# $H_\infty$ Tracking Control of Completely Unknown Continuous-Time Systems via Off-Policy Reinforcement Learning

Hamidreza Modares, Frank L. Lewis, *Fellow, IEEE*, and Zhong-Ping Jiang, *Fellow, IEEE*

***Abstract*— This paper deals with the design of an $H_\infty$ tracking controller for nonlinear continuous-time systems with completely unknown dynamics. A general bounded $L_2$-gain tracking problem with a discounted performance function is introduced for the $H_\infty$ tracking. A tracking Hamilton–Jacobi–Isaac (HJI) equation is then developed that gives a Nash equilibrium solution to the associated min–max optimization problem. A rigorous analysis of bounded $L_2$-gain and stability of the control solution obtained by solving the tracking HJI equation is provided. An upper-bound is found for the discount factor to assure local asymptotic stability of the tracking error dynamics. An off-policy reinforcement learning algorithm is used to learn the solution to the tracking HJI equation online without requiring any knowledge of the system dynamics. Convergence of the proposed algorithm to the solution to the tracking HJI equation is shown. Simulation examples are provided to verify the effectiveness of the proposed method.**

***Index Terms*— Bounded $L_2$-gain, $H_\infty$ tracking controller, reinforcement learning (RL), tracking Hamilton–Jacobi–Isaac (HJI) equation.**

## I. Introduction

THE $H_\infty$ optimal control has been extensively used in the effort to attenuate the effect of disturbances on the system performance. The $H_\infty$ control theory has mostly concentrated on designing regulators to drive the states of the system to zero in the presence of disturbance [1]–[5]. In practice, however, it is often required to force the states or outputs of the system to track a reference trajectory. Existing solutions to the $H_\infty$ tracking problem are composed of two steps [6]–[9]. First, a feedforward control input is designed to guarantee the perfect tracking. Second, a feedback control input is designed by solving a Hamilton–Jacobi–Isaacs (HJI) equation to stabilize the tracking error dynamics. These methods are suboptimal as they ignore the cost of the feedforward control input in the performance function. Moreover, in these methods, procedures for computing the feedback and feedforward terms are based on the offline solution methods that require complete knowledge of the system dynamics.

During the last few years, reinforcement learning (RL) [10]–[13] has been extensively used to solve the optimal $H_2$ [14]–[25] and $H_\infty$ [26]–[37] regulation problems, and has been successfully applied to several real-world applications [38]–[43]. Offline iterative RL algorithms [26], [27], online synchronous RL algorithms [28]–[31], and simultaneous RL algorithms [32]–[34] were proposed to approximate the solution to the HJI equation arising in the $H_\infty$ regulation problem. These mentioned methods require complete knowledge of the system dynamics. Vrabie and Lewis [35] and Li *et al.* [36] used an integral RL (IRL) algorithm [15], [16] to learn the solution to the HJI equation for systems with unknown dynamics. Although efficient, these methods require the disturbance to be adjustable. However, this is not practical in most systems, because the disturbance is independent and cannot be specified. Luo *et al.* [37], inspired by [21] and [22], proposed an efficient off-policy RL algorithm to learn the solution to the HJI equation. In the off-policy RL algorithm, the system data, which are used to learn the HJI solution, can be generated with arbitrary policies rather than the evaluating policy. Their method does not require an adjustable disturbance input. However, it requires partial knowledge of the system dynamics.

While significant progress has been achieved by the use of RL algorithms for the design of the $H_\infty$ optimal controllers, and these algorithms are limited to the case of the regulation problem. In practice, however, it is desired to make the system to follow a reference trajectory. Therefore, the $H_\infty$ optimal tracking controllers are required. Although the RL algorithms have been recently presented for solving $H_2$ optimal tracking [44]–[50], only Liu *et al.* [51] proposed an RL solution to the $H_\infty$ tracking. However, their solution is suboptimal as the cost of the feedforward control input is ignored in the performance function, and it requires complete knowledge of the system dynamics.

In this paper, an online off-policy RL algorithm is developed to find the solution to the $H_\infty$ optimal tracking problem of

Manuscript received October 27, 2014; revised April 25, 2015 and May 17, 2015; accepted May 31, 2015. Date of publication June 24, 2015; date of current version September 16, 2015. This work was supported in part by the National Science Foundation (NSF) under Grant ECCS-1405173 and Grant IIS-1208623, in part by the Office of Naval Research, Arlington, VA, USA, under Grant N00014-13-1-0562 and Grant N000141410718, and in part by the U.S. Army Research Office under Grant W911NF-11-D-0001. The work of Z.-P. Jiang was supported by NSF under Grant ECCS-1101401 and Grant ECCS-1230040.

H. Modares is with the University of Texas at Arlington Research Institute, Fort Worth, TX 76118 USA (e-mail: modares@uta.edu).

F. L. Lewis is with the University of Texas at Arlington Research Institute, Fort Worth, TX 76118 USA, and also with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004, China (e-mail: lewis@uta.edu).

Z.-P. Jiang is with the Department of Electrical and Computer Engineering with the Polytechnic School of Engineering, New York University, NY 11201 USA (e-mail: zjiang@nyu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2441749

nonlinear completely unknown systems. It is not required that the disturbance be adjustable. An augmented system is constructed from the tracking error dynamics and the command generator dynamics, and a new discounted performance function is introduced for the $H_\infty$ optimal tracking problem. This allows developing a more general version of the $L_2$-gain, where the whole control input and the tracking error energies are weighted by an exponential discount factor in the performance function. This is in contrast to the existing methods that include only the cost of the feedback part of the control input in the performance function. A tracking HJI equation associated with the discounted performance function is derived, which gives both the feedforward and feedback parts of the control input simultaneously. Stability and $L_2$-gain boundness of the solution to the tracking HJI equation are discussed. An upper-bound is obtained for the discount factor to assure local asymptotic stability of the tracking error dynamics. An off-policy RL algorithm is then developed to find the solution to the tracking HJI equation online using only the measured data and without any knowledge about the system dynamics. Convergence of this algorithm to the solution to the tracking HJI equation is shown.

## II. $H_\infty$ Tracking Problem

In this section, a new formulation for the $H_\infty$ tracking is presented. A general $L_2$-gain condition is defined. In this $L_2$-gain condition, a discounted performance index is used, which penalizes both tracking error and control effort. A solution to this problem is presented in the subsequent sections III and IV.

Consider the affine nonlinear system defined as

$$\dot{x} = f(x) + g(x)u + k(x)d \tag{1}$$

where $x \in \mathbb{R}^n$ is the state, $u = [u_1, \ldots, u_m] \in \mathbb{R}^m$ is the control input, $d = [d_1, \ldots, d_q] \in \mathbb{R}^q$ denotes the external disturbance, $f(x) \in \mathbb{R}^n$ is the drift dynamics, $g(x) \in \mathbb{R}^{n\times m}$ is the input dynamics, and $k(x) \in \mathbb{R}^{n\times q}$ is the disturbance dynamics. It is assumed that $f(x)$, $g(x)$, and $k(x)$ are unknown Lipchitz functions with $f(0) = 0$, and that the system (1) is robustly stabilizing.

*Assumption 1:* Let $r(t)$ be the bounded reference trajectory, and assume that there exists a Lipschitz continuous command generator function $h_d(.) \in \mathbb{R}^n$ with $h_d(0) = 0$ such that

$$\dot{r}(t) = h_d(r(t)). \tag{2}$$

Define the tracking error

$$e_d \triangleq x(t) - r(t). \tag{3}$$

Using (1)–(3), the tracking error dynamics is given by

$$\dot{e}_d(t) = f(x(t)) + g(x(t))\,u(t) + k(x(t))\,d(t) - h_d(r(t)). \tag{4}$$

The fictitious performance output to be controlled is defined, such that it satisfies

$$\|z(t)\|^2 = e_d^T Q\, e_d + u^T R\, u. \tag{5}$$

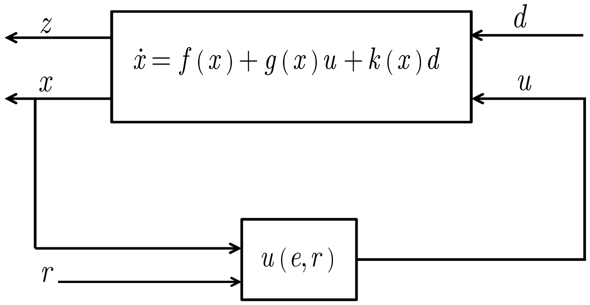


Fig. 1. State-feedback $H_\infty$ tracking control configuration.

Fig. 1 shows the system dynamics (1) and its inputs and outputs. The goal of the $H_\infty$ tracking is to attenuate the effect of the disturbance input $d$ on the performance output $z$. Before defining the $H_\infty$ tracking control problem, we define the following general $L_2$-gain or disturbance attenuation condition.

*Definition 1 (Bounded $L_2$-Gain or Disturbance Attenuation):* The nonlinear system (1) is said to have $L_2$-gain less than or equal to $\gamma$ if the following disturbance attenuation condition is satisfied for all $d \in L_2[0, \infty)$:

$$\frac{\int_t^\infty e^{-\alpha(\tau-t)}\,\|z(\tau)\|^2\,d\tau}{\int_t^\infty e^{-\alpha(\tau-t)}\,\|d(\tau)\|^2 d\tau} \leq \gamma^2 \tag{6}$$

where $\alpha > 0$ is the discount factor and $\gamma$ represents the amount of attenuation from the disturbance input $d(t)$ to the defined performance output variable $z(t)$.

*Remark 1:* The disturbance attenuation condition (6) implies that the effect of the disturbance input to the desired performance output is attenuated by a degree at least equal to $\gamma$. The minimum value of $\gamma$ for which the disturbance attenuation condition (6) is satisfied gives the so-called optimal robust control solution. However, there exists no way to find the smallest amount of the disturbance attenuation for general nonlinear systems, and a large enough value is usually predetermined for $\gamma$ [3].

Using (5) in (6), one has

$$\int_t^\infty e^{-\alpha(\tau-t)}\big(e_d^T Q e_d + u^T R u\big)d\tau \leq \gamma^2 \int_t^\infty e^{-\alpha(\tau-t)}(d^T d)d\tau. \tag{7}$$

*Definition 2 ($H_\infty$ Tracking):* The $H_\infty$ tracking control problem is to find a control policy $u = \beta(e, r)$ for some smooth function $\beta$ depending on the tracking error $e$ and the reference trajectory $r$, such that the following holds.

1) The closed-loop system $\dot{x} = f(x) + g(x)\,\beta(e, r) + k(x)\,d$ satisfies the attenuation condition (7).
2) The tracking error dynamics (4) with $d = 0$ is locally asymptotically stable.

The main difference between Definition 2 and the standard definition of $H_\infty$ tracking control problem (see [6, Definition 5.2.1]) is that a more general disturbance attenuation condition is defined here. Since the whole control input and the tracking error are penalized in the disturbance attenuation condition (7), the problem formulated in Definition 1 gives an optimal solution, in contrast to the standard definition that results in a suboptimal solution, as stated in [6].

*Remark 2:* The performance function in the left-hand side of the disturbance attenuation condition (7) represents a meaningful cost in the sense that it includes a positive penalty on the tracking error and a positive penalty on the control effort. The use of the discount factor is essential. This is because the feedforward part of the control input does not converge to zero in general, and thus, penalizing the control input in the performance function without a discount factor makes the performance function unbounded.

*Remark 3:* Previous work on the $H_\infty$ optimal tracking divides the control input into feedback and feedforward parts. First, the feedforward part is obtained separately without considering any optimality criterion. Then, the problem of optimal design of the feedback part is reduced to an $H_\infty$ optimal regulation problem. In contrast, in the new formulation, both the feedback and feedforward parts of the control input are obtained simultaneously and optimally as a result of the defined $L_2$-gain with a discount factor in (7).

The control solution to the $H_\infty$ tracking problem with the proposed attenuation condition (7) is provided in the subsequent sections III and IV. We shall see in the subsequent sections that this general disturbance attenuation condition enables us to find both the feedback and feedforward parts of the control input simultaneously, and therefore extends the method of off-policy RL for solving the problem in hand without requiring any knowledge of the system dynamics.

## III. HJI Equation for $H_\infty$ Tracking

In this section, it is first shown that the problem of solving the $H_\infty$ tracking problem can be transformed into a min–max optimization problem subject to an augmented system composed of the tracking error dynamics and the command generator dynamics. A tracking HJI equation is then developed, which gives the solution to the min–max optimization problem. The stability and $L_2$-gain boundedness of the tracking HJI control solution are discussed.

### A. Tracking HJI Equation

In this section, a tracking HJI equation is formulated, which gives the solution to the $H_\infty$ tracking problem stated in Definition 2.

Define the augmented system state

$$X(t) = [e_d(t)^T \; r(t)^T]^T \in \mathbb{R}^{2n} \tag{8}$$

where $e_d(t)$ is the tracking error defined in (3) and $r(t)$ is the reference trajectory.

Putting (2) and (4) together yields the augmented system

$$\dot{X}(t) = F(X(t)) + G(X(t))\,u(t) + K(X(t))\,d(t) \tag{9}$$

where $u(t) = u(X(t))$ and

$$F(X) = \begin{bmatrix} f(e_d + r) - h_d(r) \\ h_d(r) \end{bmatrix}, \quad G(X) = \begin{bmatrix} g(e_d + r) \\ 0 \end{bmatrix}$$
$$K(X) = \begin{bmatrix} k(e_d + r) \\ 0 \end{bmatrix}. \tag{10}$$

Using the augmented system (9), the disturbance attenuation condition (7) becomes

$$\int_t^\infty e^{-\alpha(\tau-t)}(X^T Q_T X + u^T R u)d\tau \le \gamma^2 \int_t^\infty e^{-\alpha(\tau-t)}(d^T d)d\tau \tag{11}$$

where

$$Q_T = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}. \tag{12}$$

Based on (11), define the performance function

$$J(u,d) = \int_t^\infty e^{-\alpha(\tau-t)}(X^T Q_T\, X + u^T R\, u - \gamma^2 d^T d)\, d\tau. \tag{13}$$

*Remark 4:* Note that the problem of finding a control policy that satisfies bounded $L_2$-gain condition for the optimal tracking problem is equivalent to minimizing the discounted performance function (13) subject to the augmented system (9).

It is well-known that the $H_\infty$ control problem is closely related to the two-player zero-sum differential game theory [5]. In fact, solvability of the $H_\infty$ control problem is equivalent to solvability of the following zero-sum game [5]:

$$V^*(X(t)) = J(u^*, d^*) = \min_u \max_d \; J(u,d) \tag{14}$$

where $J$ is defined in (13) and $V^*(X(t))$ is defined as the optimal value function. This two-player zero-sum game control problem has a unique solution if a game theoretic saddle point exists, i.e., if the following Nash condition holds:

$$V^*(X(t)) = \min_u \max_d \; J(u,d) = \max_d \min_u \; J(u,d). \tag{15}$$

Note that differentiating (13) and noting that $V(X(t)) = J(u(t), d(t))$ give the following Bellman equation:

$$\begin{aligned} H(V,u,d) \triangleq\; & X^T Q_T X + u^T R\, u - \gamma^2 d^T d - \alpha V \\ & + V_X^T (F + G\, u + K\, d) = 0 \end{aligned} \tag{16}$$

where $F(X) \triangleq F$, $G \triangleq G(X)$, $K \triangleq K(X)$, and $V_X = \partial V / \partial X$. Applying stationarity conditions $\partial H(V^*, u, d)/\partial u = 0$ and $\partial H(V^*, u, d)/\partial d = 0$ [52] give the optimal control and disturbance inputs as

$$u^* = -\frac{1}{2} R^{-1} G^T \; V_X^* \tag{17}$$

$$d^* = \frac{1}{2\gamma^2} K^T \; V_X^* \tag{18}$$

where $V^*$ is the optimal value function defined in (14). Substituting the control input $u$ (17) and the disturbance $d$ (18) into (16), the following tracking HJI equation is obtained:

$$\begin{aligned} H(V^*, u^*, d^*) \triangleq\; & X^T Q_T X + V_X^{*T} F - \alpha V_X \\ & - \frac{1}{4} V_X^{*T} G^T R^{-1} G\, V_X^* \\ & + \frac{1}{4\gamma^2} V_X^{*T} K K^T V_X^* = 0. \end{aligned} \tag{19}$$

In the following, it is shown that the control solution (17), which is found by solving the HJI equation (19), solves the $H_\infty$ tracking problem formulated in Definition 2.

### *B. Disturbance Attenuation and Stability of the Solution to the Tracking HJI Equation*

In this section, first, it is shown that the control solution (17) satisfies the disturbance attenuation condition (11) [part 1) of Definition 2]. Then, the stability of the tracking error dynamics (4) without the disturbance is discussed [part 2) of Definition 2]. It is shown that there exists an upper bound $\alpha^*$ such that if the discount factor is $<\alpha^*$, the control solution (17) makes the system locally asymptotically stable.

*Theorem 1 (Saddle Point Solution):* Consider the $H_\infty$ tracking control problem as a two-player zero-sum game problem with the performance function (13). Then, the pair of strategies $(u^*, d^*)$ defined in (17) and (18) provides a saddle point solution to the game.

*Proof:* See [26] for the same proof. □

*Theorem 2 ($L_2$-Gain of System for the Solution to the HJI Equation):* Assume that there exists a continuous positive-semidefinite solution $V^*(X)$ to the tracking HJI equation (19). Then, $u^*$ in (17) makes the closed-loop system (9) to have $L_2$-gain less than or equal to $\gamma$.

*Proof:* The Hamiltonian (16) for the optimal value function $V^*$, and any control policy $u$ and disturbance policy $w$ become

$$\begin{aligned}H(V^*, u, d) = {}& X^T Q_T X + u^T R\, u - \gamma^2 d^T d - \alpha V^* \\ &+ V_X^{*T}(F + G\, u + K\, d).\end{aligned} \tag{20}$$

On the other hand, using (17)–(19), one has

$$\begin{aligned}H(V^*, u, d) = {}& H(V^*, u^*, d^*) + (u - u^*)^T R\,(u - u^*) \\ &+ \gamma^2 (d - d^*)^T (d - d^*).\end{aligned} \tag{21}$$

Based on the HJI equation (19), we have $H(V^*, u^*, d^*) = 0$. Therefore, (20) and (21) give

$$\begin{aligned}&X^T Q_T X + u^T R u - \gamma^2 d^T d - \alpha V^* + V_X^{*T}(F + G\ u + K d) \\ &\quad = -(u - u^*)^T R\,(u - u^*) - \gamma^2 (d - d^*)^T (d - d^*).\end{aligned} \tag{22}$$

Substituting the optimal control policy $u = u^*$ in the above equation yields

$$\begin{aligned}&X^T Q_T X + u^{*T} R u^* - \gamma^2 d^T d - \alpha V^* \\ &\quad + V_X^{*T}(F + G u^* + K d) = -\gamma^2 (d - d^*)^T (d - d^*) \le 0.\end{aligned} \tag{23}$$

Multiplying both the sides of this equation by $e^{-\alpha t}$, and defining $\dot V^* = V_X^{*T}(F + G\, u^* + K\, d)$ as the derivative of $V^*$ along the trajectories of the closed-loop system, it gives

$$\frac{d}{dt}(e^{-\alpha t} V^*(X)) \le e^{-\alpha t}(-X^T Q_T\, X - u^{*T} R\, u^* + \gamma^2 d^T d). \tag{24}$$

Integrating from both the sides of this equation yields

$$\begin{aligned}&e^{-\alpha T} V^*(X(T)) - V^*(X(0)) \\ &\quad \le \int_0^T e^{-\alpha\tau}(-X^T Q_T X - u^{*T} R u^* + \gamma^2 d^T d)\, d\tau\end{aligned} \tag{25}$$

for every $T > 0$ and every $d \in L_2[0, \infty)$. Since $V^*(.) \ge 0$ the above equation yields

$$\begin{aligned}&\int_0^T e^{-\alpha\tau}(X^T Q_T X + u^{*T} R u^*)\, d\tau \\ &\quad \le \int_0^T e^{-\alpha\tau}(\gamma^2 d^T d)\, d\tau + V^*(X(0)).\end{aligned} \tag{26}$$

This completes the proof. □

Theorem 2 solves part 1) of the state-feedback $H_\infty$ tracking control problem given in Definition 2. In the following, we consider the problem of stability of the closed-loop system without disturbance, which is part 2) of Definition 2.

*Theorem 3 (Stability of the Optimal Solution for $\alpha \to 0$):* Suppose that $V^*(X)$ is a smooth positive-semidefinite and locally quadratic solution to the tracking HJI equation. Then, the control input given by (17) makes the error dynamics (4) with $d = 0$ asymptotically stable in the limit as the discount factor goes to zero.

*Proof:* Differentiating $V^*$ along the trajectories of the closed-loop system with $d = 0$, and using the tracking HJI equation give

$$V_X^{*T}(F + G\, u^*) = \alpha V^* - X^T Q_T\, X - u^{*T} R\, u^* + \gamma^2 d^T d \tag{27}$$

or equivalently

$$\frac{d}{dt}(e^{-\alpha t} V^*(X)) = e^{-\alpha t}(-X^T Q_T X - u^{*T} R u^* + \gamma^2 d^T d) \le 0. \tag{28}$$

If the discount factor goes to zero, then LaSalle's extension can be used to show that the tracking error is locally asymptotically stable. More specifically, if $\alpha \to 0$, based on LaSalle's extension, $X(t) = [e_d(t)^T\ r(t)^T]^T$ goes to a region wherein $\dot V = 0$. Since $X^T Q_T X = e_d(t)^T Q\, e_d(t)$, where $Q$ is the positive definite, $\dot V = 0$ only if $e_d(t) = 0$, and $u = 0$ when $d = 0$. On the other hand, $u = 0$ also requires that $e_d(t) = 0$; therefore, for $\gamma = 0$, the tracking error is locally asymptotically stable. □

Theorem 3 shows that if the discount factor goes to zero, then optimal control solution found by solving the tracking HJI equation makes the system locally asymptotically stable. However, if the discount factor is nonzero, the local asymptotic stability of the optimal control solution cannot be guaranteed by Theorem 3. In Theorem 4, it is shown that the local asymptotic stability of the optimal solution is guaranteed as long as the discount factor is smaller than an upper bound. Before presenting the proof of local asymptotic stability, the following example shows that if the discount factor is not small, the control solution obtained by solving the tracking HJI equation can make the system unstable.

*Example 1:* Consider the scalar dynamical system

$$\dot X = X + u + d. \tag{29}$$

Assume that in the HJI equation (19), we have $Q_T = R = 1$ and the attenuation level is $\gamma = 1$. For this linear system with quadratic performance, the value function is quadratic. That is, $V(X) = p\, X^2$, and therefore, the HJI equation reduces to

$$(2 - \alpha)\, p - \frac{3}{4} p^2 + 1 = 0 \tag{30}$$

and the optimal control solution becomes

$$u = -pX. \tag{31}$$

Solving this equation gives the optimal solution as

$$u = -\left(\frac{4}{3}(1-0.5\alpha) + \frac{2}{\sqrt{3}}\sqrt{\frac{4}{3}(1-0.5\alpha)^2+1}\right)X. \tag{32}$$

However, this optimal solution does not make the system stable for all values of the discount factor $\alpha$. If fact, if $\alpha > \alpha^* = 27/12$, then the system is unstable. The next theorem shows how to find an upper bound $\alpha^*$ for the discount factor to assure the stability of the system without disturbance.

Before presenting the stability theorem, note that the augmented system dynamics (9) can be written as

$$\dot{X} = F(X) + G(X)u + K(X)d = AX + Bu + Dd + \bar{F}(X) \tag{33}$$

where $AX + Bu + Kd$ is the linearized model with

$$A = \begin{bmatrix} A_{l1} & A_{l1} - A_{l2} \\ 0 & A_{l2} \end{bmatrix}, \quad B = \begin{bmatrix} B_l^T & 0^T \end{bmatrix}^T, \quad D = \begin{bmatrix} D_l^T & 0^T \end{bmatrix}^T \tag{34}$$

where $A_{l1}$ and $A_{l2}$ are the linearized models of the drift system dynamics $f$ and the command generator dynamics $h_d$, respectively, and $\bar{F}(X)$ is the remaining nonlinear term.

*Theorem 4 (Stability of the Optimal Solution and Upper Bound for $\alpha$):* Consider the system (9). Define

$$L_l = B_l R^{-1} B_l^T + \frac{1}{\gamma^2} D_l D_l^T \tag{35}$$

where $B_l$ and $D_l$ are defined in (34). Then, the control solution (17) makes the error system (4) with $d = 0$ locally asymptotically stable if

$$\alpha \leq \alpha^* = 2\|(L_l Q)^{1/2}\|. \tag{36}$$

*Proof:* Given the augmented dynamics (9) and the performance function (13), the Hamiltonian function in terms of the optimal control and disturbance is defined as [52]

$$\begin{aligned} H(\rho, u^*, d^*) = e^{-\alpha t}(X^T Q_T X + u^{*T} R u^* - \gamma^2 d^{*T} d^*) \\ + \rho^T (F + G\, u^* + Kd^*) \end{aligned} \tag{37}$$

where $\rho$ is known as the costate variable. Using Pontryagin's maximum principle, the optimal solutions $u^*$ and $d^*$ satisfy the following state and costate equations:

$$\dot{X} = H_\rho(X, \rho) \tag{38}$$

$$\dot{\rho} = -H_X(X, \rho). \tag{39}$$

Define the new variable

$$\mu = e^{\alpha t}\rho. \tag{40}$$

Based on (40), define the modified Hamiltonian function as

$$\begin{aligned} H^m = e^{-\alpha t}H = (X^T Q_T X + u^{*T} R u^* - \gamma^2 d^{*T} d^*) \\ + \mu^T (F + Gu^* + Kd^*). \end{aligned} \tag{41}$$

Then, conditions (38) and (39) become

$$\dot{X} = H_\mu^m(X, \mu) \tag{42}$$

$$\dot{\mu} = \alpha\, \mu - H_X^m(X, \mu). \tag{43}$$

Equation (42) gives the augmented system dynamics (9), and (43) is equivalent to the HJI equation (19) with $\mu = V_X^*$. In order to prove the local stability of the closed-loop system, the stability of the closed-loop linearized system is investigated. Using (33) for the system dynamics, (41) becomes

$$\begin{aligned} H^m = (X^T Q_T X + u^{*T} R u^* - \gamma^2 d^{*T} d^*) \\ + \mu^T (AX + Bu^* + Dd^* + \bar{F}(X)). \end{aligned} \tag{44}$$

Then, the costate can be written as the sum of a linear and a nonlinear term as

$$\mu = 2PX + \varphi_0(X) \equiv \mu_1 + \varphi_0(X). \tag{45}$$

Using $\partial H^m/\partial u = 0, \partial H^m/\partial d = 0$ and (45), one has

$$u^* = -R^{-1} B^T PX + \varphi_1(X) \tag{46}$$

$$d^* = \frac{1}{\gamma^2} D^T PX + \varphi_2(X) \tag{47}$$

for some $\varphi_1(X)$ and $\varphi_2(X)$ depending on $\varphi_0(X)$, $\bar{F}(X)$, and $P$. Using (44)–(47), conditions (42) and (43) become

$$\begin{aligned} \begin{bmatrix} \dot{X} \\ \dot{\mu}_1 \end{bmatrix} = \begin{bmatrix} A & -\left(B R^{-1} B^T - \frac{1}{\gamma^2} D D^T\right) \\ -Q_T & -A^T + \alpha I \end{bmatrix} \begin{bmatrix} X \\ \mu_1 \end{bmatrix} \\ + \begin{bmatrix} F_1(X) \\ F_2(X) \end{bmatrix} \triangleq W \begin{bmatrix} X \\ \mu_1 \end{bmatrix} + \begin{bmatrix} F_1(X) \\ F_2(X) \end{bmatrix} \end{aligned} \tag{48}$$

for some nonlinear functions $F_1(X)$ and $F_2(X)$. The linear part of costate is a stable manifold of $W$, and thus, based on the linear part of (48), it satisfies the following game algebraic Riccati equation (GARE):

$$Q_T + A^T P + PA - \alpha P - PBR^{-1}B^T P + \frac{1}{\gamma^2} PDD^T P = 0. \tag{49}$$

Define

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{bmatrix}.$$

Then, based on (12) and (34), the upper left-hand side of the GARE (49) becomes

$$\begin{aligned} Q + A_{l1}^T P_{11} + P_{11} A_{l1} - \alpha P_{11} - P_{11} B_l R^{-1} B_l^T\, P_{11} \\ + \frac{1}{\gamma^2} P_{11} D_l D_l^T\, P_{11} = 0. \end{aligned} \tag{50}$$

The closed-loop system dynamics for the control input (46) without the disturbance is

$$\dot{X} = (A - BR^{-1}B^T P)X + F_f(X) \tag{51}$$

for some nonlinear function $F_f(X)$ with $F_f = [F_{f1}^T, F_{f2}^T]^T$, which gives the following tracking error dynamics:

$$\dot{e}_d = \left(A_{l1} - B_l R^{-1} B_l^T P_{11}\right) e_d + F_{f1} = A_c\, e_d + F_{f1}. \tag{52}$$

The GARE (50) based on the closed-loop error dynamics $A_c$ becomes

$$\begin{aligned} Q + A_c^T P_{11} + P_{11} A_c - \alpha P_{11} + P_{11} B_l R^{-1} B_l^T\, P_{11} \\ + \frac{1}{\gamma^2} P_{11} D_l D_l^T\, P_{11} = 0. \end{aligned} \tag{53}$$

To find a condition on the discount factor to assure stability of the linearized error dynamics, assume that $\lambda$ is an eigenvalue of the closed-loop error dynamics $A_c$. That is, $A_c x = \lambda x$ with $x$, the eigenvector corresponding to $\lambda$. Then, multiplying the left- and right-hand sides of the GARE (53) by $x^T$ and $x$, respectively, one has

$$2\,(\text{Re}(\lambda) - 0.5\alpha)x^T P_{11} x = -x^T Q\, x - x^T P_{11}\big(B_l R^{-1} B_l^T + DD^T\big)P_{11} x. \quad (54)$$

Using the inequality $a^2 + b^2 \geq 2ab$ and since $P_{11} > 0$, (54) becomes

$$(\text{Re}(\lambda) - 0.5\alpha) \;\leq -\big\|\big(QP_{11}^{-1}\big)^{1/2}\big\|\,\big\|(L_l P_{11})^{1/2}\big\| \quad (55)$$

or equivalently

$$\text{Re}(\lambda) \leq -\big\|\big(\text{QP}_{11}^{-1}\big)^{1/2}\big\|\,\|(L_l P_{11})^{1/2}\| + 0.5\alpha \quad (56)$$

where $L_l$ is defined in (35). Using the fact that $\|A\|\|B\| \geq \|AB\|$ gives

$$\text{Re}(\lambda) \;\leq -\|(\text{LQ})^{1/2}\| + 0.5\alpha. \quad (57)$$

Therefore, the linear error dynamics in (52) is stable if condition (36) is satisfied, and this completes the proof. □

*Remark 5:* Note that the GARE (49) can be written as

$$Q_T + (A - 0.5\alpha I)^T P + P(A - 0.5\alpha I) - PBR^{-1}B^T P + \frac{1}{\gamma^2} PDD^T P = 0.$$

This amounts to a GARE without the discount factor and with the system dynamics given by $A - 0.5\alpha I$, $B$, and $D$. Therefore, existence of a unique solution to the GARE (30) requires $(A-0.5\alpha I, B)$ be stabilizable. Based on the definition of $A$ and $B$ in (34), this requires that $(A_{l1} - 0.5\alpha I, B_l)$ be stabilizable and $(A_{l2} - 0.5\alpha I)$ is stable. However, since $(A_{l1}, B_l)$ be stabilizable, as the system dynamics in (1) is assumed robustly stabilizing, then $(A_{l1} - 0.5\alpha I, B_l)$ is also stabilizable for any $\alpha > 0$. Moreover, since the reference trajectory is assumed bounded, the linearized model of the command generator dynamics in (2), i.e., $A_{l2}$, is marginally stable, and thus, $(A_{l2} - 0.5\alpha I)$ is stable. Therefore, the discount factor does not affect the existence of the solution to the GARE.

*Remark 6:* Theorem 4 shows that the asymptotic stability of only the first $n$ variables of $X$ is guaranteed, which are the error dynamic states. This is reasonable as the last $n$ variables of $X$ are the reference command generator variables, which are not under our control.

*Remark 7:* For Example 1, condition (35) gives the bound $\alpha < \sqrt{80}/12$ to assure the stability. This bound is very close to the actual bound obtained in Example 1. However, it is obvious that condition (35) gives a conservative bound for the discount factor to assure the stability.

*Remark 8:* Theorem 4 confirms the existence of an upper bound for the discount factor to assure the stability of the solution to the HJI tracking equation, and relates this bound to the input and disturbance dynamics, and the weighting matrices in the performance function. Condition (36) is not a restrictive condition even if the system dynamics are unknown. In fact, one can always pick a very small discount factor, and/or a large weighting matrix $Q$ (which is a design matrix) to assure that condition (36) is satisfied.

**Algorithm 1** Offline RL Algorithm

```
Initialization: Start with an admissible stabilizing control policy $u_0$
1) For a control input $u_i$ and a disturbance policy $d_i$, find $V_i$ using the following Bellman equation
     $H(V_i, u_i, d_i) = X^T Q_T X + V_{Xi}^T (F + G u_i + K d_i) - \alpha V_i + u_i^T R u_i - \gamma^2 d_i^T d_i = 0$   (58)
2) Update the disturbance using
     $d_{i+1} = \arg\max_d [H(V_i, u_i, d)] = \frac{1}{2\gamma^2} K^T V_{Xi}$   (59)
   and the control policy using
     $u_{i+1} = \arg\min_u [H(V_i, u, d)] = -\frac{1}{2} R^{-1} G^T V_{Xi}$   (60)
3) Go to 1.
```

## IV. Off-Policy IRL for Learning the Tracking HJI Equation

In this section, an offline RL algorithm is first given to solve the problem of $H_\infty$ optimal tracking by learning the solution to the tracking HJI equation. An off-policy IRL algorithm is then developed to learn the solution to the HJI equation online and without requiring any knowledge of the system dynamics. Three neural networks (NNs) on an actor–critic–disturbance structure are used to implement the proposed off-policy IRL algorithm.

### *A. Off-Policy Reinforcement Learning Algorithm*

The Bellman equation (16) is linear in the cost function $V$, while the HJI equation (19) is nonlinear in the value function $V^*$. Therefore, solving the Bellman equation for $V$ is easier than solving the HJI for $V^*$. Instead of directly solving for $V^*$, a policy iteration (PI) algorithm iterates on both the control and disturbance players to break the HJI equation into a sequence of differential equations linear in the cost. An offline PI algorithm for solving the $H_\infty$ optimal tracking problem is given in Algorithm 1.

Algorithm 1 extends the results of the simultaneous RL algorithm in [33] to the tracking problem. The convergence of this algorithm to the minimal nonnegative solution of the HJI equation was shown in [33]. In fact, similar to [33], the convergence of Algorithm 1 can be established by proving that iteration on (58) is essentially Newton's iterative sequence, which converges to the unique solution of the HJI equation (19).

Algorithm 1 requires complete knowledge of the system dynamics. In the following, the off-policy IRL algorithm, which was presented in [21] and [22] for solving the $H_2$ optimal regulation problem, is extended here to solve the $H_\infty$ optimal tracking for systems with completely unknown dynamics. To this end, the system dynamics (9) is first written as

$$\dot{X} = F + G\,u_i + K\,d_i + G\,(u - u_i) + K\,(d - d_i) \tag{61}$$

where $u_i = [u_{i,1}, \ldots, u_{i,m}] \in \mathbb{R}^m$ and $d_i = [d_{i,1}, \ldots, d_{i,q}] \in \mathbb{R}^q$ are policies to be updated. Differentiating $V_i(X)$ along with the system dynamics (61) and using (58)–(60) give

$$\begin{aligned}\dot{V}_i &= V_{Xi}^T(F + G\,u_i + K\,d_i) + V_{Xi}^T G(u - u_i) + V_{Xi}^T K\,(d - d_i)\\ &= \alpha\,V_i - X^T Q_T\,X - u_i^T R\,u_i + \gamma^2 d_i^T d_i - 2\,u_{i+1}^T R\,(u - u_i)\\ &\quad + 2\gamma^2 d_{i+1}^T(d - d_i).\end{aligned} \tag{62}$$

Multiplying both the sides of (62) by $e^{-\alpha(\tau - t)}$ and integrating from both the sides yield the following off-policy IRL Bellman equation:

$$\begin{aligned}&e^{-\alpha T} V_i(X(t+T)) - V_i(X(t))\\ &= \int_t^{t+T} e^{-\alpha(\tau - t)}\big(-X^T Q_T\,X - u_i^T R\,u_i + \gamma^2 d_i^T d_i\big)\,d\tau\\ &\quad + \int_t^{t+T} e^{-\alpha(\tau - t)}\big(-2\,u_{i+1}^T R\,(u - u_i) + 2\gamma^2 d_{i+1}^T(d - d_i)\big)\,d\tau.\end{aligned} \tag{63}$$

Note that for a fixed control policy $u$ (the policy that is applied to the system) and a given disturbance $d$ (the actual disturbance that is applied to the system), (63) can be solved for both the value function $V_i$ and the updated policies $u_{i+1}$ and $d_{i+1}$, simultaneously.

*Lemma 1:* The off-policy IRL equation (63) gives the same solution for the value function as the Bellman equation (58), and the same updated control and disturbance policies as (59) and (60).

*Proof:* Dividing both the sides of the off-policy IRL Bellman equation (63) by $T$, and taking limit results in

$$\begin{aligned}&\lim_{T\to 0} \frac{e^{-\alpha T} V_i(X(t+T)) - V_i(X(t))}{T}\\ &+ \lim_{T\to 0} \frac{\int_t^{t+T} e^{-\alpha(\tau - t)}\big(X^T Q_T X + u_i^T R\,u_i - \gamma^2 d_i^T d_i\big)\,d\tau}{T}\\ &+ \lim_{T\to 0} \frac{\int_t^{t+T} e^{-\alpha(\tau - t)}\big(2u_{i+1}^T R(u - u_i) - 2\gamma^2 d_{i+1}^T(d - d_i)\big)d\tau}{T}\\ &= 0.\end{aligned} \tag{64}$$

By L'Hopital's rule, the first term in (64) becomes

$$\begin{aligned}&\lim_{T\to 0} \frac{e^{-\alpha T} V_i(X(t+T)) - V_i(X(t))}{T}\\ &= \lim_{T\to 0}\,[-\alpha\,e^{-\alpha\ T} V_i(X(t+T)) + e^{-\alpha\ T}\dot{V}_i(X(t+T))]\\ &= -\alpha V_i + V_{Xi}(F + Gu_i + Kd_i + G(u - u_i) + K(d - d_i))\end{aligned} \tag{65}$$

where the last term in the right-hand side is obtained using $\dot{V} = V_X\,\dot{X}$. Similarly, for the second and third terms of (64), one has

$$\lim_{T\to 0} \frac{\int_t^{t+T} e^{-\alpha(\tau - t)}\big(X^T Q_T X + u_i^T Ru_i - \gamma^2 d_i^T d_i\big)d\tau}{T} = X^T Q_T X + u_i^T R\,u_i - \gamma^2 d_i^T d_i \tag{66}$$

$$\lim_{T\to 0} \frac{\int_t^{t+T} e^{-\alpha(\tau - t)}\big(2u_{i+1}^T R(u - u_i) - 2\gamma^2 d_{i+1}^T(d - d_i)\big)d\tau}{T} = 2\,u_{i+1}^T R\,(u - u_i) - 2\gamma^2 d_{i+1}^T(d - d_i). \tag{67}$$

Substituting (65)–(67) in (64) yields

$$\begin{aligned}&-\alpha\,V_i + V_{Xi}\,(F + G\,u_i + K\,d_i + G\,(u - u_i) + K\,(d - d_i))\\ &\quad + X^T Q_T X + u_i^T R\,u_i - \gamma^2 d_i^T d_i + 2u_{i+1}^T R\,(u - u_i)\\ &\quad - 2\gamma^2 d_{i+1}^T(d - d_i) = 0.\end{aligned} \tag{68}$$

Substituting the updated policies $u_{i+1}$ and $d_{i+1}$ from (59) and (60) into (68) gives the Bellman equation (58). This completes the proof. □

*Remark 9:* In the off-policy IRL Bellman equation (63), the control input $u$, which is applied to the system, can be different from the control policy $u_i$, which is evaluated and updated. The fixed control policy $u$ should be a stable and exploring control policy. Moreover, in this off-policy IRL Bellman equation, the disturbance input $d$ is the actual external disturbance that comes from a disturbance source, and is not under our control. However, $d_i$ is the disturbance, which is evaluated and updated. One advantage of this off-policy IRL Bellman equation is that, in contrast to on-policy RL-based methods, the disturbance input, which is applied to the system does not require to be adjustable.

The following algorithm uses the off-policy tracking Bellman equation (63) to iteratively solve the HJI equation (19) without requiring any knowledge of the system dynamics. The implementation of this algorithm is discussed in Section IV-B. It is shown how the data collected from a fixed control policy $u$ are reused to evaluate many updated control policies $u_i$ sequentially until convergence to the optimal solution is achieved.

*Remark 10:* Inspired by the off-policy algorithm in [21], Algorithm 2 has two separate phases. First, a fixed initial exploratory control policy $u$ is applied and the system information is recorded over the time interval $T$. Second, without requiring any knowledge of the system dynamics, the information collected in phase 1 is repeatedly used to find a sequence of updated policies $u_i$ and $d_i$ converging to $u^*$ and $d^*$. Note that (69) is a scalar equation, and can be solved in a least-square sense after collecting enough number of data samples from the system. It is shown in Section IV-B how to collect required information in phase 1 and reuse them in phase 2 in a least-square sense to solve (69) for $V_i$, $u_{i+1}$, and $d_{i+1}$ simultaneously. After the learning is done and the optimal control policy $u^*$ is found, it can then be applied to the system.

*Theorem 5 (Convergence of Algorithm 2):* The off-policy Algorithm 2 converges to the optimal control and disturbance

**Algorithm 2** Online Off-Policy RL Algorithm for Solving Tracking HJI Equation

*Phase 1 (data collection using a fixed control policy):* Apply a fixed control policy $u$ to the system and collect required system information about the state, control input and disturbance at $N$ different sampling interval $T$.

*Phase 2 (reuse of collected data sequentially to find an optimal policy iteratively):* Given $u_i$ and $d_i$, use collected information in phase 1 to Solve the following Bellman equation for $V_i$, $u_{i+1}$ and $d_{i+1}$ simultaneously:

$$\begin{aligned} &e^{-\alpha T} V_i(X(t+T)) - V_i(X(t)) \\ &= \int_t^{t+T} e^{-\alpha(\tau-t)}\big(-X^T Q_T X - u_i^T R\, u_i + \gamma^2 d_i^T d_i\big) d\tau \\ &\quad + \int_t^{t+T} e^{-\alpha(\tau-t)}\big(-2u_{i+1}^T R(u-u_i) \\ &\qquad\qquad + 2\gamma^2 d_{i+1}^T (d-d_i)\big) d\tau \end{aligned} \tag{69}$$

Stop if a stopping criterion is met, otherwise set $i = i+1$ and got to 2.

solutions given by (17) and (18), where the value function satisfies the tracking HJI equation (19).

*Proof:* It was shown in Lemma 1 that the off-policy tracking Bellman equation (69) gives the same value function as the Bellman equation (58) and the same updated policies as (59) and (60). Therefore, both Algorithms 1 and 2 have the same convergence properties. Convergence of Algorithm 1 is proved in [33]. This confirms that Algorithm 2 converges to the optimal solution. □

*Remark 11:* Although both Algorithms 1 and 2 have the same convergence properties, Algorithm 2 is a model-free algorithm, which finds an optimal control policy without requiring any knowledge of the system dynamics. This is in contrast to Algorithm 1 that requires full knowledge of the system dynamics. Moreover, Algorithm 1 is an on-policy RL algorithm, which requires the disturbance input to be specified and adjustable. On the other hand, Algorithm 2 is an off-policy RL algorithm, which obviates this requirement.

### *B. Implementing Algorithm 2 Using Neural Networks*

In order to implement the off-policy RL Algorithm 2, it is required to reuse the collected information found by applying a fixed control policy $u$ to the system to solve (69) for $V_i$, $u_{i+1}$, and $d_{i+1}$ iteratively. Three NNs, i.e., the actor NN, the critic NN, and the disturber NN, are used here to approximate the value function and the updated control and disturbance policies in the Bellman equation (69). That is, the solution $V_i$, $u_{i+1}$, and $d_{i+1}$ of the Bellman equation (69) is approximated by three NNs as

$$\hat{V}_i(X) = \hat{W}_1^T \sigma(X) \tag{70}$$

$$\hat{u}_{i+1}(X) = \hat{W}_2^T \phi(X) \tag{71}$$

$$\hat{d}_{i+1}(X) = \hat{W}_3^T \varphi(X) \tag{72}$$

where $\sigma = [\sigma_1, \ldots, \sigma_{l_1}] \in \mathbb{R}^{l_1}$, $\phi = [\phi_1, \ldots, \phi_{l_2}] \in \mathbb{R}^{l_2}$, and $\varphi = [\varphi_1, \ldots, \varphi_{l_3}] \in \mathbb{R}^{l_3}$ provide suitable basis function vectors, $\hat{W}_1 \in \mathbb{R}^{l_1}$, $\hat{W}_2 \in \mathrm{R}^{m\times l_2}$, and $\hat{W}_3 \in \mathbb{R}^{q\times l_3}$ are constant weight vectors, and $l_1$, $l_2$, and $l_2$ are the number of neurons. Define $v^1 = [v_1^1, \ldots, v_1^m]^T = u - u_i$, $v^2 = [v_1^2, \ldots, v_q^2]^T = d - d_i$ and assume $R = \mathrm{diag}(r, \ldots, r_m)$. Then, substituting (70)–(72) in (69) yields

$$\begin{aligned} e(t) &= \hat{W}_1^T\big(e^{-\alpha T}\sigma(X(t+T)) - \sigma(X(t))\big) \\ &\quad - \int_t^{t+T} e^{-\alpha(\tau-t)}\big(-X^T Q_T X - u_i^T R u_i + \gamma^2 d_i^T d_i\big)\, d\tau \\ &\quad + 2\sum_{l=1}^{m} r_l \int_t^{t+T} e^{-\alpha(\tau-t)}\ \hat{W}_{2,l}^T \phi(X(t))\, v_l^1\, d\tau \\ &\quad - 2\gamma^2 \sum_{k=1}^{q} \int_t^{t+T} e^{-\alpha(\tau-t)}\ \hat{W}_{3,k}^T \varphi(X(t))\, v_k^2\, d\tau \end{aligned} \tag{73}$$

where $e(t)$ is the Bellman approximation error, $\hat{W}_{2,l}$ is the $l$th column of $\hat{W}_2$, and $\hat{W}_{3,k}$ is the $k$th column of $\hat{W}_3$. The Bellman approximation error is the continuous-time counterpart of the temporal difference (TD) [10]. In order to bring the TD error to its minimum value, a least-squares method is used. To this end, rewrite equation (73) as

$$y(t) + e(t) = \hat{W}^T h(t) \tag{74}$$

where

$$\hat{W} = \big[\hat{W}_1^T, \hat{W}_{2,l}^T, \ldots, \hat{W}_{2,m}^T, \hat{W}_{3,1}^T, \ldots, \hat{W}_{3,q}^T\big]^T \in \mathbb{R}^{l_1 + m\times l_2 + q\times l_3} \tag{75}$$

$$h(t) = \begin{bmatrix} e^{-\alpha T}\sigma(X(t+T)) - \sigma(X(t))) \\ 2r_1 \int_t^{t+T} e^{-\alpha(\tau-t)}\ \phi(X(t))\, v_1^1\, d\tau \\ \vdots \\ 2r_m \int_t^{t+T} e^{-\alpha(\tau-t)}\ \phi(X(t))\, v_m^1\, d\tau \\ -2\gamma^2 \int_t^{t+T} e^{-\alpha(\tau-t)}\ \varphi(X(t))\, v_1^2\, d\tau \\ \vdots \\ -2\gamma^2 \int_t^{t+T} e^{-\alpha(\tau-t)}\ \varphi(X(t))\, v_q^2\, d\tau \end{bmatrix} \tag{76}$$

$$y(t) = \int_t^{t+T} e^{-\alpha(\tau-t)}\big(-X^T Q_T X - u_i^T R u_i + \gamma^2 d_i^T d_i\big) d\tau. \tag{77}$$

The parameter vector $\hat{W}$, which gives the approximated value function, actor, and disturbance (70)–(72), is found by minimizing, in the least-squares sense, the Bellman error (74). Assume that the systems state, input, and disturbance information are collected at $N \geq l_1 + m \times l_2 + q \times l_3$ (the number of independent elements in $\hat{W}$) points $t_1$ to $t_N$ in the state space, over the same time interval $T$ in phase 1. Then, for a given $u_i$ and $d_i$, one can use this information to evaluate (76) and (77) at $N$ points to form

$$H = [h(t_1), \ldots., h(t_N))] \tag{78}$$

$$Y = [y(t_1), \ldots., y(t_N)]^T. \tag{79}$$

The least-squares solution to (74) is then equal to

$$\hat{W} = (HH^T)^{-1} HY \tag{80}$$

which gives $V_i$, $u_{i+1}$, and $d_{i+1}$.

*Remark 12:* Note that, although $X(t+T)$ appears in (73), this equation is solved in a least-square sense after observing N samples $X(t),\ X(t+T),\ldots,X(t+NT)$. Therefore, the knowledge of the system is not required to predict the future state $X(t+T)$ at time $t$ to solve (73).

## V. Simulation Results

In this section, the proposed off-policy IRL method is first applied to a linear system to show that it converges to the optimal solution. Then, it is tested on a nonlinear system.

### A. Linear System: F16 Aircraft Systems

Consider the F16 aircraft system described by $\dot{x} = Ax + Bu + Dd$ with the following dynamics:

$$A = \begin{bmatrix} -1.01887 & 0.90506 & -0.00215 \\ 0.82225 & -1.07741 & -0.17555 \\ 0 & 0 & -1 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}, \quad D = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{81}$$

The system state vector is $x = [x_1\ x_2\ x_3] = [\alpha\ q\ \delta_e]$, where $\alpha$ denotes the angle of attack, $q$ is the pitch rate, and $\delta_e$ is the elevator deflection angle. The control input is the elevator actuator voltage, and the disturbance is wind gusts on angle of attack. It is assumed that the output is $y = \alpha$, and the desired value is constant. Thus, the command generator dynamics become $\dot{r} = 0$. Therefore, the augmented dynamics (9) becomes equal to (82), as shown at the bottom of this page. Since only $e_1 = x_1 - r_1$ is concerned as the tracking error, the first element of the matrix $Q_T$ in (12) is considered to be 20, and all other elements are zero. It is also assumed here that $R = 1$ and $\gamma = 10$. The offline solution to the GARE (49) and consequently the optimal control policy (46) are given in (83), as shown at the bottom of this page.

We now implement the off-policy IRL Algorithm 2. The reinforcement interval is chosen as $T = 0.05$. The initial control gain is chosen as zero. Figs. 2 and 3 show the convergence of the kernel matrix $P$ and the control gain to their optimal values. In fact, $P$ converges to

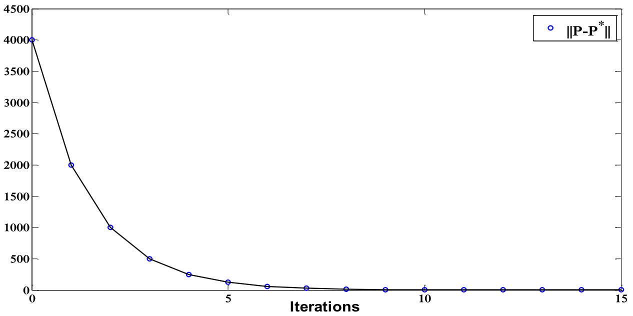


Fig. 2. Convergence of the kernel matrix $P$ to its optimal value for F-16 example.

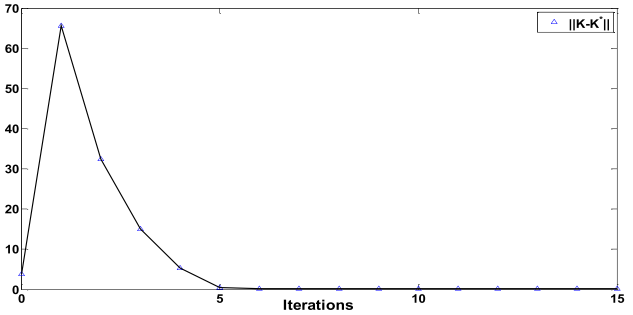


Fig. 3. Convergence of the control gain to its optimal value for F-16 example.

$$P = \begin{bmatrix} 12.675 & 5.418 & -0.432 & -7.481 & 5.424 & -0.439 \\ 5.420 & 3.412 & -0.330 & -4.985 & 3.404 & -0.329 \\ -0.427 & -0.323 & 0.042 & 0.546 & -0.333 & 0.046 \\ -7.495 & -4.973 & 0.545 & 201.408 & -4.985 & 0.527 \\ 5.419 & 3.406 & -0.328 & -4.968 & 3.405 & -0.339 \\ -0.421 & -0.347 & -0.201 & 0.036 & -0.333 & 0.046 \end{bmatrix}$$

which is very close to its optimal value given in (83). These results and Figs. 2 and 3 confirm that the proposed method converses with the optimal tracking solution without requiring the knowledge of the system dynamics. The optimal control solution found in (83) is now applied to the system to test its performance. To this end, it is assumed that the desired value for the output is $r_1 = 2$ for 0–30 s, and changes

$$\dot{X} = \begin{bmatrix} -1.01887 & 0.90506 & -0.00215 & -1.01887 & 0.90506 & -0.00215 \\ 0.82225 & -1.07741 & -0.17555 & 0.82225 & -1.07741 & -0.17555 \\ 0 & 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} X + \begin{bmatrix} 0 \\ 0 \\ 5 \\ 0 \\ 0 \\ 0 \end{bmatrix} u + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} d \tag{82}$$

$$P^* = \begin{bmatrix} 12.677 & 5.420 & -0.432 & -7.474 & 5.420 & -0.432 \\ 5.420 & 3.405 & -0.332 & -4.980 & 3.405 & -0.332 \\ -0.432 & -0.332 & 0.040 & 0.544 & -0.332 & 0.040 \\ -7.474 & -4.980 & 0.544 & 201.451 & -4.980 & 0.544 \\ 5.420 & 3.405 & -0.332 & -4.980 & 3.405 & -0.332 \\ -0.432 & -0.332 & -0.205 & 0.040 & -0.332 & 0.040 \end{bmatrix}$$

$$u^* = -[-2.1620,\quad -1.6623,\quad 0.2005,\quad 2.7198,\quad -1.6623,\quad 0.2005]X \tag{83}$$

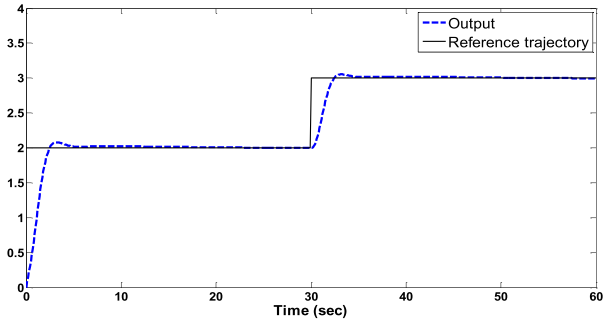


Fig. 4. Reference trajectory versus output for F-16 systems using the proposed control method.

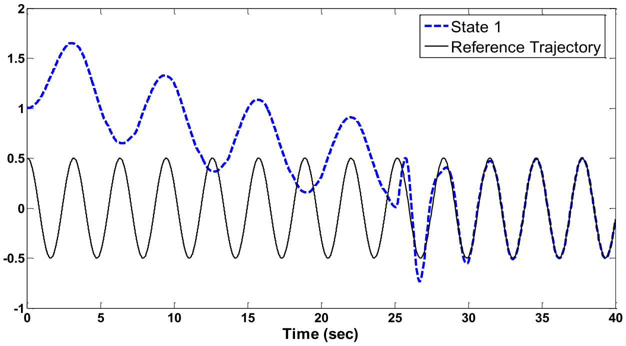


Fig. 5. Reference trajectory versus the first state of the robot manipulator systems during and after learning.

to $r_1 = 3$ at 30 s. The disturbance is assumed to be $d = 0.1e^{-0.1t}\sin(0.1t)$. Fig. 4 shows how the output converges to its desired values after the control solution (83) is applied to the system, and confirms that the proposed optimal control solution achieves suitable results.

### *B. Nonlinear System: A Two-Link Robot Manipulator*

In this section, the proposed off-policy IRL algorithm is applied to a two-link manipulator [46], [53], which is modeled using

$$M\,\ddot{q} + V_m\,\dot{q} + F_d\,\dot{q} + G(q) = u + d \tag{84}$$

where $q = [q_1\ q_2]^T$ is the vector of joint angles and $\dot{q} = [\dot{q}_1\ \dot{q}_2]^T$ is the vector of joint angular velocities, and

$$M = \begin{bmatrix} p_1 + 2p_3c_2 & p_2 + p_3c_2 \\ p_2 + p_3c_2 & p_2 \end{bmatrix}$$

$$V_m = \begin{bmatrix} -p_3s_2\dot{q}_2 & -p_3s_2(\dot{q}_1 + \dot{q}_2) \\ p_3s_2\dot{q}_1 & 0 \end{bmatrix}$$

are the inertia and Coriolis-centripetal matrices, respectively, with $c_2 = \cos(q_2)$, $s_2 = \sin(q_2)$, $p_1 = 3.473$ kgm$^2$, $p_2 = 0.196$ kgm$^2$, and $p_3 = 0.242$ kgm$^2$. Moreover, $F_d = \mathrm{diag}\,[5.3,\ 1.1]$, $G(q) = [8.45\tanh(\dot{q}_1),\ 2.35\tanh(\dot{q}_2)]^T$, $u$, and $\tau_d$ are the static friction, the dynamic friction, the control torque, and the disturbance, respectively.

Defining the state vector as $x = [q_1\ q_2\ \dot{q}_1\ \dot{q}_2]^T$, the state-space equations for (84) become (1) with [46]

$$f(x) = \left[ x_3\ x_4\ \left( M^{-1}\left(-V_m - F_d\right)\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} - G(q) \right)^T \right]^T$$

$$g(x) = k(x) = \left[ [0\ 0]^T \quad [0\ 0]^T \quad [0\ 0]^T \quad (M^{-1})^T \right]^T.$$

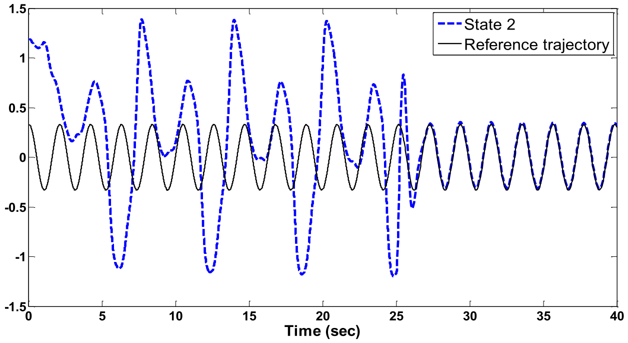


Fig. 6. Reference trajectory versus the second state of the robot manipulator systems during and after learning.

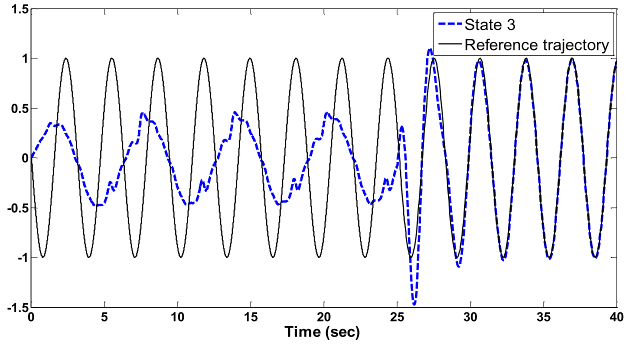


Fig. 7. Reference trajectory versus the third state of the robot manipulator systems during and after learning.

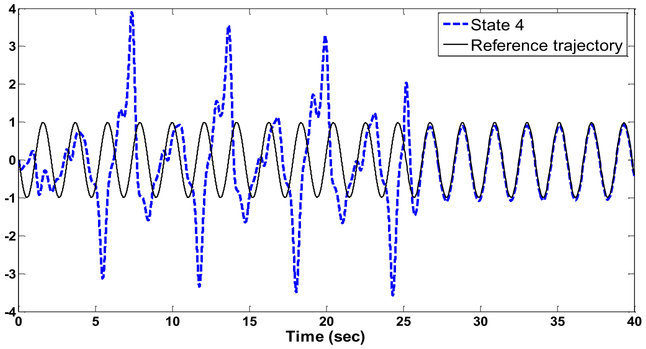


Fig. 8. Reference trajectory versus the fourth state of the robot manipulator systems during and after learning.

The objective is to find the control input $u$ to make the state follow the desired trajectory given as:

$$r = [\,0.5\cos(2t) \quad 0.33\cos(3t) \quad -\sin(2t) \quad -\sin(3t)\,]^T$$

which is generated by the command generator (2) with

$$h_d(r) = [\,r_3 \quad r_4 \quad -4r_1 \quad -9r_2\,]^T.$$

It is assumed in the disturbance attenuation condition (7) that $Q = 10\,I$, $R = 1$, and $\gamma = 20$. Based on (8), the augmented state becomes $X = [\,e_1\ e_2\ e_3\ e_4\ r_1\ r_2\ r_3\ r_4\,]^T$ with $e_i = x_i - r_i$, $i = 1, 2, 3, 4$. A power series NN containing even powers of the state variables of the system up to order four is used for the critic in (70). The activation functions for the control and disturbance policies in (71) and (72) are chosen as polynomials of all powers of the states up

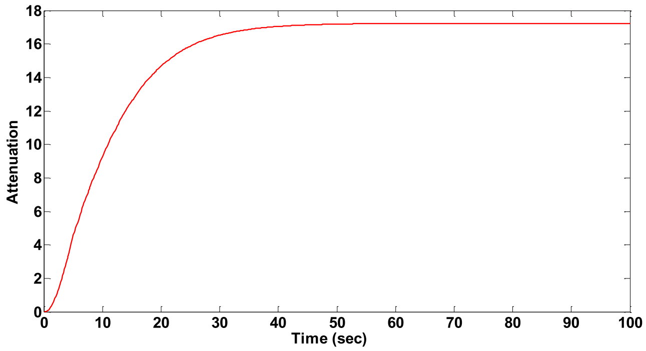


Fig. 9. Disturbance attenuation for the final controller.

to order three. We now implement Algorithm 2 to find the optimal control solution online. The reinforcement interval is chosen as $T = 0.05$. The proposed algorithm starts the learning process from the beginning of the simulation and finishes it after 25 s, when the control policy is updated. The plots of state trajectories of the closed-loop system and the reference trajectory are shown in Figs. 5–8. The disturbance is assumed to be $d = 0.1e^{-0.1t}\sin(0.1t)$ after the learning is done. From Figs. 5–8, it is obvious that the system tracks the reference trajectory after the learning is finished and the optimal controller is found. Fig. 9 shows the disturbance attenuation level of the optimal control policy found by the proposed method after the learning is done.

## VI. Conclusion

A model-free $H_\infty$ tracker was developed for nonlinear continuous-time systems in the presence of disturbance. A generalized discounted $L_2$-gain condition was proposed for obtaining the solution to this problem in which the norm of the performance output includes both the feedback and feedforward control inputs. This enables us to extend RL algorithms for solving the $H_\infty$ optimal tracking problem without requiring complete knowledge of the system dynamics. A tracking HJI equation is developed to find the solution to the problem in hand. The stability and optimality of the resulting solution were analyzed, and an upper bound for the discount factor was found to assure the stability of the control solution found by solving the tracking HJI equation. An online off-policy RL algorithm was proposed to learn the solution to the tracking HJI equation without requiring any knowledge of the system dynamics. It is shown that, using off-policy RL, the disturbance input does not require to be specified and adjusted. Simulation results confirmed the suitability of the proposed method.

## References

[1] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Trans. Autom. Control*, vol. 26, no. 2, pp. 301–320, Apr. 1981.

[2] J. C. Doyle, K. Glover, P. Khargonekar, and B. A. Francis, "State-space solutions to standard $H_2$ and $H_\infty$ control problems," *IEEE Trans. Autom. Control*, vol. 34, no. 8, pp. 831–847, Aug. 1989.

[3] A. J. Van der Schaft, "$L_2$-gain analysis of nonlinear systems and nonlinear state-feedback $H_\infty$ control," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 770–784, Jun. 1992.

[4] A. Isidori and W. Lin, "Global $L_2$-gain design for a class of nonlinear systems," *Syst. Control Lett.*, vol. 34, no. 5, pp. 295–302, 1998.

[5] T. Başar and P. Bernard, *$H_\infty$-Optimal Control and Related Minimax Design Problems*. Boston, MA, USA: Birkhäuser, 1995.

[6] M. D. S. Aliyu, *Nonlinear $H_\infty$ Control, Hamiltonian Systems and Hamilton-Jacobi Equations*. Boca Raton, FL, USA: CRC Press, 2011.

[7] S. Devasia, D. Chen, and B. Paden, "Nonlinear inversion-based output tracking," *IEEE Trans. Autom. Control*, vol. 41, no. 7, pp. 930–942, Jul. 1996.

[8] G. J. Toussaint, T. Basar, and F. Bullo, "$H_\infty$-optimal tracking control techniques for nonlinear underactuated systems," in *Proc. 39th IEEE Conf. Decision Control*, Sydney, NSW, Australia, Dec. 2000, pp. 2078–2083.

[9] J. A. Ball, P. Kachroo, and A. J. Krener, "$H_\infty$ tracking control for a class of nonlinear systems," *IEEE Trans. Autom. Control*, vol. 44, no. 6, pp. 1202–1206, Jun. 1999.

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

[11] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[12] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles* (Control Engineering). Stevenage, U.K.: IET Press, 2012.

[13] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control*, D. A. White and D. A. Sofge, Eds. Brentwood, U.K.: Multiscience Press, 1992.

[14] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[15] D. Vrabie and F. L. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.

[16] D. Vrabie, O. Pastravanu, M. Abou-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.

[17] R. Song, W. Xiao, H. Zhang, and C. Sun, "Adaptive dynamic programming for a class of complex-valued nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1733–1739, Sep. 2014.

[18] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

[19] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, 2013.

[20] T. Bian, Y. Jiang, and Z.-P. Jiang, "Adaptive dynamic programming and optimal control of nonlinear nonaffine systems," *Automatica*, vol. 50, no. 10, pp. 2624–2632, 2014.

[21] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 882–893, May 2014.

[22] 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.

[23] D. Liu, H. Li, and D. Wang, "Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1323–1334, Jun. 2015.

[24] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design," *Automatica*, vol. 50, no. 12, pp. 3281–3290, 2014.

[25] B. Luo, H.-N. Wu, and H.-X. Li, "Adaptive optimal control of highly dissipative nonlinear spatially distributed processes with neuro-dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 684–696, Apr. 2015.

[26] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Neurodynamic programming and zero-sum games for constrained control systems," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1243–1252, Jul. 2008.

[27] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207–214, 2011.

[28] K. G. Vamvoudakis and F. L. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," *Int. J. Robust Nonlinear Control*, vol. 22, no. 13, pp. 1460–1483, 2012.

[29] K. G. Vamvoudakis and F. L. Lewis, “Online gaming: Real time solution of nonlinear two-player zero-sum games using synchronous policy iteration,” in *Advances in Reinforcement Learning*, A. Mellouk, Ed. Delhi, India: Intech, 2011.

[30] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, “Online solution of nonquadratic two-player zero-sum games arising in the $H_\infty$ control of constrained input systems,” *Int. J. Adapt. Control Signal Process.*, vol. 28, nos. 3–5, pp. 232–254, 2014.

[31] H. Zhang, C. Qin, B. Jiang, and Y. Luo, “Online adaptive policy learning algorithm for $H_\infty$ state feedback control of unknown affine nonlinear discrete-time systems,” *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2706–2718, Dec. 2014.

[32] H.-N. Wu and B. Luo, “Simultaneous policy update algorithms for learning the solution of linear continuous-time $H_\infty$ state feedback control,” *Inf. Sci.*, vol. 222, pp. 472–485, Feb. 2013.

[33] H.-N. Wu and B. Luo, “Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear $H_\infty$ control,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1884–1895, Dec. 2012.

[34] B. Luo and H.-N. Wu, “Computationally efficient simultaneous policy update algorithm for nonlinear $H_\infty$ state feedback control with Galerkin’s method,” *Int. J. Robust Nonlinear Control*, vol. 23, no. 9, pp. 991–1012, 2013.

[35] D. Vrabie and F. L. Lewis, “Adaptive dynamic programming for online solution of a zero-sum differential game,” *J. Control Theory Appl.*, vol. 9, no. 3, pp. 353–360, 2011.

[36] H. Li, D. Liu, and D. Wang, “Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics,” *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 3, pp. 706–714, Jul. 2014.

[37] 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, Jan. 2015.

[38] Q. Wei, D. Liu, G. Shi, and Y. Liu, “Multibattery optimal coordination control for home energy management systems via distributed iterative adaptive dynamic programming,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 7, pp. 4203–4214, Jul. 2015.

[39] S. Jagannathan and G. Galan, “Adaptive critic neural network-based object grasping control using a three-finger gripper,” *IEEE Trans. Neural Netw.*, vol. 15, no. 2, pp. 395–407, Mar. 2004.

[40] Q. Wei, D. Liu, and G. Shi, “A novel dual iterative *Q*-learning method for optimal battery management in smart residential environments,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2509–2518, Apr. 2015.

[41] Q. Wei and D. Liu, “Data-driven neuro-optimal temperature control of water–gas shift reaction using stable iterative adaptive dynamic programming,” *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6399–6408, Nov. 2014.

[42] H. Modares, I. Ranatunga, F. L. Lewis, and D. O. Popa, “Optimized assistive human–robot interaction using reinforcement learning,” *IEEE Trans. Cybern.*, to be published.

[43] Q. Wei and D. Liu, “Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification,” *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1020–1036, Oct. 2014.

[44] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, “Reinforcement QQ-learning for optimal tracking control of linear discrete-time systems with unknown dynamics,” *Automatica*, vol. 50, no. 4, pp. 1167–1175, 2014.

[45] 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. 3051–3065, Nov. 2014.

[46] 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, Jan. 2015.

[47] 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.

[48] H. Zhang, R. Song, Q. Wei, and T. Zhang, “Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming,” *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1851–1862, Dec. 2011.

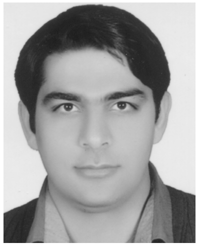
[49] B. Kiumarsi and F. L. Lewis, “Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 140–151, Jan. 2015.

[50] 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.

[51] D. Liu, Y. Huang, and Q. Wei, “Neural network $H_\infty$ tracking control of nonlinear systems using GHJI method,” in *Advances in Neural Networks*, C. Guo, Z.-G. Huo, and Z. Zeng, Eds. Dalian, China: Springer-Verlag, 2013.

[52] F. L. Lewis, D. Vrabie, and V. Syrmos, *Optimal Control*, 3rd ed. New York, NY, USA: Wiley, 2012.

[53] F. L. Lewis, D. M. Dawson, and C. T. Abdallah, *Robot Manipulator Control: Theory and Practice*, 2nd ed. New York, NY, USA: CRC Press, 2003.



**Hamidreza Modares** 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. His current research interests include cyber-physical systems, reinforcement learning, distributed control, robotics, and pattern recognition.

**Frank. L. Lewis** (S’70–M’81–SM’86–F’94) received the bachelor’s degree in physics/electrical engineering and the M.S. degree in electrical engineering from Rice University, Houston, TX, USA, the M.S. degree in aeronautical engineering from the University of West Florida, Pensacola, FL, USA, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA.

He is currently a U.K. Chartered Engineer, the IEEE Control Systems Society Distinguished Lecturer, a University of Texas at Arlington Distinguished Scholar Professor, a UTA Distinguished Teaching Professor, and a Moncrief-O’Donnell Chair with the University of Texas at Arlington Research Institute, Fort Worth, TX, USA. He is a Qian Ren Thousand Talents Consulting Professor with Northeastern University, Shenyang, China. He is involved in feedback control, reinforcement learning, intelligent systems, and distributed control systems. He has authored six U.S. patents, 301 journal papers, 396 conference papers, 20 books, 44 chapters, and 11 journal special issues.

Dr. Lewis is a member of the National Academy of Inventors. He is a fellow of the International Federation of Automatic Control, the U.K. Institute of Measurement and Control, and Professional Engineer at Texas. He was a Founding Member of the Board of Governors of the Mediterranean Control Association. He received the Fulbright Research Award, the NSF Research Initiation Grant, the ASEE Terman Award, the International Neural Network Society Gabor Award in 2009, and the U.K. Institute of Measurement and Control Honeywell Field Engineering Medal in 2009. He was a recipient of the IEEE Computational Intelligence Society Neural Networks Pioneer Award in 2012, the Distinguished Foreign Scholar from the Nanjing University of Science and Technology, the 111 Project Professor at Northeastern University, China, the Outstanding Service Award from the Dallas IEEE Section, an Engineer of the Year from the Fort Worth IEEE Section. He was listed in Fort Worth Business Press Top 200 Leaders in Manufacturing. He was also a recipient of the 2010 IEEE Region Five Outstanding Engineering Educator Award, the 2010 UTA Graduate Dean’s Excellence in Doctoral Mentoring Award, was elected to the UTA Academy of Distinguished Teachers in 2012, and the Texas Regents Outstanding Teaching Award in 2013. He served on the NAE Committee on Space Station in 1995. He also received the IEEE Control Systems Society Best Chapter Award (as a Founding Chairman of DFW Chapter), the National Sigma Xi Award for Outstanding Chapter (as a President of UTA Chapter), and the U.S. SBA Tibbets Award in 1996 (as the Director of ARRI’s SBIR Program).

**Zhong-Ping Jiang** (M'94–SM'02–F'08) received the B.Sc. degree in mathematics from the University of Wuhan, Wuhan, China, in 1988, the M.Sc. degree in statistics from the University of Paris XI, Paris, France, in 1989, and the Ph.D. degree in automatic control and mathematics from the Ecole des Mines de Paris, Paris, France, in 1993.

He is currently a Professor of Electrical and Computer Engineering with the Polytechnic School of Engineering, New York University, New York, NY, USA. His current research interests include stability theory, robust, adaptive and distributed nonlinear control, adaptive dynamic programming, and their applications to information, mechanical and biological systems. He has co-authored the book *Stability and Stabilization of Nonlinear Systems* (Springer, 2011) with Dr. I. Karafyllis, and *Nonlinear Control of Dynamic Networks* (Taylor & Francis, 2014) with Dr. T. Liu and Dr. D. J. Hill.

Prof. Jiang is a fellow of the International Federation of Automatic Control. He was a recipient of the prestigious Queen Elizabeth II Fellowship Award from the Australian Research Council, the CAREER Award from the U.S. National Science Foundation (NSF), and the Distinguished Overseas Chinese Scholar Award from the NSF of China. He was also a recipient of recent awards recognizing his research work, including the Best Theory Paper Award with Y. Wang at the 2008 World Congress on Intelligent Control and Automation, the Guan Zhao Zhi Best Paper Award, with T. Liu and D. Hill, at the 2011 Chinese Control Conference, and the Shimemura Young Author Prize with his student Y. Jiang at the 2013 Asian Control Conference in Istanbul, Turkey. He is a Deputy co-Editor-in-Chief of the *Journal of Control and Decision*, an Editor for the *International Journal of Robust and Nonlinear Control*, and has served as an Associate Editor for several journals, including the *Mathematics of Control, Signals and Systems*, *Systems and Control Letters*, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the *European Journal of Control*, and *Science China: Information Sciences*.