MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

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 continuoustime 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.

This work has been partly supported by the U.S. National Science Foundation grant ECCS-1501044 and Mitsubishi Electric Research Laboratories.

W. Gao and Z.P. Jiang are with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, 11201 USA (e-mail:{weinan.gao, zjiang}@nyu.edu).

F. Lewis is with UTA Research Institute, The University of Texas at Arlington, Texas 76118 USA and Qian Ren Consulting Professor, Northeastern University, Shenyang 110036, China. (e-mail:lewis@uta.edu)

Y. Wang is with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 USA (e-mail: yebinwang@ieee.org).

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. $\mathrm{vec}(A) = [a_1^T, a_2^T, \cdots, a_m^T]^T$, where $a_i \in \mathbb{R}^n$ are the columns of $A \in \mathbb{R}^{n\times m}$.

For a symmetric matrix $P \in \mathbb{R}^{m\times m}$, $\text{vecs}(P) = [p_{11}, 2p_{12}, \cdots, 2p_{1m}, p_{22}, 2p_{23}, \cdots, 2p_{m-1,m}, p_{mm}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)}$. $\text{vecv}(v) = [v_1^2, v_1v_2, \cdots, v_1v_m, v_2^2, v_2v_3, \cdots, v_{m-1}v_m, v_m^2]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)}$ for an arbitrary column vector $v \in \mathbb{R}^m$. $|v|_P$ denotes v^TPv.

II. Problem statement

Consider the following linear multi-agent system

$$\dot{v} = Ev, \tag{1}$$

$$\begin{aligned} \dot{x}_i &= A_ix_i + B_iu_i + D_iv, \\ e_i &= C_ix_i + F_iv, \quad i = 1, \cdots, N, \end{aligned} \tag{2}$$

where $x_i \in \mathbb{R}^{n_i}, u_i \in \mathbb{R}^{m_i}$ and $e_i \in \mathbb{R}^{p_i}$ are the state, control input and tracking error of the ith subsystem, and $v \in \mathbb{R}^q$ is the state of the exosystem (1) which generates the disturbance D_iv and the reference signal $-F_iv$ of each subsystem. Given the exosystem (1) and the plant (2), define a digraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. $\mathcal{V} = \{0, 1, \cdots, N\}$ 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). $\mathcal{E} \subset \mathcal{V}\times\mathcal{V}$ refers to the edge set. Denote $\mathcal{N}_i$ the set of all the nodes j such that $(j, i) \in \mathcal{E}$. The adjacency matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{(N+1)\times(N+1)}$ is defined such that $a_{ij} > 0$ is a positive weight if $(j, i) \in \mathcal{E}$ and otherwise $a_{ij} = 0$. Then, the corresponding Laplacian $\mathcal{L}$ of $\mathcal{G}$ is

$$\mathcal{L} = \begin{bmatrix} \sum_{j=1}^N a_{0j} & -[a_{01} \cdots, a_{0N}] \\ -\Delta\mathbf{1}_N & \mathcal{H} \end{bmatrix} \tag{3}$$

where $\Delta = \text{diag}\{a_{10}, a_{20}, \cdots, a_{N0}\}$, 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 $i = 1, \cdots, N$.

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

Assumption 3: The digraph $\mathcal{G}$ 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 $i = 1, 2, \cdots, N$.

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 $i = 1, 2, \cdots, N$. 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 $i = 1, 2, \cdots, q$ and any $j = 1, 2, \ldots, N$, $\text{Re}(\lambda_i(E) - \mu\lambda_j(\mathcal{H})) < 0$. Let $K_i, i = 1, 2, \cdots, N$, be such that $A_i - B_iK_i$ is Hurwitz and matrices $L_i = U_i + K_iX_i$ where pairs (X_i, U_i) solve the following regulator equations

$$X_iE = A_iX_i + B_iU_i + D_i, \tag{4}$$

$$0 = C_iX_i + F_i. \tag{5}$$

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

$$u_i = -K_ix_i + L_i\zeta_i, i = 1, 2, \cdots, N, \tag{6}$$

$$\dot{\zeta}_i = E\zeta_i + \mu\left[\sum_{j\in\mathcal{N}_i} a_{ij}(\zeta_j - \zeta_i) + a_{i0}(v - \zeta_i)\right]. \tag{7}$$

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:

$$\begin{aligned} &\min_{(X_i,U_i)} \text{Tr}(X_i^T\bar{Q}_iX_i + U_i^T\bar{R}_iU_i) \\ &\text{subject} \quad \text{to} \quad (4) - (5), \end{aligned} \tag{8}$$

where $\bar{Q}_i = \bar{Q}_i^T > 0, \bar{R}_i = \bar{R}_i^T > 0$.

Note that Assumption 2 ensures the solvability of regulator equations (4)-(5) with any matrices D_i and F_i, for $i = 1, 2, \ldots, N$ [1]. It has been shown in [27] that the solution (X_i^*, U_i^*) to Problem 1 is unique. Let $\bar{x}_i := x_i - X_i^*v, \bar{u}_i := u_i - U_i^*v$. The error system is obtained as follows

$$\dot{\bar{x}}_i = A_i\bar{x}_i + B_i\bar{u}_i, \tag{9}$$

$$e_i = C_i\bar{x}_i. \tag{10}$$

The optimal feedback controller $\bar{u}_i^* = -K_i^*\bar{x}_i$ is found by solving the following constrained minimization problem.

Problem 2:

$$\begin{aligned} &\min_{\bar{u}_i} \int_0^\infty (|e_i|_{Q_i} + |\bar{u}_i|_{R_i})dt \\ &\text{subject} \quad \text{to} \quad (9), \end{aligned}$$

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

$$u_i^* = -K_i^*x + L_i^*v \tag{11}$$

where, for $i = 1, 2, \cdots, N$:

1) K_i^* is computed by solving Problem 2, to arrive at

$$K_i^* = -R_i^{-1}B_i^T P_i^*$$

with $P_i^* = (P_i^*)^T > 0$ the solution to the following algebraic Riccati equation

$$A_i^T P_i^* + P_i^* A_i + C_i^T Q_i C_i - P_i^* B_i R_i^{-1} B_i^T P_i^* = 0, \tag{12}$$

and in this case, the cost of Problem 2 for each ith 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

$$\begin{aligned} 0 =&(A_i - B_i K_{ik})^T P_{ik} + P_{ik}(A_i - B_i K_{ik}) \\ &+ C_i^T Q_i C_i + K_{ik}^T R_i K_{ik} \end{aligned} \tag{13}$$

$$K_{i,k+1} = R_i^{-1} B_i^T P_{ik} \tag{14}$$

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

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

$$u_i = -K_i^* x_i + L_i^* \zeta_i, i = 1, 2, \cdots, N. \tag{15}$$

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

Theorem 2: Letting $J_i^\odot$ be the cost of Problem 2 for the ith subsystem (2) in closed-loop with controller (15), the summation of the cost error of each subsystem $J_i^\odot - J_i^*$ is

$$\sum_{i=1}^{N}(J_i^\odot - J_i^*) = \int_0^\infty |L^*(\zeta - \nu)|_R d\tau \tag{16}$$

where $\zeta = [\zeta_1^T, \zeta_2^T, \cdots, \zeta_N^T]^T$, $L^* = \text{blockdiag}\{L_1^*, L_2^*, \cdots, L_N^*\}$, $R = \text{blockdiag}\{R_1, R_2, \cdots, R_N\}$, $\nu = \mathbf{1}_N \otimes v$.

Proof: By the fact that $(\Delta \otimes I_q)(\mathbf{1}_N \otimes v) = (\mathcal{H} \otimes I_q)(\mathbf{1}_N \otimes v)$, the overall system incorporated with the dynamic compensator (7) is

$$\begin{aligned} \dot{x} =&Ax + Bu + D\nu, \\ \dot{\zeta} =&[(I_N \otimes E) - \mu(\mathcal{H} \otimes I_q)\zeta] + \mu(\mathcal{H} \otimes I_q)\nu, \\ \dot{\nu} =&(I_N \otimes E)\nu, \\ e =&Cx + Fv \end{aligned} \tag{17}$$

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

$$\dot{\bar{x}} =(A - BK^*)\bar{x} + BL^*(\zeta - \nu), \tag{18}$$

$$e =C\bar{x} \tag{19}$$

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

Denote $P^* = \text{blockdiag}\{P_1^*, P_2^*, \cdots, P_N^*\}$. Differentiating the Lyapunov function $V = \bar{x}^T P^* \bar{x}$ along the solutions of the system (18), we have

$$\begin{aligned} &\frac{d}{dt}(\bar{x}^T P^* \bar{x}) \\ =&\bar{x}^T[(A - BK)^T P^* + P^*(A - BK)]\bar{x} \\ &+ 2\bar{x}^T P^* B L^*(\zeta - \nu) \\ =&-|e|_Q - |K^*\bar{x}|_R + 2\bar{x}^T (K^*)^T R L^*(\zeta - v) \\ =&-(|e|_Q + |\bar{u}|_R) + |L^*(\zeta - \nu)|_R. \end{aligned}$$

where $\bar{u} = [\bar{u}_1^T, \bar{u}_2^T, \cdots, \bar{u}_N^T]^T$.

Integrating both sides of the last equation, we have

$$\sum_{i=1}^{N} J_i^\odot - \sum_{i=1}^{N} J_i^* = \int_0^\infty |L^*(\zeta - \nu)|_R d\tau,$$

which directly implies (16).

Moreover, from (17), one can get

$$\frac{d}{dt}(\zeta - \nu) = [(I_N \otimes E) - \mu(\mathcal{H} \otimes I_q)](\zeta - \nu). \tag{20}$$

A selection of μ 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 ith subsystem. Defining $\bar{x}_{ij} = x_i - X_{ij} v$ for $j = 0, 1, 2, \cdots, h_i + 1$, where $X_{i0} = 0_{n_i \times q}$, $X_{i1} \in \mathbb{R}^{n_i \times q}$ such that $C_i X_{i1} + F_i = 0$. $X_{ij} \in \mathbb{R}^{n_i \times q}$ for $j = 2, 3, \cdots, h_i + 1$ are selected such that all the vectors $\text{vec}(X_{ij})$ 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,

$$\begin{aligned} \dot{\bar{x}}_{ij} &= A_i x_i + B_i u_i + (D_i - X_{ij} E)v \\ &= A_{ik}\bar{x}_{ij} + B_i(K_{ik}\bar{x}_{ij} + u_i) + (D_i - \mathcal{S}_i(X_{ij}))v, \end{aligned} \tag{21}$$

where $A_{ik} = A_i - B_i K_{ik}$ and $\mathcal{S}_i : \mathbb{R}^{n_i \times q} \to \mathbb{R}^{n_i \times q}$ is a Sylvester map $\mathcal{S}_i(X) = XE - A_i X, X \in \mathbb{R}^{n_i \times q}$.

The motivation to introduce $\bar{x}_{ij}$ 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

$$\begin{aligned}
&|\bar{x}_{ij}(t+\delta t)|_{P_{ik}} - |\bar{x}_{ij}(t)|_{P_{ik}} \\
=&\int_t^{t+\delta t} [|\bar{x}_{ij}|_{(A_{ik}^T P_{ik}+P_{ik}A_{ik})} + 2(u_i + K_{ik}\bar{x}_{ij})^T B_i^T P_{ik}\bar{x}_{ij} \\
&+ 2v^T(D_i - \mathcal{S}_i(X_{ij}))^T P_{ik}\bar{x}_{ij}]d\tau \\
=&\int_t^{t+\delta t} [-|\bar{x}_{ij}|_{(Q_i+K_{ik}^T R_i K_{ik})} + 2(u_i + K_{ik}\bar{x}_{ij})^T R_i K_{i,k+1}\bar{x}_{ij} \\
&+ 2v^T(D_i - \mathcal{S}_i(X_{ij}))^T P_{ik}\bar{x}_{ij}]d\tau. \qquad (22)
\end{aligned}$$

By Kronecker product representation, we obtain

$$\begin{aligned}
&|\bar{x}_{ij}|_{(Q_i+K_{ik}^T R_i K_{ik})} = (\bar{x}_{ij}^T \otimes \bar{x}_{ij}^T)\mathrm{vec}(Q_i + K_{ik}^T R K_{ik}), \\
&v^T(D_i - \mathcal{S}(X_{ij}))^T P_{ik}\bar{x}_{ij} = \\
&(\bar{x}_{ij}^T \otimes v^T)\mathrm{vec}((D_i - \mathcal{S}_i(X_{ij}))^T P_{ik}), \\
&(u_i + K_{ik}\bar{x}_{ij})^T R_i K_{i,k+1}\bar{x}_{ij} = [(\bar{x}_{ij}^T \otimes \bar{x}_{ij}^T)(I_{n_i} \otimes K_{ik}^T R_i) \\
&+ (\bar{x}_{ij}^T \otimes u_i^T)(I_{n_i} \otimes R_i)]\mathrm{vec}(K_{i,k+1}). \qquad (23)
\end{aligned}$$

Moreover, for any two vectors a, b and a sufficiently large number $s > 0$, define

$$\begin{aligned}
\delta_a =&[\mathrm{vecv}(a(t_1)) - \mathrm{vecv}(a(t_0)), \cdots, \\
&\mathrm{vecv}(a(t_s)) - \mathrm{vecv}(a(t_{s-1}))]^T, \\
\Gamma_{a,b} =&[\int_{t_0}^{t_1} a \otimes b d\tau, \int_{t_1}^{t_2} a \otimes b d\tau, \cdots, \int_{t_{s-1}}^{t_s} a \otimes b d\tau]^T. \qquad (24)
\end{aligned}$$

Equations (22)-(24) imply the following linear equation

$$\Psi_{ijk}\begin{bmatrix} \mathrm{vecs}(P_{ik}) \\ \mathrm{vec}(K_{i,k+1}) \\ \mathrm{vec}((D_i - \mathcal{S}_i(X_{ij}))^T P_{ik}) \end{bmatrix} = \Phi_{ijk}, \qquad (25)$$

where

$$\begin{aligned}
\Psi_{ijk} =&[\delta_{\bar{x}_{ij}\bar{x}_{ij}}, -2\Gamma_{\bar{x}_{ij}\bar{x}_{ij}}(I_{n_i} \otimes K_{ik}^T R_i) - 2\Gamma_{\bar{x}_{ij}u_i}(I_{n_i} \otimes R_i), \\
&- 2\Gamma_{\bar{x}_{ij}v}], \\
\Phi_{ijk} =&-\Gamma_{\bar{x}_{ij}\bar{x}_{ij}}\mathrm{vec}(Q_i + K_{ik}^T R_i K_{ik}).
\end{aligned}$$

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 $t_0 < t_1 < \cdots < t_s$,

$$\mathrm{rank}([\Gamma_{\bar{x}_{ij}\bar{x}_{ij}}, \Gamma_{\bar{x}_{ij}u_i}, \Gamma_{\bar{x}_{ij}v}]) = \frac{n_i(n_i+1)}{2} + (m_i + q)n_i, \qquad (26)$$

then the matrix Ψ_{ijk} for the ith 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

$$\begin{aligned}
X_i =&X_{i1} + \sum_{j=2}^{h_i+1} \alpha_{ij} X_{ij}. \\
\mathcal{S}_i(X_i) =&\mathcal{S}_i(X_{i1}) + \sum_{j=2}^{h_i+1} \alpha_{ij}\mathcal{S}_i(X_{ij}) = B_i U_i + D_i
\end{aligned}$$

which is rewritten as

$$\mathcal{A}_i \chi_i = b_i, \qquad (27)$$

where

$$\begin{aligned}
\mathcal{A}_i =&[\mathcal{A}_{i1} \quad \mathcal{A}_{i2}], \\
\mathcal{A}_{i1} =&\begin{bmatrix} \mathrm{vec}(\mathcal{S}_i(X_{i2})) & \cdots & \mathrm{vec}(\mathcal{S}_i(X_{i,h+1})) \\ \mathrm{vec}(X_{i2}) & \cdots & \mathrm{vec}(X_{i,h+1}) \end{bmatrix}, \\
\mathcal{A}_{i2} =&\begin{bmatrix} 0 & -I_q \otimes (P_{ik}^{-1} K_{i,k+1}^T R_i) \\ -I_{n_i q} & 0 \end{bmatrix}, \\
\chi_i =&[\alpha_{i2}, \cdots, \alpha_{i,h+1}, \mathrm{vec}(X_i)^T, \mathrm{vec}(U_i)^T]^T, \\
b_i =&\begin{bmatrix} \mathrm{vec}(-\mathcal{S}_i(X_{i1}) + D_i) \\ -\mathrm{vec}(X_{i1}) \end{bmatrix}.
\end{aligned}$$

One can observe that matrices $P_{ik}, K_{i,k+1}$ and $\mathcal{S}_i(X_{i1})$, $\cdots, \mathcal{S}_i(X_{i,h_i+1})$ in (27) are obtainable by (25). Also, D_i can be computed from (25) by $\mathrm{vec}(D_i) = \mathrm{vec}(D_i - \mathcal{S}_i(X_{i0}))$.

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

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

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

Proof: Letting $P_{ik} = P_{ik}^T > 0$ be the solution to (13). $K_{i,k+1}$ is uniquely determined by (14) and $T_{i,k} = (D_i - \mathcal{S}_i(X_{ij}))^T P_{ik}$. On the other hand, letting $\hat{P}$, $\hat{K}$ and $\hat{T}$ solve (25), condition (26) ensures that $P_{ik} = \hat{P}$, $K_{i,k+1} = \hat{K}$ and $T_{ik} = \hat{T}$ are uniquely determined. By [28], we have $\lim_{k\to\infty} K_{ik} = K_i^*$, $\lim_{k\to\infty} P_{ik} = P_i^*$. The convergence of sequences $\{\bar{P}_{ik}\}_{k=0}^{\infty}$ and $\{\bar{K}_{ik}\}_{k=1}^{\infty}$ obtained by non-model-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 mininal polynomial of matrix E such that there is an unknown matrix $G \in \mathbb{R}^{q \times q_m}$ such that $v(t) = Gw(t)$. Interestingly, we convert the problem of unmeasurability of $v(t)$ into that of an unknown matrix G.

Algorithm 2 ADP Algorithm for COORP

$i \leftarrow 1$
repeat
  Compute matrices $X_{i0}, X_{i1}, \cdots, X_{i,h_i+1}$
  Apply an initial policy $u_i^0 = -K_{i0}x + \xi_i$ on $[t_0, t_s]$ with exploration noise ξ_i and $A_i - B_i K_{i0}$ a Hurwitz matrix
  $j \leftarrow 0$
  repeat
    Compute $\Gamma_{\bar{x}_{ij}\bar{x}_{ij}}, \Gamma_{\bar{x}_{ij}u_i}, \Gamma_{\bar{x}_{ij}v}$ s.t. (26) holds
    $j \leftarrow j + 1$
  until $j = h_i + 2$
  $j \leftarrow 0, k \leftarrow 0$
  repeat
    Solve P_{ik} and $K_{i,k+1}$ from (25)
    $k \leftarrow k + 1$
  until $|P_{ik} - P_{i,k-1}| < \epsilon_i$ with ϵ_i a small positive constant.
  $k^* \leftarrow k, j \leftarrow 1$
  repeat
    Solve $\mathcal{S}_i(X_{ij})$ from (25)
    $j \leftarrow j + 1$
  until $j = h_i + 2$
  Solve (X_i^*, U_i^*) from Problem 1
  $L_{i,k^*} \leftarrow U_i^* + K_{i,k^*} X_i^*$
  Obtain the following suboptimal controller

$$u_i^* = -K_{i,k^*}x_i + L_{i,k^*}\zeta_i \tag{28}$$

  $i \leftarrow i + 1$
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 ith vehicle for the purpose of simulation [32],

$$\begin{aligned} \dot{s}_i =& \nu_i, \\ \dot{\nu}_i =& a_i, \\ \dot{a}_i =& \tau_i^{-1} a_i + \tau_i^{-1} u_i + d_i \end{aligned} \tag{29}$$

where s_i, ν_i, a_i, τ_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 τ_i and the mass of vehicle $\#i$. The values of τ_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,

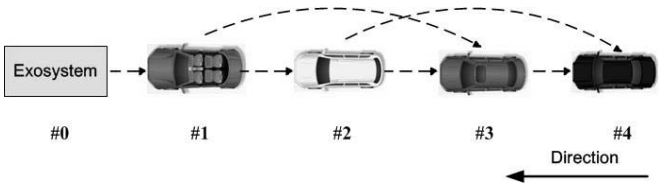

Fig. 1. The structure of the vehicular platoon

TABLE I
System parameters

Parameter	Value	Parameter	Value	Parameter	Value
$\tau_1[s]$	0.1	$\tau_2[s]$	0.11	$\tau_3[s]$	0.12
$\tau_4[s]$	0.13	$d_1[m/s^3]$	3	$d_2[m/s^3]$	3.1
$d_3[m/s^3]$	3.2	$d_4[m/s^3]$	3.3		

both desired trajectory s_i^* and disturbance d_i are generated by the following exosystem:

$$\begin{aligned} \dot{v}_1 =& v_2, \\ \dot{v}_2 =& 0, \\ d_i =& d_i v_2, \\ s_i^* =& -15v_1 - 20(5-i)v_2 \end{aligned} \tag{30}$$

with the initial value $v = [0 \quad 1]^T$.

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.

$$X_1 = \begin{bmatrix} 15.0000 & 80.0000 \\ 0.0016 & 15.0084 \\ -0.0068 & -0.0045 \end{bmatrix}, X_1^* = \begin{bmatrix} 15 & 80 \\ 0 & 15 \\ 0 & 0 \end{bmatrix},$$
$$U_1 = \begin{bmatrix} -0.0019 & -0.2982 \end{bmatrix}, U_1^* = \begin{bmatrix} 0 & -0.3 \end{bmatrix}.$$

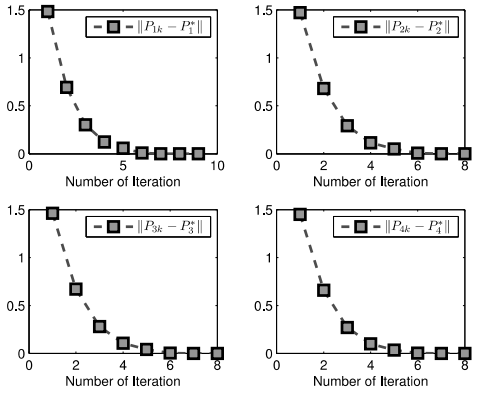

Fig. 2. The comparison of P_{ik} at kth iteration of ith subsystem and their optimal values

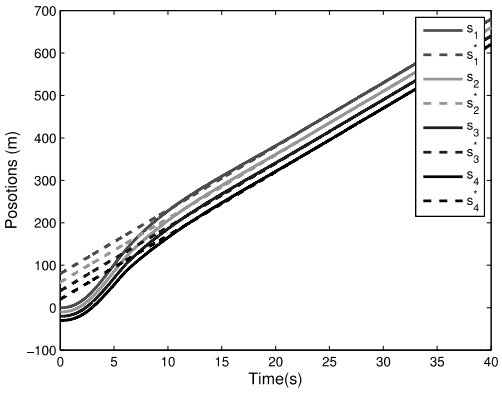

Fig. 3. The positions of vehicles and their desired values

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

[1] J. Huang, *Nonlinear Output Regulation: Theory and Applications*. Philadelphia, PA: SIAM, 2004.

[2] A. Isidori, L. Marconi, and A. Serrani, *Robust Autonomous Guidance: An Internal Model Approach*. London: UK: Springer-Verlag, 2003.

[3] R. Marino and P. Tomei, "Output regulation for linear systems via adaptive internal model," *IEEE Transactions on Automatic Control*, vol. 48, no. 12, pp. 2199–2202, 2003.

[4] W. Ren and R. Beard, *Distributed Consensus in Multi-vehicle Cooperative Control*, ser. Communications and Control Engineering. London, U.K.: Springer-Verlag London, 2008.

[5] M. Porfiri, D. Roberson, and D. Stilwell, "Tracking and formation control of multiple autonomous agents: A two-level consensus approach," *Automatica*, vol. 43, no. 8, pp. 1318 – 1328, 2007.

[6] Y. Su and J. Huang, "Cooperative output regulation of linear multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 4, pp. 1062–1066, 2012.

[7] X. Wang, Y. Hong, J. Huang, and Z.-P. Jiang, "A distributed control approach to a robust output regulation problem for multi-agent linear systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 12, pp. 2891–2895, 2010.

[8] L. Liu, "Adaptive cooperative output regulation for a class of nonlinear multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 6, pp. 1677–1682, 2015.

[9] Z. Ding, "Consensus output regulation of a class of heterogeneous nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 10, pp. 2648–2653, 2013.

[10] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. Cambridge, MA: MIT Press, 1998.

[11] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard University, 1974.

[12] Y. Jiang and Z. P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[13] ——, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.

[14] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[15] W. Gao, Y. Jiang, Z. P. Jiang, and T. Chai, "Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming," *Automatica*, vol. 72, pp. 37–45, 2016.

[16] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[17] W. Gao, Z. P. Jiang, and K. Ozbay, "Data-driven adaptive optimal control of connected vehicles," to appear in *IEEE Transactions on Intelligent Transportation Systems*, 2016.

[18] W. Gao, M. Huang, Z. P. Jiang, and T. Chai, "Sampled-data-based adaptive optimal output-feedback control of a 2-degree-of-freedom helicopter," *IET Control Theory and Applications*, vol. 10, no. 12, pp. 1440–1447, 2016.

[19] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, 2015.

[20] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3051–3056, 2014.

[21] W. Gao and Z. P. Jiang, "Nonlinear and adaptive suboptimal control of connected vehicles: A global adaptive dynamic programming approach," to appear in *Journal of Intelligent & Robotic Systems*, 2016.

[22] ——, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 4164–4169, 2016.

[23] M. I. Abouheaf, F. L. Lewis, K. G. Vamvoudakis, S. Haesaert, and R. Babuska, "Multi-agent discrete-time graphical games and reinforcement learning solutions," *Automatica*, vol. 50, no. 12, pp. 3038 – 3053, 2014.

[24] Y. Jiang and Z. P. Jiang, "Robust adaptive dynamic programming for large-scale systems with an application to multimachine power systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 10, pp. 693–697, 2012.

[25] Z. P. Jiang, A. R. Teel, and L. Praly, "Small-gain theorem for ISS systems and applications," *Mathematics of Control, Signals and Systems*, vol. 7, no. 2, pp. 95–120, 1994.

[26] J. Xiang, W. Wei, and Y. Li, "Synchronized output regulation of linear networked systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1336–1341, 2009.

[27] A. J. Krener, "The construction of optimal linear and nonlinear regulators," in *Systems, Models and Feedback: Theory and Applications*, A. Isidori and T. J. Tarn, Eds. Birkhauser Boston, 1992, vol. 12, pp. 301–322.

[28] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[29] G. Tao, *Adaptive Control of Systems with Actuator Failures*. London: UK: Springer, 2004.

[30] S. Shladover, D. Su, and X.-Y. Lu, "Impacts of cooperative adaptive cruise control on freeway traffic flow," *Transportation Research Record*, vol. 2324, pp. 63–70, 2012.

[31] J. Ploeg, D. P. Shukla, N. van de Wouw, and H. Nijmeijer, "Controller synthesis for string stability of vehicle platoons," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 854–865, 2014.

[32] S. S. Stankovic, M. J. Stanojevic, and D. D. Siljak, "Decentralized overlapping control of a platoon of vehicles," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 5, pp. 816–832, 2000.