IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 August 2022, accepted 6 September 2022, date of publication 20 September 2022, date of current version 28 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3208058

RESEARCH ARTICLE

Regulation With Guaranteed Convergence Rate for Continuous-Time Systems With Completely Unknown Dynamics in the Presence of Disturbance

ALI RAHDARIAN[1], D. SADRIAN ZADEH[2], (Student Member, IEEE), SAEED SHAMAGHDARI[1], BEHZAD MOSHIRI[2,3], (Senior Member, IEEE), AND ALLAHYAR MONTAZERI[4]

[1]Electrical Engineering Department, Iran University of Science and Technology, Tehran 16844, Iran
[2]School of Electrical and Computer Engineering, University of Tehran, Tehran 1439957131, Iran
[3]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[4]School of Engineering, Lancaster University, LA1 4YW Lancaster, U.K.

Corresponding author: Allahyar Montazeri (a.montazeri@lancaster.ac.uk)

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/V027379/1 and Grant EP/R02572X/1, and in part by the National Centre for Nuclear Robotics (NCNR).

ABSTRACT This paper presents the design of a novel H_∞-based control framework for state regulation of continuous-time linear systems with completely unknown dynamics. The proposed method solves the regulation problem with the desired convergence rate and simultaneously seeks to attenuate the adverse effect of disturbance on the system. The H_∞ regulation problem assumes a cost function that considers regulation with a guaranteed rate of convergence as well as disturbance attenuation. The problem is then turned into a two-player zero-sum game optimization problem that can be solved by solving the associated algebraic Riccati equation (ARE), which provides a model-based solution. To solve this problem in a model-free way, a novel integral reinforcement learning (IRL) algorithm is designed to learn the solution online without requiring any prior knowledge of the system dynamics. It is shown that the model-free method (i.e., IRL-based method) provides the same solution as the model-based method (i.e., ARE). The effectiveness of the proposed method is ascertained through simulation examples; it is shown that the proposed method effectively addresses the problem for both stable and unstable systems.

INDEX TERMS Convergence rate, disturbance attenuation, H_∞ control, regulation problem, integral reinforcement learning.

I. INTRODUCTION

Optimal regulation involves developing a controller that ensures the system states optimally converge to zero, balancing the system costs and the control efforts. Solving optimal control problems requires solving the Hamilton-Jacobi-Bellman (HJB) equation. For the linear systems, this can be achieved through solving the linear quadratic regulation (LQR) problem using the algebraic Riccati equation (ARE), as the simplified version of the HJB equation [1], [2], [3], [4]; however, for the nonlinear systems, due to lack of analytical solution, the HJB equation should be solved numerically. The disadvantage of this approach is the need for an explicit knowledge of the system model; therefore, model-free methods working based on the concept of reinforcement learning (RL) have been proposed in the literature. This method has been effectively used to solve the regulation problem for the nonlinear systems and learn the optimal control solution in real-time while cutting the need for the complete knowledge of the system dynamics [5], [6], [7], [8], [9], [10], [11], [12].

Reinforcement learning has been widely employed as a machine learning technique for solving complex optimization problems [13]. As it is commonly known, RL has been used as a tool to solve optimal control problems by solving the HJB equations iteratively using either the policy iteration (PI) [5],

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

[14], [15] or value iteration (VI) [2], [16], [17] techniques. Hence, RL can provide a real-time solution to the HJB equation, by optimizing the control cost RL iteratively, and under the assumption of unknown system dynamic [18], [19], [20]. In both PI and VI algorithms the policy evaluation and policy improvement steps are carried out iteratively until an optimal solution is achieved. Another merit of leveraging RL for feedback control problems is its ability to address the ''curse of dimensionality'' in such problems [21].

A. LITERATURE REVIEW

The off-policy integral reinforcement learning (IRL) algorithm is first introduced in [22] and [23] to avoid the time derivatives in continuous-time (CT) systems and to design an optimal state-feedback controller. The authors of [24] present an online IRL algorithm to find the solution to the tracking-constrained HJB equation for partially-unknown systems with a bounded control input. This algorithm produces observations using a policy that differs from the evaluated one, thus the term off-policy is coined to this method. In [25], the authors apply the IRL algorithm, which provides the solution to the HJB equation, to learn the CT optimal control solution for nonlinear systems with an infinite horizon cost and incomplete knowledge of the system dynamics. IRL is also used in [26] to solve the linear quadratic tracking (LQT) problem. A PI-based IRL algorithm is presented in [27] to solve the Nash equilibrium for a two-player zero-sum differential game. The authors of [28] focused on output regulation with the help of the IRL algorithm without the need for a discounting factor to design a model-free controller for the linear systems. Integral RL algorithm is also employed in [29] for adaptive control of high-order multivariable nonlinear systems with unknown control coefficients.

A discussion on Q-learning and IRL algorithms for discrete-time (DT) and CT systems has been provided in [30]. The authors of [31] developed a model-free off-policy RL algorithm to solve ARE for robustly stabilizing a DT linear system affected by bounded and mismatched uncertainties. In [32], the optimal tracking control of DT nonlinear systems is studied under the condition of unknown system drift dynamics; the tracking problem is turned into a regulation problem by augmentation, and the associated HJB equation is solved by applying a new RL-based scheme using an actor-critic neural network structure. Discrete-time LQR problem is also studied in [33], in which the focus is on achieving the robustness utilising a new off-policy model-free RL algorithm, called optimistic least-squares policy iteration, for a system with additive stochastic noise.

Reinforcement learning and adaptive dynamic programming (ADP) techniques are studied in [34] to achieve an optimal output regulation controller for linear systems with unmeasurable disturbances and unknown dynamics. An ADP-based dynamic output feedback scheme is also developed in [35] for the linear quadratic regulation of CT systems; the solution is model-free and uses measurable input-output data to find the optimal control parameters.

In [36], the authors developed an off-policy model-free IRL algorithm to learn the optimal output-feedback (OPFB) solution for CT linear systems without the knowledge of system dynamics; the algorithm is applied to both regulation and tracking problems based on a discounted performance function and a discounted ARE. In addition, the authors of [37] presented an IRL-based online algorithm for learning a sub-optimal OPFB H_∞ control law to address CT linear tracking problems under disturbances; the IRL solves the game ARE online, which gives a Nash equilibrium solution related to the optimization problem. The same scheme is also developed in [38] but for OPFB control of CT linear systems with input delay. Furthermore, the authors of [39] developed an H_∞ tracking controller for CT nonlinear systems by developing a tracking Hamilton-Jacobi-Isaac (HJI) equation and solving it through an online off-policy RL algorithm without the need for the system dynamic; they show that the algorithm does not require a specified disturbance input. In [40], the authors employed an RL algorithm to solve the H_∞ tracking problem in real-time for a nonlinear system without the need for the system dynamics. This is also the case in [41], in which the authors used an event-triggered RL algorithm to solve the tracking HJI equation. Moreover, IRL-based event-triggered ADP is introduced in [42] to cut the need for drift dynamics and to control a CT nonlinear system with saturated input. In [43], the authors considered a CT neural network parameter update law based on variable gain gradient descent augmented with robust terms for model-free IRL-based H_∞ optimal tracking control problem of a CT nonlinear system with unknown dynamics for disturbance rejection.

The convergence rate or speed of regulation is an important consideration, and it is hoped that the convergence rate is as fast as possible. In [44], an RL-based method for solving the regulation problem with a guaranteed convergence rate for CT linear systems is developed. Also, an off-policy model-free RL-based solution for solving the LQR problem is developed in [45] for a DT linear system with a guaranteed convergence rate of the state variables. However, the adverse effect of disturbance is ignored in both works.

B. MAIN CONTRIBUTIONS

Reviewing the literature shows that the majority of previous studies have focused on either guaranteeing the convergence rate or attenuating the disturbance, but not both concurrently. Hence, the contributions of this paper are as follows:

- This paper proposes an H_∞-based control method for solving the regulation problem with a guaranteed convergence rate for CT linear systems through leveraging a novel model-free IRL algorithm; therefore, no knowledge of system dynamics is required. The poof of regulation with the desired convergence rate is also provided.
- This paper considers the system to be adversely affected by an unknown time-varying disturbance and shows that the proposed method not only attenuates the disturbance but also ensures regulation with a desired convergence

rate. Simulation results will also demonstrate the effectiveness of the proposed method for various unknown harsh disturbances.

C. MANUSCRIPT LAYOUT

The remainder of this manuscript is organized as follows. Section II formulates the regulation problem with guaranteed convergence rate under the effect of disturbance. Section III presents the novel model-based solution to this problem with proofs. The proposed model-free method of this paper is proposed in Section IV and Section V provides some examples to support the proposed method. Finally, the manuscript is concluded in Section VI, and some related future research directions are provided.

II. PROBLEM FORMULATION

This section of the manuscript presents the basics of the state-feedback regulation control strategy. The CT state-evolution equation of the system can be written as

$$\dot{x}(t) = Ax(t) + Bu(t) + D\omega(t), \tag{1}$$

where $x(t) \in \mathbb{R}^{n\times 1}$ is the vector of system states, $x(0) = x_0$ is the vector of initial states, $u(t) \in \mathbb{R}^{m\times 1}$ is the vector of control input, and $\omega(t) \in \mathbb{R}^{p\times 1}$ is the vector of disturbance.

Assumption 1: The pair (A, B) is stabilizable.

The objective of the regulation problem is to design a control input $u(t)$ such that the trajectories of system dynamics tend to zero, i.e., $\lim_{t\to\infty} x(t) = 0$. The control input is represented by

$$u(t) = K_u x(t), \tag{2}$$

in which $K_u \in \mathbb{R}^{m\times n}$ is the gain matrix of the feedback loop.

Definition 1: The system is called normal if it is not directly affected by disturbance, i.e., $\omega(t) = 0$.

Considering (2) as the control input and letting (1) be normal, (1) can be rewritten as

$$\dot{x}(t) = (A + BK_u)x(t), \tag{3}$$

which is called the closed-loop equation of the system.

It should be noted that the feedback gain K_u guarantees the regulation of the system dynamics at a convergence rate faster than $e^{-\alpha t}$ if all eigenvalues of the closed-loop system (3) are located on the left-hand side of the line $s = -\alpha$ in the s-plane, i.e., $\max\{Re[\lambda(A + BK_u)]\} < -\alpha$, thus guaranteeing the convergence rate of the regulation problem.

Given α and γ as predetermined values, the problem is to devise a control strategy (2) for the defined CT linear system (1) such that:

1) Regulation is achieved at least with the rate of $e^{-\alpha t}$ as t approaches to ∞, i.e., $\lim_{t\to\infty} e^{\alpha t}x(t) = 0$ when $\omega(t) = 0$.
2) The following bounded L_2-gain condition should be satisfied when $\omega(t) \in L_2[0, \infty)$:

$$\frac{\int_0^\infty e^{\beta\tau}||Z(\tau)||^2\, d\tau}{\int_0^\infty e^{\beta\tau}||\omega(\tau)||^2\, d\tau} \le \gamma^2, \tag{4}$$

where $||Z(t)||^2 = x^T(t)Qx(t) + u^T(t)Ru(t)$ is the performance output, $Q = Q^T \ge 0$ and $R = R^T > 0$ are weight matrices, α is the minimum degree of stability, β is the rate of disturbance attenuation, and γ is the level of disturbance attenuation.

III. H_∞ REGULATION CONTROL WITH GUARANTEED CONVERGENCE RATE

The above-mentioned H_∞ regulation problem can be regarded as a two-player zero-sum game. Therefore, with regard to the second condition of the problem, the cost function for this problem can be defined as

$$J\big(x(t), u(t), \omega(t)\big) = \int_0^\infty e^{\beta\tau}\big[x^T(\tau)Qx(t) + u^T(\tau)Ru(\tau) - \gamma^2\omega^T(\tau)\omega(\tau)\big]\, d\tau. \tag{5}$$

Since the satisfaction of the second condition of H_∞ regulation problem is equivalent to minimizing the cost function (5), the optimization problem turns into the following:

$$\min J\big(x(t), u(t), \omega(t)\big) \text{ s.t. (1).} \tag{6}$$

By choosing $\beta = 2\alpha$, the cost function (5) can be written in the following form:

$$\begin{aligned}&\big(x(t), u(t), \omega(t)\big)\\ &= \int_0^\infty \Big[\big(e^{\alpha\tau}x(\tau)\big)^T Q\big(e^{\alpha\tau}x(\tau)\big) + \big(e^{\alpha\tau}u(\tau)\big)^T R\big(e^{\alpha\tau}u(\tau)\big)\\ &\quad - \gamma^2\big(e^{\alpha\tau}\omega(\tau)\big)^T\big(e^{\alpha\tau}\omega(\tau)\big)\Big]\, d\tau.\end{aligned} \tag{7}$$

In order to proceed with the solution, $\bar{x}(t)$, $\bar{u}(t)$, $\bar{\omega}(t)$, and $\bar{A}$ are introduced as

$$\begin{aligned}\bar{x}(t) &= e^{\alpha t}x(t), \quad \bar{u}(t) = e^{\alpha t}u(t),\\ \bar{\omega}(t) &= e^{\alpha t}\omega(t), \quad \bar{A} = A + \alpha I,\end{aligned} \tag{8}$$

and substitute them in (7); consequently, the cost function can be written as

$$J\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) = \int_0^\infty \big[\bar{x}^T(\tau)Q\bar{x}(\tau) + \bar{u}^T(\tau)R\bar{u}(\tau) - \gamma^2\bar{\omega}^T(\tau)\bar{\omega}(\tau)\big]\, d\tau. \tag{9}$$

Now, by taking the derivative of $\bar{x}(t)$,

$$\begin{aligned}\frac{d}{dt}\big(\bar{x}(t)\big) &= \frac{d}{dt}\big(e^{\alpha t}x(t)\big) = \alpha e^{\alpha t}x(t) + e^{\alpha t}\dot{x}(t)\\ &= \alpha e^{\alpha t}x(t) + e^{\alpha t}\big(Ax(t) + Bu(t) + D\omega(t)\big)\\ &= (A + \alpha I)e^{\alpha t}x(t) + B\big(e^{\alpha t}u(t)\big) + D\big(e^{\alpha t}\omega(t)\big)\\ &= \bar{A}\bar{x}(t) + B\bar{u}(t) + D\bar{\omega}(t),\end{aligned} \tag{10}$$

the optimization problem turns into the following:

$$\min J\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) \text{ s.t. (10).} \tag{11}$$

As mentioned previously, the H_∞ regulation problem can be regarded as a two-player zero-sum game in the sense that the minimizing player is the control input $\bar{u}(t)$ and the maximizing player is the disturbance $\bar{\omega}(t)$. In other words,

$\bar{\omega}(t)$ attempts to maximize the cost function (9) while $\bar{u}(t)$ aims to minimize it. Besides, due to the linearity of the system, the value function is quadratic with the following form:

$$V\big(\bar{x}(t)\big) = J\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) = \bar{x}^T(t)P\bar{x}(t). \tag{12}$$

Consequently, the solution to the optimization problem (11) is equivalent to the solution to the min-max optimization problem

$$\begin{aligned} V^*\big(\bar{x}(t)\big) &= J\big(\bar{x}(t), \bar{u}^*(t), \bar{\omega}^*(t)\big) \\ &= \min_{\bar{u}(t)} \max_{\bar{\omega}(t)} J\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big), \end{aligned} \tag{13}$$

where $V^*\big(\bar{x}(t)\big)$ is the optimal value of $V\big(\bar{x}(t)\big)$.

In order for this optimization problem to have a unique solution, a game-theoretic saddle point must exist, i.e., the following condition must hold:

$$\begin{aligned} V^*\big(\bar{x}(t)\big) &= \min_{\bar{u}(t)} \max_{\bar{\omega}(t)} J\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) \\ &= \max_{\bar{\omega}(t)} \min_{\bar{u}(t)} J\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big). \end{aligned} \tag{14}$$

Now, taking (10), (11), and (12) into account, the Hamiltonian function can be written as

$$\begin{aligned} &H\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) \\ &\quad = \left(\frac{dV\big(\bar{x}(t)\big)}{d\bar{x}(t)}\right)^T \frac{d}{dt}\big(\bar{x}(t)\big) \\ &\qquad + \bar{x}^T(t)Q\bar{x}(t) + \bar{u}^T(t)R\bar{u}(t) - \gamma^2\bar{\omega}^T(t)\bar{\omega}(t), \end{aligned} \tag{15}$$

and continuing by differentiating from (12), the following Bellman equation is achieved:

$$\begin{aligned} &H\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) \\ &\quad = \big(\bar{A}\bar{x}(t) + B\bar{u}(t) + D\bar{\omega}(t)\big)^T P\bar{x}(t) \\ &\qquad + \bar{x}(t)^T P\big(\bar{A}\bar{x}(t) + B\bar{u}(t) + D\bar{\omega}(t)\big) + \bar{x}^T(t)Q\bar{x}(t) \\ &\qquad + \bar{u}^T(t)R\bar{u}(t) - \gamma^2\bar{\omega}^T(t)\bar{\omega}(t) = 0. \end{aligned} \tag{16}$$

The minimizing (optimal) control input and the maximizing (worst-case) disturbance can be achieved by applying the stationary conditions $\partial H(\cdot)/\partial\bar{u}(t) = 0$ and $\partial H(\cdot)/\partial\bar{\omega}(t) = 0$, which results in:

$$\bar{u}^*(t) = -R^{-1}B^T P\bar{x}(t) = K_u^*\bar{x}(t), \tag{17}$$

$$\bar{\omega}^*(t) = \frac{1}{\gamma^2}D^T P\bar{x}(t) = K_\omega^*\bar{x}(t). \tag{18}$$

The substitution of (17) and (18) into (16) produces the following equality

$$\bar{x}^T(t)\big[\bar{A}^T P + P\bar{A} + Q - PBR^{-1}B^T P + \frac{1}{\gamma^2}PDD^T P\big]\bar{x}(t) = 0,$$

which can be then simplified to an ARE as shown below:

$$\begin{aligned} &(A+\alpha I)^T P + P(A+\alpha I) + Q - PBR^{-1}B^T P \\ &\qquad + \frac{1}{\gamma^2}PDD^T P = 0. \end{aligned} \tag{19}$$

Theorem 1: Consider the CT linear system (1). By means of the control input signal (2), the H_∞ regulation problem is solved if

$$K_u^* = -R^{-1}B^T P, \tag{20}$$

where $P = P^T > 0$ is the solution to the ARE (19).

The proof is required to show that the optimal control input signal (17) satisfies both conditions of the problem; therefore, the proof is separated into two parts.

Proof of Theorem 1 (part 1): The first condition of the problem disregards the disturbance, i.e., considers the system normal. Since if $\omega(t) = 0$ then $\bar{\omega}(t) = 0$, (10) can be written as

$$\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + B\bar{u}(t), \tag{21}$$

and by substituting (17) in (21), the dynamic representation of the system changes into

$$\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) - BR^{-1}B^T P\bar{x}(t) = A_c\bar{x}(t), \tag{22}$$

where $A_c = \bar{A} - BR^{-1}B^T P$. Now, considering the quadratic value function (12) as a Lyapunov candidate, its derivative yields to

$$\begin{aligned} \dot{V}\big(\bar{x}(t)\big) &= \dot{\bar{x}}^T(t)P\bar{x}(t) + \bar{x}^T(t)P\dot{\bar{x}}(t) \\ &= \big[(\bar{A} - BR^{-1}B^T P)\bar{x}(t)\big]^T P\bar{x}(t) \\ &\quad + \bar{x}^T(t)P\big[(\bar{A} - BR^{-1}B^T P)\bar{x}(t)\big] \\ &= \bar{x}^T(t)(\bar{A} - BR^{-1}B^T P)^T P\bar{x}(t) \\ &\quad + \bar{x}^T(t)P(\bar{A} - BR^{-1}B^T P)\bar{x}(t) \\ &= \bar{x}^T(t)(\bar{A}^T P + P\bar{A} - 2PBR^{-1}B^T P)\bar{x}(t). \end{aligned} \tag{23}$$

Since the system is considered normal, the ARE (19) reduces to

$$\bar{A}^T P + P\bar{A} + Q - PBR^{-1}B^T P = 0, \tag{24}$$

and the substitution of (24) in (23) produces

$$\dot{V}\big(\bar{x}(t)\big) = \bar{x}^T(t)(-Q - PBR^{-1}B^T P)\bar{x}(t), \tag{25}$$

for which $\dot{V}\big(\bar{x}(t)\big) < 0$ due to $Q \geq 0$, $P > 0$, and $R > 0$; thus $\lim_{t\to\infty}\bar{x}(t) = 0$. Now, since $\bar{x}(t) = e^{\alpha t}x(t)$, it can be concluded that $\lim_{t\to\infty} e^{\alpha t}x(t) = 0$. Hence, the first part of the proof is completed.

Proof of Theorem 1 (part 2): Considering the optimal values $\bar{u}^*(t)$ and $\bar{\omega}^*(t)$, the Hamiltonian function (15) can be represented as follows:

$$\begin{aligned} &H\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) \\ &\quad = \big(\bar{u}(t) - \bar{u}^*(t)\big)^T R\big(\bar{u}(t) - \bar{u}^*(t)\big) \\ &\qquad - \gamma^2||\bar{\omega}(t) - \bar{\omega}^*(t)||^2 + H\big(\bar{x}(t), \bar{u}^*(t), \bar{\omega}^*(t)\big). \end{aligned} \tag{26}$$

Regarding (16), (17), and (18), it is known that $H\big(\bar{x}(t), \bar{u}^*(t), \bar{\omega}^*(t)\big) = 0$. Consequently, in order to complete the proof, it is sufficient to show that (17) is the solution

to the H_∞ regulation problem. To continue the proof, (15) can be rewritten as

$$H\big(\bar{x}(t), \bar{u}(t), \bar{\omega}(t)\big) = \frac{dV\big(\bar{x}(t)\big)}{dt} \bar{x}^T(t)Q\bar{x}(t) + \bar{u}^T(t)R\bar{u}(t) - \gamma^2\bar{\omega}^T(t)\bar{\omega}(t). \tag{27}$$

Considering (26) and (27), the following equality holds:

$$\bar{x}^T(t)Q\bar{x}(t) + \bar{u}^T(t)R\bar{u}(t) - \gamma^2\bar{\omega}^T(t)\bar{\omega}(t) + \frac{dV\big(\bar{x}(t)\big)}{dt} = \big(\bar{u}(t) - \bar{u}^*(t)\big)^T R\big(\bar{u}(t) - \bar{u}^*(t)\big) - \gamma^2||\bar{\omega}(t) - \bar{\omega}^*(t)||^2. \tag{28}$$

Considering $\bar{u}^*(t) = K_u^*\bar{x}(t)$ with (20), (28) produces

$$\frac{dV\big(\bar{x}(t)\big)}{dt} + \bar{x}^T(t)[Q + K_u^{*T}RK_u^*]\bar{x}(t) - \gamma^2\bar{\omega}^T(t)\bar{\omega}(t) = -\gamma^2||\bar{\omega}(t) - \bar{\omega}^*(t)||^2 \leq 0. \tag{29}$$

Taking into account (8) and (12), the first term of (29) can be written as

$$\frac{dV\big(\bar{x}(t)\big)}{dt} = \frac{d}{dt}\big(\bar{x}^T(t)P\bar{x}(t)\big) = \frac{d}{dt}\big(e^{2\alpha t}x^T(t)Px(t)\big) = \frac{d}{dt}\Big(e^{2\alpha t}V\big(x(t)\big)\Big), \tag{30}$$

and by considering (8), (29), and (30), the following inequality can be introduced:

$$\frac{d}{dt}\Big(e^{2\alpha t}V\big(x(t)\big)\Big) + e^{2\alpha t}x^T(t)[Q + K_u^{*T}RK_u^*]x(t) - e^{2\alpha t}\gamma^2\omega^T(t)\omega(t) \leq 0. \tag{31}$$

Applying integration to both sides of the inequality (31) results in

$$\int_0^{t_f} \frac{d}{d\tau}\Big(e^{2\alpha\tau}V\big(x(\tau)\big)\Big)\,d\tau + \int_0^{t_f} e^{2\alpha\tau}x^T(\tau)[Q + K_u^{*T}RK_u^*]x(\tau)\,d\tau - \gamma^2\int_0^{t_f} e^{2\alpha\tau}\omega^T(\tau)\omega(\tau)\,d\tau \leq 0, \tag{32}$$

which leads to

$$e^{2\alpha t_f}V\big(x(t_f)\big) - V\big(x(0)\big) + \int_0^{t_f} e^{2\alpha\tau}x^T(\tau)[Q + K_u^{*T}RK_u^*]x(\tau)\,d\tau - \gamma^2\int_0^{t_f} e^{2\alpha\tau}\omega^T(\tau)\omega(\tau)\,d\tau \leq 0. \tag{33}$$

Since $e^{2\alpha t_f}V\big(x(t_f)\big) \geq 0$ for every $t_f > 0$ and $\omega(t) \in L_2[0, \infty)$, one can conclude that

$$\int_0^{t_f} e^{2\alpha\tau}x^T(\tau)[Q + K_u^{*T}RK_u^*]x(\tau)\,d\tau \leq \gamma^2\int_0^{t_f} e^{2\alpha\tau}\omega^T(\tau)\omega(\tau)\,d\tau + V\big(x(0)\big), \tag{34}$$

in which $\beta = 2\alpha$, $t_f = \infty$, and $V\big(x(0)\big)$ can regarded zero without the loss of generality; hence, the second part of the proof is also completed.

Remark 1: Up until now, the validity of Theorem 1 has been proved, and it has been demonstrated that the ARE (19) solves the H_∞ regulation problem. Therefore, all eigenvalues of the closed-loop system will be located on the left-hand side of the line $s = -\alpha$ in the s-plane, i.e., $\max\{Re[\lambda(A + BK)]\} < -\alpha$, thus guaranteeing the convergence rate of the regulation problem. However, (19) depends on A and B, i.e., the system dynamics. In this regard, a method will be proposed in the next section that is independent of the system dynamics.

IV. ONLINE OFF-POLICY MODEL-FREE INTEGRAL REINFORCEMENT LEARNING

This section focuses on developing an online off-policy model-free integral reinforcement learning (IRL) algorithm. This algorithm enables the ARE to be solved using measured data in real time without needing prior knowledge of the system dynamics.

In order to develop the IRL algorithms, (10) can be rewritten as

$$\begin{aligned}\frac{d}{dt}\big(\bar{x}(t)\big) &= \bar{A}\bar{x}(t) + B\bar{u}(t) + D\bar{\omega}(t)\\ &= \bar{A}\bar{x}(t) + B\bar{u}_k(t) + D\bar{\omega}_k(t)\\ &\quad + B\big(\bar{u}(t) - \bar{u}_k(t)\big) + D\big(\bar{\omega}(t) - \bar{\omega}_k(t)\big),\end{aligned} \tag{35}$$

where $\bar{u}(t)$ is the behavior policy applied to the system for data generation, $\bar{\omega}(t)$ is the actual disturbance affecting the system, $\bar{u}_k(t) = K_{u,k}\bar{x}(t)$ is the control policy, $\bar{\omega}_k(t) = K_{\omega,k}\bar{x}(t)$ is the disturbance policy, and, lastly, k indicates the iteration of the algorithm as it is a recursive algorithm. Data generated by the behavior policy is used to evaluate and update control and disturbance policies.

In addition, the value function can be rewritten as

$$V_k\big(\bar{x}(t)\big) = \bar{x}^T(t)P_k\bar{x}(t), \tag{36}$$

and its differentiation can be written as

$$\frac{dV_k\big(\bar{x}(t)\big)}{d\bar{x}(t)} = 2P_k\bar{x}(t), \tag{37}$$

$$\frac{dV_k\big(\bar{x}(t)\big)}{d\bar{x}(t)}\frac{d\bar{x}(t)}{dt} = \dot{V}_k\big(\bar{x}(t)\big) = 2P_k\bar{x}(t)\frac{d\bar{x}(t)}{dt}, \tag{38}$$

which can be further represented as

$$\begin{aligned}\dot{V}_k\big(\bar{x}(t)\big) &= \big(2P_k\bar{x}(t)\big)^T\big(A\bar{x}(t) + B\bar{u}_k(t) + D\bar{\omega}_k(t)\big)\\ &\quad + \big(2P_k\bar{x}(t)\big)^T B\big(\bar{u}(t) - \bar{u}_k(t)\big)\\ &\quad + \big(2P_k\bar{x}(t)\big)^T D\big(\bar{\omega}(t) - \bar{\omega}_k(t)\big).\end{aligned} \tag{39}$$

The Bellman equation (16) can be rewritten as

$$H\big(P_k, \bar{u}_k(t), \bar{\omega}_k(t)\big) = \big(2P_k\bar{x}(t)\big)^T\big(\bar{A}\bar{x}(t) + B\bar{u}_k(t) + D\bar{\omega}_k(t)\big) + \bar{x}^T(t)Q\bar{x}(t) + \bar{u}_k^T(t)R\bar{u}_k(t) - \gamma^2\bar{\omega}_k^T(t)\bar{\omega}_k(t) = 0, \tag{40}$$

Algorithm 1 Online off-Policy Model-Free Integral Reinforcement Learning for Solving the H_∞ Regulation Problem With a Guaranteed Convergence Rate

Definitions:

- $F_k = ||P_k - P_{k-1}|| + ||\bar{u}_{k+1}(t) - \bar{u}_k(t)|| + ||\bar{\omega}_{k+1}(t) - \bar{\omega}_k(t)||$
- η = predetermined error bound

Initialization: $k \leftarrow 0$

Step 1. Regarding (10), apply an initial stabilizing control input $u_0(t)$ for collecting the required data (system states) in N sample times.

Step 2. Given $u_k(t)$ and $\omega_k(t)$, solve the Bellman equation (45) for P_k, $K_{u,k+1}$, and $K_{\omega,k+1}$ concurrently.

Step 3. If $F_k < \eta$, stop; otherwise, set $k = k + 1$ and return to Step 2.

Step 4. Set $u^*(t) = u_k(t)$ and $\omega^*(t) = \omega_k(t)$ on convergence.

and with regard to (17) and (18), $\bar{u}_{k+1}(t)$ and $\bar{\omega}_{k+1}(t)$ can be determined as

$$\bar{u}_{k+1}(t) = -R^{-1}B^T P_k \bar{x}(t), \tag{41}$$

$$\bar{\omega}_{k+1}(t) = \frac{1}{\gamma^2} D^T P_k \bar{x}(t). \tag{42}$$

Therefore, (39) can be rewritten as

$$\begin{aligned}\dot{V}_k\big(\bar{x}(t)\big) = &-\bar{x}^T(t)Q\bar{x}(t) - \bar{u}_k^T(t)R\bar{u}(t) \\ &+ \gamma^2\bar{\omega}_k^T(t)\bar{\omega}(t) - 2\bar{u}_{k+1}^T(t)R\big(\bar{u}(t) - \bar{u}_k(t)\big) \\ &+ 2\gamma^2\bar{\omega}_{k+1}^T(t)\big(\bar{\omega}(t) - \bar{\omega}_k(t)\big).\end{aligned} \tag{43}$$

In order to determine the Bellman equation for the IRL algorithm, the integral of (43) over the interval $[t, t + \Delta]$ is calculated as follows:

$$\begin{aligned}&V_k\big(\bar{x}(t+\Delta)\big) - V_k\big(\bar{x}(t)\big) \\ &= -\int_t^{t+\Delta} \Big(\bar{x}^T(\tau)Q\bar{x}(\tau) + \bar{u}_k^T(\tau)R\bar{u}(\tau) \\ &\quad - \gamma^2\bar{\omega}_k^T(\tau)\bar{\omega}(\tau)\Big)\,d\tau \\ &\quad - 2\int_t^{t+\Delta} \bar{u}_{k+1}^T(\tau)R\big(\bar{u}(\tau) - \bar{u}_k(\tau)\big)\,d\tau \\ &\quad + 2\gamma^2\int_t^{t+\Delta} \bar{\omega}_{k+1}^T(\tau)\big(\bar{\omega}(\tau) - \bar{\omega}_k(\tau)\big)\,d\tau.\end{aligned} \tag{44}$$

Now, with the consideration of (8), (12), (34), (40), (41), and (42), one can rewrite (44) as

$$\begin{aligned}&e^{2\alpha(t+\Delta)}x^T(t+\Delta)P_k x(t+\Delta) - e^{2\alpha t}x^T(t)P_k x(t) \\ &= -\int_t^{t+\Delta} e^{2\alpha\tau}\Big(x^T(\tau)Qx(\tau) + \big(K_{u,k}x(\tau)\big)^T R\big(K_{u,k}x(\tau)\big) \\ &\quad - \gamma^2\big(K_{\omega,k}x(\tau)\big)^T\big(K_{\omega,k}x(\tau)\big)\Big)\,d\tau \\ &\quad - 2\int_t^{t+\Delta} e^{2\alpha\tau}\Big(\big(K_{u,k+1}x(\tau)\big)^T R\big(u(\tau) - K_{u,k}x(\tau)\big)\Big)\,d\tau \\ &\quad + 2\gamma^2\int_t^{t+\Delta} e^{2\alpha\tau}\Big(\big(K_{\omega,k+1}x(\tau)\big)^T R\big(\omega(\tau) \\ &\quad - K_{\omega,k}x(\tau)\big)\Big)\,d\tau.\end{aligned} \tag{45}$$

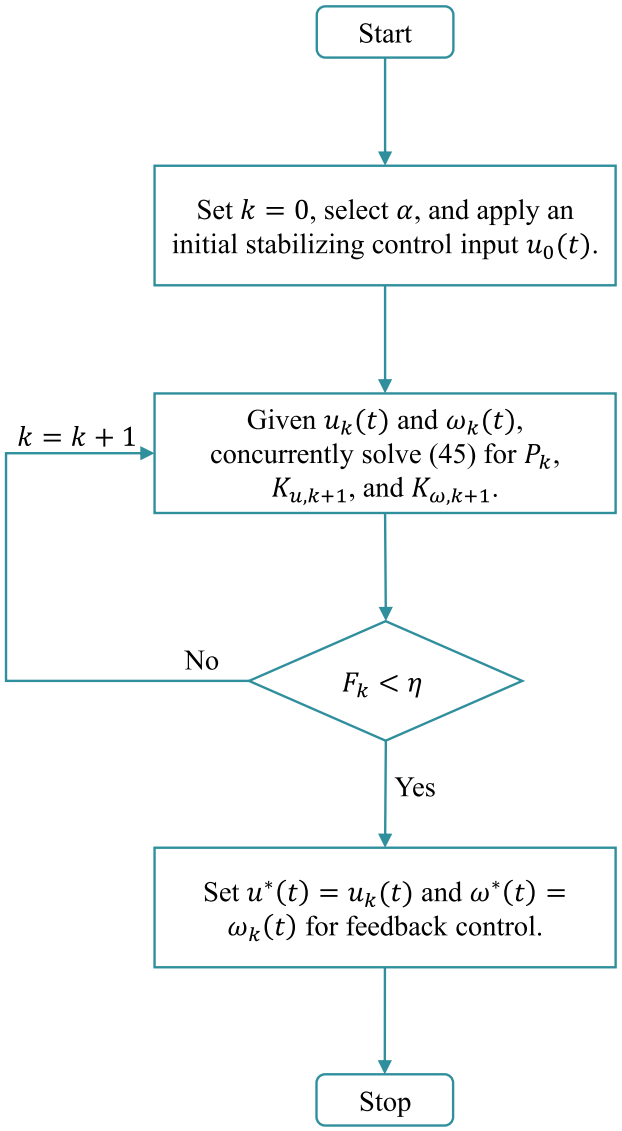

FIGURE 1. Flowchart diagram of Algorithm 1.

Now that the preliminaries of the online off-policy model-free IRL algorithm are completed, the step-by-step procedure of the algorithm can be demonstrated in Algorithm 1. A graphic demonstration of the Algorithm 1 is also depicted in Fig. 1 as a flowchart diagram. Algorithm 1 employs the IRL Bellman equation (45) to iteratively solve the Bellman equation (16). The online implementation of the Algorithm 1 employs least squares and is similar to the practice described in [28], thus omitted.

Theorem 2: Algorithm 1 ensures the convergence of $\{P_k, K_{u,k+1}, K_{\omega,k+1}\}_{k=1}^{\infty}$ to $\{P^*, K_u^*, K_\omega^*\}$, in which P^* is the unique solution to ARE (19).

Proof of Theorem 2: The proof follows the practice described in [39] and is therefore omitted.

V. NUMERICAL EVALUATIONS

In this section, the effectiveness of the proposed approach is illustrated through two examples: a four-state stable system

TABLE 1. Comparison of the effect of two α values for Example 1.

α	Model-Based	Model-Free (IRL)	Eigenvalues
1	$P^* = \begin{bmatrix} 1.1132 & 1.1240 & 0.1521 & 2.6890 \\ 1.1240 & 1.5443 & 0.2389 & 2.4261 \\ 0.1521 & 0.2389 & 0.0708 & 0.3151 \\ 2.6890 & 2.4261 & 0.3151 & 13.1956 \end{bmatrix}$ $K_u^* = \begin{bmatrix} -2.0896 & -3.2815 & -0.9724 & -4.3284 \end{bmatrix}$	$P_{10} = \begin{bmatrix} 1.0507 & 1.0563 & 0.1329 & 2.4659 \\ 1.0563 & 1.5092 & 0.2234 & 2.1750 \\ 0.1329 & 0.2234 & 0.0353 & 0.2766 \\ 2.4659 & 2.1750 & 0.2766 & 12.1641 \end{bmatrix}$ $K_{u,10} = \begin{bmatrix} -2.0090 & -3.2379 & -0.8550 & -4.0319 \end{bmatrix}$	$\begin{bmatrix} -19.9798 + 0.0000i \\ -2.0955 + 0.0000i \\ -4.3385 + 3.4478i \\ -4.3385 - 3.4478i \end{bmatrix}$
3	$P^* = \begin{bmatrix} 5.5833 & 4.4777 & 0.5502 & 25.1274 \\ 4.4777 & 4.3641 & 0.5954 & 17.9009 \\ 0.5502 & 0.5954 & 0.1210 & 2.0896 \\ 25.1274 & 17.9009 & 2.0896 & 146.7013 \end{bmatrix}$ $K_u^* = \begin{bmatrix} -7.5569 & -8.1784 & -1.6616 & -28.7027 \end{bmatrix}$	$P_{10} = \begin{bmatrix} 5.3867 & 4.3450 & 0.5364 & 23.9493 \\ 4.3450 & 4.2687 & 0.5849 & 17.1507 \\ 0.5364 & 0.5849 & 0.1198 & 2.0155 \\ 23.9493 & 17.1507 & 2.0155 & 138.7026 \end{bmatrix}$ $K_{u,10} = \begin{bmatrix} -7.3433 & -8.0170 & -1.6424 & -27.5309 \end{bmatrix}$	$\begin{bmatrix} -20.5426 + 0.0000i \\ -5.7569 + 0.0000i \\ -6.9597 + 3.4883i \\ -6.9597 - 3.4883i \end{bmatrix}$

and a two-state unstable system. Algorithm 1, which is a model-free algorithm, is used to solve the H_∞ regulation problem, and the results are compared with the model-based solution obtained by (19) and (20).

This paper considers two different types of disturbance, (i) sawtooth waveform, (ii) sinusoidal waveform (both are plotted in Fig. 2). The sawtooth disturbance has a bounded amplitude range and constant frequency, and it can be presented by the following equation

$$\omega_1(t) = \begin{cases} 0, & 0 \le t < 4 \\ t-4, & 4 \le t < 6 \\ t-6, & 6 \le t < 8 \\ t-8, & 8 \le t < 10. \end{cases} \tag{46}$$

The sinusoidal waveform, on the other hand, has increasing amplitude and frequency, and is demonstrated through the following equation:

$$\omega_2(t) = \begin{cases} 0, & 0 \le t < 4 \\ \sin(\pi t), & 4 \le t < 6 \\ 2\sin(2\pi t), & 6 \le t < 8 \\ 3\sin(5\pi t), & 8 \le t < 10. \end{cases} \tag{47}$$

A. AN EXAMPLE OF STABLE SYSTEMS

The first example is a stable system with its state-space matrices given as follows:

$$A = \begin{bmatrix} 0 & 8 & 0 & 0 \\ 0 & -3.66 & 3.66 & 0 \\ -6.86 & 0 & -13.736 & -13.736 \\ 0.6 & 0 & 0 & 0 \end{bmatrix},$$

$$x_0 = \begin{bmatrix} 1 \\ -1 \\ -1.5 \\ 1.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 13.736 \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

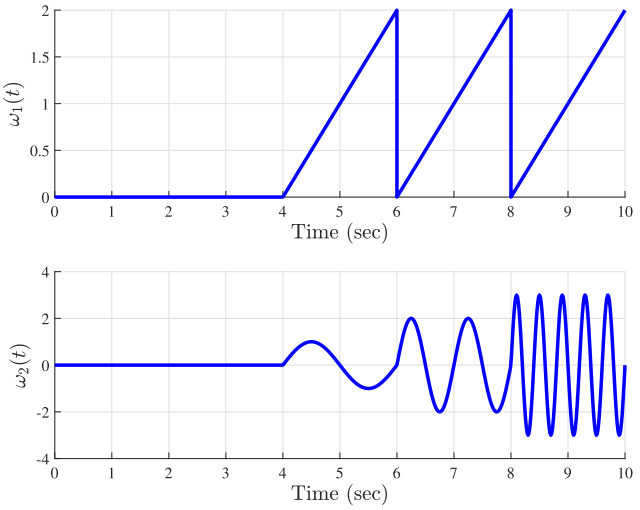

FIGURE 2. The disturbance waveforms.

The eigenvalues of this system are -14.8474, $-0.5260 \pm 3.2531i$, and -1.4967. Also, the weight matrices for this example are $Q = I_{4\times4}$ and $R = I_{1\times1}$ with $\gamma = 10$.

In order to show the efficacy of the proposed IRL algorithm, a comparison between the model-based method and the model-free method is given by comparing the solutions that both methods provide to the H_∞ regulation problem. Besides, two values for α are considered for this system in order to compare the effect of α on the speed of regulation and disturbance attenuation.

Fig. 3 shows the convergence of the model-free solutions to the model-based solutions for the two values of α. More details are provided in Table 1. According to this table, the model-free method achieves the same solution as the model-based (the table shows the solution achieved in the 10th iteration). Fig. 4 and Fig. 5 depict the response of the states

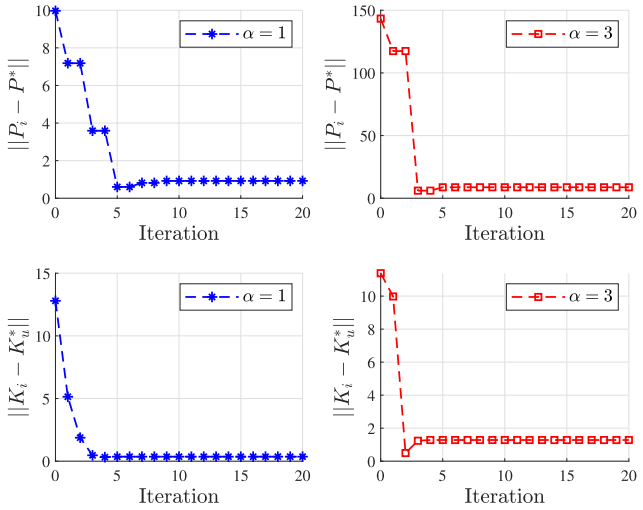

FIGURE 3. Convergence of P_i and $K_{u,i}$ to their optimal values for $\alpha = 1$ and $\alpha = 3$ for Example 1.

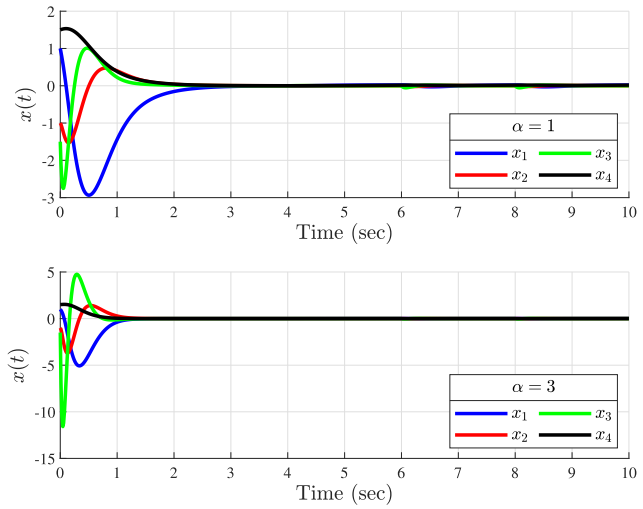

FIGURE 4. System states for Example 1 under the effect of $\omega_1(t)$ for $\alpha = 1$ and $\alpha = 3$.

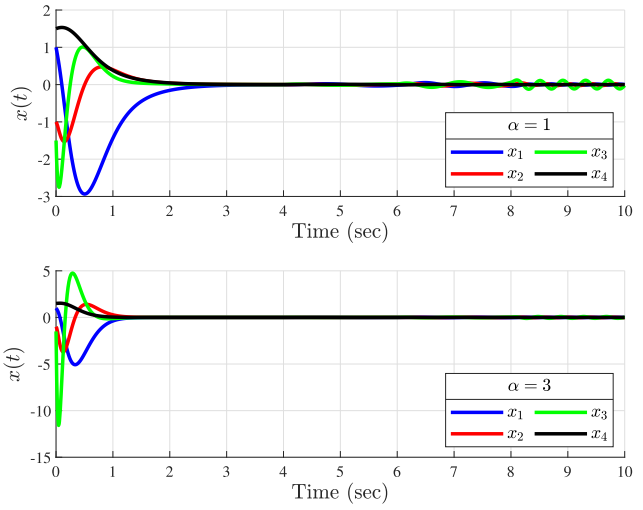

FIGURE 5. System states for Example 1 under the effect of $\omega_2(t)$ for $\alpha = 1$ and $\alpha = 3$.

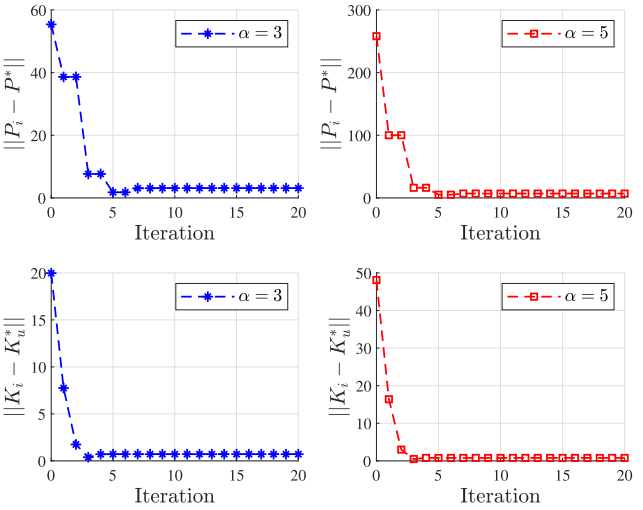

FIGURE 6. Convergence of P_i and $K_{u,i}$ to their optimal values for $\alpha = 3$ and $\alpha = 5$ for Example 2.

for different values of α under both disturbances. Regarding Table 1, Fig. 4, and Fig. 5, it can be seen that the eigenvalues are always on the left half of the line $s = -\alpha$, and by increasing α, the speed of regulation increases and also disturbance is attenuated in a better way.

In addition, note that $K_{u,0}$ values for $\alpha = 1$ and $\alpha = 3$ have been selected as $\begin{bmatrix}-6 & -10 & -10 & -9\end{bmatrix}$ and $\begin{bmatrix}-8 & -7 & -1 & -40\end{bmatrix}$, respectively.

B. AN EXAMPLE OF UNSTABLE SYSTEMS

The second example is an unstable system with its state-space matrices given as follows:

$$A = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}, \quad x_0 = \begin{bmatrix}-2\\ 3\end{bmatrix}, \quad B = \begin{bmatrix}0\\ 2\end{bmatrix}, \quad D = \begin{bmatrix}0\\ 1\end{bmatrix}.$$

The eigenvalues of this system are -1 and $+1$. Also, the weight matrices for this example are $Q = I_{2\times 2}$ and $R = I_{1\times 1}$ with $\gamma = 10$.

Similar to the previous example, the same comparison between the model-based and model-free methods is provided here with two different values for α.

Fig. 6 shows the convergence of the model-free solutions to the model-based solutions for the two values of α. Model-based and model-free solutions are also compared in Table 2 (the table shows the solution achieved in the 10th iteration). Fig. 7 and Fig. 8 depict the response of the states for different values of α under both disturbances. Furthermore, it can be observed that the eigenvalues always lie within the left half of the line $s = -\alpha$, and by increasing α, the speed of regulation increases and the disturbance is attenuated in a better way.

In addition, note that $K_{u,0}$ values for $\alpha = 3$ and $\alpha = 5$ have been selected as $\begin{bmatrix}-40 & -8\end{bmatrix}$ and $\begin{bmatrix}-100 & -14\end{bmatrix}$, respectively.

TABLE 2. Comparison of the effect of two α values for Example 2.

α	Model-Based	Model-Free (IRL)	Eigenvalues
3	$P^* = \begin{bmatrix} 63.9060 & 10.0696 \\ 10.0696 & 3.1733 \end{bmatrix}$ $K_u^* = \begin{bmatrix} -20.1392 & -6.3466 \end{bmatrix}$	$P_{10} = \begin{bmatrix} 60.6824 & 9.7371 \\ 9.7371 & 3.1338 \end{bmatrix}$ $K_{u,10} = \begin{bmatrix} -19.3843 & -6.2545 \end{bmatrix}$	$\begin{bmatrix} -5.3816 \\ -7.2800 \end{bmatrix}$
5	$P^* = \begin{bmatrix} 266.1029 & 26.0816 \\ 26.0816 & 5.1125 \end{bmatrix}$ $K_u^* = \begin{bmatrix} -52.1632 & -10.2249 \end{bmatrix}$	$P_{10} = \begin{bmatrix} 258.5788 & 25.6577 \\ 25.6577 & 5.0759 \end{bmatrix}$ $K_{u,10} = \begin{bmatrix} -51.2530 & -10.1469 \end{bmatrix}$	$\begin{bmatrix} -9.2188 \\ -11.1800 \end{bmatrix}$

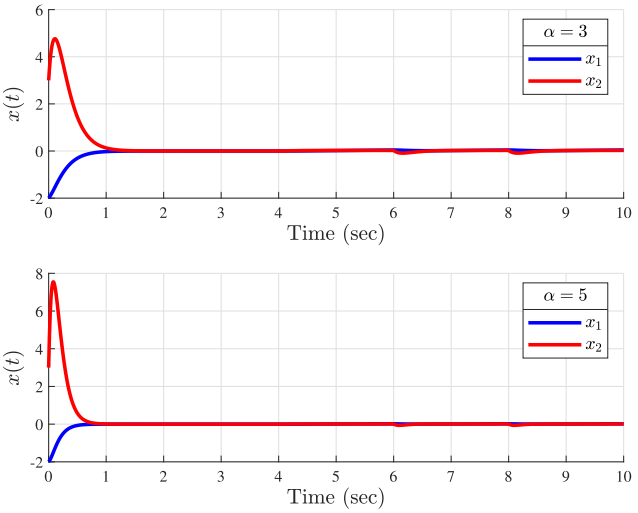

FIGURE 7. System states for Example 2 under the effect of $\omega_1(t)$ for $\alpha = 3$ and $\alpha = 5$.

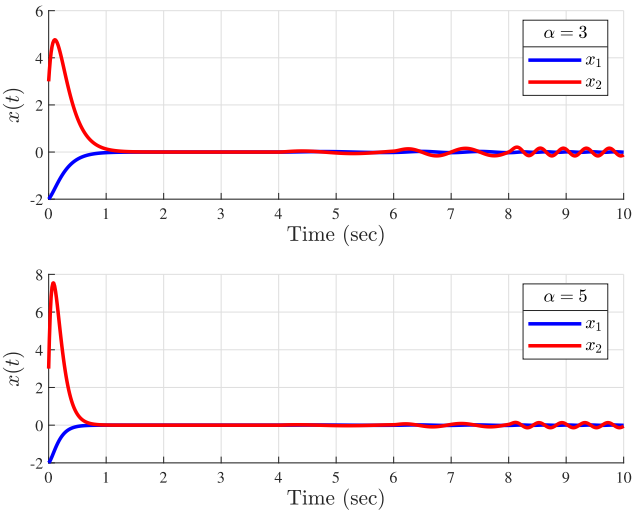

FIGURE 8. System states for Example 2 under the effect of $\omega_2(t)$ for $\alpha = 3$ and $\alpha = 5$.

VI. CONCLUSION AND FUTURE WORK

In this paper the problem of designing a model-free H_∞ controller for the state regulation of continuous-time linear systems with a guaranteed convergence rate in the presence of disturbance is studied. Both regulation and disturbance attenuation problems are addressed in a single unified cost function by formulating it as a two-player zero-sum game optimization problem. Then the optimization problem is solved using the associated algebraic Riccati equation, which provides a model-based solution. A novel model-free integral reinforcement learning algorithm was developed to learn the solution in real-time using no prior knowledge of the system dynamics. The results show that the algorithm substantially attenuates the adverse effect of the disturbance on the system performance and also guarantees a predefined rate of regulation. The efficacy of the proposed method for both a stable and unstable systems is verified numerically. The approach proposed in this paper may serve as an effective tool to study the optimal control design problem with a guaranteed convergence rate for a wide range of applications such as robotics, industrial manufacturing systems, process control, and so forth. In future, this development could be extended for systems with input delay. In addition, developing a similar method for systems with bounded inputs or constrained states could be an important area of research.

REFERENCES

[1] C. Xu, M. Li, and F. Pan, "The system design and LQR control of a two-wheels self-balancing mobile robot," in *Proc. Int. Conf. Electr. Control Eng.*, Sep. 2011, pp. 2786–2789.
[2] 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 Syst.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.
[3] H. Jiang, H. Zhang, Y. Luo, and X. Cui, "H_∞ control with constrained input for completely unknown nonlinear systems using data-driven reinforcement learning method," *Neurocomputing*, vol. 237, pp. 226–234, May 2017.
[4] S. A. A. Rizvi and Z. Lin, "Output feedback Q-learning control for the discrete-time linear quadratic regulator problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1523–1536, May 2019.
[5] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 32, no. 2, pp. 140–153, May 2002.
[6] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, May 2005.
[7] H. Li and D. Liu, "Optimal control for discrete-time affine non-linear systems using general value iteration," *IET Control Theory Appl.*, vol. 6, no. 18, pp. 2725–2736, Dec. 2012.

[8] X. Yang, D. Liu, and Q. L. Wei, "Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming," *IET Control Theory Appl.*, vol. 8, no. 16, pp. 1676–1688, Nov. 2014.
[9] D. Zhao and Y. Zhu, "MEC—A near-optimal online reinforcement learning algorithm for continuous deterministic systems," *IEEE Trans. Neural Netw. Learn. Sys.*, vol. 26, no. 2, pp. 346–356, Feb. 2015.
[10] Y. Zhu, D. Zhao, and D. Liu, "Convergence analysis and application of fuzzy-HDP for nonlinear discrete-time HJB systems," *Neurocomputing*, vol. 149, pp. 124–131, Feb. 2015.
[11] I. Sanusi, A. Mills, T. Dodd, and G. Konstantopoulos, "Online optimal and adaptive integral tracking control for varying discrete-time systems using reinforcement learning," *Int. J. Adapt. Control Signal Process.*, vol. 34, no. 8, pp. 971–991, Aug. 2020.
[12] L. Guo, S. A. A. Rizvi, and Z. Lin, "Optimal control of a two-wheeled self-balancing robot by reinforcement learning," *Int. J. Robust Nonlinear Control*, vol. 31, no. 6, pp. 1885–1904, Apr. 2021.
[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
[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, May 2010.
[15] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.
[16] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.
[17] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 840–853, Mar. 2016.
[18] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.
[19] H.-G. Zhang, X. Zhang, Y.-H. Luo, and J. Yang, "An overview of research on adaptive dynamic programming," *Acta Autom. Sinica*, vol. 39, no. 4, pp. 303–311, Apr. 2013.
[20] A. Perrusquía and W. Yu, "Robust control under worst-case uncertainty for unknown nonlinear systems using modified reinforcement learning," *Int. J. Robust Nonlinear Control*, vol. 30, no. 7, pp. 2920–2936, 2020.
[21] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Aug. 2009.
[22] D. Vrabie and F. 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, Apr. 2009.
[23] 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, Feb. 2009.
[24] 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, Jul. 2014.
[25] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis, "Online adaptive algorithm for optimal control with integral reinforcement learning," *Int. J. Robust Nonlinear Control*, vol. 24, no. 17, pp. 2686–2710, Nov. 2014.
[26] 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, Nov. 2014.
[27] 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.
[28] F. A. Yaghmaie, S. Gunnarsson, and F. L. Lewis, "Output regulation of unknown linear systems using average cost reinforcement learning," *Automatica*, vol. 110, Dec. 2019, Art. no. 108549.
[29] Q. Wang, "Integral reinforcement learning control for a class of high-order multivariable nonlinear dynamics with unknown control coefficients," *IEEE Access*, vol. 8, pp. 86223–86229, 2020.
[30] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.
[31] Y. Yang, Z. Guo, H. Xiong, D. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3735–3747, Dec. 2019.
[32] J. Zhao and P. Vishal, "Neural network-based optimal tracking control for partially unknown discrete-time non-linear systems using reinforcement learning," *IET Control Theory Appl.*, vol. 15, no. 2, pp. 260–271, Jan. 2021.
[33] B. Pang and Z.-P. Jiang, "Robust reinforcement learning: A case study in linear quadratic regulation," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 10, pp. 9303–9311. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17122
[34] 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, Dec. 2016.
[35] S. A. A. Rizvi and Z. Lin, "Reinforcement learning-based linear quadratic regulation of continuous-time systems using dynamic output feedback," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4670–4679, Nov. 2020.
[36] H. Modares, F. L. Lewis, and Z.-P. Jiang, "Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2401–2410, Nov. 2016.
[37] R. Moghadam and F. L. Lewis, "Output-feedback H_∞ quadratic tracking control of linear systems using reinforcement learning," *Int. J. Adapt. Control Signal Process.*, vol. 33, no. 2, pp. 300–314, Feb. 2019.
[38] G. Wang, B. Luo, and S. Xue, "Integral reinforcement learning-based optimal output feedback control for linear continuous-time systems with input delay," *Neurocomputing*, vol. 460, pp. 31–38, Oct. 2021.
[39] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2550–2562, Oct. 2015.
[40] H. Modares, B. Kiumarsi, K. G. Vamvoudakis, and F. L. Lewis, "Adaptive H_∞ tracking control of nonlinear systems using reinforcement learning," in *Adaptive Learning Methods for Nonlinear System Modeling*, D. Comminiello and J. C. Príncipe, Eds. Oxford, U.K. Butterworth-Heinemann, 2018, pp. 313–333.
[41] L. Cui, W. Qu, L. Wang, Y. Luo, and Z. Wang, "Event-triggered H_∞ tracking control of nonlinear systems via reinforcement learning method," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[42] S. Xue, B. Luo, and D. Liu, "Integral reinforcement learning based event-triggered control with input saturation," *Neural Netw.*, vol. 131, pp. 144–153, Nov. 2020.
[43] A. Mishra and S. Ghosh, "H_∞ tracking control via variable gain gradient descent-based integral reinforcement learning for unknown continuous time non-linear system," *IET Control Theory Appl.*, vol. 14, no. 20, pp. 3476–3489, Dec. 2020.
[44] K. Zhang and S.-L. Ge, "Adaptive optimal control with guaranteed convergence rate for continuous-time linear systems with completely unknown dynamics," *IEEE Access*, vol. 7, pp. 11526–11532, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8610005
[45] S. E. Razavi, M. A. Moradi, S. Shamaghdari, and M. B. Menhaj, "Adaptive optimal control of unknown discrete-time linear systems with guaranteed prescribed degree of stability using reinforcement learning," *Int. J. Dyn. Control*, vol. 10, no. 3, pp. 870–878, Jun. 2022.

ALI RAHDARIAN received the B.Sc. degree in electrical engineering from the Jundi Shapur University of Technology, Dezful, Iran, in 2018, and the M.Sc. degree in electrical engineering from the Iran University of Science and Technology, Tehran, in 2020. He is currently a Research Assistant with the Department of Electrical Engineering, Iran University of Science and Technology. His research interests include reinforcement learning, game theory, and control theory.

D. SADRIAN ZADEH (Student Member, IEEE) received the B.Sc. degree in control engineering from the Petroleum University of Technology, Ahvaz, Iran, in 2018, and the M.Sc. degree in control engineering from the University of Tehran, Tehran, Iran, in 2021. He is currently pursuing the Ph.D. degree with the University of Waterloo, ON, Canada. His research interests include data fusion, machine learning, control theory, microgrids, and autonomous vehicles.

SAEED SHAMAGHDARI received the Ph.D. degree in electrical engineering from the Amirkabir University of Technology, Tehran, Iran. He is currently an Associate Professor at the School of Electrical Engineering, Iran University of Science and Technology, Tehran. His research interests include control systems, such as model predictive control, convex optimization, and reinforcement learning.

BEHZAD MOSHIRI (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in control systems engineering from the University of Manchester Institute of Science and Technology (UMIST), in 1987 and 1991, respectively. He is currently a Full Professor in control systems engineering at the School of ECE, University of Tehran. He has also been an Adjunct Professor with the Department of ECE, University of Waterloo, Canada, since 2014. He has been working as the Chairperson of the IEEE Control System Chapter in the IEEE Iran Section, since December 2018. He has been a member of the International Society of Information Fusion (ISIF), since 2002, and has also been a member of the Waterloo AI Institute, since 2019. He is the author/coauthor of more than 360 articles, including more than 130 journal articles and more than 20 book chapters. His research interests include advanced industrial control, advanced instrumentation systems, data fusion theory, and its applications in areas, such as robotics, process control, mechatronics, information technology (IT), intelligent transportation systems (ITS), bioinformatics, and financial engineering.

ALLAHYAR MONTAZERI received the B.S. degree in electrical engineering from Tehran University, Tehran, Iran, in 2000, and the M.Sc. and Ph.D. degrees in electrical engineering from the Iran University of Science and Technology, Tehran, in 2002, and 2009, respectively. From 2010 to 2011, he was a Research Fellow with the Fraunhofer Institute, Germany, and then carried on his research with Fraunhofer IDMT and Control Engineering Group, Ilmenau University, Germany, from 2011 to 2013. He has been the Visiting Research Scholar with the Control Engineering Group, ETH Zurich, Switzerland, and the Chemical Engineering Group, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. Since 2013, he has been an Assistant Professor with the Engineering Department, Lancaster University, U.K. His research interests include the areas on control theory and digital signal processing. Particularly, he is interested in adaptive signal processing and control, robust control, linear and nonlinear system identification, estimation theory, and evolutionary computing and optimization with applications in active noise and vibration control systems, robotics. He was a recipient of the European Research Consortium on Informatics and Mathematic (ERCIM) and Humboldt research awards, in 2010 and 2011, respectively. He is also a fellow of Higher Education Academy. His research is funded by different councils and industries in U.K., such as Engineering and Physical Research Council, Sellafield Ltd., National Nuclear Laboratory, and Nuclear Decommissioning Authority. He is currently working on IFAC Technical Committees ''Adaptive and Learning Systems'' and ''Modeling, Identification, and Signal Processing''.