Policy Iteration Algorithm [28]

1. Find a $K_{i0}$ such that $A_i - B_i K_{i0}$ is a Hurwitz matrix. $k \leftarrow 0$. Select a sufficiently small constant $\epsilon > 0$. 
2. repeat 
3. Solve $P_{ik}$ and $K_{i,k+1}$ from
$$0 = (A_i - B_i K_{ik})^TP_{ik} + P_{ik}(A_i - B_i K_{ik})$$
$$+ C_i^TQ_iC_i + K_{ik}^TR_iK_{ik}$$
$$K_{i,k+1} = R_i^{-1}B_i^TP_{ik}$$

4. $k \leftarrow k + 1.$
5. until $|P_{ik} - P_{i,k-1}| < \epsilon$

However, the optimal controller (11) is not implementable since $\nu$ might not be accessible by some agents instantly. To this end, we develop a suboptimal distributed controller that subjects to the required communication topology.

$$u_i = -K^*_i x_i + L_i^* \zeta_i, i = 1, 2, \cdots, N.$$  
(15)

The suboptimality of system (1)-(2) in closed-loop with (15) is characterized in the following Theorem.

Theorem 2: Letting $J_{i}^\circ$ be the cost of Problem 2 for the $i$th subsystem (2) in closed-loop with controller (15), the summation of the cost error of each subsystem $J_{i}^\circ - J_i^*$ is
$$\sum_{i=1}^{N} (J_{i}^\circ - J_i^*) = \int_{0}^{\infty} [L^*(\zeta - \nu)] R d\tau$$
where $\zeta = [\zeta_1^T, \zeta_2^T, \cdots, \zeta_N^T]^T, L^* = \text{blockdiag}\{L_1^*, L_2^*, \cdots, L_N^*\}, R = \text{blockdiag}\{R_1, R_2, \cdots, R_N\}, \nu = 1_N \otimes v$.

Proof: By the fact that $(\Delta \otimes I_q)(1_N \otimes v) = (H \otimes I_q)(1_N \otimes v)$, the overall system incorporated with the dynamic compensator (7) is
$$\dot{x} = Ax + Bu + Dv,$$
$$\dot{\zeta} = [I_N \otimes E] - \mu(H \otimes I_q)\zeta + \mu(H \otimes I_q)\nu,$$
$$\dot{\nu} = (I_N \otimes E)\nu,$$
$$\dot{e} = C\bar{x} + Fv$$
(17)

where $x = [x_1^T, x_2^T, \cdots, x_N^T]^T$, for $M = A, B, C, D, F, M = \text{blockdiag}\{M_1, M_2, \cdots, M_N\}$.

Then, the closed-loop error system is
$$\dot{\zeta} = (A - BK^*)\zeta + BL^*(\zeta - \nu),$$
(18)
$$\dot{e} = C\bar{x}$$
(19)

where $\bar{z} = [\bar{z}_1^T, \bar{z}_2^T, \cdots, \bar{z}_N^T]^T$ and $K^* = \text{blockdiag}\{K_1^*, K_2^*, \cdots, K_N^*\}$.

Denote $\bar{P}^* = \text{blockdiag}\{P_1^*, P_2^*, \cdots, P_N^*\}$. Differentiating the Lyapunov function $V = \bar{z}^TP^*\bar{z}$ along the solutions of the system (18), we have
$$\frac{d}{dt}(\bar{z}^TP^*\bar{z}) = \bar{z}^T[(A - BK^*)^TP^* + P^*(A - BK)]\bar{z}$$
$$+ 2\bar{z}^TP^*BL^*\zeta $$
$$= -|\bar{e}|_Q - |K^*\zeta|_R + 2\bar{z}^T(K^*)^TRL^*(\zeta - \nu)$$
$$= -|\bar{e}|_Q + |u|_R + |L^*(\zeta - \nu)|_R,$$
where $u = [u_1^T, u_2^T, \cdots, u_N^T]^T$.

Integrating both sides of the last equation, we have
$$\sum_{i=1}^{N} J_i^\circ - \sum_{i=1}^{N} J_i^* = \int_{0}^{\infty} [L^*(\zeta - \nu)] R d\tau,$$
which directly implies (16).

Moreover, from (17), one can get
$$\frac{d}{dt}(\zeta - \nu) = [(I_N \otimes E) - \mu(H \otimes I_q)](\zeta - \nu).$$
(20)

A selection of $\mu$ in Theorem 1 ensures the exponential convergence of signal $\zeta - \nu$, which, in turn, guarantees the boundedness of the cost error in (16). The proof is thus completed.

IV. DATA-DRIVEN COOPERATIVE OPTIMAL OUTPUT REGULATION CONTROLLER DESIGN

In this section, we develop a data-driven suboptimal controller design approach for COORP via ADP. Interestingly, the developed approach is able to approximate the control gains $K^*$ and $L^*$ without relying on the knowledge of system dynamics $A, B$ and $D$.

Consider the $i$th subsystem. Defining $\bar{z}_{ij} = x_i - X_{ij}v$ for $j = 0, 1, 2, \cdots, h_i + 1$, where $X_{i0} = 0_{n_i \times q}, X_i \in \mathbb{R}^{n_i \times q}$ such that $C_iX_{i1} + F_i = 0$. $X_i \in \mathbb{R}^{n_i \times q}$ for $j = 2, 3, \cdots, h_i + 1$ are selected such that all the vectors vec$(X_{ij})$ form a basis for ker$(I_q \otimes C_i)$, where $h_i = (n_i - p_i)q$ is the dimension of the null space of $I_q \otimes C_i$. Then,
$$\hat{z}_{ij} = A_i x_i + B_i u_i + (D_i - X_{ij}E)v$$
$$= A_i x_i + B_i K_{ik} \bar{z}_{ij} + u_i + (D_i - S_i(X_{ij}))v$$
(21)

where $A_i = A - B K_{ik}$ and $S_i : \mathbb{R}^{n_i \times q} \rightarrow \mathbb{R}^{n_i \times q}$ is a Sylvester map $S_i(X) = X E - A_i X, X \in \mathbb{R}^{n_i \times q}$.

The motivation to introduce $\hat{z}_{ij}$ is that we hope to solve not only $P_{ik}$ and $K_{ik}$, but also the Sylvester map of trail matrices $X_{ij}$ which is a crucial term for solving regulator equations without accurate knowledge of $A, B$ and $D$. 
Then, along the solutions (21) by (13)-(14), we have
\[
|\bar{x}_{ij}(t+\delta t)|_{P_{ik}} - |\bar{x}_{ij}(t)|_{P_{ik}} \\
= \int_{t}^{t+\delta t} (-|\bar{x}_{ij}|)(A_{ik} P_{ik} + P_{ik} A_{ik}) + 2(u_i + K_{ik} \bar{x}_{ij})^T B_i^T P_{ik} \bar{x}_{ij} \\
+ 2\nu^T(D_i - S_i(X_{ij}))^T P_{ik} \bar{x}_{ij}] \, \text{d}t \\
= \int_{t}^{t+\delta t} [-|\bar{x}_{ij}|(Q_i + K_{ik}^T R_i K_{ik}) + 2(u_i + K_{ik} \bar{x}_{ij})^T R_i K_{i,k+1} \bar{x}_{ij} \\
+ 2\nu^T(D_i - S_i(X_{ij}))^T P_{ik} \bar{x}_{ij}] \, \text{d}t. \quad (22)
\]

By Kronecker product representation, we obtain
\[
|\bar{x}_{ij}|(Q_i + K_{ik}^T R_i K_{ik}) = (\bar{x}_{ij}^T \otimes \bar{x}_{ij}^T) \text{vec}(Q_i + K_{ik}^T R_i K_{ik}), \\
\nu^T(D_i - S_i(X_{ij}))^T P_{ik} \bar{x}_{ij} = \\
(\bar{x}_{ij}^T \otimes \nu^T) \text{vec}((D_i - S_i(X_{ij}))^T P_{ik}), \\
(u_i + K_{ik} \bar{x}_{ij})^T R_i K_{i,k+1} \bar{x}_{ij} = [(\bar{x}_{ij}^T \otimes \bar{x}_{ij}^T)(I_n \otimes K_{ik}^T R_i) \\
+ (\bar{x}_{ij}^T \otimes u_i^T)(I_n \otimes R_i)] \text{vec}(K_{i,k+1}). \quad (23)
\]

Moreover, for any two vectors $a, b$ and a sufficiently large number $s > 0$, define
\[
\delta_a = [\text{vec}(a(t_1)) - \text{vec}(a(t_0)), \ldots, \\
\text{vec}(a(s))) - \text{vec}(a(t_{s-1})),]^T, \\
\Gamma_{a,b} = \left[ \int_{t_0}^{t_1} a \otimes \text{b} \, \text{d}t, \int_{t_1}^{t_2} a \otimes \text{b} \, \text{d}t, \ldots, \int_{t_{s-1}}^{t_s} a \otimes \text{b} \, \text{d}t \right]^T. \quad (24)
\]

Equations (22)-(24) imply the following linear equation
\[
\Psi_{ijk} \begin{bmatrix} \\
\text{vec}(P_{ik}) \\
\text{vec}(K_{i,k+1}) \\
\text{vec}((D_i - S_i(X_{ij}))^T P_{ik}) \\
\end{bmatrix} = \Phi_{ijk}. \quad (25)
\]

where
\[
\Psi_{ijk} = [\delta_{x_{ij}}, \text{vec}(Q_i + K_{ik}^T R_i)], \\
\Phi_{ijk} = -2 \Gamma_{x_{ij},u_i} (I_{n_i} \otimes K_{ik}^T R_i) - 2 \Gamma_{x_{ij},u_i} (I_{n_i} \otimes R_i), \\
-2 \Gamma_{x_{ij},x_{ij}}. \\
\]

The uniqueness of solution to (25) is guaranteed under some rank condition as shown below. This rank condition is like the condition of persistent excitation (PE) in adaptive control [29]. Similar as other ADP algorithms, we add an exploration noise to the input to satisfy the rank condition. However, a major difference from traditional adaptive control is that one can directly approximate the solution to the Riccati equations and regulator equations.

Lemma 4.1: For all $j \in \mathbb{Z}_+$, if there exists a $s^* \in \mathbb{Z}_+$ such that for all $s > s^*$, for any sequence $t_0 < t_1 < \cdots < t_s$,
\[
\text{rank}([\Gamma_{x_{ij},x_{ij}}, \Gamma_{x_{ij},u_i}, \Gamma_{x_{ij},v}]) = \frac{n_i(n_i + 1)}{2} + (m_i + q)n_i, \\
\]
then the matrix $\Psi_{ijk}$ for the $i$th subsystem has full column rank for all $k \in \mathbb{Z}_+$.

A general solution to (4)-(5) can be described by a sequence of $\alpha_{ij} \in \mathbb{R}$ as
\[
X_i = X_{i1} + \sum_{j=2}^{h_{i+1}} \alpha_{ij} X_{ij}, \\
S_i(X_i) = S_i(X_{i1}) + \sum_{j=2}^{h_{i+1}} \alpha_{ij} S_i(X_{ij}) = B_i U_i + D_i
\]
which is rewritten as
\[
A_i X_i = b_i, \quad (27)
\]
where
\[
A_i = [A_{i1} A_{i2}], \\
A_{i1} = \begin{bmatrix} \text{vec}(S_i(X_{i2})) & \cdots & \text{vec}(S_i(X_{i,h_{i+1}})) \end{bmatrix}, \\
A_{i2} = \begin{bmatrix} 0 & \cdots & \text{vec}(S_i(X_{i,h_{i+1}})) \end{bmatrix}, \\
\chi_i = [\alpha_i, \cdots, \alpha_i, h_{i+1}, \text{vec}(X_{i2})^T, \text{vec}(U_i)^T]^T, \\
b_i = \begin{bmatrix} -S_i(X_{i1}) + D_i \end{bmatrix}.
\]

One can obtain that matrices $P_{ik}, K_{i,k+1}$ and $S_i(X_{i1}), \cdots, S_i(X_{i,h_{i+1}})$ in (27) are obtainable by (25). Also, $D_i$ can be computed from (25) by $\text{vec}(D_i) = \text{vec}(D_i - S_i(X_{i0}))$.

Now, we are ready to present the non-mode-based ADP Algorithm 2 to solve COORP.

Remark 4: Albeit some cannot get instant information from the leader, by Assumption 3, all the nodes in $G$ are reachable from node 0. There always exists a $\Delta t > 0$ such that $\nu(t)$ in the period $[t_i, t_{i+1}]$ is received by all the other subsystems at $t = t_i + \Delta t$.

Theorem 3: If (26) is satisfied, for $i = 1, 2, \cdots, N$, sequences $\{\bar{P}_{ik}\}_{k=0}^\infty$ and $\{\bar{K}_{ik}\}_{k=0}^\infty$ computed by Algorithm 2 converge to $\bar{P}_{ik}$ and $\bar{K}_{ik}$. Furthermore, the multi-agent system (1)-(2) in closed-loop with the learned controller (28) achieves cooperative output regulation.

Proof: Letting $\bar{P}_{ik} = \bar{P}_{ik}^\infty > 0$ be the solution to (13). $K_{i,k+1}$ is uniquely determined by (14) and $T_{ik} = (D_i - S_i(X_{ij}))^T P_{ik}$. On the other hand, letting $\bar{P}, \bar{K}$ and $\bar{T}$ solve (25), condition (26) ensures that $P_{ik} = \bar{P}$, $K_{i,k+1} = \bar{K}$ and $T_{ik} = \bar{T}$ are uniquely determined. By (28), we have $\lim_{k \to \infty} K_{ik} = K_{i,k+1} = \bar{K}$, $\lim_{k \to \infty} P_{ik} = P_{ik}^\infty$. The convergence of sequences $\{\bar{B}_{ik}\}_{k=0}^\infty$ and $\{\bar{K}_{ik}\}_{k=0}^\infty$ obtained by non-mode-based Algorithm 2 is thus ensured. Moreover, we have $\{A_i - B_i K_{i,k}\}$ is a Hurwitz matrix for small threshold $\epsilon_i > 0$.
Based on Theorem 1, we observe that the tracking error of each subsystem is guaranteed asymptotically converging to 0 if the learned controller (28) is applied.

Remark 5: Note that if $v$ is not measurable by any nodes, one can refer to [22] to generate a signal $w \in \mathbb{R}^{q_m}$ by the knowledge of the minimal polynomial of matrix $E$ such that there is an unknown matrix $G \in \mathbb{R}^{q \times q_m}$ such that $v(t) = G w(t)$. Interestingly, we convert the problem of unmeasurability of $v(t)$ into that of an unknown matrix $G$. 

Algorithm 2 ADP Algorithm for COORP

1: \( i \leftarrow 1 \)
2: repeat
3: \( \text{Compute matrices } X_{i0}, X_{i1}, \ldots, X_{ih+1} \)
4: \( \text{Apply an initial policy } u_i^0 = -K_{i0}x + \xi_i \text{ on } [t_0, t_2] \)
5: \( \text{with exploration noise } \xi_i \text{ and } A_i - B_iK_{i0} \text{ a Hurwitz matrix} \)
6: \( j \leftarrow 0 \)
7: repeat
8: \( j \leftarrow j + 1 \)
9: until \( j = h_i + 2 \)
10: \( j \leftarrow 0, k \leftarrow 0 \)
11: repeat
12: \( k \leftarrow k + 1 \)
13: until \( |P_{ik} - P_{i,k-1}| < \epsilon_i \) with \( \epsilon_i \) a small positive constant.
14: \( k^* \leftarrow k, j \leftarrow 1 \)
15: repeat
16: \( \text{Solve } S_i(X_{ij}) \text{ from } (25) \)
17: \( j \leftarrow j + 1 \)
18: until \( j = h_i + 2 \)
19: \( \text{Solve } (X_i^*, U_i^*) \text{ from Problem 1} \)
20: \( L_{ik} \leftarrow U_i^* + K_{ik}X_i^* \)
21: \( \text{Obtain the following suboptimal controller} \)
22: \( u_i^* = -K_{ik}x_i + L_{ik}x_i \) \( \xi_i \)
23: \( i \leftarrow i + 1 \)
24: until \( i = N + 1 \)

V. APPLICATION TO CONNECTED AND AUTONOMOUS VEHICLES

In this section, we apply Algorithm 2 to the longitudinal cooperative adaptive cruise control (CACC) of a platoon of connected vehicles. CACC is an intelligent autonomous driving strategy based on wireless vehicle-to-vehicle (V2V) communication that is believed to be realizable in the near future. Different from the existing CACC approaches [30], [31], this paper designs a suboptimal distributed controller while the mathematical models of the vehicles are unknown.

For \( i = 1, 2, 3, 4 \), we utilized the following model of the \( i \)-th vehicle for the purpose of simulation [32],

\( \dot{s}_i = v_i \),
\( \dot{v}_i = a_i \),
\( \dot{a}_i = \tau_i^{-1}a_i + \tau_i^{-1}u_i + d_i \)

where \( s_i, v_i, a_i, \tau_i \) are the position, velocity, acceleration and time constant of the engine of vehicle \( \#i \). The constant \( d_i \) is the ratio of the mechanical drag to the product of \( \tau_i \) and the mass of vehicle \( \#i \). The values of \( \tau_i \) and \( d_i \) are illustrated in Tab. I. The topology of this platoon is depicted in Fig. 1, where all the followers can receive motional data from its preceding vehicle and the leader, while the exosystem is only accessed by the leader. In this example, both desired trajectory \( s_i^* \) and disturbance \( d_i \) are generated by the following exosystem:

\( \begin{align*}
    v_1 &= v_2, \\
    v_2 &= 0, \\
    d_1 &= d_2, \\
    s_i^* &= -15v_1 - 20(5-i)v_2 \\
\end{align*} \)

with the initial value \( v = [0 \ 1]^T \).

For \( t < 6s \), we apply initial admissible control policy added by an exploration noise, which is a summation of sinusoidal signals with different frequencies. Then, we follow Algorithm 2 to iteratively learn matrices \( K_{ik} \) and \( P_{ik} \). The comparisons of \( P_{ik} \) and their optimal values \( P_i^* \) are shown in Fig. 2. After that, the optimal solution to the regulator equations is obtained which can be employed to get the feedforward gain. To be more specific, we give the learned approximated solution \((X_1, U_1)\) to the regulator equation of vehicle \#1 and its optimal one as follows.

\( \begin{align*}
    X_1 &= \begin{bmatrix} 15.0000 & 80.0000 \\ 0.0016 & 15.0084 \end{bmatrix}, X_1^* &= \begin{bmatrix} 15 & 80 \end{bmatrix}, \\
    U_1 &= \begin{bmatrix} -0.0019 & -0.2982 \end{bmatrix}, U_1^* &= \begin{bmatrix} 0 & -0.3 \end{bmatrix}. \\
\end{align*} \)

![Fig. 1. The structure of the vehicular platoon](image)

![Fig. 2. The comparison of \( P_{ik} \) at \( k \)th iteration of \( i \)th subsystem and their optimal values](image)
We utilize the learned distributed controller (28) to regulate the motion of connected vehicles. From Fig. 3, one can observe that all the vehicles in the platoon can effectively achieve asymptotic tracking while rejecting unknown disturbance.

VI. CONCLUSIONS

This paper has studied the cooperative optimal output regulation problem (COORP) of multi-agent linear systems from a perspective of adaptive dynamic programming (ADP). In the presence of unknown system dynamics, a novel data-driven, non-model-based approach to COORP is proposed by reinforcement learning and ADP. Both theoretical analysis and simulations indicate that the closed-loop multi-agent systems can achieve asymptotic tracking and disturbance rejection with desired optimality properties.

REFERENCES


