Leader-Follower Based Online Reinforcement Learning Algorithm in Problem with Hierarchy Decision Makers

Xiaoqian Li, Zonglei Jing, and Peijun Ju

***Abstract*—This paper presents an online integral reinforcement learning (RL) solution for problems with hierarchy decision makers. Specifically, we reformulate this model as a leader-follower game, in which control input and deterministic disturbance act as decision makers at different levels of hierarchy: The control input plays the role of the leader, while the disturbance plays the role of the follower. The main contributions of this paper can be summarized as follows. First, we introduce online RL to deal with systems that have partially unknown information, meaning that accurate dynamic information is not required. Second, we solve the leader-follower coupled Hamilton-Jacobi (HJ) and Riccati equations approximately online using the derived algorithm. Third, we provide turning laws for cost functions and controllers, which ensure closed-loop stability simultaneously.**

***Index Terms*—Online reinforcement learning algorithm, leader-follower game, hierarchy decision makers**

I. INTRODUCTION

Games have been widely used in complex systems involving multiple decision-making players [1]. The behaviors of these players depend on the information that each has compiled about the others, and each seeks to optimize its individual cost. Depending on their roles in practical applications, the involved players may have cooperative or non-cooperative objectives in scenarios that include economics [2, 3] and engineering [4, 5]. Multi-player games can be classified into Nash games or Stackelberg games based on the number of decision-making hierarchies involved. In Nash games, different players select decisions simultaneously without knowing those made by others [6]. Conversely, Stackelberg games involve multi-level decision making, with lower hierarchy players constrained by higher hierarchy players. The mentioned players in Stackelberg game are called the leaders and the followers [7]. Leaders, who are in the higher hierarchy, have complete information on decisions of others and make decisions first. Followers, who are in the lower hierarchy, react to the decisions of leaders. Control methods for both Nash and Stackelberg games have been applied to dynamical systems to study coordination, conflict, and control among different players [8]. The Stackelberg game has been used in various fields, including multi-input dynamic systems and network communication [9, 10]. In Refs. [11, 12], Stackelberg games with open-loop strategies were introduced to handle multi-layered optimization problems with one leader and one follower. These problems are typically formulated as mixed H_2/H_∞ control, where incomplete system information is considered for the reason of the complex uncertain environments.

Various approaches have been employed to address the Stackelberg game. Xu et al. [13] presented sufficient conditions and necessary conditions for leader-follower games. An algorithm named off-policy reinforcement learning (RL) was applied in Ref. [14] to derive solutions for nonzero-sum games with multi-player. The approach established in Ref. [14] is offline, which places greater demands on system dynamics. However, due to the high uncertainty of the environment, obtaining accurate information on system dynamics poses a challenge. Moreover, players under this control policy cannot adjust their aims in real time. These issues highlight the challenges involved in studying the Stackelberg game.

Reinforcement learning is a promising method for studying multi-player games [15–17], allowing agents to gradually improve their performances by learning new behaviors online with little information about other players. Researchers in control systems have been motivated by the advantages of reinforcement learning and have applied it to find optimal solutions for systems in differential or difference norms. In Ref. [18], online policy was introduced to handle optimization problem with unknown information. Adaptive feedback form controllers were derived by using natural decision methods in Refs. [19, 20]. An optimal control policy for constrained input systems with unknown information was established using policy iteration methods in Ref. [21]. Integral RL was also applied to address tracking problems with constrained inputs [22].

Motivated by previous research on the advantages and outcomes of reinforcement learning, we introduce an online integral RL method to find optimal solutions for mixed H_2/H_∞ problems. In this type of control problem, both control input and deterministic disturbance shape the Stackelberg game, with the control input serving as leader and the disturbance as follower. This paper presents several contributions. First, online RL is introduced to address systems with partially unknown information, eliminating the

Manuscript received: 6 November 2023; revised: 21 December 2023; accepted: 9 January 2024. *(Corresponding author: Xiaoqian Li.)*

Citation: X. Li, Z. Jing, and P. Ju, Leader-follower based online reinforcement learning algorithm in problem with hierarchy decision makers, *IJICS*, 2024, 29(1), 48–58.

Xiaoqian Li and Peijun Ju are with School of Mathematics and Statistics, Taishan University, Taian 271000, China (e-mail: xiaoqianli0324@126.com; jpj615@163.com).

Zonglei Jing is with State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China (e-mail: jingzonglei@163.com).

Digital Object Identifier 10.62678/IJICS202403.10114

need for accurate dynamic information. Second, the leader-follower coupled Hamilton-Jacobi (HJ) and Riccati equations are solved approximately using our derived algorithm. Third, turning laws for cost functions and controllers are provided, ensuring the closed-loop stability.

The rest of this paper is organized as follows. In Section II, we formulate the Stackelberg game for mixed H_2/H_∞ control problems. Section III presents policy iteration methods for Stackelberg games, and Section IV provides value function approximations for these games. We then introduce an online solution for Stackelberg games using RL in Section V. In Section VI, the simulation results verify the validity and feasibility of the proposed method. Finally, we conclude this paper in Section VII.

II. Problem Statement

In this section, we discuss a two-player differential game.

$$\dot{x}(t) = Ax(t) + B_1u(t) + B_2w(t) \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state and $x(t_0) = x_0$, t_0 is the initial time and x_0 is the state at initial time; $u(t) \in \mathbb{R}^s$ is the control input; $w(t) \in \mathbb{R}^m$ is the deterministic noise; and t indicates time. $\mathbb{R}^n$ is the family of n-dimensional column vectors, which is the same as $\mathbb{R}^s$ and $\mathbb{R}^m$, n, s, and m are finite positive integers. A, B_1, and B_2 are constant matrices with appropriate dimensions.

Consider the cost functions

$$J_1(x_0,u) = \int_0^\infty (x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru)\mathrm{d}t = \int_0^\infty r_1(x(t),u(t))\mathrm{d}t \tag{2}$$

$$J_2(x_0,u,w) = \int_0^\infty (x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru - \gamma^2\|w\|^2)\mathrm{d}t = \int_0^\infty r_2(x(t),u(t),w(t))\mathrm{d}t \tag{3}$$

where Q is the weighting matrix of state variables, $Q \geqslant 0$, and R is the weighting matrix of control variables, $R = R^{\mathrm{T}} > 0$. γ is the perturbation factor and γ^* means perturbation factor when the optimal control law $u(t)$ is invoked, and satisfy $\gamma \geqslant \gamma^* \geqslant 0$. $\|w\|^2$ means H2-norm. For the sake of simplicity, x is equal to $x(t)$, u is equal to $u(t)$, similar treatment is provided elsewhere below.

Definition 1 The existence of the adversarial input $w(t)$ is measured by the perturbed operator L with its H_∞ norm defined by

$$\|L\|_{H_\infty} := \sup_{w\in\mathbb{R}^m} \frac{\left[\int_0^\infty (x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru)\mathrm{d}t\right]^{1/2}}{\left[\int_0^\infty (w^{\mathrm{T}}w)\mathrm{d}t\right]^{1/2}} < \gamma \tag{4}$$

In this paper, we define the mixed H_2/H_∞ control problem in infinite horizon, where the designed control input $u(t)$ minimizes its corresponding cost function in Eq. (2) under the H_∞-constraint in Eq. (3). In some practical applications, the adversarial input $w(t)$ maximizes energy gain.

We also provide a general definition of Stackelberg games.

Definition 2 For a Stackelberg game, the sets of admissible strategies for leader and follower are $\mathcal{U}$ and $\mathcal{W}$, respectively. The objective of leader is to minimize J_1, while follower aims to maximize J_2.

(1) For a fixed policy $u \in \mathcal{U}$ and a given initial state $x_0 \in \mathbb{R}^n$, the follower chooses an optimal strategy $w^* \in \mathcal{W}$ such that $J_2(x_0,u,w^*) \geqslant J_2(x_0,u,w)$.

(2) Assuming an optimal controller w^* exists, which generally depends on the choice of controller u and the initial state x_0, the leader will select the optimal controller $u \in \mathcal{U}$ to minimize the cost function $J_1(x_0,u,w^*)$ over all $u \in \mathcal{U}$.

Problem 1 We aim to find an optimal feedback strategy pair (u^*,w^*) associated with Eq. (1), for that the following relationships with admissible control pair (u,w) are satisfied.

$$J_1(u^*,w^*) \leqslant J_1(u,w^*) \tag{5}$$

$$J_2(u,w^*) \geqslant J_2(u,w) \tag{6}$$

Remark 1 Formula (6) implies that the follower requests the leader to specify its choice based on the follower's optimal policy. In other words, the leader selects a reaction that is identified as the best response. The leader then chooses an input policy from the admissible policies that minimizes its function index, given the follower's designed policy.

III. Policy Iteration for Stackelberg Game

Different hierarchical levels exist in the system due to the denotations of leader and follower. The discussion will begin with the follower.

The optimal cost function of the follower is defined as

$$V_2^*(x_0,u,w) = \max_w J_2(x_0,u,w) = \max_w \int_t^\infty (x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru - \gamma^2\|w\|^2)\mathrm{d}\tau = \max_w \int_t^\infty r_2(x(t),u(t),w(t))\mathrm{d}\tau \tag{7}$$

The Hamiltonian of the follower is

$$H_2 = x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru - \gamma^2\|w\|^2 + \frac{\partial V_2}{\partial x}^{\mathrm{T}}(Ax + B_1u + B_2w) \tag{8}$$

where $\dfrac{\partial V_2}{\partial x}$ denotes the partial derivative of the value function V_2 with respect to x. To enhance visualization, we introduce the alternative operator notation $\nabla V_i \in \mathbb{R}^n$, $i = 1, 2$ to replace $\dfrac{\partial V_i}{\partial x}$, $i = 1, 2$.

The optimal controller and costate state $\dot{\lambda}_2$ satisfy the following relations

$$w^* = \frac{1}{2}(\gamma^2)^{-1}B_2^{\mathrm{T}}\nabla V_2 \tag{9}$$

$$\dot{\lambda}_2 = -A^{\mathrm{T}}\nabla V_2 - 2Qx \tag{10}$$

The focus now shifts to the leaders. The leader's objective is to minimize Eq. (2) subject to the constraints in Eqs. (1) and (10), which means that the leader should consider the costate of the follower. Then, we have

$$V_1^*(x,u,w^*) = \min_u \int_t^\infty (x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru + y^{\mathrm{T}}\dot{\lambda}_2)\mathrm{d}\tau = \min_u \int_t^\infty (r_1(x(t),u(t)) + y^{\mathrm{T}}\dot{\lambda}_2)\mathrm{d}\tau \tag{11}$$

where V_1^* means the optimal cost function of the leader and y is the costate.

The Hamiltonian of the leader is

$$H_1 = x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru + (\nabla V_1)^{\mathrm{T}}(Ax + B_1u + B_2w) + y^{\mathrm{T}}\dot{\lambda}_2 \quad (12)$$

Similarly, the equilibrium equation and leader's costate equation are derived by Hamilton-Jacobi equation

$$u^* = -\frac{1}{2}R^{-1}B_1^{\mathrm{T}}\nabla V_1 \quad (13)$$

$$\dot{\lambda}_1 = -A^{\mathrm{T}}\nabla V_1 - 2Qx + 2Qy \quad (14)$$

The following Bellman equations are presented in terms of ∇V_1 and ∇V_2 to provide differential equivalents of the value functions for the leader in Eq. (11) and the follower in Eq. (7), where the optimal control policies are inserted into the Hamiltonian.

$$0 = r_1(x, u^*, w^*) + (\nabla V_1)^{\mathrm{T}}(Ax + B_1u^* + B_2w^*) + y^{\mathrm{T}}\dot{\lambda}_2 \quad (15)$$

$$0 = r_2(x, u^*, w^*) + (\nabla V_2)^{\mathrm{T}}(Ax + B_1u^* + B_2w^*) \quad (16)$$

By using the infinite integrals in Eqs. (7) and (11), values evaluated along the system trajectories are solutions to Eqs. (15) and (16). The leader's costate is given by

$$\dot{y} = -(2\gamma^2)^{-1}B_2B_2^{\mathrm{T}}\nabla V_1 + Ay \quad (17)$$

Substituting Eqs. (9) and (13) into Eqs. (15) and (16) yields

$$\begin{aligned} 0 = {} & x^{\mathrm{T}}Qx + \frac{1}{4}(\nabla V_1)^{\mathrm{T}}B_1R^{-1}B_1^{\mathrm{T}}\nabla V_1 - \\ & \frac{1}{4}(\nabla V_2)^{\mathrm{T}}B_2(\gamma^2)^{-1}B_2^{\mathrm{T}}\nabla V_2 + (\nabla V_2)^{\mathrm{T}}(Ax - \\ & \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}\nabla V_1 + \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}\nabla V_2) \end{aligned} \quad (18)$$

$$\begin{aligned} 0 = {} & x^{\mathrm{T}}Qx + \frac{1}{4}(\nabla V_1)^{\mathrm{T}}B_1R^{-1}B_1^{\mathrm{T}}\nabla V_1 + (\nabla V_1)^{\mathrm{T}}(Ax - \\ & \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}\nabla V_1 + \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}\nabla V_2) - \\ & y^{\mathrm{T}}(A^{\mathrm{T}}\nabla V_2 + 2Qx) \end{aligned} \quad (19)$$

The following denotations are introduced to rewrite costate variables

$$\nabla V_1 = P_1x,\ \nabla V_2 = P_2x,\ y = P_3x \quad (20)$$

where P_1, P_2, and P_3 are positive definite symmetric matrices with compatible dimensions. Equations (17)–(19) are rewritten as the following forms

$$\begin{aligned} 0 = {} & Q + \frac{1}{4}P_1^{\mathrm{T}}B_1R^{-1}B_1^{\mathrm{T}}P_1 - \frac{1}{4}P_2^{\mathrm{T}}B_2(\gamma^2)^{-1}B_2^{\mathrm{T}}P_2 + \\ & P_2^{\mathrm{T}}(A - \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}P_1 + \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}P_2) \end{aligned} \quad (21)$$

$$\begin{aligned} 0 = {} & Q + \frac{1}{4}P_1^{\mathrm{T}}B_1R^{-1}B_1^{\mathrm{T}}P_1 + P_1^{\mathrm{T}}(A - \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}P_1 + \\ & \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}P_2) - P_3^{\mathrm{T}}(A^{\mathrm{T}}P_2 + 2Q) \end{aligned} \quad (22)$$

$$\begin{aligned} 0 = {} & -P_3^{\mathrm{T}}(A - \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}P_1 + \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}P_2) - \\ & \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}P_1 + AP_3 \end{aligned} \quad (23)$$

However, solving the Hamilton-Jacobi equations shown in Eqs. (18) and (19) for the leader and follower can be a significant challenge, and in some cases, obtaining solutions may be impossible. To address this issue, a policy iteration algorithm is introduced to approximate the HJ solution by using iteration algorithm on Bellman Eqs. (15) and (16).

The strategy iteration steps of the Stackelberg game are as follows:

(i) Give initial policies $u^0(x)$ and $w^0(x)$.

(ii) Given $u^k(x)$ and $w^k(x)$, where k represents the k-th step after discretization means, solve for the costs $V_1^k(x)$ and $V_2^k(x)$ by using

$$\begin{aligned} 0 = {} & r_1(x, u^k(x)) + (\nabla V_1^k)^{\mathrm{T}}(Ax + B_1u^k + \\ & \frac{1}{2\gamma^2}B_2B_2^{\mathrm{T}}\nabla V_2^k) + y^{\mathrm{T}}(-A^{\mathrm{T}}\nabla V_2^k - 2Qx) \end{aligned} \quad (24)$$

$$\begin{aligned} 0 = {} & r_2(x, u^k(x), w^k(x)) + \\ & (\nabla V_2)^{\mathrm{T}}(Ax + B_1u^k + B_2w^k) \end{aligned} \quad (25)$$

(iii) Update the control policies using

$$u^{k+1}(x) = -\frac{1}{2}R^{-1}B_1^{\mathrm{T}}\nabla V_1^k \quad (26)$$

$$w^{k+1}(x) = \frac{1}{2\gamma^2}B_2^{\mathrm{T}}\nabla V_2^k \quad (27)$$

Remark 2 The policy iteration algorithm is an offline algorithm that requires a significant amount of dynamic information. This limitation motivates the development of a new online algorithm that can update policies and value functions online while requiring less information about the system dynamics.

IV. Value Function Approximation for Stackelberg Game

Based on previous consideration, when the dynamics are unknown, the integral reinforcement learning Bellman equation is presented as follows

$$V_1(x(t-T)) = \int_{t-T}^{t} r_1(x(t), u(t))\mathrm{d}t + V_1(x(t)) \quad (28)$$

$$V_2(x(t-T)) = \int_{t-T}^{t} r_2(x(t), u(t), w(t))\mathrm{d}t + V_2(x(t)) \quad (29)$$

where $t \geqslant 0$ and the time interval $T > 0$.

It should be noted that $V_1^k(x(t))$ can be obtained with the solution of Eq. (24) and $V_2^k(x(t))$ in Eq. (25) can be derived incidentally. Let $V_1'(x)$ and $V_2'(x)$ be admissible feedback control policies of $V_1^k(x(t))$ and $V_2^k(x(t))$, and assume that the value functions are smooth and form a compact set. Thus, $\forall x \in \Omega$, where set Ω is included in $\mathbb{R}^n$, $V_1'(x) \geqslant 0$ and $V_2'(x) \geqslant 0$. Constant matrix weights W_1 and W_2 are introduced here to approximate the value $V_i(x)$.

$$V_1'(x) = W_1^{\mathrm{T}}\phi_1(x) + \varepsilon_1(x) \quad (30)$$

$$V_2'(x) = W_2^{\mathrm{T}}\phi_2(x) + \varepsilon_2(x) \quad (31)$$

where $\phi_i(x) : \mathbb{R}^n \to \mathbb{R}^N$, $i = 1, 2$ is one vector of the activation function, the number of hidden layers is N, and $\varepsilon_i(x)$, $i = 1, 2$ is the approximation error, respectively.

Activation functions are chosen to provide a complete and independent basis set, allowing for the uniform approximation of both the value functions and their derivatives.

$$\frac{\partial V_i'}{\partial x} = \nabla V_i' = (\frac{\partial \phi_i(x)}{\partial x})^{\mathrm{T}} W_i + \frac{\partial \varepsilon_i}{\partial x} = \nabla \phi_i^{\mathrm{T}} W_i + \nabla \varepsilon_i,\ i = 1,2 \tag{32}$$

As demonstrated by the higher-order approximation theorem, with the number of hidden layer neurons N increasing (e.g., $N \to +\infty$), approximation error $\varepsilon(x)$ and its gradient $\nabla\varepsilon$ both approach zero. Additionally, if the number of layers is fixed, then ε and $\nabla\varepsilon$ are bounded on a compact set with certain constants.

Considering the value function approximation, the Hamiltonians in Eqs. (8) and (12) become

$$\varepsilon_{H_2} = x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u - \gamma^2 \|w\|^2 + W_2^{\mathrm{T}} \nabla \phi_2(x)(Ax + B_1 u + B_2 w) \tag{33}$$

$$\varepsilon_{H_1} = x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u + W_1^{\mathrm{T}} \nabla \phi_1(x) \times (Ax + B_1 u + B_2 w) + y^{\mathrm{T}} \dot{\lambda}_2 \tag{34}$$

where ε_{H_2} and ε_{H_1} are the residual errors for the follower and the leader, respectively.

Under fixed policies u and w, we have

$$\varepsilon_{H_i} = -(\nabla \varepsilon_i)^{\mathrm{T}}(Ax + B_1 u + B_2 w),\ i = 1,2 \tag{35}$$

Substituting Eq. (32) into Eqs. (18) and (19), one obtains

$$x^{\mathrm{T}} Q x + \frac{1}{4} W_1^{\mathrm{T}} \nabla \phi_1(x) B_1 R^{-1} B_1^{\mathrm{T}} \nabla \phi_1^{\mathrm{T}}(x) W_1 + \frac{1}{4} W_2^{\mathrm{T}} \nabla \phi_2(x) B_2 (\gamma^2)^{-1} B_2^{\mathrm{T}} \nabla \phi_2^{\mathrm{T}}(x) W_2 + W_2^{\mathrm{T}} \nabla \phi_2(x) \times (Ax - \frac{1}{2} B_1 R^{-1} B_1^{\mathrm{T}} \nabla \phi_1^{\mathrm{T}}(x) W_1) = \varepsilon_{\mathrm{HJ}_2} \tag{36}$$

$$x^{\mathrm{T}} Q x - \frac{1}{4} W_1^{\mathrm{T}} \nabla \phi_1(x) B_1 R^{-1} B_1^{\mathrm{T}} \nabla \phi_1^{\mathrm{T}}(x) W_1 + W_1^{\mathrm{T}} \nabla \phi_1(x)(Ax + \frac{1}{2\gamma^2} B_2 B_2^{\mathrm{T}} \nabla \phi_2^{\mathrm{T}}(x) W_2) - y^{\mathrm{T}}(A^{\mathrm{T}} \nabla \phi_2^{\mathrm{T}}(x) W_2 + 2Qx) = \varepsilon_{\mathrm{HJ}_1} \tag{37}$$

where subscripts HJ_2 and HJ_1 mean the second and the first Hamilton-Jacobi equations, respectively. And we have

$$\varepsilon_{\mathrm{HJ}_2} = -\frac{1}{2} W_1^{\mathrm{T}} \nabla \phi_1(x) B_1 R^{-1} B_1^{\mathrm{T}} \nabla \varepsilon_1 - \frac{1}{2\gamma^2} W_2^{\mathrm{T}} \nabla \phi_2(x) \times B_2 B_2^{\mathrm{T}} \nabla \varepsilon_2 - \frac{1}{4\gamma^2} \nabla \varepsilon_2{}^{\mathrm{T}} B_2 B_2^{\mathrm{T}} \nabla \varepsilon_2 - \frac{1}{4} \nabla \varepsilon_1{}^{\mathrm{T}} B_1 \times R^{-1} B_1^{\mathrm{T}} \nabla \varepsilon_1 - \nabla \varepsilon_2{}^{\mathrm{T}} A x + \frac{1}{2} \nabla \varepsilon_2{}^{\mathrm{T}} B_1 R^{-1} B_1^{\mathrm{T}} \times \nabla \phi_1{}^{\mathrm{T}} W_1 + \frac{1}{2} \nabla \varepsilon_2{}^{\mathrm{T}} B_1 R^{-1} B_1^{\mathrm{T}} \nabla \varepsilon_1 + \frac{1}{2} W_2^{\mathrm{T}} \nabla \phi_2(x) B_1 R^{-1} B_1^{\mathrm{T}} \nabla \varepsilon_1 \tag{38}$$

$$\varepsilon_{\mathrm{HJ}_1} = \frac{1}{2} W_1^{\mathrm{T}} \nabla \phi_1(x) B_1 R^{-1} B_1^{\mathrm{T}} \nabla \varepsilon_1 + \frac{1}{4} \nabla \varepsilon_1{}^{\mathrm{T}} B_1 \times R^{-1} B_1^{\mathrm{T}} \nabla \varepsilon_1 - \nabla \varepsilon_1{}^{\mathrm{T}} A x - \frac{1}{2\gamma^2} \nabla \varepsilon_1{}^{\mathrm{T}} B_2 B_2^{\mathrm{T}} \times \nabla \phi_2(x)^{\mathrm{T}} W_2 - \frac{1}{2\gamma^2} \nabla \varepsilon_1{}^{\mathrm{T}} B_2 B_2^{\mathrm{T}} \nabla \varepsilon_2 - \frac{1}{2\gamma^2} W_1^{\mathrm{T}} \nabla \phi_1(x) B_2 B_2^{\mathrm{T}} \nabla \varepsilon_2 - y^{\mathrm{T}} A^{\mathrm{T}} \nabla \varepsilon_2 \tag{39}$$

Lemma 1 For the admissible policy sets $u(x(t))$ and $w(x(t))$, the solution to Eqs. (18) and (19) with least-squares pair exists and changes as N changes. Give the following denotation

$$V_i(x) = W_i^{\mathrm{T}} \phi_i(x),\ i = 1,2 \tag{40}$$

Then, as $N \to +\infty$, we have

(i) $\sup_{x\in\Omega} |\varepsilon_{H_i}| \to 0$,

(ii) $\sup_{x\in\Omega} |\varepsilon_{\mathrm{HJ}_i}| \to 0$,

(iii) $\sup_{x\in\Omega} |V_i - V_i'| \to 0$,

(iv) $\sup_{x\in\Omega} |\nabla V_i - \nabla V_i'| \to 0$.

V. Online Solution of Stackelberg Game Using RL

In this section, the issue of developing the approximate optimal adaptive control algorithm is considered, which differs from traditional adaptive control approaches. In the latter, the first step is to design an observer for the cost function and denote it as the value function. The observer obtained is then used to design the feedback strategy. By contrast, the approximate optimal adaptive approach derives the solution of the Stackelberg game online. The outputs of the two players game are as follows

$$\hat{V}_1(x) = \hat{W}_1^{\mathrm{T}} \phi_1(x) \tag{41}$$

$$\hat{V}_2(x) = \hat{W}_2^{\mathrm{T}} \phi_2(x) \tag{42}$$

where $\hat{V}_1$ and $\hat{V}_2$ mean the estimated values of leaders' cost function and followers' cost function, respectively. $\hat{W}_i$ is the current estimated value of the value critic constant matrix weights W_i, $i = 1, 2$.

Considering Eqs. (41) and (42), Hamiltonian equations in Eqs. (8) and (12) become

$$H_2(x, \hat{W}_2, u, w) = x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u - \gamma^2 \|w\|^2 + \hat{W}_2^{\mathrm{T}} \nabla \phi_2(x)(Ax + B_1 u + B_2 w) = e_2 \tag{43}$$

$$H_1(x, \hat{W}_1, u, w) = x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u + \hat{W}_1^{\mathrm{T}} \nabla \phi_1(x)(Ax + B_1 u + B_2 w) + y^{\mathrm{T}}(-A^{\mathrm{T}} \nabla \phi_2^{\mathrm{T}}(x) \hat{W}_2 - 2Qx) = e_1 \tag{44}$$

where e_2 and e_1 represent the estimated values of ε_{H_2} and ε_{H_1}, respectively.

Define the weight estimation error $\tilde{W}_i = W_i - \hat{W}_i$, $i = 1, 2$, then we get

$$\begin{aligned} e_1 &= -\tilde{W}_1^{\mathrm{T}} \nabla \phi_1(x)(Ax + B_1 u + B_2 w) + \varepsilon_{H_1}, \\ e_2 &= -\tilde{W}_2^{\mathrm{T}} \nabla \phi_2(x)(Ax + B_1 u + B_2 w) + \varepsilon_{H_2} \end{aligned} \tag{45}$$

For admissible control policy pair, the selection of $\hat{W}_1$ and $\hat{W}_2$, which aims to minimize the square residual error E, is required.

$$E = \frac{1}{2}\sum_{i=1}^{2} e_i^{\mathrm{T}} e_i,\ i = 1,2 \tag{46}$$

By calculating, the weights of both follower and leader are selected, and the tuning laws of them are given in normalized gradient descent algorithm

$$\begin{aligned}\dot{\hat{W}}_2 = &-\alpha_2 \frac{\partial E}{\partial \hat{W}_2} = \\ &-\alpha_2 \frac{\sigma_w}{(1+\sigma_w^{\mathrm{T}}\sigma_w)^2}[\sigma_w^{\mathrm{T}}\hat{W}_2 + x^{\mathrm{T}}Qx + \\ &u^{\mathrm{T}}Ru - \gamma^2\|w\|^2] = \\ &-\alpha_2 \frac{\sigma_w}{(1+\sigma_w^{\mathrm{T}}\sigma_w)^2} e_2\end{aligned} \tag{47}$$

$$\begin{aligned}\dot{\hat{W}}_1 = &-\alpha_1 \frac{\partial E}{\partial \hat{W}_1} = \\ &-\alpha_1 \frac{\sigma_u}{(1+\sigma_u^{\mathrm{T}}\sigma_u)^2}[\sigma_u^{\mathrm{T}}\hat{W}_1 + x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru + \\ &y^{\mathrm{T}}(-A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 - 2Qx)] = \\ &-\alpha_1 \frac{\sigma_u}{(1+\sigma_u^{\mathrm{T}}\sigma_u)^2} e_1\end{aligned} \tag{48}$$

where $\dot{\hat{W}}_2$ and $\dot{\hat{W}}_1$ are the estimated values of $\hat{W}_2$ and $\hat{W}_1$, α_2 and α_1 are Levenberg-Marquardt algorithm coefficients, $\sigma_u = \nabla\phi_1(x)(Ax + B_1u + B_2w)$, and $\sigma_w = \nabla\phi_2(x)(Ax + B_1u + B_2w)$. The previous algorithm is modified for the reason that $(1+\sigma_w^{\mathrm{T}}\sigma_w)^2$ and $(1+\sigma_u^{\mathrm{T}}\sigma_u)^2$ are introduced in place of $1+\sigma_w^{\mathrm{T}}\sigma_w$ and $1+\sigma_u^{\mathrm{T}}\sigma_u$, respectively. The modified algorithm guarantees the bounded $\frac{\sigma_w}{1+\sigma_w^{\mathrm{T}}\sigma_w}$ and $\frac{\sigma_u}{1+\sigma_u^{\mathrm{T}}\sigma_u}$ appearing in Eqs. (47) and (48), respectively.

Equations (33) and (34) can be rewritten as follows

$$\begin{aligned}x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru - \gamma^2\|w\|^2 = \\ -W_2^{\mathrm{T}}\nabla\phi_2(x)(Ax + B_1u + B_2w) + \varepsilon_{H_2}\end{aligned} \tag{49}$$

$$\begin{aligned}x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru = \\ -W_1^{\mathrm{T}}\nabla\phi_1(x)(Ax + B_1u + B_2w) - y^{\mathrm{T}}\dot{\lambda}_2 + \varepsilon_{H_1}\end{aligned} \tag{50}$$

Substituting Eqs. (49) and (50) into Eqs. (47) and (48), respectively, one has

$$\dot{\hat{W}}_2 = -\alpha_2 \frac{\sigma_w}{(1+\sigma_w^{\mathrm{T}}\sigma_w)^2}[\sigma_w^{\mathrm{T}}(\hat{W}_2 - W_2) + \varepsilon_{H_2}] \tag{51}$$

$$\begin{aligned}\dot{\hat{W}}_1 = &-\alpha_1 \frac{\sigma_u}{(1+\sigma_u^{\mathrm{T}}\sigma_u)^2}[\sigma_u^{\mathrm{T}}(\hat{W}_1 - W_1) + \varepsilon_{H_1} - \\ &y^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)(\hat{W}_2 - W_2)]\end{aligned} \tag{52}$$

Recalling the denotation that $\tilde{W}_i = W_i - \hat{W}_i$, $i = 1, 2$, the weight estimation error of the dynamic can be written as

$$\dot{\tilde{W}}_2 = \alpha_2 \frac{\sigma_w}{(1+\sigma_w^{\mathrm{T}}\sigma_w)^2}[-\sigma_w^{\mathrm{T}}\tilde{W}_2 + \varepsilon_{H_2}] \tag{53}$$

$$\dot{\tilde{W}}_1 = \alpha_1 \frac{\sigma_u}{(1+\sigma_u^{\mathrm{T}}\sigma_u)^2}[-\sigma_u^{\mathrm{T}}\tilde{W}_1 + \varepsilon_{H_1} + y^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2] \tag{54}$$

Assumption 1 Persistence of excitation (PE) assumption [23] Denote that $\bar{\sigma}_w = \frac{\sigma_w}{1+\sigma_w^{\mathrm{T}}\sigma_w}$ and $\bar{\sigma}_u = \frac{\sigma_u}{1+\sigma_u^{\mathrm{T}}\sigma_u}$ are exciting over the interval $[t,\ t+T]$, $T > 0$ in persisten way, for example, constants β_i^w and β_i^u, $i = 1, 2$ exist such that for $\forall t$

$$\beta_1^w I \leqslant S_w,\ S_w = \int_t^{t+T} \bar{\sigma}_w\bar{\sigma}_w^{\mathrm{T}}\mathrm{d}\tau \leqslant \beta_2^w I \tag{55}$$

$$\beta_1^u I \leqslant S_u,\ S_u = \int_t^{t+T} \bar{\sigma}_u\bar{\sigma}_u^{\mathrm{T}}\mathrm{d}\tau \leqslant \beta_2^u I \tag{56}$$

where I indicates the identity matrix with appropriate dimension.

The PE assumption is necessary when using RL to derive system identification in adaptive control. It plays an important role in identifying the critic parameters required to approximate solutions for $V_1(x)$ and $V_2(x)$.

Lemma 2 Given a set of admissible bounded feedback policies $u(x(t))$ and $w(x(t))$, the critics' tuning laws are provided by Eqs. (47) and (48), assuming that $\bar{\sigma}_w$ and $\bar{\sigma}_u$ are persistently exciting. With bounded rationalities $\varepsilon_{\max_1}$ and $\varepsilon_{\max_2}$, the residual errors in Eqs. (33) and (34) are bounded by $\|\varepsilon_{H_i}\| < \varepsilon_{\max_i}$. Therefore, the critic parameter error converges exponentially.

$$\tilde{W}_2(t) \leqslant \frac{\sqrt{\beta_2^w T}}{\beta_1^w}\{[1 + 2\delta_2\beta_2^w\alpha_2]\varepsilon_{\max_2}\} \tag{57}$$

$$\tilde{W}_1(t) \leqslant \frac{\sqrt{\beta_2^u T}}{\beta_1^u}\{[1 + 2\delta_1\beta_2^u\alpha_1]\varepsilon_{\max_1}\} \tag{58}$$

where δ_1 and δ_2 are positive constants of the order of 1 [24].

Lemma 3 Based on the definition of $V_i(x)$, $i = 1, 2$ in Eq. (40), the control input and adversarial input in Eqs. (9) and (13), respectively, can be written as

$$w^*(x) = \frac{1}{2\gamma^2}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)W_2 \tag{59}$$

$$u^*(x) = -\frac{1}{2}R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)W_1 \tag{60}$$

Then, with the number in the hidden layer tending to infinity, $\sup_{x\in\Omega}\|u' - u\| \to 0$ and $\sup_{x\in\Omega}\|w' - w\| \to 0$ exist.

The algorithm for updating the desired control policies is shown in Eqs. (59) and (60), with unknown matrices W_1 and W_2. Therefore, denoting the control policies as $\hat{w}(x)$ and $\hat{u}(x)$, we can write them compactly in structured form as follows

$$\hat{w}(x) = \frac{1}{2\gamma^2}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_w \tag{61}$$

$$\hat{u}(x) = -\frac{1}{2}R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\hat{W}_u \tag{62}$$

where $\hat{W}_w$ is the current estimated value of W_2 and $\hat{W}_u$ is the current estimated value of W_1. Actor estimation errors are defined as $\tilde{W}_w = W_2 - \hat{W}_w$ and $\tilde{W}_u = W_1 - \hat{W}_u$.

Definition 3 For given bounds B and T, if relationship $\|x(t)\| \leqslant B$ for all $t \geqslant t_0 + T$ holds for any initial value $x_0 \in \mathbb{R}^n$, then $x(t)$ satisfies the speciality of uniformly ultimately bounded.

Definition 4 Given a compact set $\Omega \in \mathbb{R}^n$:

(1) The approximation errors ε_1 and ε_2 and their gradients

$\nabla\varepsilon_1$ and $\nabla\varepsilon_2$ are bounded, and the following relationships hold

$$\|\varepsilon_1\| \leqslant b_{\varepsilon_1},\ \|\nabla\varepsilon_1\| \leqslant b_{\varepsilon_1,x},\ \|\varepsilon_2\| \leqslant b_{\varepsilon_2},\ \|\nabla\varepsilon_2\| \leqslant b_{\varepsilon_2,x} \tag{63}$$

where b_{ε_1}, $b_{\varepsilon_1,x}$, b_{ε_2}, and $b_{\varepsilon_2,x}$ are corresponding boundaries.

(2) The activation functions $\phi_1(x)$ and $\phi_2(x)$ and their gradients $\nabla\phi_1(x)$ and $\nabla\phi_2(x)$ are bounded, and the following relationships hold

$$\begin{aligned}&\|\phi_1(x)\| \leqslant b_{\phi_1},\ \|\nabla\phi_1(x)\| \leqslant b_{\phi_1,x},\\ &\|\phi_2(x)\| \leqslant b_{\phi_2},\ \|\nabla\phi_2(x)\| \leqslant b_{\phi_2,x}\end{aligned} \tag{64}$$

where b_{ϕ_1}, $b_{\phi_1,x}$, b_{ϕ_2}, and $b_{\phi_2,x}$ are corresponding boundaries.

(3) The critic weights W_1 and W_2 are bounded by known constants $W_{1,\max}$ and $W_{2,\max}$

$$\begin{aligned}&\|W_1\| \leqslant W_{1,\max},\\ &\|W_2\| \leqslant W_{2,\max}\end{aligned} \tag{65}$$

The main results of this paper concern the tuning laws for mixed H_2/H_∞ control based on a real-time reinforcement learning algorithm. These tuning laws ensure convergence of the game.

Theorem 1 Considering the dynamics given in Eq. (1) and the Stackelberg game formulated in Section II, the outputs of the two players game are presented in Eqs. (41) and (42), and the control input and disturbance input are given in Eqs. (61) and (62). Let tuning laws be provided by

$$\begin{aligned}\dot{\hat{W}}_2 = &-\alpha_2\frac{\sigma_{\hat{w}}}{(1+\sigma_{\hat{w}}^{\mathrm{T}}\sigma_{\hat{w}})^2}[\sigma_{\hat{w}}^{\mathrm{T}}\hat{W}_2 + x^{\mathrm{T}}Qx + \\ &\hat{u}^{\mathrm{T}}R\hat{u} - \gamma^2\|\hat{w}\|^2]\end{aligned} \tag{66}$$

$$\begin{aligned}\dot{\hat{W}}_1 = &-\alpha_1\frac{\sigma_{\hat{u}}}{(1+\sigma_{\hat{u}}^{\mathrm{T}}\sigma_{\hat{u}})^2}[\sigma_{\hat{u}}^{\mathrm{T}}\hat{W}_1 + x^{\mathrm{T}}Qx + \hat{u}^{\mathrm{T}}R\hat{u} + \\ &y^{\mathrm{T}}(-A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 - 2Qx)]\end{aligned} \tag{67}$$

where $\sigma_{\hat{w}} = \nabla\phi_2(x)(Ax + B_1\hat{u} + B_2\hat{w})$ and $\sigma_{\hat{u}} = \nabla\phi_1(x)(Ax + B_1\hat{u} + B_2\hat{w})$.

Tuning laws $\dot{\hat{W}}_w$ and $\dot{\hat{W}}_u$ for follower's and leader's actor are designed as

$$\dot{\hat{W}}_w = -\alpha_4\left[(\kappa_4\hat{W}_w - \kappa_3\bar{\sigma}_{\hat{w}}^{\mathrm{T}}\hat{W}_2) - \frac{1}{4}\bar{D}_2(x)\hat{W}_w m_w^{\mathrm{T}}\hat{W}_2\right] \tag{68}$$

$$\begin{aligned}\dot{\hat{W}}_u = &-\alpha_3\Bigg[(\kappa_2\hat{W}_u - \kappa_1\bar{\sigma}_{\hat{u}}^{\mathrm{T}}\hat{W}_1) - \frac{1}{4}(\bar{D}_1(x)\hat{W}_u m_w^{\mathrm{T}}\hat{W}_2 + \\ &\bar{D}_1(x)\hat{W}_u m_u^{\mathrm{T}}\hat{W}_1) + \frac{1}{2\gamma^2}\nabla\phi_1 B_2 B_2^{\mathrm{T}}P_2 y\Bigg]\end{aligned} \tag{69}$$

where $\bar{D}_1(x) = \nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)$, $\bar{D}_2(x) = \nabla\phi_2(x)B_2(\gamma^2)^{-1}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)$, $m_w = \dfrac{\sigma_{\hat{w}}}{(1+\sigma_{\hat{w}}^{\mathrm{T}}\sigma_{\hat{w}})^2}$, $m_u = \dfrac{\sigma_{\hat{u}}}{(1+\sigma_{\hat{u}}^{\mathrm{T}}\sigma_{\hat{u}})^2}$, $\bar{\sigma}_{\hat{w}} = \dfrac{\sigma_{\hat{w}}}{1+\sigma_{\hat{w}}^{\mathrm{T}}\sigma_{\hat{w}}}$, $\bar{\sigma}_{\hat{u}} = \dfrac{\sigma_{\hat{u}}}{1+\sigma_{\hat{u}}^{\mathrm{T}}\sigma_{\hat{u}}}$, respectively. $\kappa_i > 0$ and $\alpha_i > 0, \forall i = 1, 2, 3, 4$ are tuning parameters, which are detailedly presented in the proof of Theorem 1 in Appendix A. Suppose that $\bar{\sigma}_{\hat{w}}$ and $\bar{\sigma}_{\hat{u}}$ are persistently exciting, a constant number N_0 exists such that, for any $N > N_0$, the following mentioned notions, such as the system state, the critic errors $\tilde{W}_i$, $i = 1, 2$, and the actor errors $\tilde{W}_w$ and $\tilde{W}_u$, are uniformly ultimately bounded.

Theorem 2 If the hypotheses proposed in Theorem 1 are met, the following results hold

(i) $H_1(x, \hat{W}_1, \hat{w}'(x), \hat{u}'(x))$ and $H_2(x, \hat{W}_2, \hat{w}'(x), \hat{u}'(x))$ are uniformly ultimately bounded.

$$\hat{w}'(x) = \frac{1}{2\gamma^2}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 \tag{70}$$

$$\hat{u}'(x) = -\frac{1}{2}R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\hat{W}_1 \tag{71}$$

where $\hat{w}'$ and $\hat{u}'$ are approximate values of $\hat{w}$ and $\hat{u}$, respectively. $\hat{W}_1$ and $\hat{W}_2$ converge to the approximate coupled HJ solution.

(ii) $\hat{w}(x)$ and $\hat{u}(x)$ in Eqs. (61) and (62) converge to the approximate solution of the Stackelberg game.

The proof of Theorem 2 is presented in Appendix B.

VI. Simulation

The 2-degree of freedom (2-DOF) robot serves as a basic control platform for the manipulator. The system is a planar 2-DOF with two drives and three unactuated swivel joints. Two servomotors, mounted at a fixed distance on the rotary motion servo platform, control a four-bar system. The goal of 2-DOF robot trajectory tracking is to regulate the $x-y$ position through a 4-link end-actuator. To achieve this, a controller must be designed to regulate the plane coordinates of the end-actuator and to drive the two arms connected to the motor to realize the motion of the other two arms. The trajectory is represented by the end-effector.

Based on the output angular displacements θ_A and θ_B of the two servo systems detected by the encoder, the corresponding control voltage V_m is calculated at each moment. The power module then controls the servo system. The principle diagram of the control system is shown in Fig. 1.

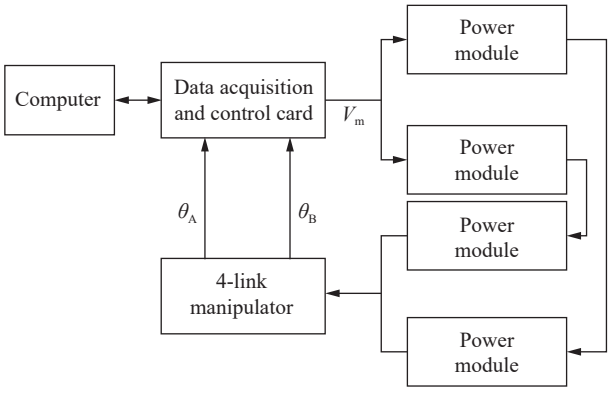

Figure 1 Schematic diagram of 2-DOF robot control system.

Considering the system with the form shown in Eq. (1), we have

$$\begin{aligned}\dot{x}(t) = &\begin{bmatrix}0.000 & 1.000\\ -835.210 & -39.882\end{bmatrix}x(t) + \\ &\begin{bmatrix}0.000\\ -835.210\end{bmatrix}u(t) + \begin{bmatrix}0.000\\ 1.000\end{bmatrix}w(t)\end{aligned} \tag{72}$$

with matrices $Q = R = I$, $\gamma = 0.09$, and tuning gains $\alpha_1 = 5$, $\alpha_2 = 1$, $\alpha_3 = 1$, and $\alpha_4 = 1$.

The computer provides the expected trajectory of the state. Figure 2 compares the state obtained from the online reinforcement learning algorithm with the ideal state. It is clear that the estimated state can quickly follow the ideal, demonstrating the effectiveness of the algorithm discussed in this paper.

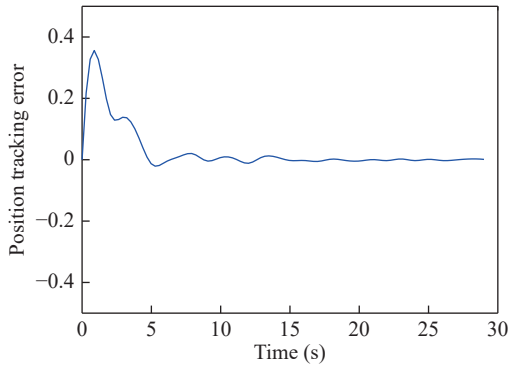

Figure 2 Tracking error.

VII. Conclusion

This paper presents an online integral RL algorithm for finding the optimal solution to a problem with hierarchical decision makers. In this model, the leader is the control input and the follower is the disturbance. Both the control input and the deterministic disturbance shape the Stackelberg game. Full dynamic information of the system is not required, as online solution strategies are applied, making explicit solutions unnecessary. The designed approximators for the leader and follower are used to learn decision policies online for the purpose of team decision making. This online learning approach is suitable for varying team dynamics. As shown in this paper, the algorithms presented here offer many advantages in dealing with dynamic systems, particularly mixed H_2/H_∞ problems.

Appendix A
Proof of Theorem 1

Proof The convergence proof is presented by introducing Lyapunov analysis.

Specifically, we consider the Lyapunov function $L(x)$

$$L(x) = V_1(x) + V_2(x) + \frac{1}{2}\tilde{W}_1^{\mathrm{T}}\alpha_1^{-1}\tilde{W}_1 + \frac{1}{2}\tilde{W}_2^{\mathrm{T}}\alpha_2^{-1}\tilde{W}_2 + \frac{1}{2}\tilde{W}_w^{\mathrm{T}}\alpha_3^{-1}\tilde{W}_w + \frac{1}{2}\tilde{W}_u^{\mathrm{T}}\alpha_4^{-1}\tilde{W}_u \tag{A1}$$

where $V_1(x)$ and $V_2(x)$ are approximate solutions of Eqs. (18) and (19) and are given in Lemma 1.

It is easy to obtain the time derivative as follows

$$\dot{L}(x) = \dot{V}_1(x) + \dot{V}_2(x) + \tilde{W}_1^{\mathrm{T}}\alpha_1^{-1}\dot{\tilde{W}}_1 + \tilde{W}_2^{\mathrm{T}}\alpha_2^{-1}\dot{\tilde{W}}_2 + \tilde{W}_w^{\mathrm{T}}\alpha_3^{-1}\dot{\tilde{W}}_w + \tilde{W}_u^{\mathrm{T}}\alpha_4^{-1}\dot{\tilde{W}}_u \tag{A2}$$

The key step is calculating the terms in $\dot{L}(x)$. For the first term, differentiate Eq. (40) when $i = 1$, then add and subtract $\frac{1}{2}W_1^{\mathrm{T}}\bar{D}_1(x)W_1$ and $\frac{1}{2}W_2^{\mathrm{T}}\bar{E}_2(x)W_2$, we have

$$\dot{V}_1(x) = W_1^{\mathrm{T}}\sigma_1 + \frac{1}{2}W_1^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u - \frac{1}{2}W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w + \dot{\varepsilon}_1(x) \tag{A3}$$

where

$$\begin{aligned}
\sigma_1 &= \nabla\phi_1(x)(Ax + B_1u' + B_2w'),\\
\bar{E}_2(x) &= \nabla\phi_1(x)B_2(\gamma^2)^{-1}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x),\\
\dot{\varepsilon}_1(x) &= \nabla\varepsilon_1^{\mathrm{T}}(x)[Ax - \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}\hat{W}_u + \frac{1}{2}B_2(\gamma^2)^{-1}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}\hat{W}_w],\\
W_1^{\mathrm{T}}\sigma_1 &= -x^{\mathrm{T}}Qx - \frac{1}{4}W_1^{\mathrm{T}}\bar{D}_1W_1 + y^{\mathrm{T}}(A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)W_2 + 2Qx) + \varepsilon_{\mathrm{HJ}_1}
\end{aligned} \tag{A4}$$

Using similar method to deal with the second term, we have

$$\dot{V}_2(x) = W_2^{\mathrm{T}}\sigma_2 + \frac{1}{2}W_2^{\mathrm{T}}\bar{E}_1(x)\tilde{W}_w + \frac{1}{2}W_2^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w + \dot{\varepsilon}_2(x) \tag{A5}$$

where

$$\begin{aligned}
\sigma_2 &= \nabla\phi_2(x)(Ax + B_1u' + B_2w'),\\
\bar{E}_1(x) &= \nabla\phi_2(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x),\\
\dot{\varepsilon}_2(x) &= \nabla\varepsilon_2^{\mathrm{T}}(x)[Ax - \frac{1}{2}B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}\hat{W}_u + \frac{1}{2}B_2(\gamma^2)^{-1}B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}\hat{W}_w],\\
W_2^{\mathrm{T}}\sigma_2 &= -x^{\mathrm{T}}Qx - \frac{1}{4}W_1^{\mathrm{T}}\bar{D}_1W_1 + \frac{1}{4}W_2^{\mathrm{T}}\bar{D}_2W_2 + \varepsilon_{\mathrm{HJ}_2}
\end{aligned} \tag{A6}$$

Adding $\dot{V}_1(x)$ and $\dot{V}_2(x)$, thus we get

$$\dot{L}_v = \dot{\tilde{L}}_v + \frac{1}{2}W_1^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u - \frac{1}{2}W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w + \frac{1}{2}W_2^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w + \frac{1}{2}W_2^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w + \dot{\varepsilon}_1(x) + \dot{\varepsilon}_2(x) \tag{A7}$$

for the sake of simplicity

$$\dot{\tilde{L}}_v = -x^{\mathrm{T}}Qx - \frac{1}{4}W_1^{\mathrm{T}}\bar{D}_1W_1 + y^{\mathrm{T}}(A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)W_2 + 2Qx) + \varepsilon_{\mathrm{HJ}_1} - x^{\mathrm{T}}Qx - \frac{1}{4}W_1^{\mathrm{T}}\bar{D}_1W_1 + \frac{1}{4}W_2^{\mathrm{T}}\bar{D}_2W_2 + \varepsilon_{\mathrm{HJ}_2} \tag{A8}$$

Combining the tuning law in Eq. (67) and the error definitions, we can obtain the third term

$$\begin{aligned}
\dot{L}_3 = \tilde{W}_1^{\mathrm{T}}\alpha_1^{-1}\dot{\tilde{W}}_1 &= \\
&\tilde{W}_1^{\mathrm{T}}m_u[\sigma_{\hat{u}}^{\mathrm{T}}\hat{W}_1 + x^{\mathrm{T}}Qx + \hat{u}^{\mathrm{T}}R\hat{u} + y^{\mathrm{T}}(-A^{\mathrm{T}}\phi_2^{\mathrm{T}}(x)\hat{W}_2 - 2Qx)] =\\
&\dot{\tilde{L}}_3 + \tilde{W}_1^{\mathrm{T}}m_uy^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 + \frac{1}{4}\tilde{W}_1^{\mathrm{T}}m_u\times\\
&[\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u + 2W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w]
\end{aligned} \tag{A9}$$

where

$$\dot{\tilde{L}}_3 = \tilde{W}_1^{\mathrm{T}}\bar{\sigma}_{\hat{u}}\Big[-\bar{\sigma}_{\hat{u}}\tilde{W}_1 + \frac{\varepsilon_{\mathrm{HJ}_1}}{1+\sigma_{\hat{u}}^{\mathrm{T}}\sigma_{\hat{u}}}\Big] \tag{A10}$$

By using the similar method, the forth term becomes

$$\begin{aligned}\dot{L}_4 = \dot{\tilde{L}}_4 + \frac{1}{4}\tilde{W}_2^{\mathrm{T}} m_w[\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w + \tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u + \\ 2\tilde{W}_u^{\mathrm{T}}\bar{E}_1(x)W_2 - 2\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1]\end{aligned} \tag{A11}$$

where

$$\dot{\tilde{L}}_4 = \tilde{W}_2^{\mathrm{T}}\bar{\sigma}_{\hat{w}}\Big[-\bar{\sigma}_{\hat{w}}\tilde{W}_2 + \frac{\varepsilon_{\mathrm{HJ}_2}}{1+\sigma_{\hat{w}}^{\mathrm{T}}\sigma_{\hat{w}}}\Big] \tag{A12}$$

In order to obtain the tuning laws of actors of follower and leader, these two terms $\tilde{W}_u$ and $\tilde{W}_w$ are grouped by multiplying with estimated values. This yields

$$\begin{aligned}\dot{L}(x) = \dot{\tilde{L}}_v + \dot{\tilde{L}}_3 + \dot{\tilde{L}}_4 + \dot{\varepsilon}_1(x) + \dot{\varepsilon}_2(x) + \frac{1}{2}W_1^{\mathrm{T}}\bar{D}_1(x)\times \\ \tilde{W}_u - \frac{1}{2}W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w + \frac{1}{2}W_2^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w + \\ \frac{1}{2}W_2^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w + \tilde{W}_1^{\mathrm{T}}m_u y^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 + \\ \frac{1}{2}W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w m_u\tilde{W}_1 + \frac{1}{2}W_2^{\mathrm{T}}\bar{E}_1(x)\tilde{W}_u m_w\times \\ \tilde{W}_2 - \frac{1}{2}W_1^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u m_w\tilde{W}_2 + \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\times \\ W_1 m_u\tilde{W}_1 - \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1 m_u W_1 + \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\times \\ \bar{D}_1(x)\tilde{W}_u m_u W_1 + \frac{1}{4}\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)W_2 m_w\tilde{W}_2 - \\ \frac{1}{4}\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)W_2 m_w W_2 + \frac{1}{4}\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w m_w\times \\ W_2 + \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1 m_w\tilde{W}_2 - \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\times \\ W_1 m_w W_2 + \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u m_w W_2 - \tilde{W}_w^{\mathrm{T}}[\alpha_4^{-1}\times \\ \dot{\hat{W}}_w - \frac{1}{4}\bar{D}_2(x)\hat{W}_w m_w\hat{W}_2] - \tilde{W}_u^{\mathrm{T}}[\alpha_3^{-1}\dot{\hat{W}}_u - \\ \frac{1}{4}\bar{D}_1(x)\hat{W}_u m_u\hat{W}_1 - \frac{1}{4}\bar{D}_1(x)\hat{W}_u m_w\hat{W}_w + \\ \frac{1}{2\gamma^2}\nabla\phi_1 B_2 B_2^{\mathrm{T}}P_2 y]\end{aligned} \tag{A13}$$

Considering the last two terms, the following tuning laws for the leader and follower can be designed to simplify the equation

$$\dot{\hat{W}}_w = -\alpha_4[(\kappa_4\hat{W}_w - \kappa_3\bar{\sigma}_{\hat{w}}^{\mathrm{T}}\hat{W}_2) - \frac{1}{4}\bar{D}_2(x)\hat{W}_w m_w^{\mathrm{T}}\hat{W}_2] \tag{A14}$$

$$\begin{aligned}\dot{\hat{W}}_u = -\alpha_3[(\kappa_2\hat{W}_u - \kappa_1\bar{\sigma}_{\hat{u}}^{\mathrm{T}}\hat{W}_1) - \frac{1}{4}(\bar{D}_1(x)\hat{W}_u m_w^{\mathrm{T}}\hat{W}_2 + \\ \bar{D}_1(x)\hat{W}_u m_u^{\mathrm{T}}\hat{W}_1) + \frac{1}{2\gamma^2}\nabla\phi_1 B_2 B_2^{\mathrm{T}}P_2 y]\end{aligned} \tag{A15}$$

Then, the time derivative of Lyapunov function can be rewritten as

$$\begin{aligned}\dot{L}(x) = -x^{\mathrm{T}}Qx - \frac{1}{4}W_1^{\mathrm{T}}\bar{D}_1 W_1 + y^{\mathrm{T}}(A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)W_2 + \\ 2Qx) + \varepsilon_{\mathrm{HJ}_1} - x^{\mathrm{T}}Qx - \frac{1}{4}W_1^{\mathrm{T}}\bar{D}_1 W_1 + \frac{1}{4}W_2^{\mathrm{T}}\times \\ \bar{D}_2 W_2 + \varepsilon_{\mathrm{HJ}_2} + \tilde{W}_1^{\mathrm{T}}m_u y^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 + \\ \tilde{W}_1^{\mathrm{T}}\bar{\sigma}_{\hat{u}}\Big[-\bar{\sigma}_{\hat{u}}\tilde{W}_1 + \frac{\varepsilon_{\mathrm{HJ}_1}}{1+\sigma_{\hat{u}}^{\mathrm{T}}\sigma_{\hat{u}}}\Big] + \dot{\varepsilon}_1(x)\tilde{W}_2^{\mathrm{T}}\times \\ \bar{\sigma}_{\hat{w}}\Big[-\bar{\sigma}_{\hat{w}}\tilde{W}_2 + \frac{\varepsilon_{\mathrm{HJ}_2}}{1+\sigma_{\hat{w}}^{\mathrm{T}}\sigma_{\hat{w}}}\Big] + \dot{\varepsilon}_2(x) + \frac{1}{2}W_1^{\mathrm{T}}\times \\ \bar{D}_1(x)\tilde{W}_u - \frac{1}{2}W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w + \frac{1}{2}W_2^{\mathrm{T}}\bar{E}_2(x)\times \\ \tilde{W}_w + \frac{1}{2}W_2^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w + \frac{1}{2}W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w m_u\tilde{W}_1 + \\ \frac{1}{2}W_2^{\mathrm{T}}\bar{E}_1(x)\tilde{W}_u m_w\tilde{W}_2 - \frac{1}{2}W_1^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u m_w\tilde{W}_2 + \\ \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1 m_u\tilde{W}_1 - \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1 m_u W_1 + \\ \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u m_u W_1 + \frac{1}{4}\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)W_2 m_w\tilde{W}_2 - \\ \frac{1}{4}\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)W_2 m_w W_2 + \frac{1}{4}\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w m_w W_2 + \\ \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1 m_w\tilde{W}_2 - \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1 m_w W_2 + \\ \frac{1}{4}\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u m_w W_2 + \tilde{W}_w^{\mathrm{T}}\kappa_4 W_2 - \tilde{W}_w^{\mathrm{T}}\kappa_4\tilde{W}_w - \\ \tilde{W}_w^{\mathrm{T}}\kappa_3\bar{\sigma}_{\hat{w}}^{\mathrm{T}}W_2 + \tilde{W}_w^{\mathrm{T}}\kappa_3\bar{\sigma}_{\hat{w}}^{\mathrm{T}}\tilde{W}_w + \tilde{W}_u^{\mathrm{T}}\kappa_2 W_1 - \\ \tilde{W}_u^{\mathrm{T}}\kappa_2\tilde{W}_u - \tilde{W}_u^{\mathrm{T}}\kappa_1\bar{\sigma}_{\hat{u}}^{\mathrm{T}}W_1 + \tilde{W}_u^{\mathrm{T}}\kappa_1\bar{\sigma}_{\hat{u}}^{\mathrm{T}}\tilde{W}_u\end{aligned} \tag{A16}$$

Here comes the norm bounds. Based on the definition of uniform ultimate boundedness, Eqs. (A4) and (A6) become

$$\begin{aligned}\|\dot{\varepsilon}_1(x)\| < b_{\varepsilon_1 x}A\|x\| + \frac{1}{2}b_{\varepsilon_1 x}B_1^2 b_{\phi_1 x}\sigma_{\min}(R)(\|W_1\| + \\ \|\tilde{W}_u\|) + \frac{1}{2}b_{\varepsilon_1 x}B_2^2 b_{\phi_2 x}\sigma_{\min}(\gamma^2)(\|W_2\| + \\ \|\tilde{W}_w\|)\end{aligned} \tag{A17}$$

$$\begin{aligned}\|\dot{\varepsilon}_2(x)\| < b_{\varepsilon_2 x}A\|x\| + \frac{1}{2}b_{\varepsilon_2 x}B_1^2 b_{\phi_1 x}\sigma_{\min}(R)(\|W_1\| + \\ \|\tilde{W}_u\|) + \frac{1}{2}b_{\varepsilon_2 x}B_2^2 b_{\phi_2 x}\sigma_{\min}(\gamma^2)(\|W_2\| + \\ \|\tilde{W}_w\|)\end{aligned} \tag{A18}$$

where $\sigma_{\min}$ means the inversion of matrix R.

Given $\bar{\varepsilon}_i > 0$, $i = 1, 2$, $K_0(\bar{\varepsilon}_1)$, and $K_0(\bar{\varepsilon}_2)$, where K_0 is the hidden layer unit, there are $\sup_{x\in\Omega}\|\varepsilon_{\mathrm{HJ}_1}\| < \bar{\varepsilon}_1$ and $\sup_{x\in\Omega}\|\varepsilon_{\mathrm{HJ}_2}\| < \bar{\varepsilon}_2$. Then assuming $N > K_0$, define

$$\tilde{\Xi} = [x \quad \bar{\sigma}_{\hat{u}}^{\mathrm{T}}\tilde{W}_1 \quad \bar{\sigma}_{\hat{w}}^{\mathrm{T}}\tilde{W}_2 \quad \tilde{W}_u \quad \tilde{W}_w \quad y]^{\mathrm{T}} \tag{A19}$$

there is

$$\dot{L}(x) = \frac{1}{2}\|W_1\|^2\|\bar{D}_1\| + \frac{1}{4}\|W_2\|^2\|\bar{D}_2\| + \bar{\varepsilon}_1 + \bar{\varepsilon}_2\frac{1}{2}b_{\varepsilon_1 x}B_1^2 b_{\phi_1 x}\sigma_{\min}(R)\|W_1\| + \frac{1}{2}b_{\varepsilon_1 x}B_2^2\times b_{\phi_2 x}\sigma_{\min}(\gamma^2)\|W_2\| + \frac{1}{2}b_{\varepsilon_2 x}B_1^2 b_{\phi_1 x}\sigma_{\min}(R)\times\|W_1\| + \frac{1}{2}b_{\varepsilon_2 x}B_2^2 b_{\phi_2 x}\sigma_{\min}(\gamma^2)\|W_2\| + \tilde{\Xi}^{\mathrm{T}}\begin{bmatrix} -2Q & 0 & 0 & 0 & 0 & 0 \\ 0 & -I & 0 & m_{24} & m_{25} & m_{26} \\ 0 & 0 & -I & m_{34} & m_{35} & 0 \\ 0 & m_{42} & m_{43} & m_{44} & 0 & 0 \\ 0 & m_{52} & m_{53} & 0 & m_{55} & 0 \\ 0 & m_{62} & 0 & 0 & 0 & 0 \end{bmatrix}\tilde{\Xi} + \tilde{\Xi}\begin{bmatrix} (b_{\varepsilon_1 x} + b_{\varepsilon_2 x})A \\ \dfrac{\varepsilon_{\mathrm{HJ}_1}}{m_{s1}} \\ \dfrac{\varepsilon_{\mathrm{HJ}_2}}{m_{s2}} \\ \bar{m}_4 \\ \bar{m}_5 \\ A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)W_2 \end{bmatrix} = \bar{\varepsilon}_1 + \bar{\varepsilon}_2 + \bar{M} + \tilde{\Xi}^{\mathrm{T}}M\tilde{\Xi} + \tilde{\Xi}\tilde{M} \tag{A20}$$

where M is a matrix, $\bar{M}$ and $\tilde{M}$ are bounded in $\bar{M}_{\max}$ and $\tilde{M}_{\max}$, and

$$\begin{aligned}
m_{24} &= \frac{1}{2}\kappa_1 + \frac{1}{8m_{s1}\bar{D}_1(x)W_1} = m_{42}^{\mathrm{T}},\\
m_{25} &= \frac{1}{4m_{s1}}\bar{E}_2(x)W_1 = m_{52}^{\mathrm{T}},\\
m_{26} &= \frac{1}{2m_{s1}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 = m_{62}^{\mathrm{T}},\\
m_{34} &= \frac{1}{4m_{s2}}\bar{E}_1(x)W_2 - \frac{1}{4m_{s2}}\bar{D}_1(x)W_1 + \frac{1}{4m_{s2}}\bar{D}_1(x)W_2 = m_{43}^{\mathrm{T}},\\
m_{35} &= \frac{1}{2}\kappa_3 + \frac{1}{8m_{s1}}\bar{D}_2(x)W_2 = m_{53}^{\mathrm{T}},\\
m_{44} &= -\kappa_2 + \frac{1}{4}\bar{D}_1(x)m_u W_1 + \frac{1}{4}\bar{D}_1(x)m_w W_2,\\
m_{55} &= -\kappa_4 + \frac{1}{4}\bar{D}_2(x)m_w W_2,\\
m_{s1} &= 1 + \sigma_{\hat{u}}^{\mathrm{T}}\sigma_{\hat{u}},\\
m_{s2} &= 1 + \sigma_{\hat{w}}^{\mathrm{T}}\sigma_{\hat{w}},\\
\bar{m}_4 &= \kappa_2 W_1 - \kappa_1\bar{\sigma}_{\hat{u}}^{\mathrm{T}}W_1 + \frac{1}{2}\bar{D}_1(x)W_1 + \frac{1}{4}\bar{D}_1(x)W_1 m_u W_1 - \frac{1}{4}\bar{D}_1(x)W_1 m_w W_2,\\
\bar{m}_5 &= \kappa_4 W_2 - \kappa_3\bar{\sigma}_{\hat{w}}^{\mathrm{T}}W_2 + \frac{1}{2}\bar{E}_2(x)W_1 + \frac{1}{2}\bar{E}_2(x)W_2 + \frac{1}{2}\bar{D}_2(x)W_2 - \frac{1}{4}\bar{D}_2(x)W_2 m_w W_2,\\
\bar{M} &= \frac{1}{2}\|W_1\|^2\|\bar{D}_1\| + \frac{1}{4}\|W_2\|^2\|\bar{D}_2\| + \frac{1}{2}b_{\varepsilon_1 x}B_1^2 b_{\phi_1 x}\times\sigma_{\min}(R)\|W_1\| + \frac{1}{2}b_{\varepsilon_1 x}B_2^2 b_{\phi_2 x}\sigma_{\min}(\gamma^2)\|W_2\| + \frac{1}{2}b_{\varepsilon_2 x}B_1^2 b_{\phi_1 x}\sigma_{\min}(R)\|W_1\| + \frac{1}{2}b_{\varepsilon_2 x}B_2^2 b_{\phi_2 x}\sigma_{\min}(\gamma^2)\|W_2\|
\end{aligned} \tag{A21}$$

According to the bounds given in Definition 4, $\bar{M}$ and $\tilde{M}$ are bounded in $\bar{M}_{\max}$ and $\tilde{M}_{\max}$, respectively.

Now, the matrix M can be compacted and rewritten as

$$M = \begin{bmatrix} -2Q & 0 & 0 & 0 \\ 0 & -I_2 & M'_{23} & M'_{24} \\ 0 & M'_{32} & M'_{33} & 0 \\ 0 & M'_{42} & 0 & 0 \end{bmatrix} = \begin{bmatrix} -2Q & 0 & 0 \\ 0 & -I_2 & M_{23} \\ 0 & M_{32} & M_{33} \end{bmatrix} \tag{A22}$$

where

$$\begin{aligned}
-I_2 &= \begin{bmatrix} -I & 0 \\ 0 & -I \end{bmatrix},\\
M'_{23} &= \begin{bmatrix} m_{24} & m_{25} \\ m_{34} & m_{35} \end{bmatrix},\\
M'_{24} &= \begin{bmatrix} m_{26} \\ 0 \end{bmatrix},\\
M'_{32} &= \begin{bmatrix} m_{42} & m_{43} \\ m_{52} & m_{53} \end{bmatrix},\\
M'_{33} &= \begin{bmatrix} m_{44} & 0 \\ 0 & m_{55} \end{bmatrix},\\
M'_{42} &= \begin{bmatrix} m_{62} \\ 0 \end{bmatrix},\\
M_{23} &= \begin{bmatrix} M'_{23} & M'_{24} \end{bmatrix},\\
M_{32} &= \begin{bmatrix} M'_{32} \\ M'_{42} \end{bmatrix},\\
M_{33} &= \begin{bmatrix} M'_{33} & 0 \\ 0 & 0 \end{bmatrix}
\end{aligned} \tag{A23}$$

in which the following conditions should hold in order to guarantee the negative value of M

(i) $-2Q < 0$,

(ii) $-2I_2 < 0$,

(iii) Shur complement for $-I_2$ is $D_{-I_2} = -I_2 - M_{23}M_{33}^{-1}M_{32} < 0$ with proper parameter κ_i, $i = 1, 2, 3, 4$,

(iv) Shur complement for M_{33} is $D_{M_{33}} = M_{33} + M_{32}I_2^{-1}M_{23} < 0$ with proper matrix κ_i, $i = 1, 2, 3, 4$.

Equation (A20) satisfies the following relationship

$$\dot{L} < \|\tilde{\Xi}\|^2\sigma_{\min}(M) + \tilde{M}_{\max}\|\tilde{\Xi}\| + \bar{M}_{\max} + \bar{\varepsilon}_1 + \bar{\varepsilon}_2 \tag{A24}$$

and derives negative Lyapunov derivative by completing the squares

$$\|\tilde{\Xi}\| > -\frac{\tilde{M}_{\max}}{2\sigma_{\min}(M)} + \sqrt{\frac{\tilde{M}_{\max}^2}{4\sigma_{\min}^2(M)} - \frac{\bar{M}_{\max} + \bar{\varepsilon}_1 + \bar{\varepsilon}_2}{\sigma_{\min}(M)}} \tag{A25}$$

which implies that $\dot{L}$ is negative if the norm $\|\tilde{\Xi}\|$ exceeds a certain bound.

The above result is obtained using the standard Lyapunov extension theorem, which demonstrates that both the state and the weights are uniformly ultimately bounded.

The error dynamics for $\tilde{W}_1$ and $\tilde{W}_2$ are given by

$$\dot{\tilde{W}}_1 = -\alpha_1\bar{\sigma}_{\hat{u}}^{\mathrm{T}}\bar{\sigma}_{\hat{u}}\tilde{W}_1 + \alpha_1\bar{\sigma}_{\hat{u}}\frac{\varepsilon_{\mathrm{HJ}_1}}{m_{s1}} + \frac{\alpha_1}{m_{s1}^2}y^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\times \tilde{W}_2 + \frac{\alpha_1}{4m_{s1}^2}\left[\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u + 2W_1^{\mathrm{T}}\bar{E}_2(x)\tilde{W}_w\right] \tag{A26}$$

$$\dot{\tilde{W}}_2 = -\alpha_2\bar{\sigma}_{\hat{w}}^{\mathrm{T}}\bar{\sigma}_{\hat{w}}\tilde{W}_2 + \alpha_2\bar{\sigma}_{\hat{w}}\frac{\varepsilon_{\mathrm{HJ}_2}}{m_{s2}} + \frac{\alpha_2}{4m_{s2}^2}\left[\tilde{W}_w^{\mathrm{T}}\bar{D}_2(x)\tilde{W}_w + \tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)\tilde{W}_u + 2\tilde{W}_u^{\mathrm{T}}\bar{E}_1(x)W_2 - 2\tilde{W}_u^{\mathrm{T}}\bar{D}_1(x)W_1\right] \tag{A27}$$

The following gives the proof of effective bounds of $\|\bar{\sigma}_{\hat{u}}\tilde{W}_1\|$ and $\|\bar{\sigma}_{\hat{w}}\tilde{W}_2\|$

$$\|\bar{\sigma}_{\hat{u}}\tilde{W}_1\| > \varepsilon_{\max_1} > \bar{\varepsilon}_1\|\frac{1}{m_{s1}}\| + \|y\|\|\frac{A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)}{m_{s1}}\|\|\tilde{W}_2\| + \frac{1}{4}\|\tilde{W}_u\|^2\|\frac{\bar{D}_1(x)}{m_{s1}}\| + \frac{1}{2}\|W_1\|\times \|\frac{\bar{E}_2(x)}{m_{s1}}\|\|\tilde{W}_w\| \tag{A28}$$

$$\|\bar{\sigma}_{\hat{w}}\tilde{W}_2\| > \varepsilon_{\max_2} > \bar{\varepsilon}_2\|\frac{1}{m_{s2}}\| + \frac{1}{4}\|\tilde{W}_w\|^2\|\frac{\bar{D}_2(x)}{m_{s2}}\| + \frac{1}{4}\times \|\tilde{W}_u\|^2\|\frac{\bar{D}_1(x)}{m_{s2}}\| + \frac{1}{2}\|W_2\|\|\frac{\bar{E}_1(x)}{m_{s2}}\|\times \|\tilde{W}_u\| - \frac{1}{2}\|W_1\|\|\frac{\bar{D}_1(x)}{m_{s2}}\|\|\tilde{W}_u\| \tag{A29}$$

Thus, the proof is complete. ∎

Appendix B
Proof of Theorem 2

Proof The approximate coupled HJ equations are

$$H_2(x,\hat{W}_2,\hat{w}',\hat{u}') = H_2(x,\hat{W}_1,\hat{W}_2) = x^{\mathrm{T}}Qx + \frac{1}{4}\hat{W}_1^{\mathrm{T}}\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\hat{W}_1 + \frac{1}{4\gamma^2}\hat{W}_2^{\mathrm{T}}\nabla\phi_2(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 + \hat{W}_2^{\mathrm{T}}\times \nabla\phi_2(x)Ax - \frac{1}{2}\hat{W}_2^{\mathrm{T}}\nabla\phi_2(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\times \hat{W}_1 - \varepsilon_{\mathrm{HJ}_2} \tag{A30}$$

$$H_1(x,\hat{W}_1,\hat{w}',\hat{u}') = H_1(x,\hat{W}_1,\hat{W}_2) = x^{\mathrm{T}}Qx - \frac{1}{4}\hat{W}_1^{\mathrm{T}}\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\hat{W}_1 + \hat{W}_1^{\mathrm{T}}\nabla\phi_1(x)Ax + \frac{1}{2\gamma^2}\hat{W}_1^{\mathrm{T}}\nabla\phi_1(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 - y^{\mathrm{T}}(A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 + 2Qx) - \varepsilon_{\mathrm{HJ}_1} \tag{A31}$$

Combining Eqs. (36) and (37), Eqs. (A30) and (A31) can be rewritten as

$$H_2(x,\hat{W}_1,\hat{W}_2) = \tilde{W}_2^{\mathrm{T}}\nabla\phi_2(x)Ax - \frac{1}{4}\hat{W}_1^{\mathrm{T}}\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\hat{W}_1 + \frac{1}{4}W_1^{\mathrm{T}}\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)W_1 + \frac{1}{2}\hat{W}_2^{\mathrm{T}}\nabla\phi_2(x)\times B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\hat{W}_1 - \frac{1}{2}W_2^{\mathrm{T}}\nabla\phi_2(x)B_1R^{-1}\times B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)W_1 - \frac{1}{4\gamma^2}\tilde{W}_2^{\mathrm{T}}\nabla\phi_2(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 + \frac{1}{2\gamma^2}W_2^{\mathrm{T}}\nabla\phi_2(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 - \varepsilon_{\mathrm{HJ}_2} \tag{A32}$$

$$H_1(x,\hat{W}_1,\hat{W}_2) = \tilde{W}_1^{\mathrm{T}}\nabla\phi_1(x)Ax - \frac{1}{2\gamma^2}\hat{W}_1^{\mathrm{T}}\nabla\phi_1(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\hat{W}_2 + \frac{1}{2\gamma^2}W_1^{\mathrm{T}}\nabla\phi_1(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)W_2 - \frac{1}{2}W_1^{\mathrm{T}}\nabla\phi_1(x)\times B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\tilde{W}_1 + \frac{1}{4}\tilde{W}_1^{\mathrm{T}}\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\times \nabla\phi_1^{\mathrm{T}}(x)\tilde{W}_1 - y^{\mathrm{T}}A^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\tilde{W}_2 - \varepsilon_{\mathrm{HJ}_1} \tag{A33}$$

Letting $\|W_1\| \leqslant W_{1,\max}$ and $\|W_2\| \leqslant W_{2,\max}$, thus we have

$$\|\hat{W}_1\| = \|-\tilde{W}_1 + W_1\| \leqslant \|\tilde{W}_1\| + \|W_1\| \leqslant \|\tilde{W}_1\| + W_{1,\max} \tag{A34}$$

$$\|\hat{W}_2\| = \|-\tilde{W}_2 + W_2\| \leqslant \|\tilde{W}_2\| + \|W_2\| \leqslant \|\tilde{W}_2\| + W_{2,\max} \tag{A35}$$

By taking into account Eqs. (A34) and (A35) and $\sup_{x\in\Omega}\|\varepsilon_{\mathrm{HJ}_1}\| \leqslant \bar{\varepsilon}_1$ and $\sup_{x\in\Omega}\|\varepsilon_{\mathrm{HJ}_2}\| \leqslant \bar{\varepsilon}_2$, Eqs. (A32) and (A33) become

$$\|H_2(x,\hat{W}_1,\hat{W}_2)\| \leqslant b_{\phi_2,x}\|Ax\|\|\tilde{W}_2\| + \frac{1}{4}(\|\tilde{W}_1 + W_{1,\max}\|)^2\|\nabla\phi_1(x)B_1\times R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\| + \frac{1}{4}W_{1,\max}^2\|\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\times \nabla\phi_1^{\mathrm{T}}(x)\| + \frac{1}{2}(\|\tilde{W}_1 + W_{1,\max}\|)(\|\tilde{W}_2 + W_{2,\max}\|)\times \|\nabla\phi_2(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\| + \frac{1}{2}W_{1,\max}W_{2,\max}\times \|\nabla\phi_2(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\| + \frac{1}{4\gamma^2}\|\tilde{W}_2\|^2\|\nabla\phi_2(x)\times B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\| + \frac{1}{2\gamma^2}W_{2,\max}\|\tilde{W}_2\|\times \|\nabla\phi_2(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\| + \bar{\varepsilon}_2 \tag{A36}$$

$$\|H_1(x,\hat{W}_1,\hat{W}_2)\| = b_{\phi_1,x}\|Ax\|\|\tilde{W}_1\| + \frac{1}{2\gamma^2}(\|\tilde{W}_1 + W_{1,\max}\|)(\|\tilde{W}_2 + W_{2,\max}\|)\|\nabla\phi_1(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\| + \frac{1}{2\gamma^2}W_{1,\max}\times W_{2,\max}\|\nabla\phi_1(x)B_2B_2^{\mathrm{T}}\nabla\phi_2^{\mathrm{T}}(x)\| - \frac{1}{2}W_{1,\max}\|\tilde{W}_1\|\times \|\nabla\phi_1(x)B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\| + \frac{1}{4}\|\tilde{W}_1\|^2\|\nabla\phi_1(x)\times B_1R^{-1}B_1^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}(x)\| - \|\tilde{W}_2\|b_{\phi_2,x}\|Ay\| + \bar{\varepsilon}_1 \tag{A37}$$

The terms on the right-hand side of Eqs. (A36) and (A37) are uniformly ultimately bounded, so both $H_2(x, \hat{W}_2, \hat{w}'(x), \hat{u}'(x))$ and $H_1(x, \hat{W}_1, \hat{w}'(x), \hat{u}'(x))$ are uniformly ultimately bounded. Furthermore, $\hat{W}_1$ and $\hat{W}_2$ converge to an approximate coupled Hamilton-Jacobi solution.

Based on Theorems 1 and 2 and Eqs. (59)–(62), it follows that $\|\hat{w} - w'\|$ and $\|\hat{u} - u'\|$ are uniformly ultimately bounded, since $\|\hat{W}_w - W_2\|$ and $\|\hat{W}_u - W_1\|$ are uniformly ultimately bounded.

Therefore, the policy pair $(\hat{w}(x), \hat{u}(x))$ provides an approximate equilibrium solution for the Stackelberg game.

This completes the proof. ■

ACKNOWLEDGMENT

This work was supported by the Support Plan on Science and Technology for Youth Innovation of Universities in Shandong Province (No. 2021KJ086), the National Natural Science Foundation of China (Nos. 62003234 and 61873179), and the National Natural Science Foundation of Shandong Province (No. ZR2020QF048).

REFERENCES

[1] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory, 2nd Ed*. Philadelphia, PA, USA: SIAM, 1999.

[2] M. O. Rabin, Effective computability of winning strategies, in *Contributions to the Theory of Games*. Princeton, NJ, USA: Princeton University Press, 1957, 147–157.

[3] U. Schwalbe and P. Walker, Zermelo and the early history of game theory, *Games Econ. Behav*., 2001, 34(1), 123–137.

[4] S. M. LaValle and S. A. Hutchinson, Optimal motion planning for multiple robots having independent goals, *IEEE Trans. Rob. Autom*., 1998, 14(6), 912–925.

[5] Z. J. Ma, D. S. Callaway, and I. A. Hiskens, Decentralized charging control of large populations of plug-in electric vehicles, *IEEE Trans. Control Syst. Technol*., 2013, 21(1), 67–78.

[6] D. J. N. Limebeer, B. D. O. Anderson, and B. Hendel, A Nash game approach to mixed H_2/H_∞ control, *IEEE Trans. Autom. Control*, 1994, 39(1), 69–82.

[7] K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, Open-loop Stackelberg learning solution for hierarchical control problems, *Int. J. Adapt. Control Signal Process*., 2019, 33(2), 285–299.

[8] M. Li, J. H. Qin, Q. C. Ma, W. X. Zheng, and Y. Kang, Hierarchical optimal synchronization for linear systems via reinforcement learning: A Stackelberg-Nash game perspective, *IEEE Trans. Neural Netw. Learn. Syst*., 2021, 32(4), 1600–1611.

[9] I. Atzeni, L. G. Ordóñez, G. Scutari, D. P. Palomar, and J. R. Fonollosa, Noncooperative and cooperative optimization of distributed energy generation and storage in the demand-side of the smart grid, *IEEE Trans. Signal Process.*, 2013, 61(10), 2454–2472.

[10] C. Dextreit and I. V. Kolmanovsky, Game theory controller for hybrid electric vehicles, *IEEE Trans. Control Syst. Technol*., 2014, 22(2), 652–663.

[11] X. Q. Li, W. Wang, J. J. Xu, and H. S. Zhang, LQG control on mixed H_2/H_∞ problem: The discrete-time case, *Int. J. Syst. Sci*., 2020, 51(1), 191–201.

[12] X. Q. Li, J. J. Xu, W. Wang, and H. S. Zhang, Mixed H_2/H_∞ control for discrete-time systems with input delay, *IET Control Theory Appl*., 2018, 12(16), 2221–2231.

[13] J. J. Xu, H. S. Zhang, and T. Y. Chai, Necessary and sufficient condition for two-player Stackelberg strategy, *IEEE Trans. Autom. Control*, 2015, 60(5), 1356–1361.

[14] R. Z. Song, F. L. Lewis, and Q. L. Wei, Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games, *IEEE Trans. Neural Netw. Learn. Syst*., 2017, 28(3), 704–713.

[15] A. Odekunle, W. N. Gao, M. Davari, and Z. P. Jiang, Reinforcement learning and non-zero-sum game output regulation for multi-player linear uncertain systems, *Automatica*, 2020, 112, 108672.

[16] Q. C. Zhang and D. B. Zhao, Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics, *IEEE Trans. Cybern*., 2019, 49(8), 2874–2885.

[17] H. Kebriaei and L. Iannelli, Discrete-time robust hierarchical linear-quadratic dynamic games, *IEEE Trans. Autom. Control*, 2018, 63(3), 902–909.

[18] D. Vrabie and F. Lewis, Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems, *Neural Netw*, 2009, 22(3), 237–246.

[19] 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. Mag*., 2012, 32(6), 76–105.

[20] F. L. Lewis and D. Vrabie, Reinforcement learning and adaptive dynamic programming for feedback control, *IEEE Circ. Syst. Mag*., 2009, 9(3), 32–50.

[21] H. Modares, F. 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*., 2013, 24(10), 1513–1525.

[22] H. Modares and F. L. Lewis, Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning, *Automatica*, 2014, 50(7), 1780–1792.

[23] P. Ioannou and B. Fidan, *Advances in Design and Control: Adaptive Control Tutorial*. Philadelphia, PA, USA: SIAM, 2006.

[24] K. G. Vamvoudakis and F. L. Lewis, Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem, *Automatica*, 2010, 46(5), 878–888.

Xiaoqian Li received the ME degree in control engineering from Northeastern University, Shenyang, China, in 2013, and the PhD degree in control science and engineering from Shandong University, Jinan, China, in 2019. She is currently an associate professor at Taishan University, Taian, China. Her research interests include optimal control, reinforcement learning, and game theory.

Zonglei Jing received the ME degree from Beihang University, Beijing, China, in 2024. He is currently pursuing the PhD degree at State Key Laboratory of Software Development Environment, Beihang University, China. His research interests include reinforcement learning, adversarial robustness, and optimal control.

Peijun Ju received the BE degree in mathematics from Qufu Normal University, Jining, China, in 2000, and the ME degree in mathematics and the PhD degree in control science and engineering from Shandong University, Jinan, China, in 2004 and 2017, respectively. He is currently a professor at Taishan University, China. His research interests include time delay systems, optimal control, and game theory.