This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



# Reinforcement Learning-based Adaptive Optimal Exponential Tracking Control of Linear Systems with Unknown Dynamics

Ci Chen, Hamidreza Modares, Kan Xie, Frank L. Lewis, *Fellow, IEEE*, Yan Wan, *Senior Member, IEEE*, and Shengli Xie, *Fellow, IEEE*

***Abstract*—Reinforcement learning has been successfully employed as a powerful tool in designing adaptive optimal controllers. Recently, off-policy learning has emerged to design optimal controllers for systems with completely unknown dynamics. However, current approaches for optimal *tracking* control design either result in bounded tracking error, rather than zero tracking error, or require partial knowledge of the system dynamics. Moreover, they usually require to collect a large set of data to learn the optimal solution. To obviate these limitations, this paper applies a combination of off-policy learning and experience-replay for output regulation tracking control of continuous-time linear systems with completely unknown dynamics. To this end, the off-policy integral reinforcement learning-based technique is first used to obtain the optimal control feedback gain, and to explicitly identify the involved system dynamics using the same data. Secondly, a data-efficient based experience replay method is developed to compute the exosystem dynamics. Finally, the output regulator equations are solved using data measured online. It is shown that the proposed control method stabilizes the closed-loop tracking error dynamics, and gives an explicit exponential convergence rate for the output tracking error. Simulation results show the effectiveness of the proposed approach.**

***Index Terms*—Adaptive optimal control, exponential tracking control, reinforcement learning, output regulation.**

## I. Introduction

LAST decades have witnessed reinforcement learning (RL) [1]–[3] or approximate/adaptive dynamic programming (ADP) [4]–[6] as a promising and powerful technique in designing non-model based/data-driven control protocols. RL, inspired by biological systems, finds an optimal control policy by optimizing accumulated rewards [1]. RL algorithms are usually implemented on an actor-critic structure in which the critic evaluate the performance the current policy based on measured data and the actor finds an improved policy using the evaluated policy [1], [2]. The state-of-art RL algorithms include [7]–[10], in which [7] solves a constrained optimization problem to update the policy in each step. This algorithm is validated on a variety of applications. Compared with classical dynamic programming [11], RL provides a feasible way to avoid the curse of dimensionality and requiring a system model [5], [6]. On the other hand, in contrast to traditional adaptive controllers [12], [13], which only consider stabilizing the tracking error dynamics, RL approaches learn the solution to the optimal tracking controller in an adaptive fashion, resulting in an adaptive optimal controller that minimizes the transient response of the error while assuring stability of the system [14].

Model-based RL algorithms have been widely used for continuous-time (CT) systems to find an optimal control solution assuming that the knowledge of the system dynamics is known a priori [15], [16]. This knowledge might not be available in many applications. To obviate this issue, the *Integral Reinforcement Learning* (IRL) algorithm was first presented for CT systems to relax the requirement of knowing the drift dynamics. The IRL-based method has a feature of learning the optimal feedback gains directly from input-output data along the system trajectories [2]. This is different from system identification based methods, where the optimal gains are only calculated after the system dynamics are identified. The details of the IRL algorithm and its implementation for solving the optimal regulator problem can be found in [2], [4], [17]. A suboptimal output-feedback controller using IRL is developed in [18] for CT systems. Partial knowledge of the system dynamics is required in the aforementioned IRL-based methods [19]–[21]. To deal with completely unknown system dynamics, an adaptive optimal algorithm is proposed for CT linear systems in [17], and is successfully extended to control of nonlinear systems [22], [23], output-feedback systems [24], and stochastic systems [25]. Experience replay [26], also called concurrent learning [27]–[29], has also been used in [26], [30] for CT linear systems to speed up the learning process. Although elegant, most of these learning-based algorithms are aimed at achieving adaptive optimal stabilization with respect to a set-point or an equilibrium of interest [4], [31]. Most of control problems in real world are, however, tracking problems for which the system state

This work was supported in part by the National Natural Science Foundation of China under Grant 61703112, Grant 61703113, Grant 61333013, Grant 61727810, and Grant 61633007, in part by the U.S. National Science Foundation under Grant 1839804, and in part by the Office of Naval Research under Grant N00014-17-1-2239 and Grant N00014-18-1-2221. (*Corresponding author: Shengli Xie.*)

C. Chen and K. Xie are with School of Automation, Guangdong University of Technology, Guangdong Key Laboratory of IoT Information Technology, Guangzhou, 510006 China, and also with UTA Research Institute, The University of Texas at Arlington, Fort Worth 76118 USA (e-mail: gdutc-c@gmail.com, kanxiegdut@gmail.com).

H. Modares is with Mechanical Engineering Department, Michigan State University, East Lansing, USA (e-mail: modaresh@mst.edu).

S. Xie is with School of Automation, Guangdong University of Technology, and also with Guangdong Key Laboratory of IoT Information Technology, Guangzhou, 510006 China (e-mail: shlxie@gdut.edu.cn).

F. L. Lewis, and Y. Wan are with UTA Research Institute, The University of Texas at Arlington, Fort Worth 76118, USA. F. L. Lewis is also Foreign Consulting Professor, Guangdong University of Technology, Guangzhou, 510006 China. (e-mail: lewis@uta.edu, yan.wan@uta.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



of output trajectory is required to track a desired trajectory.

Extension of designing RL-based adaptive optimal regulators to adaptive optimal trackers has received numerous attention due to its broader applications. An IRL-based optimal learning algorithm to solve the linear quadratic tracking problem is proposed in [32] by introducing a discounted performance function. In [33], an ADP-based online optimal control is proposed for linear systems, assuming that the desired trajectory is generated by an asymptotically stable exosystem. Several neural-network based optimal tracking controllers are developed for nonlinear uncertain systems, see [34]–[39] for instance. In [40], ultimately bounded tracking result is achieved for CT nonlinear affine systems. The $H_\infty$ tracking control problem has also been studied in [41], where an off-policy IRL is proposed for nonlinear systems with unknown dynamics. Most recently, optimal synchronization of heterogeneous nonlinear systems with unknown dynamics is studied in [42]. The above mentioned results might achieve a bounded tracking error if some conditions on the discount factor or weighting matrices are not satisfied. Moreover, they assume that eigenvalues of the linearized command generator dynamics are located in a certain predesign region.

For solving the linear optimal output regulation problem (LOORP), which is the problem of interest in this paper, model-based solutions that require the knowledge of the system dynamics are presented in [43], [44]. In order to solve LOORP in means of adaptive optimal control, output regulation theory and ADP are brought together in [45], where an asymptotic trajectory tracking control is ensured. However, this solution requires partial knowledge of the system dynamics, i.e. the output dynamics of the system and the command generator (exosystem) dynamics, which might not be available in real world applications. Moreover, it first uses a data set to find the optimal feedback gain and then collects another set of data, as well as the feedback gain found in the first step, to find the optimal feedforward gain. This might lead to a computationally expensive learning algorithm.

This paper aims to design adaptive optimal tracking controllers for completely unknown CT linear uncertain systems. An adaptive IRL-based feedback algorithm is proposed to achieve online optimal control. The off-policy IRL technique is first employed to obtain the optimal control feedback gain, and to explicitly identify the system dynamics using the same collected data. Experience replay is then employed to identify the exosystem dynamics. Finally, the output regulator equations are solved online, and consequently to find the feedforward control input. Our learning structure allows to use the same data set for both finding the optimal feedback gain as well as solving the output regulator equations. This allows us to build a unified adaptive optimal exponential tracking control scheme, wherein no prior knowledge of the system dynamics is required. This is in contrast to [45], which requires the knowledge of the output dynamics of the system and the exosystem. Moreover, the proposed method leads to an asymptotic tracking control, which improves the bounded tracking result in [32]. The work of [7] uses a technique to find an approximated optimal control solution, while this paper provides an exact solution.

The remainder of the paper is organized as follows. In Section II, we formulate the trajectory tracking optimal control problem, and introduce preliminaries on output regulators and optimal control. In Section III, we combine an off-policy IRL method and an experience replay method to achieve adaptive optimal control design. In Section IV, we present numerical experiments to testify the effectiveness of the proposed algorithm. In Section V, results are concluded.

*Notations:* Throughout this paper, $\mathbb{Z}_+$ denotes the set of nonnegative integers. $\mathbb{C}^-/\mathbb{C}^+$ stands for the open left/right-half complex plane. For a matrix $X$, $||X||_F$ denotes its Frobenius norm, $\lambda(X)$ denotes its spectrum, $\sigma(X)$ is a set of $X$'s singular values with $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ being the minimum and maximin singular values, respectively, and $\rho(X)$ denotes its spectral radius. The notation $\otimes$ indicates the Kronecker product operator and $vec(X) = [x_1^T, x_2^T, \dots, x_n^T]^T$ denotes a vector-valued function of a matrix $X$ with $x_i \in \mathbb{R}^m$ being the $i$th column of $X$. For a symmetric matrix $X \in \mathbb{R}^{m\times m}$, $vecs(X) = [x_{11}, 2x_{12}, ..., 2x_{1m}, x_{22}, 2x_{23}, ..., 2x_{m-1,m}, x_{m,m}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)}$. For a column vector $v \in \mathbb{R}^n$, $vecv(v) = [v_1^2, v_1v_2, ..., v_1v_n, v_2^2, v_2v_3, ..., v_{n-1}v_n, v_n^2]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)}$. The notation $1_n \in \mathbb{R}^n$ is defined as a column vector with all the entries being ones, i.e., $1_n = [1, 1, ..., 1]^T$. The notation $O_{m\times n} \in \mathbb{R}^{m\times n}$ denotes an $m$-by-$n$ matrix with all the entries being zeroes. The notation $I_n$ denotes an $n$-by-$n$ identity matrix.

## II. Problem Formulation and Preliminaries

In this section, we first formulate the linear optimal output regulation problem and review its standard solution. A policy iteration technique is also provided to solve the corresponding algebraic Riccati equations (ARE) derived from the optimal control problem.

### A. Problem Formulation

Consider a class of continuous-time linear systems given by

$$\dot{x} = Ax + Bu + Dw, \tag{1}$$

$$y = Cx, \tag{2}$$

$$\dot{w} = Ew, \tag{3}$$

$$y_d = -Fw, \tag{4}$$

$$e = y + y_d, \tag{5}$$

where $x \in \mathbb{R}^n$ is the state vector; $u \in \mathbb{R}^m$ is the control input; $w \in \mathbb{R}^{q_m}$ is the state of the exosystem (3); $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{r\times n}$, $D \in \mathbb{R}^{n\times q_m}$, $E \in \mathbb{R}^{q_m\times q_m}$, and $F \in \mathbb{R}^{r\times q_m}$ are constant matrices; $y \in \mathbb{R}^r$ is the output of the plant; $y_d \in \mathbb{R}^r$ is the reference signal; and $e \in \mathbb{R}^r$ is the tracking error. The control objective is to design an optimal control protocol $u$ in terms of the system state $x$ and the exosystem state $w$, so that the tracking error $e$ converges to zero.

Some standard and common assumptions made on the system (1)–(5) are as follows.

***Assumption* 1:** $(A, B)$ is stabilizable.

***Assumption* 2:** For all $\lambda_i(E) \in \lambda(E)$, for $i = 1, 2, \dots, q_m$, $\text{rank}\begin{pmatrix} \begin{bmatrix} A - \lambda_i(E)I_n & B \\ C & O_{r\times m} \end{bmatrix} \end{pmatrix} = n + r.$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



*Assumption 1* ensures the controllability of the system [2]. As shown in [46], *Assumption 2* is introduced to ensure the solvability of output regulator equations, which are reviewed in the next subsection.

Note that we do not make any assumptions on stability of the exosystem matrix $E$. Some results on the optimal tracking control are provided in [32], [34], where the exosystem dynamics are required to be marginally stable. In our case, we extend these results, and allow the desired trajectory be generated by an unstable exosystem. In fact, it has the importance to allow unstable leader's dynamics. Take the satellite control for example, where tracking a ramp angle command is an important task.

### *B. Preliminaries on output regulator equations and optimal control*

In this subsection, we present some preliminaries on output regulator equations and optimal control. For linear systems, the output regulator control problem aims to design a class of controllers as

$$u = -Kx + Lw, \tag{6}$$

such that the system output tracking error $e$ from (5) globally exponentially converges to the zero with $\lambda(A - BK) \subset \mathbb{C}^-$, $K \in \mathbb{R}^{m\times n}$ is a feedback control gain matrix, and $L \in \mathbb{R}^{m\times q_m}$ is a feedforward control gain matrix. To solve the linear output regulation problem (LORP), a sufficient condition for designing gains $K$ and $L$ from (6) is needed:

***Theorem* 1:** ( [47]) Under *Assumption 1*, choose a $K$ such that $\lambda(A - BK) \subset \mathbb{C}^-$. Under *Assumption 2*, the regulator equations

$$XE = AX + BU + D, \tag{7a}$$
$$0 = CX + F, \tag{7b}$$

have a solution pair $(X, U)$ with $X \in \mathbb{R}^{n\times q_m}$ and $U \in \mathbb{R}^{m\times q_m}$. Moreover, the LORP is solvable by the controller (6) with $L = U + KX$, so that one has $\lim_{t\to\infty} u(t) - Uw(t) = 0$ and $\lim_{t\to\infty} x(t) - Xw(t) = 0$. □

From *Theorem 1*, we could see that the prior knowledge of system dynamics is a necessary condition for solving output regulator equations (7).

As for the optimal control of linear systems, defining

$$\bar{x} = x - Xw, \tag{8}$$
$$\bar{u} = u - Uw, \tag{9}$$

then, the system dynamics (1)–(5) become

$$\dot{\bar{x}} = A\bar{x} + B\bar{u}, \tag{10}$$
$$e = C\bar{x}. \tag{11}$$

The optimal feedback controller is found as

$$\bar{u} = -K^*\bar{x}, \tag{12}$$

where the feedback gain $K^*$ is found by solving the following constrained optimization problem.

*Problem 1:*

$$\min_{(\bar{x},\bar{u})} \int_0^\infty (\bar{x}^T Q\bar{x} + \bar{u}^T R\bar{u})dt \tag{13}$$
$$\text{subject to } (10) \tag{14}$$

where $Q = Q^T \geq 0$, $R = R^T > 0$, with $(A, \sqrt{Q})$ observable.

Following the optimal control theory [2], to solve *Problem 1* is equivalent to solving a linear quadratic regulator problem. Thus, in presence of (13) and (14), the optimal control is achieved by designing the feedback gain $K^*$ in (12) as

$$K^* = R^{-1}B^T P^*, \tag{15}$$

where $P^* = P^{*T} > 0$ is a unique solution to ARE as

$$A^T P^* + P^* A + Q - P^* BR^{-1}B^T P^* = 0. \tag{16}$$

As a result, the optimal control is solved if we design [2]

$$u = -K^* x + L^* w, \tag{17}$$

where $K^*$ is computed by solving *Problem 1*, and the feedforward gain is found by

$$L^* = U + K^* X, \tag{18}$$

with $(X, U)$ a solution pair of (7).

As shown in (15) and (16), the realization of the optimal controller (17) is based on the complete knowledge of the system dynamics. Since (16) is nonlinear in $P$, it is usually difficult to directly solve $P^*$ from (16). A model-based policy iteration algorithm for solving ARE is proposed in [48], and is recalled below.

***Lemma* 1:** ( [48]) Let $K_0 \in \mathbb{R}^{m\times n}$ be any stabilizing feedback gain matrix. $P_j = P_j^T > 0$ is the solution to the Lyapunov equation

$$(A - BK_j)^T P_j + P_j(A - BK_j) = -Q - K_j^T RK_j, \tag{19}$$

where for each $j = 1, 2, \ldots,$

$$K_j = R^{-1}B^T P_{j-1}. \tag{20}$$

Then, the following properties hold

1) $\lambda(A - BK_j) \subset \mathbb{C}^-$;
2) $P^* \leq P_{j+1} \leq P_j$;
3) $\lim_{j\to\infty} K_j = K^*$, $\lim_{j\to\infty} P_j = P^*$. □

In this paper, we present a data-driven adaptive optimal approach to solve the output regulation equations (7), and consequently the feedforward gain (18), as well as to solve *Problem 1*, which gives the optimal feedback control policy, without requiring any prior knowledge about the system dynamics. A partially model-free solution to this problem is presented in [45], which requires the knowledge of the output dynamics of the system and the exosystem.

## III. Data-Driven Adaptive Optimal Control Design

In this section, we give a data-driven algorithm to achieve an optimal control in presence of unknown system dynamics. Our algorithm consists of three parts. The first part is to obtain optimal control matrices $P_j$ and $K_{j+1}$ in *Lemma 1* using the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



off-policy IRL method. The second part is to employ off-policy IRL method to explicitly identify the system dynamics based on the same data set in the first part. The third part uses the experience replay method for solving the output regulator equations in (7) without relying on any system dynamics, including $C$ or $F$.

### *A. Solving the optimal gain $K^*$ using IRL based off-policy method*

In this subsection, we show the first part of our algorithm, which focuses on solving $K^*$ in *Problem 1* based on off-policy IRL method [3], [4], [17], [45], [49].

Considering the system (1), adding and subtracting $K_jx$ to the control input gives

$$\dot{x} = A_jx + B(K_jx + u) + Dw, \tag{21}$$

where $A_j = A - BK_j$. The control input $u$ acts as the behavior policy and generates the data required for learning, and the control policy $K_jx$ acts as the target policy and is being learned using the generated data from the behavior policy. Taking the time derivative of $x$ with respect to (21), (19), and (20) yields the off-policy Bellman equation

$$\begin{aligned} &x(t+\delta t)^T P_j x(t+\delta t) - x(t)^T P_j x(t) \\ &= \int_t^{t+\delta t} \Big[x^T(A_j^T P_j + P_j A_j)x + 2(u + K_jx)^T B^T P_j x \\ &\qquad + 2w^T D^T P_j x\Big] d\tau \\ &= -\int_t^{t+\delta t} x^T(Q + K_j^T R K_j)x d\tau + 2\int_t^{t+\delta} (u + K_jx)^T \\ &\qquad \times RK_{j+1}x d\tau + 2\int_t^{t+\delta} w^T D^T P_j x d\tau, \end{aligned} \tag{22}$$

where $R$ and $Q$ are known matrices given by designers. It is clear that the terms at the right hand side of (22) satisfy

$$x^T(Q + K_j^T R K_j)x = (x^T \otimes x^T)vec(Q + K_j R K_j), \tag{23}$$
$$x^T K_j^T R K_{j+1} x = (x^T \otimes x^T)(I_n \otimes K_j^T R)vec(K_{j+1}), \tag{24}$$
$$u^T R K_{j+1} x = (x^T \otimes u^T)(I_n \otimes R)vec(K_{j+1}), \tag{25}$$
$$w^T D^T P_j x = (x^T \otimes w^T)vec(D^T P_j). \tag{26}$$

For convenience, we define

$$\begin{aligned} \delta_{xx} = [&vecv(x(t_1)) - vecv(x(t_0)), vecv(x(t_2)) - vecv(x(t_1)), \\ &\ldots, vecv(x(t_s)) - vecv(x(t_{s-1}))]^T, \end{aligned} \tag{27}$$

$$\begin{aligned} \Gamma_{xx} = \Big[&\int_{t_0}^{t_1} x(\tau) \otimes x(\tau) d\tau, \int_{t_1}^{t_2} x(\tau) \otimes x(\tau) d\tau, \\ &\ldots, \int_{t_{s-1}}^{t_s} x(\tau) \otimes x(\tau) d\tau\Big]^T, \end{aligned} \tag{28}$$

$$\begin{aligned} \Gamma_{xu} = \Big[&\int_{t_0}^{t_1} x(\tau) \otimes u(\tau) d\tau, \int_{t_1}^{t_2} x(\tau) \otimes u(\tau) d\tau, \\ &\ldots, \int_{t_{s-1}}^{t_s} x(\tau) \otimes u(\tau) d\tau\Big]^T, \end{aligned} \tag{29}$$

$$\begin{aligned} \Gamma_{xw} = \Big[&\int_{t_0}^{t_1} x(\tau) \otimes w(\tau) d\tau, \int_{t_1}^{t_2} x(\tau) \otimes w(\tau) d\tau, \\ &\ldots, \int_{t_{s-1}}^{t_s} x(\tau) \otimes w(\tau) d\tau\Big]^T, \end{aligned} \tag{30}$$

where $t_0 < t_1 < t_2 < \ldots < t_{s-1} < t_s$ with $s$ being a positive integer. Substituting (23)–(30) into (22) results in the following compact linear form

$$\wp_j \begin{bmatrix} vecs(P_j) \\ vec(K_{j+1}) \\ vec(D^T P_j) \end{bmatrix} = \zeta_j, \tag{31}$$

where $\wp_j = [\delta_{xx}, -2\Gamma_{xx}(I_n \otimes K_j^T R) - 2\Gamma_{xu}(I_n \otimes R), -2\Gamma_{xw}]$, and $\zeta_j = -\Gamma_{xx} vec(Q + K_j^T R K_j)$. The uniqueness of solution to (31) is guaranteed under a rank condition given as follows.

***Lemma 2:*** [45] If there exists an integer $s^*$ such that for all $s > s^*$, for any sequence $t_0 < t_1 < \ldots < t_s$,

$$\begin{aligned} &rank([\Gamma_{xx}, \Gamma_{xu}, \Gamma_{xw}]) \\ &= \frac{n(n+1)}{2} + (m + q_m)n, \end{aligned} \tag{32}$$

then $\wp_j$ has full column rank for all $j \in \mathbb{Z}_+$. □

If the full column rank condition (32) is satisfied, (31) is uniquely solved as

$$\begin{bmatrix} vecs(P_j) \\ vec(K_{j+1}) \\ vec(D^T P_j) \end{bmatrix} = (\wp_j^T \wp_j)^{-1} \wp_j^T \zeta_j. \tag{33}$$

From (33), we directly obtain $P_j$ and $K_{j+1}$ using data generated by the behavior policy. From (19), $P_j$ is always ensured to be a positive definite matrix. For a given stabilizing gain $K_j$, if (19) is uniquely solved for $P_j = P_j^T$, then $K_{j+1}^j = R^{-1}B^T P_j$. Let $V_j = D^T P_j \in \mathbb{R}^{q_m \times n}$. By (22), it is clear that $P_j$, $K_{j+1}$, and $V_j$ satisfy (33). In addition, suppose $P = P^T \in \mathbb{R}^{n \times n}$, $K \in \mathbb{R}^{m \times n}$, and $V \in \mathbb{R}^{q_m \times n}$ solve (33). Then, one obtains $P_j = P$, $K_{j+1} = K$, and $V_j = V = D^T P$. The full column rank condition (32) guarantees that (33) is uniquely solved for the solution pair $(P, K, V)$. Therefore, computing (33) is equivalent to solving (19) and (20). By *Lemma 1*, the convergence of $P_j$ and $K_j$ is proved. Associated with the system data, the optimality $K^*$ in (15) is adaptively obtained by increasing the integer $j$ and repeatedly computing $P_j$ and $K_{j+1}$.

### *B. Identifying the system dynamics A, B, and D using off-policy IRL method*

We now compute the dynamics $A$, $B$, and $D$ in (1) by further exploiting the off-policy IRL method.

From the previous subsection, $P_j$ is obtained under the condition (32). Once $P_j$ is found, the system dynamics $D$ can be computed from the last row of (33). Moreover, given $P_j$ and $K_{j+1}$, the system dynamics $B$ are solvable by following (20). Hence, it remains to solve the dynamics $A$.

To this end, we define

$$\bar{x}_{mi} = X_{mi}x, \tag{34}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



where $X_{mi} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the $i$th entry on the diagonal being nonzero. Note that $X_{mi}$ is defined as a mask to map $x$ into $\bar{x}_{mi}$. Without loss of generosity, we assume that nonzero entry in $X_{mi}$ is one. Let $x_i$ be the $i$th entry of $x$, and $(A)_i$ be the $i$th row of $A$. Taking the time derivative of $\bar{x}_{mi}$ with respect to (1) yields

$$\begin{aligned}&\bar{x}_{mi}(t+\delta t)^T\bar{x}_{mi}(t+\delta t) - \bar{x}_{mi}(t)^T\bar{x}_{mi}(t)\\ &= \int_t^{t+\delta t}\Big[2\bar{x}_{mi}^T X_{mi}Ax + 2u^TB^T\bar{x}_{mi} + 2w^TD^T\bar{x}_{mi}\Big]d\tau\\ &= \int_t^{t+\delta t}\Big[2x_i(A)_i x + 2u^TB^T\bar{x}_{mi} + 2w^TD^T\bar{x}_{mi}\Big]d\tau. \quad (35)\end{aligned}$$

Moreover, we define

$$\begin{aligned}\Gamma_{xx_i} = \Big[&\int_{t_0}^{t_1} x(\tau)x_i(\tau)d\tau, \int_{t_1}^{t_2} x(\tau)x_i(\tau)d\tau,\\ &\ldots, \int_{t_{s-1}}^{t_s} x(\tau)x_i(\tau)d\tau\Big]^T. \quad (36)\end{aligned}$$

$$\begin{aligned}\delta_{\bar{x}_{mi}\bar{x}_{mi}} = [&vecv(\bar{x}_{mi}(t_1)) - vecv(\bar{x}_{mi}(t_0)),\\ &vecv(\bar{x}_{mi}(t_2)) - vecv(\bar{x}_{mi}(t_1)),\\ &\ldots, vecv(\bar{x}_{mi}(t_s)) - vecv(\bar{x}_{mi}(t_{s-1}))]^T, \quad (37)\end{aligned}$$

$$\begin{aligned}\Gamma_{\bar{x}_{mi}u} = \Big[&\int_{t_0}^{t_1} \bar{x}_{mi}(\tau)\otimes u(\tau)d\tau, \int_{t_1}^{t_2} \bar{x}_{mi}(\tau)\otimes u(\tau)d\tau,\\ &\ldots, \int_{t_{s-1}}^{t_s} \bar{x}_{mi}(\tau)\otimes u(\tau)d\tau\Big]^T, \quad (38)\end{aligned}$$

$$\begin{aligned}\Gamma_{\bar{x}_{mi}w} = \Big[&\int_{t_0}^{t_1} \bar{x}_{mi}(\tau)\otimes w(\tau)d\tau, \int_{t_1}^{t_2} \bar{x}_{mi}(\tau)\otimes w(\tau)d\tau,\\ &\ldots, \int_{t_{s-1}}^{t_s} \bar{x}_{mi}(\tau)\otimes w(\tau)d\tau\Big]^T. \quad (39)\end{aligned}$$

Based on the above definitions, (35) is thus rewritten as

$$\begin{aligned}\Gamma_{xx_i}(A)_i =&\frac{1}{2}\delta_{\bar{x}_{mi}\bar{x}_{mi}}vecs(I_n) - \Gamma_{\bar{x}_{mi}u}vec(B^T)\\ &- \Gamma_{\bar{x}_{mi}w}vec(D^T). \quad (40)\end{aligned}$$

The uniqueness of solution to (40) is guaranteed as follows.

***Lemma* 3:** If the rank condition (32) holds, then (40) yields

$$\begin{aligned}(A)_i =&(\Gamma_{xx_i}^T\Gamma_{xx_i})^{-1}\Gamma_{xx_i}^T\Big(\frac{1}{2}\delta_{\bar{x}_{mi}\bar{x}_{mi}}vecs(I_n)\\ &- \Gamma_{\bar{x}_{mi}u}vec(B^T) - \Gamma_{\bar{x}_{mi}w}vec(D^T)\Big), \quad (41)\end{aligned}$$

such that the dynamics $A$ in (1) is reconstructed as $A = [(A)_1^T, (A)_2^T, \ldots, (A)_n^T]^T$. □

***Proof***: The rank condition in (32) implies that the matrix $\big[\Gamma_x, 2\Gamma_{xu}, 2\Gamma_{xw}\big]$ has the full column rank [17], [45], where

$$\begin{aligned}\Gamma_x \triangleq \Big[&\int_{t_0}^{t_1} vecv(x(\tau))d\tau, \int_{t_1}^{t_2} vecv(x(\tau))d\tau,\\ &\ldots, \int_{t_{s-1}}^{t_s} vecv(x(\tau))d\tau\Big]^T. \quad (42)\end{aligned}$$

This means $\Gamma_x$ also has the full column rank. From the definitions of $\Gamma_x$ in (42) and $\Gamma_{xx_i}$ in (36), it is clear that $\Gamma_{xx_i}$, for $i = 1, 2, \ldots, n$, are subparts of $\Gamma_x$. Thus, under the rank condition in (32), the matrix $\Gamma_{xx_i}$ must have full column rank. Therefore, the system dynamics $A$, as well as $(A)_i$, is uniquely solved from (41). ■

***Remark* 1:** As shown in (34), a mask mapping based technique is proposed to identify the system dynamics $A$ using the collected system data. Technically, there are two advantages of using the proposed technique. One is to distribute the computation burdens of computing $A$ into $n$ channels by changing the order of the nonzero entry on the diagonal of $X_{mi}$. This distributed structure reveals that our method may be scalable to a large system. The other is to reveal that finding $A$ does not come at the expense of introducing extra rank assumptions, since the rank condition required to compute $A$ has been guaranteed by the same condition (32) in the previous subsection. •

***Remark* 2:** We find that IRL based off-policy method not only solves the optimal gain, but also explicitly identifies all the involved system dynamics (see solving the dynamics $A$, together with $B$ and $D$). •

### *C. Identifying the exosystem dynamics E, C, and F by experience-replay based method*

In this subsection, we solve the dynamics $E$, $C$, and $F$ by employing the experience replay method [26]–[30] to speed up the convergence and present a data-efficient algorithm.

Now, we further exploit the signal $w$ to reconstruct the dynamics $E$ from the collected data. To this end, the exosystem (3) is changed into a compact linear form as

$$\dot{w} = G(t)\chi, \quad (43)$$

where $G(t) \triangleq w^T \otimes I_{q_m} \in \mathbb{R}^{q_m\times q_m^2}$, and $\chi \triangleq vec(E) \in \mathbb{R}^{q_m^2}$. Let $\dot{w}$ and $G(t)$ in (43) pass through the following filters in terms of $\varrho(t) \in \mathbb{R}^{q_m}$ and $\Omega(t) \in \mathbb{R}^{q_m\times q_m^2}$, respectively,

$$\dot{\varrho}(t) = -\beta\varrho(t) + \dot{w}, \quad (44)$$

$$\dot{\Omega}(t) = -\beta\Omega(t) + G(t), \quad (45)$$

where $\beta \in \mathbb{R}$ is a positive design gain, and $g(0) = \Omega(0) = 0$. In presence of (43)–(44), the filtered signal $\Omega(t)$ in (45) is rewritten as $\Omega(t) = w_s^T(t) \otimes I_{q_m}$, where $w_s \in \mathbb{R}^{q_m}$ is computed by

$$\dot{w}_s(t) = -\beta w_s(t) + w, \quad (46)$$

with $w_s(0) = 0$. Solving (44) and (45), their solutions are given as

$$\varrho(t) = e^{-\beta t}\int_0^t e^{\beta\tau}\dot{w}(\tau)d\tau, \quad (47)$$

$$\Omega(t) = e^{-\beta t}\int_0^t e^{\beta\tau}G(\tau)d\tau. \quad (48)$$

Therefore, the system dynamics (3) can be rewritten as

$$\varrho(t) = Ew_s(t). \quad (49)$$

Considering (43), (47), and (48), (49) is further changed to

$$\varrho(t) = \Omega(t)\chi. \quad (50)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



Moreover, based on integration by parts, $\varrho(t)$ is expressed in terms of known variables $w(t)$ and $w_s(t)$ as

$$\varrho(t) = w(t) - e^{-\beta t}w(0) - \beta w_s(t). \tag{51}$$

Using (50) and (51), we define a prediction error $\psi \in \mathbb{R}^{q_m}$ as

$$\psi(t) = \varrho(t) - \Omega(t)\hat{\chi}(t), \tag{52}$$

where $\hat{\chi}(t) \in \mathbb{R}^{q_m^2}$ denotes an estimate of $\chi$. Accordingly, we define $\hat{E}(t)$ is an estimate of $E$. Hence, $\hat{\chi}(t) = vec(\hat{E})$.

In what follows, the system parameter $\chi$ is estimated using experience replay based computation. To this end, two memory stacks $\{\varrho_i\}_{i=1}^{p}$ and $\{\Omega_i\}_{i=1}^{p}$ with $\varrho_i = \varrho(t_i)$ and $\Omega_i = \Omega(t_i)$ are created that store $\varrho(t)$ and $\Omega(t)$ respectively at different time instants $t = t_i$ with $t_1 > t_2 > \dots t_p$, where $p$ is the length of the stored data.

Similar to (52), a prediction error driven by the stored data is defined as

$$\psi_i(t) = \varrho_i - \Omega_i\hat{\chi}(t), \tag{53}$$

where $i \leq p$. Using (52) and (53), an update law for the parameter estimate $\hat{\chi}$ is defined as

$$\dot{\hat{\chi}} = \beta_{\chi 1}\Omega^T(t)\psi(t) + \beta_{\chi 12}(||\hat{\chi}|| + \beta_{\chi 2})\sum_{i=1}^{p}\Omega_i^T\psi_i(t), \tag{54}$$

where $\beta_{\chi 1}$, $\beta_{\chi 12}$, and $\beta_{\chi 2}$ are positive scalar gains. The second term at the right hand side of (54) uses past store data in the learning process, inspired by experience replay technique [26]–[30]. The convergence of $\hat{\chi}$ to $\chi$ is guaranteed under a rank condition given as follows.

***Lemma* 4:** Consider the dynamics (54). If there exists a $p^*$ such that for all $p > p^*$, for any sequence $t_1 < t_2 < \dots < t_p$,

$$rank([\Omega_1^T, \Omega_2^T, \dots, \Omega_p^T]) = q_m^2, \tag{55}$$

then, the system parameter estimation $\hat{\chi}(t)$ is bounded $\forall t \geq 0$, and the estimated system dynamics $\hat{E}$ given by reshaping (54) exponentially converge to the actual dynamics $E$, $\forall t \geq t_p$. □

***Proof***: Consider the Lyapunov function candidate $V_\chi : \mathbb{R}^{q_m^2} \to \mathbb{R}$ defined as

$$V_\chi = \frac{1}{2}\tilde{\chi}^T\tilde{\chi}, \tag{56}$$

where $\tilde{\chi} = \chi - \hat{\chi} \in \mathbb{R}^{q_m^2}$. Using the update law (54), the error dynamics for $\tilde{\chi}(t)$ is given as

$$\dot{\tilde{\chi}} = -\beta_{\chi 12}(||\hat{\chi}|| + \beta_{\chi 2})\sum_{i=1}^{p}\Omega_i^T\Omega_i\tilde{\chi} - \beta_{\chi 1}\Omega^T(t)\Omega(t)\tilde{\chi}. \tag{57}$$

The time differentiation of (56) along (57) yields

$$\begin{aligned}\dot{V}_\chi = &-\beta_{\chi 12}(||\hat{\chi}|| + \beta_{\chi 2})\tilde{\chi}^T\sum_{i=1}^{p}\Omega_i^T\Omega_i\tilde{\chi} \\ &- \beta_{\chi 1}\tilde{\chi}^T\Omega^T(t)\Omega(t)\tilde{\chi} \leq 0.\end{aligned} \tag{58}$$

From (58), $\tilde{\chi}$ is bounded for $\forall t \geq 0$ according to the Lyapunov stability theory. Under the rank condition (55), the matrix $\sum_{i=1}^{p}\Omega_i^T\Omega_i$ in (58) is ensured positive definite, which guarantees that $\sigma_{\min}(\sum_{i=1}^{p}\Omega_i^T\Omega_i) > 0$. This allows us to define $\beta_{\chi 12}\sigma_{\min}\Big(\sum_{i=1}^{p}\Omega_i^T\Omega_i\Big) = 1$. Moreover, if (55) holds, then, using (56), (58) is further rewritten as

$$\dot{V}_\chi \leq -2(||\hat{\chi}|| + \beta_{\chi 2})V_\chi, \quad \forall t \geq t_p. \tag{59}$$

Therefore, (59) reveals that $V_\chi$, as well as $\tilde{\chi}(t)$, is exponentially stable $\forall t \geq t_p$. As a result, the estimated system dynamics $\hat{E}$ given by (59) exponentially converge to the actual dynamics $E$. Furthermore, there exists a finite time $t_\chi$ such that $||\hat{\chi}|| - ||\chi|| < \beta_{\chi 2} - \alpha_\chi$, where $\alpha_\chi$ is a certain positive constant. Therefore, (59) yields

$$\dot{V}_\chi \leq -2(||\chi|| + \alpha_\chi)V_\chi, \quad \forall t \geq t_\chi, \tag{60}$$

which reveals the convergence rate of $\tilde{\chi}$ is bigger than $||\chi||$, $\forall t \geq t_\chi$. The proof is completed. ■

We are now are ready to reconstruct the dynamics $C$ and $F$ using the current data and the past collected data as done for $E$. Recalling from (2) and (4), one has

$$y = Y_x\chi_C, \tag{61}$$

$$y_d = Y_w\chi_F, \tag{62}$$

where $Y_x \in \mathbb{R}^{r\times rn}$ and $Y_w \in \mathbb{R}^{r\times rq_m}$ are the regressor matrices with $Y_x \triangleq x^T \otimes I_r$ and $Y_w \triangleq -w^T \otimes I_r$, $\chi_C \triangleq vec(C)$, and $\chi_F \triangleq vec(F)$. Accordingly, we define $\hat{C}(t)$ and $\hat{F}(t)$ as estimates of $C$ and $F$, respectively. Hence, $\hat{\chi}_C(t) = vec(\hat{C})$ and $\hat{\chi}_F(t) = vec(\hat{F})$. Referring to (61) and (62), instantaneous prediction errors $\psi_y \in \mathbb{R}^r$ and $\psi_{y_d} \in \mathbb{R}^r$ are defined as

$$\psi_y = y(t) - Y_x(t)\hat{\chi}_C(t), \tag{63}$$

$$\psi_{y_d} = y_d(t) - Y_w(t)\hat{\chi}_F(t), \tag{64}$$

where $\hat{\chi}_C(t) \in \mathbb{R}^{rn}$ and $\hat{\chi}_F(t) \in \mathbb{R}^{rq_m}$ are the estimates of $\chi_C$ and $\chi_F$, respectively. Moreover, prediction errors driven by the stored data are defined as

$$\psi_{y_i}(t) = y_i - Y_{xi}\hat{\chi}_C(t), \tag{65}$$

$$\psi_{y_{di}}(t) = y_{di} - Y_{wi}\hat{\chi}_F(t), \tag{66}$$

where $y_i \triangleq y(t_i) \in \{y_i\}_{i=1}^{p_C}$, $y_{di} \triangleq y_d(t_i) \in \{y_{di}\}_{i=1}^{p_F}$, $Y_{xi} \triangleq Y_x(t_i) \in \{Y_{xi}\}_{i=1}^{p_C}$, and $Y_{wi} \triangleq Y_w(t_i) \in \{Y_{wi}\}_{i=1}^{p_F}$ are created at different time instants $t = t_i$ with $t_1 > t_2 > \dots t_{p_C}(t_{p_F})$ with $p_C$ and $p_F$ being the memory stack lengths. Thus, update laws for the parameter estimates $\hat{\chi}_C$ and $\hat{\chi}_F$ are given as

$$\dot{\hat{\chi}}_C = \beta_{\chi x1}{Y_x}^T\psi_y + \beta_{\chi x12}(||\hat{\chi}|| + \beta_{\chi x2})\sum_{i=1}^{p_C}{Y_{xi}}^T\psi_{y_i}, \tag{67}$$

$$\dot{\hat{\chi}}_F = \beta_{\chi w1}{Y_w}^T\psi_{y_d} + \beta_{\chi w12}(||\hat{\chi}|| + \beta_{\chi w2})\sum_{i=1}^{p_F}{Y_{wi}}^T\psi_{y_{di}}, \tag{68}$$

where $\beta_{\chi x1}$, $\beta_{\chi x2}$, $\beta_{\chi w1}$, $\beta_{\chi w2}$, $\beta_{\chi x12}$, and $\beta_{\chi w12}$ are positive scalar gains. The convergence of (67) and (68) is guaranteed under a rank condition given as follows.

***Lemma* 5:** Consider the dynamics (67) and (68). If there exist $p_C^*$ and $p_F^*$ such that for all $p_C > p_C^*$ and $p_F > p_F^*$, for any sequences $t_1 < t_2 < \dots < t_{p_C}$ and $t_1 < t_2 < \dots < t_{p_F}$,

$$rank([{Y_{x1}}^T, {Y_{x2}}^T, \dots, {Y_{xp_C}}^T]) = rn, \tag{69}$$

$$rank([{Y_{w1}}^T, {Y_{w2}}^T, \dots, {Y_{wp_F}}^T]) = rq_m, \tag{70}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



then, system parameter estimations $\hat{\chi}_C$ and $\hat{\chi}_F$, as well as $\hat{C}$ and $\hat{F}$, are bounded $\forall t \geq 0$, and $\hat{C}$ and $\hat{F}$ exponentially converge to $C$ and $F$ for $\forall t \geq t_{P_{CF}} \triangleq \max\{t_{pC}, t_{pF}\}$. □

***Proof***: Consider the Lyapunov function candidate $V_{\chi CF}$ defined as

$$V_{\chi CF} = \frac{1}{2}\tilde{\chi}_C^T\tilde{\chi}_C + \frac{1}{2}\tilde{\chi}_F^T\tilde{\chi}_F, \tag{71}$$

where $\tilde{\chi}_C = \chi_C - \hat{\chi}_C$ and $\tilde{\chi}_F = \chi_F - \hat{\chi}_F$. Taking the differentiation of (71) by using (67) and (68) yields

$$\begin{aligned}\dot{V}_{\chi CF} = &-\beta_{\chi x1}\tilde{\chi}_C^T {Y_x}^T Y_x\tilde{\chi}_C - (||\hat{\chi}|| + \beta_{\chi x2})\tilde{\chi}_C \\ &- \beta_{\chi w1}\tilde{\chi}_F^T {Y_w}^T Y_w\tilde{\chi}_F - (||\hat{\chi}|| + \beta_{\chi w2})\tilde{\chi}_F.\end{aligned} \tag{72}$$

From (72), it is clear that $\tilde{\chi}_C(t)$ and $\tilde{\chi}_F(t)$ are bounded for $\forall t \geq 0$. If (69) and (70) hold, then $\sum_{i=1}^{p_C} {Y_{xi}}^T Y_{xi}$ and $\sum_{i=1}^{p_F} {Y_{wi}}^T Y_{wi}$ are positive definite, based on which we have $\sigma_{\min}(\sum_{i=1}^{p} {Y_{xi}}^T Y_{xi}) > 0$ and $\sigma_{\min}(\sum_{i=1}^{p} {Y_{wi}}^T Y_{wi}) > 0$. Furthermore, let $\beta_{\chi x12}\sigma_{\min}\Big(\sum_{i=1}^{p_C} Y_{xi}^T Y_{xi}\Big) = 1$, and $\beta_{\chi w12}\sigma_{\min}\Big(\sum_{i=1}^{p_F} Y_{wi}^T Y_{wi}\Big) = 1$. Thus, $\forall t \geq t_{P_{CF}}$, (72) is changed to

$$\dot{V}_{\chi CF} \leq -2\min\big\{(||\hat{\chi}|| + \beta_{\chi x2}), (||\hat{\chi}|| + \beta_{\chi w2})\big\}V_{\chi CF}. \tag{73}$$

As a result, $V_{\chi CF}$ exponentially converges to the zero, $\forall t \geq t_{pCF}$. Moreover, there exists a finite time $t_{\chi CF}$ such that $||\hat{\chi}|| - ||\chi|| < \min\{\beta_{\chi x2}, \beta_{\chi w2}\} - \alpha_{\chi CF}$, where $\alpha_{\chi CF}$ is a certain positive constant. Therefore, (73) is changed to

$$\dot{V}_{\chi CF} \leq -2(||\chi|| + \alpha_{\chi CF})V_{\chi CF}, \quad \forall t \geq t_{\chi CF}, \tag{74}$$

which means that the convergence rate of $\tilde{\chi}_C$ or $\tilde{\chi}_F$ is bigger than $||\chi||$, $\forall t \geq t_{\chi CF}$. This completes the proof. ■

## IV. Overall Control Design

In this section, we aim to bring the previous results together and to present main results of our data-driven control scheme. To this end, we first solve the output regulation equations (7) by using the estimated dynamics obtained in the previous subsections, and then construct the optimal control in (17).

To solve the regulator equations (7), we rewrite them into a compact form

$$\begin{aligned}&\begin{bmatrix} I_n & O_{n\times m} \\ O_{r\times n} & O_{r\times m}\end{bmatrix}\begin{bmatrix} X \\ U\end{bmatrix}E \\ &= \begin{bmatrix} A & B \\ C & O_{r\times m}\end{bmatrix}\begin{bmatrix} X \\ U\end{bmatrix} + \begin{bmatrix} D \\ F\end{bmatrix}.\end{aligned} \tag{75}$$

We further change (75) to

$$Q vec\begin{pmatrix} X \\ U\end{pmatrix} = vec\begin{pmatrix} D \\ F\end{pmatrix}, \tag{76}$$

where $Q = E^T \otimes \begin{bmatrix} I_n & O_{n\times m} \\ O_{r\times n} & O_{r\times m}\end{bmatrix} - I_q \otimes \begin{bmatrix} A & B \\ C & O_{r\times m}\end{bmatrix}$. With the real-time estimated leader dynamics $\hat{C}$, $\hat{E}$, and $\hat{F}$ obtained in the previous subsections, the output regulator equations are rewritten as

$$\hat{X}\hat{E} = A\hat{X} + B\hat{U} + D, \tag{77a}$$

$$0 = \hat{C}\hat{X} + \hat{F}, \tag{77b}$$

where $\hat{X}(t)$ and $\hat{U}(t)$ denote the solutions matrices, and $\hat{E}$, $\hat{C}$, and $\hat{F}$ are reconstructed by (54), (67) and (68), respectively. Rewriting (77) yields

$$\hat{Q} vec\begin{pmatrix} \hat{X} \\ \hat{U}\end{pmatrix} = vec\begin{pmatrix} D \\ \hat{F}\end{pmatrix}, \tag{78}$$

where $\hat{Q} = \hat{E}^T \otimes \begin{bmatrix} I_n & O_{n\times m} \\ O_{r\times n} & O_{r\times m}\end{bmatrix} - I_q \otimes \begin{bmatrix} A & B \\ \hat{C} & O_{r\times m}\end{bmatrix}$. Then, a standard form of linear equations in (78) can be reformulated as

$$\hat{Q}\hat{\Psi} = \hat{\Phi}, \tag{79}$$

where $\hat{\Psi} = vec\begin{pmatrix} \hat{X} \\ \hat{U}\end{pmatrix} \in \mathbb{R}^{(n+m)q_m}$, and $\hat{\Phi} = vec\begin{pmatrix} D \\ \hat{F}\end{pmatrix} \in \mathbb{R}^{(n+r)q_m}$. Accordingly, for a actual value case, we define $\Psi = vec\begin{pmatrix} X \\ U\end{pmatrix} \in \mathbb{R}^{(n+m)q_m}$, and $\Phi = vec\begin{pmatrix} D \\ F\end{pmatrix} \in \mathbb{R}^{(n+r)q_m}$. The estimated solutions to the output regulator equations (79) are adaptively solved by

$$\dot{\hat{\Psi}} = -\frac{2\|\hat{E}\|_F + 2\aleph}{\sigma_{\min}^{\emptyset}(\hat{Q}^T\hat{Q}) + \hbar(t)}\hat{Q}^T(\hat{Q}\hat{\Psi} - \hat{\Phi}), \tag{80}$$

where $\aleph$ is any positive constant, $\sigma_{\min}^{\emptyset}$ is the minimum nonzero singular value, and $\hbar(t) = \begin{cases} 0 & if \quad \sigma_{\min}(\hat{Q}^T\hat{Q}) \neq 0 \\ \varsigma, & else\end{cases}$ is a bounded switching signal defined to avoid the singularity with $\varsigma$ any positive constant.

Now, we are ready to state the main result of this paper. The data-driven algorithm for dealing with *Problem 1* is summarized as *Algorithm 1*. As in policy iteration algorithms [3], [4], [17], [45], [49], an initial stabilizing control gain $K_0$ is required. We use the behavior policy $u = -K_0x + \xi$ to generate behavior from $t_0$ to $t_s$. The use of $\xi$ is same with [17], [45], [50] to excite the system, and consequently make the collected system data satisfy certain rank conditions. Examples of such noise are random noises [50], sinusoidal signals [17], [45], and decayed signals [4]. Then, we evaluate and improve estimation policy in (17) by using the input/partial-state information on $[t_0, t_s]$.

***Remark* 3:** The proposed optimal controller shown in *Algorithm 1* allows us to learn all the system dynamics, as well as the feedback gain i.e., $K^*$ in (12), simultaneously, compared with the existing literature. Moreover, in contrast to the work in [45], our optimal controller creates a distinguishing structure that the feedforward gains, i.e., $X$ and $U$ in (7), are online obtained after the feedback gain is updated. •

***Theorem* 2:** Given a class of continuous-time linear systems (1)–(5), the adaptive optimal controller is designed as

$$u = -K_{j^*}x + \hat{L}_{j^*}w, \tag{81}$$

where $K_{j^*}$ and $\hat{L}_{j^*} \triangleq \hat{U} + K_{j^*}\hat{X}$ are solved by following *Algorithm 1*. Then, the tracking error $e$ defined in (5) exponentially converges to the zero. □

***Proof:*** The proof has two parts. The first part is to show the estimated solutions to the output regulator equations $(\hat{X}, \hat{U})$ converge to the actual solutions $(X, U)$. The second part is to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



**Algorithm 1** *Data-Driven Adaptive Algorithm for Optimal Control with Completely Unknown Dynamics*

**Initialize:**
Utilizing $u = -K_0 x + \xi$ on $[t_0, t_s]$ with exploration noise $\xi$ and $\lambda(A - BK_0) \subset \mathbb{C}^-$. Let $j = 0$.
Compute $\Gamma_{xx}$, $\Gamma_{xu}$, and $\Gamma_{xw}$.
**if** the rank condition (32) holds **then**
  **for** $j$ **do**
    Solve (33) to obtain $P_j$ and $K_{j+1}$.
    **if** $\|P_j - P_{j-1}\| \leqslant \varepsilon$, where the constant $\varepsilon > 0$ is a predefined small threshold, **then**
      The optimal matrix $K^*$ is found by letting $j^* \leftarrow j$.
    **end if**
    Solve (41), (20), and (33) to find the system dynamics $A$, $B$, and $D$, respectively.
  **end for**
**else**
  Find $s > s^*$, and collect more data till (32) is satisfied, and then repeat Steps 5-10.
**end if**
**if** the rank condition (55) holds **then**
  Solve (54) to reconstruct the dynamics $\hat{E}$.
**else**
  Reset $s > p^*$, collect more data to make (55), and then solve $\hat{E}$.
**end if**
**if** the rank conditions (69) and (70) hold **then**
  Solve (67) and (68) to obtain the estimated dynamics $\hat{C}$ and $\hat{F}$.
**else**
  Collect more data till (69) and (70) are satisfied, and then do Step 20.
**end if**
Return control matrices $K_{j^*}$, $A$, $B$, $D$, $\hat{E}$, $\hat{C}$, and $\hat{F}$.
Solve (80) to find the estimated solutions to output regulator equations, $(\hat{X}, \hat{U})$, and form the optimal controller as (81).

show how the output tracking control is achieved based on the first part.

In order to better show the main result, we present the first part proof in APPENDIX A. Using a differential equation to solve linear systems is presented in the literature such as [51], [52]. Here, we design the convergence rate to satisfy (122), which plays a key role in achieving the stability of the closed-loop system, and in promoting the output tracking error to converge to the zero.

In what follows, we present the second part proof. From *Lemma 1*, it is clear that $K_{j^*}$ obtained from our data-driven approach satisfies $\lambda(A - BK_{j^*}) \subset \mathbb{R}^-$. Substituting the controller (81) into the system dynamics (1)–(5) yields

$$\dot{\bar{x}} = (A - BK_{j^*})\bar{x} - BK_{j^*}\tilde{X}w - B\tilde{U}w, \tag{82}$$

$$e = C\bar{x} + C\tilde{X}w, \tag{83}$$

where $\bar{x}$ is given in (8). Since $\lambda(A - BK_{j^*}) \subset \mathbb{R}^-$, it is clear that there exist positive-definite matrices $P_{j^*}$ and $Q_{j^*}$ such that $(A - BK_{j^*})^T P_{j^*} + P_{j^*}(A - BK_{j^*}) = -Q_{j^*}$. To analyze the stability of (82), we consider the Lyapunov function candidate $V_\chi : \mathbb{R}^n \to \mathbb{R}$ defined as

$$V_{\bar{x}} = \bar{x}^T P_{j^*} \bar{x}. \tag{84}$$

Taking the time differentiation of (84) along (83) yields

$$\begin{aligned}\dot{V}_{\bar{x}} =& \bar{x}^T \Big( P_{j^*}(A - BK_{j^*}) + (A - BK_{j^*})^T P_{j^*} \Big)\bar{x} \\ & - 2\bar{x}^T P_{j^*} \big( BK_{j^*}\tilde{X}w + B\tilde{U}w \big) \\ \leq & - \bar{x}^T Q_{j^*}\bar{x} - 2\bar{x}^T P_{j^*} \big( BK_{j^*}\tilde{X}w + B\tilde{U}w \big) \\ \leq & - \frac{\sigma_{\min}(Q_{j^*})}{2}\bar{x}^T\bar{x} + \frac{4\|P_{j^*}BK_{j^*}\|_F^2}{\sigma_{\min}(Q_{j^*})}\|\tilde{X}\|_F^2\|w\|^2 \\ & + \frac{4\|P_{j^*}B\|_F^2}{\sigma_{\min}(Q_{j^*})}\|\tilde{U}\|_F^2\|w\|^2, \end{aligned} \tag{85}$$

where Young's inequality is used to obtain the last inequality. From (122), (85) is further changed to

$$\dot{V}_{\bar{x}} \leq -\frac{\sigma_{\min}(Q_{j^*})}{2}\bar{x}^T\bar{x} + (\alpha_{c_1} + \alpha_{c_2}) \exp\big( - 2\alpha_\Psi^* t\big), \tag{86}$$

where $\alpha_{c_1} = \frac{4\|P_{j^*}BK_{j^*}\|_F^2 \bar{V}_{\Psi M}}{\sigma_{\min}(Q_{j^*})}$ and $\alpha_{c_2} = \frac{4\|P_{j^*}B\|_F^2 \bar{V}_{\Psi M}}{\sigma_{\min}(Q_{j^*})}$.

Taking the integration of (86) over the time $[0, t)$ yields

$$\begin{aligned} V_{\bar{x}}(t) \leq & - \int_0^t \frac{\sigma_{\min}(Q_{j^*})}{2}\bar{x}^T(\tau)\bar{x}(\tau)d\tau + \bar{\alpha}_{c1c2} \\ \leq & - \int_0^t \frac{\sigma_{\min}(Q_{j^*})}{2\sigma_{\max}(P_{j^*})}V_{\bar{x}}(\tau)d\tau + \bar{\alpha}_{c1c2}, \end{aligned} \tag{87}$$

where $\bar{\alpha}_{c1c2} \triangleq \bar{x}^T(0)P_{j^*}\bar{x}(0) + \alpha_{c1c2}$ with $\alpha_{c1c2} \triangleq \frac{\alpha_{c_1}+\alpha_{c_2}}{2\alpha_\Psi^*}$. Considering *Bellman-Gronwall Lemma* [53], one has the following inequality holds for $t \geq 0$

$$||\bar{x}(t)|| \leq \sqrt{\frac{\bar{\alpha}_{c1c2}}{\sigma_{\min}(P_{j^*})}} \exp\Big( - \frac{\sigma_{\min}(Q_{j^*})}{4\sigma_{\max}(P_{j^*})}t\Big), \tag{88}$$

which implies that $\bar{x}$ is exponentially convergent to zero. From (122) and (88), it is clear that (83) is changed to

$$\begin{aligned}\|e\| \leq & \|C\|_F\|\bar{x}\| + \|C\|_F\|\tilde{X}\|_F\|w\|_F \\ \leq & \ \|C\|_F\sqrt{\frac{\bar{\alpha}_{c1c2}}{\sigma_{\min}(P_{j^*})}} \exp\Big( - \frac{\sigma_{\min}(Q_{j^*})}{4\sigma_{\max}(P_{j^*})}t\Big) \\ & + \|C\|_F\sqrt{\bar{V}_{\Psi M}} \exp\big( - \alpha_\Psi^* t\big). \end{aligned} \tag{89}$$

As a result, the exponential convergence of the tracking error $e$ is proved. This completes the proof. ■

***Remark* 4:** We here show more insights in our proposed control design. Note that $w$ is a time-varying signal satisfying

$$\|w\| \leq \exp(\|E\|t). \tag{90}$$

Thus, one may ask how to guarantee the solution of *Problem 1* in presence of the dynamical signal generated by an exosystem in (3), especially for a case $\lambda_i(E) \in \mathbb{C}^+$ with $i \leq q_m$ (This means the right hand side of (90) will be unbounded as the time approaches the infinity). An answer to this problem is that our adaptive optimal control design (80) makes $\|\tilde{X}\|$ and $\|\tilde{U}\|$ exponentially converge to zero at the rate of $\|E\|_F + \alpha_\Psi^*$ shown in (122) such that $\|\tilde{X}\|_F^2\|w\|^2$ and $\|\tilde{U}\|_F^2\|w\|^2$ in (85) can be enveloped by an negative exponential function $\exp\big( - 2\alpha_\Psi^* t\big)$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



in (86). Such an negative exponential function finally ensures the stability of the closed-loop system, and drives the system output to converge to the desired reference even in a form of unbounded trajectories. •

## V. Numerical Experiments

In this section, we validate the effectiveness of the proposed algorithm by implementing it on motion control of a two-mass-spring system, which can be used to model a large number of practical systems, including deformable objects' movement and vibration of mechanical systems [54]. Here, a two-mass-spring system is shown in Fig. 1, where $m_1$ and $m_2$ denote masses, $k_1$ and $k_2$ are spring constants, $u_1$ denotes the input force for mass 1, and $y_1$ and $y_2$ are the displacements of two masses.

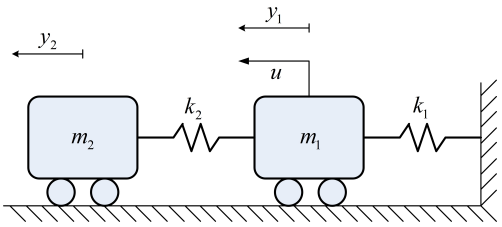



Fig. 1. Two-mass-spring system

The system dynamics are given as

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{-(k_1+k_2)}{m_1} & 0 & \frac{k_2}{m_1} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{-k_2}{m_1} & 0 & \frac{-k_2}{m_1} & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{m_1} \\ 0 \\ 0 \end{bmatrix} u + Dw, \tag{91}$$

$$\dot{w} = \begin{bmatrix} 0 & -1.5 \\ 1.5 & 0 \end{bmatrix} w, \tag{92}$$

$$e = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} x - \begin{bmatrix} 0 & 1 \end{bmatrix} w, \tag{93}$$

where $x = [y_1, \dot{y}_1, y_2, \dot{y}_2]$ and the non-zero matrix $D = \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 0 & -1 \\ 1 & 0 \end{bmatrix}$ is the disturbance matrix. The system dynamics in (91)–(93) are assumed unknown for the control design. To implement our method, we use MATLAB 7.11 as a test bed. The CPU for computing is Intel Core i3-3110M Processor with 3M Cache and 2.40 GHz. We choose $Q = 10I_4$ and $R = 1$, for which the optimal kernel matrix and optimal gain $P^*$ and $K^*$, respectively, become

$$(K^*)^T = \begin{bmatrix} 5.226851493044 \\ 4.522577029315 \\ -2.009860927015 \\ 3.398910084133 \end{bmatrix}^T, \tag{94}$$

$$P^* = \begin{bmatrix} 30.696643466203 & 5.226851493044 & -9.363690284787 & 17.381693598222 \\ 5.226851493044 & 4.522577029315 & -2.009860927015 & 3.398910084133 \\ -9.363690284787 & -2.009860927015 & 25.396317086420 & 0.776294880010 \\ 17.381693598222 & 3.398910084133 & 0.776294880010 & 28.501250994473 \end{bmatrix}. \tag{95}$$

The initial stabilizing control gain is chosen as $K_0 = [5, 5, 5, 5]$. The input, state and exogenous signals are collected from time 0 to 2 seconds. During the IRL learning period, we choose the exploration noise as $\xi = 100\sin(100t)$.

The evolutions of $P_j$ and $K_j$ to obtain the optimal gains $P^*$ and $K^*$ are depicted in Figs. 2 and 3, which clearly show that both $P_j$ and $K_j$ converge to the optimal gains $P^*$ and $K^*$ by increasing iteration step $j$. After 15 iterations, the following optimal control gains and Kernel matrix are obtained.

$$(K_{15})^T = \begin{bmatrix} 5.226848717976 \\ 4.522576405481 \\ -2.009861874422 \\ 3.398906478476 \end{bmatrix}^T,$$

$$P_{15} = \begin{bmatrix} 30.696630437957 & 5.226848717890 & -9.363699094235 & 17.381676208673 \\ 5.226848717890 & 4.522576405442 & -2.009861874032 & 3.398906478326 \\ -9.363699094235 & -2.009861874032 & 25.396295019753 & 0.776282223549 \\ 17.381676208673 & 3.398906478326 & 0.776282223549 & 28.501227919696 \end{bmatrix}.$$

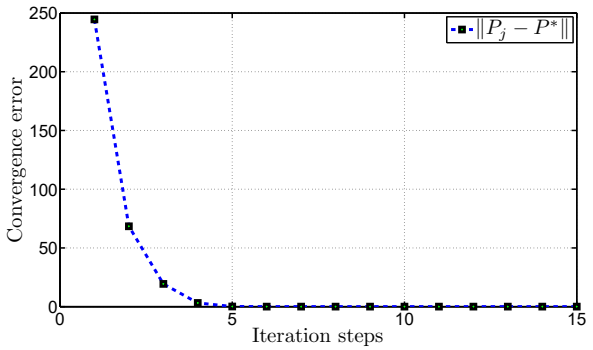



Fig. 2. Parameter evolution to obtain the optimal gain: $P^*$

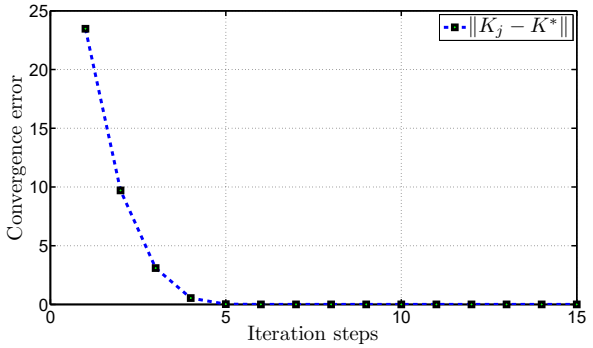



Fig. 3. Parameter evolution to obtain the optimal gain: $K^*$

Using the same collected system data, we solve for the system dynamics $A$, $B$, $D$ using the proposed off-policy IRL

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



method in Section III-B. The reconstructed system dynamics are given as

$$A = \begin{bmatrix} -0.000000248060 & 0.999999950277 \\ -2.500000007927 & -0.000000049098 \\ -0.000000094981 & -0.000000058663 \\ 1.250000107848 & 0.000000076873 \end{bmatrix}$$
$$\begin{bmatrix} 0.000001380223 & -0.000000597634 \\ 1.000000746405 & -0.000000122536 \\ 0.000000035178 & 0.999999631545 \\ -1.250000403401 & 0.000000392548 \end{bmatrix},$$

$$B = \begin{bmatrix} -0.000000000010 \\ 1.000000000003 \\ -0.000000000019 \\ 0.000000000012 \end{bmatrix},$$

$$D = \begin{bmatrix} 0.000000489285 & -1.499999817841 \\ 1.499999967739 & -0.000000261562 \\ 0.000000708626 & -1.499999874717 \\ 1.499999572274 & -0.000000096780 \end{bmatrix}.$$

Here, we compare (91) with the computed dynamics by calculating the sum of the distances between the actual and computed dynamics, which is $2.78 \times 10^{-6}$. This reveals that our off-policy IRL method efficiently and accurately recovers the system dynamics $A$, $B$, and $D$.

To solve for the dynamics $E$, $C$, and $F$, we use the proposed experience replay method in Section III-C. The data collected in the first two seconds are used to construct (54), (67), and (68), while rank conditions (55), (69), and (70) are satisfied. Trajectories of estimates $\hat{E}$, $\hat{C}$ and $\hat{F}$ are shown in Figs. 4–6, which indicate that the estimates quickly converge to the desired ones. To be more specific, from Fig. 4, the steady learning states converge to the actual values 1.5, 0, and $-1.5$, which are given in (92). Note that we can further improve the convergence performance by choosing relatively larger gains of $\beta_{\chi 2}$, $\beta_{\chi x 2}$, and $\beta_{\chi w 2}$, which are proportional to the convergence rate as shown in *Lemmas 4* and *5*.

Using the computed system dynamics $A$, $B$, $D$, $\hat{E}$, $\hat{C}$, and $\hat{F}$, we solve the output regulator equations, and their solutions are presented in Figs. 7 and 8. In our example, the solution to output regulator, driven by (91)–(93) is unique. Therefore, we use the prior knowledge of the system dynamics to calculate the actual solution as

$$X = \begin{bmatrix} 0.6 & 0.8 \\ -1.2 & 0.6 \\ 0 & 1 \\ 1.5 & 1.5 \end{bmatrix}, \quad U = [0.9 \quad 1.2]. \tag{96}$$

Comparing the actual solution (96) and our computed one in Figs. 7 and 8, it is seen that our data-driven adaptive control protocol finds the solution to output regulator equations accurately. Note that a prior knowledge of output dynamics $C$ and $F$ is required in [45], while such an requirement is now removed in our control protocol.

Considering the optimal control gains obtained in Fig. 3 and the solutions to output regulator equations shown in Figs. 7 and 8, we now implement the proposed adaptive optimal controller (81) to the system dynamics (91). The trajectories of the output tracking are shown in Fig. 9, based on which the tracking error is computed in Fig. 10. The optimal gain $K^*$ is updated at time 2 second, which is now marked as a dashed black line in Figs. 9 and 10. From Figs. 9 and 10, one can conclude that the proposed adaptive optimal controller ensures the system output converge to the desired trajectory.

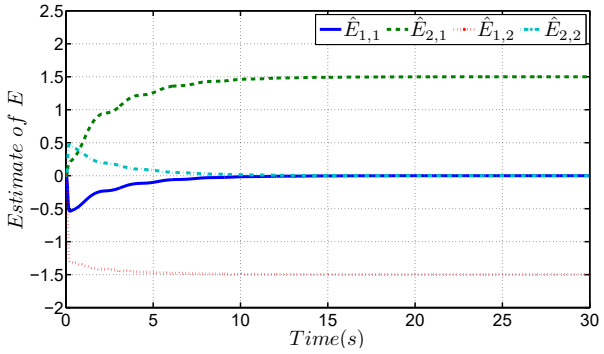



Fig. 4. Evolution of $E$ when obtaining the system dynamics

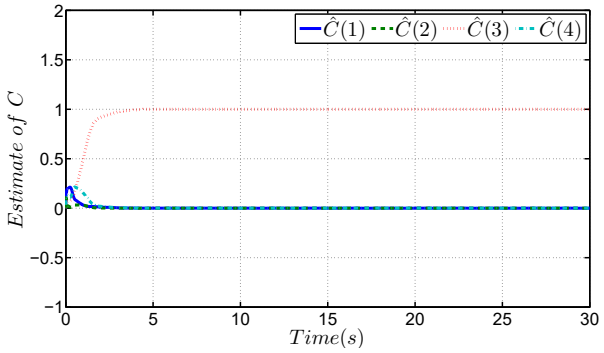



Fig. 5. Evolution of $C$ when obtaining the system dynamics

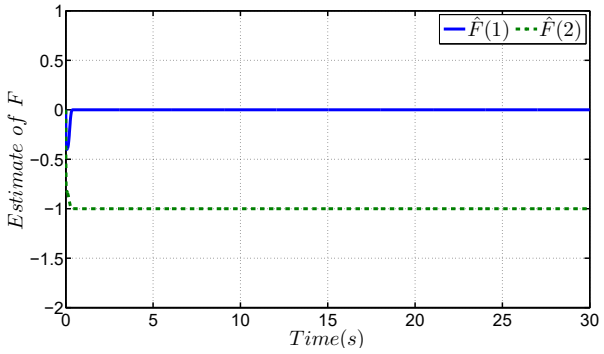



Fig. 6. Evolution of $F$ when obtaining the system dynamics

In what follows, we further apply our method to track a desired trajectory with unstable system dynamics. In our case, the reference is generated by replacing (92) with a new $E$ given as

$$E = \begin{bmatrix} 0.1 & 0.1 \\ -0.1 & 0.1 \end{bmatrix}, \tag{97}$$

while the other dynamics are the same as those in (91)–(93). From (97), it is clear that the eigenvalues of $E$ are $0.1 \pm 0.1i$, where $i$ is the imaginary unit. To be in line with the previous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



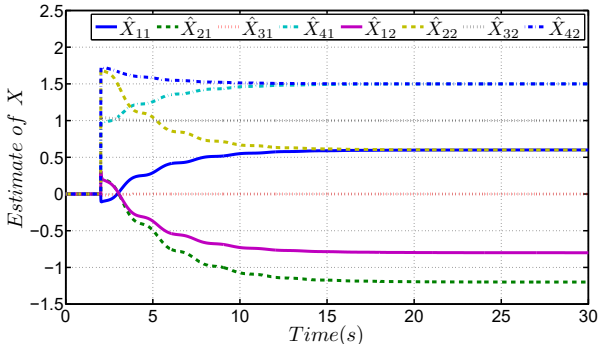



Fig. 7. Evolution of $X$ when obtaining the output regulator solution

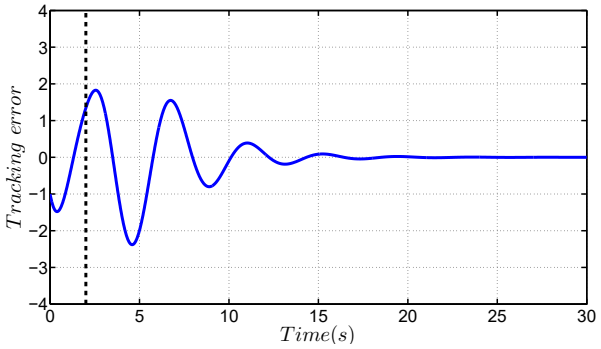



Fig. 10. Output tracking error: $e$

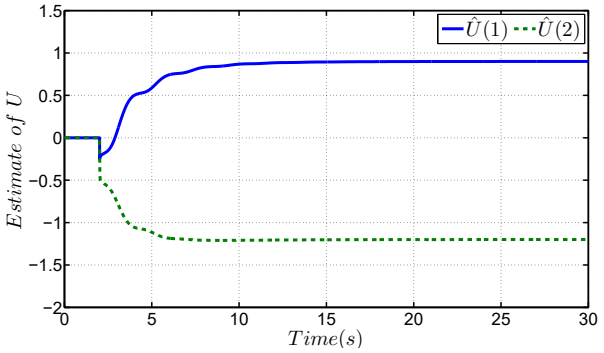



Fig. 8. Evolution of $U$ when obtaining the output regulator solution

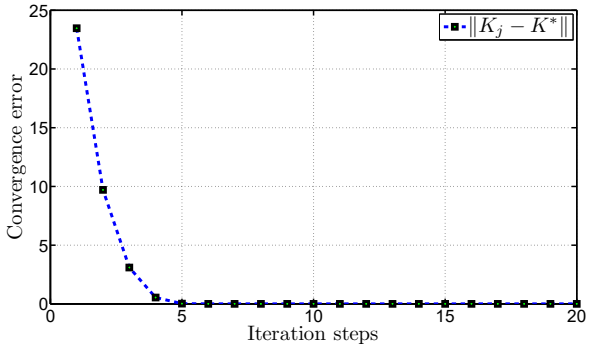



Fig. 11. Parameter evolution to obtain the optimal gain: $K^*$

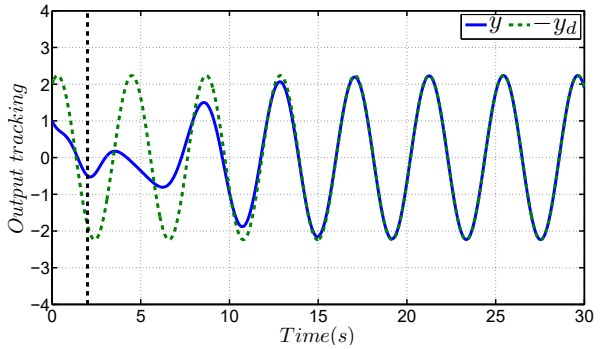



Fig. 9. Output tracking performance: $y$ and $-y_d$

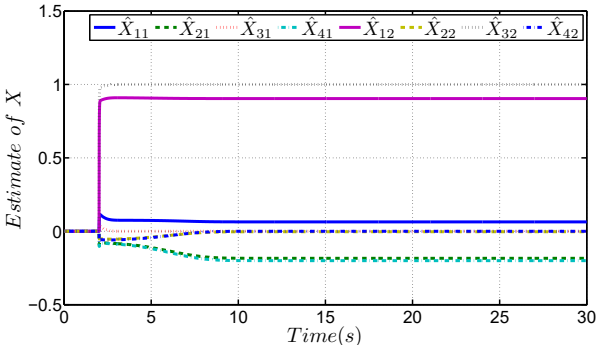



Fig. 12. Evolution of $X$ when obtaining the output regulator solution

simulation, we use two seconds to collect the system data. Considering the same system dynamics $A$, $B$, and the same design gains $Q$, $R$, the optimal gains $P^*$ and $K^*$ are the same as (94) and (95). For completeness, we show the approximated optimal gain $K^*$ in Fig. 11. Again, following Section III, we plot the evolutions of solutions to output regulator equations in Figs. 12 and 13, from which it is clear that the estimated solutions $\hat{X}$ and $\hat{U}$, respectively, converge to the desired values $X = \begin{bmatrix} 0.064 & 0.904 \\ -0.184 & -0.0032 \\ 0 & 1 \\ -0.2 & 0 \end{bmatrix}$ and $U = [0.24192 \quad 1.14128]$. After applying the above mentioned results into (91)–(93) with the unstable dynamics $E$ in (97), we present the output tracking performance and its tracking error in Figs. 14 and 15, which further validate the effectiveness of the proposed adaptive optimal control algorithm.

In what follows, a comparison study between the proposed algorithm and that in [32] is presented. For the simulation, we choose the system dynamics to be the same as (91)–(93), while setting the matrix $D$ in (91) to be zero. By doing this, the considered system dynamics are within the range of consideration of [32]. Referring to [32], the optimal controller is defined as

$$u = K^1 X, \tag{98}$$

where $K^1 = -R^{-1}B_1^T P$ and $P$ satisfies the following ARE

$$T^T P + PT - \gamma P - PB_1 R_1^{-1} B_1^T P + C_1^T Q_1 C_1 = 0$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



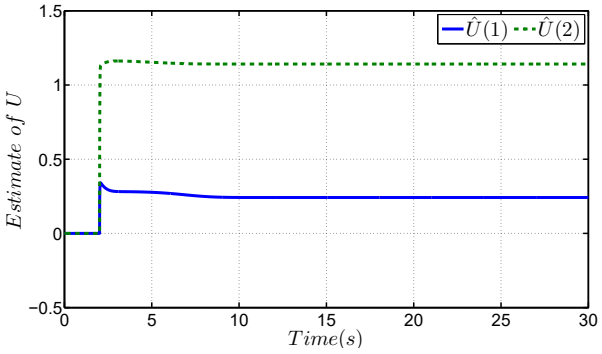



Fig. 13. Evolution of $U$ when obtaining the output regulator solution

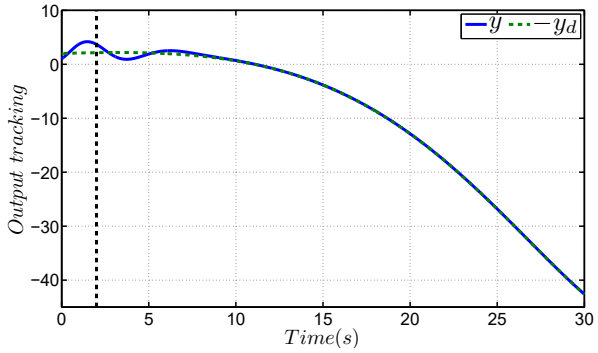



Fig. 14. Output tracking performance: $y$ and $-y_d$

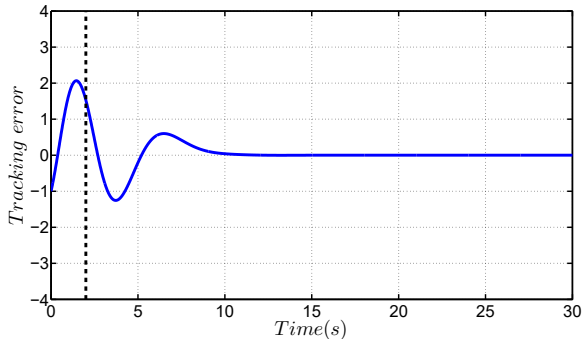



Fig. 15. Output tracking error: $e$

with the augmented system dynamics defined as $T = \begin{bmatrix} A & O \\ O & E \end{bmatrix}$, $B_1 = \begin{bmatrix} B \\ O \end{bmatrix}$, and $C_1 = \begin{bmatrix} C, & F \end{bmatrix}$. For simplification, we assume that the off-policy iteration algorithm in [32] works well to converge to the optimal gain. Therefore, we directly use the system dynamics to calculate the optimal gain as

$$(K^1)^T = \begin{bmatrix} 1.141402234042 \\ 0.951144060015 \\ -0.470202952770 \\ 1.219112974792 \\ -1.844442579921 \\ 1.759245154208 \end{bmatrix}^T \tag{99}$$

where parameters are set as $\gamma = 0.1$, $R_1 = 1$, and $Q_1 = 6$.

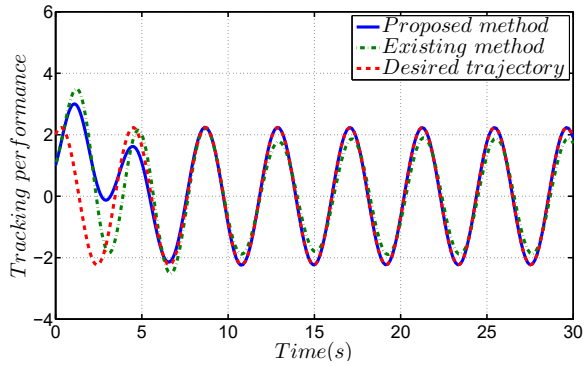



Fig. 16. Output tracking performance

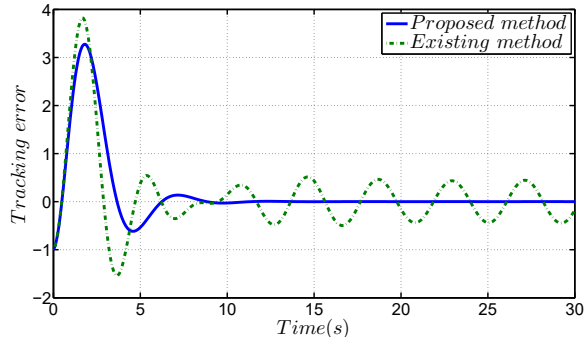



Fig. 17. Output tracking error: $e$

Applying the existing optimal controller in (98) with the gain in (99) to (91)–(93), we obtain the tracking performance and its tracking errors in Figs. 16 and 17. For comparison, we use the proposed algorithm to obtain the optimal controller, and apply it to the same simulated system. The corresponding tracking performance and its tracking error in presence of our algorithm are also plotted in Figs. 16 and 17. It can be seen that, compared to the existing work [32], our algorithm is more effective in tracking the reference trajectory.

## VI. Conclusion

This paper addresses the adaptive exponentially optimal control of a class linear systems with unknown system dynamics. We use IRL off-policy control method to obtain the optimal gains, and to explicitly identify the involved system dynamics using the same data. Moreover, experience replay method is employ to identify the exosystem dynamics. Finally, we ensure the output regulators are online accurately learned based on the estimated system dynamics. A unified adaptive optimal control protocol is proposed to force the controlled system output exponentially converge to a predefined reference. The proposed method uses the linear output regulation theory, based on which our result is limited to control a linear system. Extension of our result to control a nonlinear system will be considered in future. Moreover, implementation of the propose approach to control of real-world systems and its experimental validation is a future work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



## APPENDIX A
## PROOF OF THE FIRST PART OF *Theorem 2*

To prove the first part, we define

$$\tilde{\Psi} \triangleq \Psi - \hat{\Psi}, \tag{100}$$
$$\Psi_0 \triangleq W^T\Psi, \tag{101}$$
$$\hat{\Psi}_0 \triangleq W^T\hat{\Psi}, \tag{102}$$
$$\tilde{\Psi}_0 \triangleq \Psi_0 - \hat{\Psi}_0, \tag{103}$$

where $W \in \mathbb{R}^{(n+m)\times(n+m)}$ is a matrix satisfying

$$QW = [\bar{Q} \quad O_{(n+r)\times(n+m-k)}], \tag{104}$$

with $k$ being the rank of $Q$. Based on the singular value decomposition method, there must exist an orthogonal matrix $W$ given by (104) such that $\bar{Q}$ has full column rank. From (80), one has

$$\begin{aligned}\dot{\hat{\Psi}}_0 =& -\frac{2\|\hat{E}\|_F + 2\aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Big(W^T\hat{Q}^T(\hat{Q}\hat{\Psi} - \Phi) + W^T\hat{Q}^T\tilde{\Phi}\Big)\\ =& -\frac{2\|\hat{E}\|_F + 2\aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Big(W^TQ^T(Q\hat{\Psi} - \Phi - \tilde{Q}\hat{\Psi})\\ &- W^T\tilde{Q}^T(\hat{Q}\hat{\Psi} - \tilde{\Phi} - \hat{\Phi}) + W^T(Q^T - \tilde{Q}^T)\tilde{\Phi}\Big)\\ =& -\frac{2\|\hat{E}\|_F + 2\aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Big(-W^TQ^TQ\tilde{\Psi} - W^TQ^T\tilde{Q}\hat{\Psi}\\ &- W^T\tilde{Q}^T\hat{Q}\hat{\Psi} + W^T\tilde{Q}^T\hat{\Phi} + W^TQ^T\tilde{\Phi}\Big)\\ =& -\frac{2\|\hat{E}\|_F + 2\aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Big(-W^TQ^TQW\tilde{\Psi}_0\\ &- W^TQ^T\tilde{Q}\hat{\Psi} + W^TQ^T\tilde{\Phi}\Big),\end{aligned} \tag{105}$$

where $\tilde{Q} = Q - \hat{Q}$, and $\tilde{\Phi} = \Phi - \hat{\Phi}$.

From (104) and (105), one has

$$\begin{aligned}\dot{\hat{\Psi}}_0 =& -\frac{2\|\hat{E}\|_F + 2\aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Bigg(-\begin{bmatrix}\bar{Q}^T\bar{Q} & 0\\ 0 & 0\end{bmatrix}\tilde{\Psi}_0\\ &- W^TQ^T\tilde{Q}\hat{\Psi} + W^TQ^T\tilde{\Phi}\Bigg).\end{aligned} \tag{106}$$

If we redefine $\hat{\Psi}_0 \triangleq [\hat{\Psi}_a^T, \hat{\Psi}_b^T]^T$, (106) means $\hat{\Psi}$ is bounded for all the time, and $\hat{\Psi}_b$ converges to a certain steady vector at the exponential convergence rate of

$$\alpha_{CF} \triangleq ||\chi|| + \min\{\alpha_\chi, \alpha_{\chi CF}\}, \tag{107}$$

for $t \geq \max\{t_\chi, t_{\chi CF}\}$.

To analyze the stability of $\hat{\Psi}_a$, consider the Lyapunov function candidate

$$V_{\Psi_a} = \frac{1}{2}\tilde{\Psi}_a^T\tilde{\Psi}_a, \tag{108}$$

where $\tilde{\Psi}_a = \Psi_a - \hat{\Psi}_a$ with $\Psi_a$ being a unique solution of $\bar{Q}\Psi_a = \Phi$. Taking the time derivative of (108) along with (105) yields

$$\begin{aligned}\dot{V}_{\Psi_a} \leq& \frac{2\|\hat{E}\|_F + 2\aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Big(-\sigma_{\min}(\bar{Q}^T\bar{Q})\tilde{\Psi}_a^T\tilde{\Psi}_a\\ &+ \|\tilde{\Psi}_0\|\|W^TQ^T\|_F\|\tilde{Q}\|_F\|\hat{\Psi}\|\\ &+ \|\tilde{\Psi}_0\|\|W^TQ^T\|_F\|\tilde{\Phi}\|\Big).\end{aligned} \tag{109}$$

Considering Young's inequality, one has

$$\begin{aligned}&\|\tilde{\Psi}_0\|\|W^TQ^T\|_F(\|\tilde{Q}\|_F\|\hat{\Psi}\| + \|\tilde{\Phi}\|)\\ &\leq \frac{\|W^TQ^T\|_F^2(\|\tilde{Q}\|_F\|\hat{\Psi}\| + \|\tilde{\Phi}\|)^2}{2\sigma_{\min}(\bar{Q}^T\bar{Q})}\\ &\quad + \frac{1}{2}\sigma_{\min}(\bar{Q}^T\bar{Q})\|\tilde{\Psi}_0\|^2.\end{aligned} \tag{110}$$

Let $\sigma_k(\cdot)$ denote the $k$th largest singular, and thus one has

$$\sigma_{\min}(\bar{Q}^T\bar{Q}) = \sigma_k(Q^TQ) > 0, \tag{111}$$

where $rank(\bar{Q}) = k$ is used to ensure its positiveness. Substituting (110) and (111) in (109) yields

$$\begin{aligned}\dot{V}_{\Psi_a} \leq& \frac{\|\hat{E}\|_F + \aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}\Big(-\sigma_k(Q^TQ)||\tilde{\Psi}_a||^2\\ &+ \frac{\|W^TQ^T\|_F^2(\|\tilde{Q}\|_F\|\hat{\Psi}\| + \|\tilde{\Phi}\|)^2}{\sigma_{\min}(\bar{Q}^T\bar{Q})}\Big),\end{aligned} \tag{112}$$

where $\tilde{E}$ and $\tilde{\Phi}$ are exponentially stable as shown in the proof of Section III-C; $Q$ is a matrix with constant elements; and $\tilde{Q}$ is exponentially stable. Hence, one has

$$\begin{aligned}&\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) - \sigma_k(Q^TQ)\\ &= \sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) - \sigma^{\emptyset}_{\min}(Q^TQ).\end{aligned} \tag{113}$$

From (113), each singular value of $\hat{Q}^T\hat{Q}$ exponentially converges to that of $Q^TQ$. This means that (113) eventually approaches zero. Moreover, there exists a finite time $T_1 > 0$ such that $\sigma_k(\hat{Q}^T\hat{Q}) > 0$ if $t \geq T_1$. According to the definition in (80), $\hbar(t)$ can be zero if $t \geq T_1$. Note that $\hat{Q}$ and $\hat{E}$ are bounded using the proposed approach in Section III. Hence, there exist lower and upper bounds for $\frac{\|\hat{E}\|_F + \aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}) + \hbar(t)}$, which are defined as $c_{\min}$ and $c_{\max}$, respectively.

Therefore, based on (112), (109) is changed to

$$\begin{aligned}\dot{V}_{\Psi_a} \leq& -\frac{\|\hat{E}\|_F + \aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})}\sigma_k(Q^TQ)||\tilde{\Psi}_a||^2\\ &+ c_{\max}\frac{\|W^TQ^T\|_F^2(\|\tilde{Q}\|_F\|\hat{\Psi}\| + \|\tilde{\Phi}\|)^2}{\sigma_{\min}(\bar{Q}^T\bar{Q})},\end{aligned} \tag{114}$$

for $t \geq T_1$. Moreover, the first term at the right hand side of (114) satisfies

$$\begin{aligned}&-\frac{\|\hat{E}\|_F + \aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})}\sigma_k(Q^TQ)||\tilde{\Psi}_a||^2\\ &= -\frac{\|\hat{E}\|_F + \aleph}{\sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})}\Big(\sigma_k(Q^TQ) - \sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})\Big)||\tilde{\Psi}_a||^2\\ &\quad - (\|\hat{E}\|_F + \aleph)||\tilde{\Psi}_a||^2\\ &\leq c_{\max}|\sigma_k(Q^TQ) - \sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})|||\tilde{\Psi}_a||^2\\ &\quad - (\|E\|_F + \aleph)||\tilde{\Psi}_a||^2 - (\|\hat{E}\|_F - \|E\|_F)||\tilde{\Psi}_a||^2.\end{aligned} \tag{115}$$

Substituting (115) into (114) yields, for $t \geq T_1$,

$$\begin{aligned}\dot{V}_{\Psi_a} \leq& -(\|E\|_F + \aleph)||\tilde{\Psi}_a||^2 + (|\|\hat{E}\|_F - \|E\|_F|)||\tilde{\Psi}_a||^2\\ &+ c_{\max}|\sigma_k(Q^TQ) - \sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})|||\tilde{\Psi}_a||^2\\ &+ c_{\max}\frac{\|W^TQ^T\|_F^2(\|\tilde{Q}\|_F\|\hat{\Psi}\| + \|\tilde{\Phi}\|)^2}{\sigma_{\min}(\bar{Q}^T\bar{Q})}.\end{aligned} \tag{116}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2905215, IEEE Transactions on Automatic Control



One selects a constant $\alpha_\Psi$ satisfying $0 < \alpha_\Psi < \aleph$. Considering that $(\sigma_k(Q^TQ) - \sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q}))$ and $(\|\hat{E}\|_F - \|E\|_F)$ decay to the zero, then there exists a finite time $T_2 > 0$, such that for all $t \geq T_2$, one has

$$\aleph - \alpha_\Psi \geq d_c, \tag{117}$$

where $\aleph$ is given in (80), and

$$\begin{aligned} d_c \triangleq & c_{\max}|\sigma_k(Q^TQ) - \sigma^{\emptyset}_{\min}(\hat{Q}^T\hat{Q})| \\ & + |\|\hat{E}\|_F - \|E\|_F|. \end{aligned} \tag{118}$$

Substituting (117) into (112) yields

$$\begin{aligned} \dot{V}_{\Psi_a} \leq & - \|E\|_F||\tilde{\Psi}_a||^2 - \alpha_\Psi||\tilde{\Psi}_a||^2 \\ & + c_{\max}\frac{\|W^TQ^T\|_F^2(\|\tilde{Q}\|_F\|\hat{\Psi}\| + \|\tilde{\Phi}\|)^2}{\sigma_{\min}(\bar{Q}^T\bar{Q})}, \end{aligned} \tag{119}$$

must hold for $t \geq T_3$ with $T_3 \triangleq \max\{T_1, T_2\}$. Since both $\tilde{Q}$ and $\tilde{\Phi} = vec\begin{pmatrix} O_{n\times q_m} \\ \tilde{F} \end{pmatrix}$ exponentially converge to the zero at the convergence rate of $\alpha_{CF}$ in (107), there must exist a positive constant $V_\Psi$ such that $V_\Psi \exp(-\alpha_{CF}t)$ is an upper function of the last term at the right hand side of (119). Moreover, the time integration of last two terms at the right hand side of (119) are bounded, and its upper bound is defined as $V_\Psi$. In this sense, taking the integration of (119) over the time interval $t \geq T_3$ yields

$$||\tilde{\Psi}_a||^2 \leq - 2\int_{T_3}^{t}(\|E\|_F + \alpha_\Psi)||\tilde{\Psi}_a||^2 d\tau + \bar{V}_\Psi, \tag{120}$$

where $\bar{V}_\Psi = 2V_{\Psi_a}(T_3) + 2V_\Psi$. Considering *Bellman-Gronwall Lemma* [53], the following inequality holds for $t \geq T_3$

$$||\tilde{\Psi}_a|| \leq \sqrt{\bar{V}_\Psi}\exp\Big(-(\|E\|_F + \alpha_\Psi)t\Big). \tag{121}$$

Considering that all the elements in (105) are bounded for $\forall t \geq 0$, then one obtains that $\tilde{\Psi}_0$, as well as $\bar{V}_\Psi$ and $V_{\Psi_a}(t)$, is bounded for $0 \leq t \leq T_4$ with $T_4 \equiv \max\{T_3, t_\chi, t_{\chi CF}\}$ being a bounded constant. From (106), (107), and (121), there exists a bounded value $V_{\Psi M}$ such that

$$\begin{aligned} ||\tilde{\Psi}|| &= ||\tilde{\Psi}_0|| \\ &\leq \sqrt{\bar{V}_{\Psi M}}\exp\Big(-(\|E\|_F + \alpha^*_\Psi)t\Big), \ for\ t \geq 0, \end{aligned} \tag{122}$$

where $\bar{V}_{\Psi M} \equiv \sqrt{V_{\Psi M}}\exp\big((\|E\|_F + \alpha^*_\Psi)T_4\big)$ is bounded, and $\alpha^*_\Psi = \min\{\alpha_\Psi, \alpha_{CF}\}$. It reveals the estimated pair $(\hat{X}, \hat{U})$ converges to the actual one $(X, U)$ at the rate of $\|E\|_F + \alpha^*_\Psi$. This completes the first part proof of *Theorem 2*. ■

## ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and reviewers for the constructive comments that have contributed to improving the quality of this paper. The first author would like to thank Dr. Yi Jiang, from State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, China, for inspiring discussions on the topic during his stay at The University of Texas at Arlington.

## REFERENCES

[1] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT Press Cambridge, 1998, vol. 135.
[2] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.
[3] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal adaptive control and differential games by reinforcement learning principles*. IET, 2013, vol. 2.
[4] 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*, vol. 32, no. 6, pp. 76–105, 2012.
[5] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.
[6] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.
[7] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, 2015, pp. 1889–1897.
[8] J. Peters and S. Schaal, "Reinforcement learning by reward-weighted regression for operational space control," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 745–750.
[9] J. Peters, K. Mülling, and Y. Altun, "Relative entropy policy search." in *AAAI*. Atlanta, 2010, pp. 1607–1612.
[10] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
[11] R. Bellman, *Dynamic programming*. Courier Corporation, 2013.
[12] P. A. Ioannou and J. Sun, *Robust adaptive control*. PTR Prentice-Hall Upper Saddle River, NJ, 1996, vol. 1.
[13] M. Krstic, P. V. Kokotovic, and I. Kanellakopoulos, *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995.
[14] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE circuits and systems magazine*, vol. 9, no. 3, 2009.
[15] K. Doya, "Reinforcement learning in continuous time and space," *Neural computation*, vol. 12, no. 1, pp. 219–245, 2000.
[16] J. Murray, C. Cox, R. Saeks, and G. Lendaris, "Globally convergent approximate dynamic programming applied to an autolander," in *American Control Conference, 2001. Proceedings of the 2001*, vol. 4. IEEE, 2001, pp. 2901–2906.
[17] 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.
[18] L. M. Zhu, H. Modares, G. O. Peen, F. L. Lewis, and B. Yue, "Adaptive suboptimal output-feedback control for linear systems using integral reinforcement learning," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 1, pp. 264–273, 2015.
[19] D. Vrabie, F. Lewis, and M. Abu-Khalaf, "Biologically inspired scheme for continuous-time approximate dynamic programming," *Transactions of the Institute of Measurement and Control*, vol. 30, no. 3-4, pp. 207–223, 2008.
[20] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.
[21] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.
[22] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 25, no. 5, pp. 882–893, 2014.
[23] ——, "Global adaptive dynamic programming for continuous-time nonlinear systems," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2917–2929, 2015.
[24] 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.
[25] T. Bian, Y. Jiang, and Z.-P. Jiang, "Adaptive dynamic programming for stochastic systems with state and control dependent noise," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4170–4175, 2016.
[26] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[27] G. Chowdhary and E. Johnson, "Concurrent learning for convergence in adaptive control without persistency of excitation," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 3674–3679.

[28] ——, "A singular value maximizing data recording algorithm for concurrent learning," in *American Control Conference (ACC), 2011*. IEEE, 2011, pp. 3547–3552.

[29] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *International Journal of Adaptive Control and Signal Processing*, vol. 27, no. 4, pp. 280–301, 2013.

[30] S. K. Jha, S. B. Roy, and S. Bhasin, "Data-driven adaptive LQR for completely unknown lti systems," in *IFAC World Congress, Toulouse, France*. IFAC, 2017, pp. 4224–4229.

[31] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. John Wiley & Sons, 2017.

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

[33] C. Qin, H. Zhang, and Y. Luo, "Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming," *Int. J. Control*, vol. 87, no. 5, pp. 1000–1009, 2014.

[34] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, 2014.

[35] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive dynamic programming for control: algorithms and stability*. Springer Science & Business Media, 2012.

[36] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for $H_\infty$ control design," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 65–76, 2015.

[37] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, "Optimal control of unknown continuous-time nonaffine nonlinear systems," in *Adaptive Dynamic Programming with Applications in Optimal Control*. Springer, 2017, pp. 309–344.

[38] Y. Jiang, J. Fan, T. Chai, J. Li, and F. L. Lewis, "Data-driven flotation industrial process operational optimal control based on reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 1974–1989, May 2018.

[39] Y. Jiang, J. Fan, T. Chai, F. L. Lewis, and J. Li, "Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4607–4620, Oct 2018.

[40] 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.

[41] H. Modares, F. L. Lewis, and Z.-P. Jiang, "$H_\infty$ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 10, pp. 2550–2562, 2015.

[42] H. Modares, F. L. Lewis, W. Kang, and A. Davoudi, "Optimal synchronization of heterogeneous nonlinear systems with unknown dynamics," *IEEE Trans. Autom. Control*, vol. 63, no. 1, pp. 117–131, 2018.

[43] A. J. Krener, "The construction of optimal linear and nonlinear regulators," *Systems, Models and Feedback: Theory and Applications*, vol. 12, pp. 301–322, 1992.

[44] A. Saberi, A. A. Stoorvogel, P. Sannuti, and G. Shi, "On optimal output regulation for linear systems," *International Journal of Control*, vol. 76, no. 4, pp. 319–333, 2003.

[45] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, 2016.

[46] J. Huang, *Nonlinear output regulation: theory and applications*. SIAM, 2004, vol. 8.

[47] B. A. Francis, "The linear multivariable regulator problem," *SIAM Journal on Control and Optimization*, vol. 15, no. 3, pp. 486–505, 1977.

[48] D. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Trans. Autom. Control*, vol. 13, no. 1, pp. 114–115, 1968.

[49] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Systems*, vol. 37, no. 1, pp. 33–52, 2017.

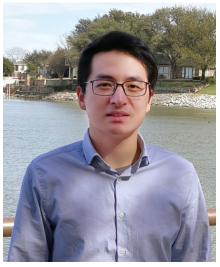
[50] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to h-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[51] A. Cichocki and R. Unbehauen, "Neural networks for solving systems of linear equations and related problems," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 39, no. 2, pp. 124–138, 1992.

[52] H. Cai, F. L. Lewis, G. Hu, and J. Huang, "The adaptive distributed observer approach to the cooperative output regulation of linear multi-agent systems," *Automatica*, vol. 75, pp. 299–305, 2017.

[53] F. L. Lewis, D. M. Dawson, and C. T. Abdallah, *Robot manipulator control: theory and practice*. CRC Press, 2003.

[54] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative control of multi-agent systems: optimal and adaptive design approaches*. Springer Science & Business Media, 2013.

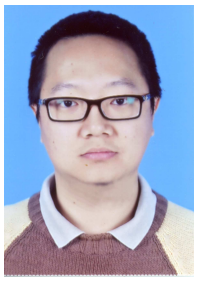



**Ci Chen** received the B.E. and Ph.D. degrees from the School of Automation, Guangdong University of Technology, Guangzhou, China, in 2011 and 2016, respectively.

He was a research assistant in School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2015 to 2016. From 2016 to 2018, he has been with The University of Texas at Arlington and The University of Tennessee at Knoxville as a Research Associate. He is now with the School of Automation, Guangdong University of Technology and also with School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include reinforcement learning, nonlinear system control, resilient control and computational intelligence. He is an Editor for International Journal of Robust and Nonlinear Control and an Associate Editor for Advanced Control for Applications: Engineering and Industrial Systems.

**Hamidreza Modares** (M'15) received the B.S. degree from the University of Tehran, Tehran, Iran, in 2004, the M.S. degree from the Shahrood University of Technology, Shahrood, Iran, in 2006, and the Ph.D. degree from The University of Texas at Arlington, Arlington, TX, USA, in 2015. He was a Senior Lecturer with the Shahrood University of Technology, from 2006 to 2009 and a Faculty Research Associate with the University of Texas at Arlington, from 2015 to 2016.

He is currently an Assistant Professor in Mechanical Engineering Department, Michigan State University, USA. His current research interests include cyber-physical systems, reinforcement learning, distributed control, robotics, and machine learning. He is an Associate Editor for the IEEE Transactions on Neural Networks and Learning Systems. He has received best paper award from 2015 IEEE International Symposium on Resilient Control Systems.

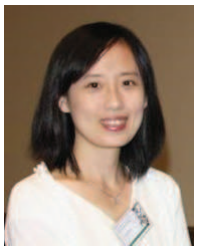


**Kan Xie** received the M.S. degree in software engineering from the South China University of Technology, Guangzhou, China, in 2009, and received the Ph.D. degree in intelligent signal and information processing at the Guangdong University of Technology, Guangzhou, China in 2016.

Currently, he is with School of Automation, Guangdong University of Technology. His research interests include machine learning, adaptive control, blind signal processing, and biomedical signal processing.

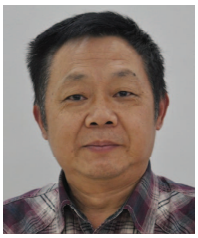
**Frank L. Lewis** (S'70–M'81–SM'86–F'94) Member, National Academy of Inventors. Fellow IEEE, Fellow IFAC, Fellow AAAS, Fellow U.K. Institute of Measurement & Control, PE Texas, U.K. Chartered Engineer. UTA Distinguished Scholar Professor, UTA Distinguished Teaching Professor, and Moncrief-O'Donnell Chair at the University of Texas at Arlington Research Institute. Qian Ren Thousand Talents Consulting Professor, Northeastern University, Shenyang, China.

He obtained the Bachelor's Degree in Physics/EE and the MSEE at Rice University, the MS in Aeronautical Engineering from Univ. W. Florida, and the Ph.D. at Ga. Tech. He works in feedback control, intelligent systems, cooperative control systems, and nonlinear systems. He is author of 7 U.S. patents, numerous journal special issues, journal papers, and 20 books, including *Optimal Control*, *Aircraft Control, Optimal Estimation*, and *Robot Manipulator Control* which are used as university textbooks worldwide. He received the Fulbright Research Award, NSF Research Initiation Grant, ASEE *Terman Award*, Int. Neural Network Soc. *Gabor Award*, U.K. Inst Measurement & Control *Honeywell Field Engineering Medal*, IEEE Computational Intelligence Society *Neural Networks Pioneer Award*, AIAA *Intelligent Systems Award*. Received Outstanding Service Award from Dallas IEEE Section, selected as Engineer of the year by Ft. Worth IEEE Section. Was listed in Ft. Worth Business Press Top 200 Leaders in Manufacturing. Texas Regents Outstanding Teaching Award 2013. He is Distinguished Visiting Professor at Nanjing University of Science & Technology and Project 111 Professor at Northeastern University in Shenyang, China. Founding Member of the Board of Governors of the Mediterranean Control Association.



**Yan Wan** (S'08-M'09-SM'17) received the B.S. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China in 2001, the M.S. degree from The University of Alabama, Tuscaloosa, AL in 2004, and the Ph.D. degree from Washington State University, Pullman, WA in 2009.

She is currently an Associate Professor with the Department of Electrical Engineering, University of Texas at Arlington. Before that, she worked as a postdoctoral scholar in the Control Systems program at the University of California at Santa Barbara, and then an Assistant and Associate Professor at the University of North Texas. Her research interest lies in decision-making tasks in large-scale networks, with applications to air traffic management, airborne networks, sensor networking, biological systems, etc. She was the recipient of the RTCA William E. Jackson Award (Excellence in aviation electronics and communication) in 2009, and the NSF CAREER Award in 2015.



**Shengli Xie** (M'00–SM'02–F'18) received the M.S. degree in mathematics from Central China Normal University, Wuhan, China, in 1992, and the Ph.D. degree in automatic control from the South China University of Technology, Guangzhou, China, in 1997.

He was a vice dean with the School of Electronics and Information Engineering, South China Univeristy of Technology, China, from 2006 to 2010. Currently, he is the Director of both the Institute of Intelligent Information Processing, and Guangdong Key Laboratory of IoT Information Technology, and also a professor with the School of Automation, Guangdong University of Technology, Guangzhou, China. He has authored or co-authored four monographs and more than 100 scientific papers published in journals and conference proceedings, and was granted more than 30 patents.