Approximate $N-$Player Nonzero-Sum Game Solution for an Uncertain Continuous Nonlinear System

Marcus Johnson, Rushikesh Kamalapurkar, Shubhendu Bhasin, and Warren E. Dixon

***Abstract*—An approximate online equilibrium solution is developed for an N-player nonzero-sum game subject to continuous-time nonlinear unknown dynamics and an infinite horizon quadratic cost. A novel actor-critic-identifier (ACI) structure is used, wherein a robust dynamic neural network (DNN) is used to asymptotically identify the uncertain system with additive disturbances, and a set of critic and actor NNs are used to approximate the value functions and equilibrium policies, respectively. The weight update laws for the actor NNs are generated using a gradient-descent method, and the critic NNs are generated by least square regression, which are both based on the modified Bellman error that is independent of the system dynamics. A Lyapunov-based stability analysis shows that UUB tracking is achieved and a convergence analysis demonstrates that the approximate control policies converge to a neighborhood of the optimal solutions. The actor, critic and identifier structures are implemented in real-time, continuously and simultaneously. Simulations on two and three player games illustrate the performance of the developed method.**

***Index Terms*—adaptive dynamic programming, differential games, actor-critic methods, optimal control, adaptive control**

I. Introduction

Noncooperative game theory [1]–[3] can be used to provide a solution to a number of control engineering applications. In a differential game formulation, the controlled system is influenced by a number of different inputs, computed by different players that are individually trying to optimize a performance function. The control objective is to determine a set of policies that are admissible [4], i.e. control policies that guarantee the stability of the dynamic system and minimize individual performance functions to yield an equilibrium. A Nash differential game consists of multiple players making simultaneous decisions where each player has an outcome that cannot be unilaterally improved from a change in strategy. Players are committed to following a predetermined strategy based on knowledge of the initial state, the system model, and the cost functional to be minimized. Solution techniques to the Nash equilibrium are classified depending on the amount of information available to the players (e.g. open-loop, feedback), the objectives of each player (zero-sum or nonzero-sum), the planning horizon (infinite horizon or finite horizon), and the nature of the dynamic constraints (e.g. continuous, discrete, linear, nonlinear).

A unique Nash equilibrium is generally not expected. Non-uniqueness issues with Nash equilibria are discussed for a nonzero-sum differential game in [5]. For an open-loop nonzero-sum game, where every player knows the initial state x_0 at time $t \in [0, T]$, conditions for the existence of a unique Nash equilibrium can be established [6]. For a closed-loop perfect state information, where every player knows the complete history of the state at time $t \in [0, T]$, there are potentially an infinite number of Nash equilibria. In this case, it is possible to restrict the Nash equilibrium to a subset of feedback solutions, which is known as the (sub)game perfect Nash equilibria (or feedback Nash equilibria). Results in [7] and [8] indicate that (sub)game perfect Nash equilibria are (at least heuristically) given by feedback strategies and that their corresponding value functions are the solution to a system of Hamilton–Jacobi equations. These concepts have been successfully applied to linear-quadratic (LQ) differential games [5], [7]. A special case of the Nash game is the min-max saddle point equilibrium, which is widely used to minimize control effort under a *worst-case* level of uncertainty. The saddle point equilibrium has been heavily exploited in H_∞ control theory [9], which considers finding the smallest gain $\gamma \geq 0$ under which the upper value of the cost function

$$J_\gamma(u, v) = \int_0^\infty Q(x) + u(x)^2 - \gamma^2 \|v(x)\|^2 d\tau, \quad (1)$$

is bounded and finding the corresponding controller that achieves this upper bound. H_∞ control theory relates to LQ dynamic games in the sense that the worst-case H_∞ design problems have equal upper and lower bounds of the objective function in (1), which results in the saddle-point solution to the LQ game problem. In both the H_∞ control problem and the LQ problem, the underlying dynamic optimization is a two player zero-sum game with the controller being the minimizing player and the disturbance being the maximizing player. In a zero-sum game with linear dynamics and an infinite horizon quadratic cost function, the Nash equilibrium solution is equivalent to solving the generalized game algebraic Riccati equation (GARE). However for nonlinear dynamics or a nonzero-sum game, analytic solutions may not be tractable for the Hamilton-Jacobi-Bellman (HJB) partial

Marcus Johnson, Rushikesh Kamalapurkar, and Warren Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA. Email: {rkamalapurkar, walters8, wdixon}@ufl.edu.

Shubhendu Bhasin is with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, India. email: sbhasin@ee.iitd.ac.in.

This research is supported in part by NSF award numbers 0547448, 0901491 and 1161260 and ONR grant number N00014-13-1-0151. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agency.

differential equation.

Due to the difficulty involved in determining a solution to the HJB equation, dynamic programming [10]–[14] is used to approximate a solution to the optimal control problem. Reinforcement learning (RL) is typically employed to implement dynamic programming online and forward in time. RL is a method wherein appropriate actions are learned based on evaluative feedback from the environment. A widely used RL method is based on the actor-critic (AC) architecture, where an actor performs certain actions by interacting with its environment, the critic evaluates the actions and gives feedback to the actor, leading to an improvement in the performance of subsequent actions. AC algorithms are pervasive in machine learning and are used to learn the optimal policy online for finite-space discrete-time Markov decision problems [15]–[17].

The machine learning community [15], [17]–[20] provides an approach to determining the solution of an optimal control problem using Approximate Dynamic Programming (ADP) through RL-based adaptive critics [10]–[14]. The discrete/iterative nature of the ADP formulation naturally leads into the design of discrete-time optimal controllers [13], [21]–[25].

Some results have also been developed for continuous time problems. Baird [26] proposed Advantage Updating, an extension of the Q-learning algorithm which could be implemented in continuous-time and provided fast convergence. A HJB-based framework is used in [27] and [28], and Galerkin's spectral method is used to approximate the generalized HJB solution in [29].

The aforementioned approaches for continuous-time nonlinear systems are computed offline and/or require complete knowledge of system dynamics. A contribution in [30] is the requirement of only partial knowledge of the system and a hybrid continuous-time/discrete-time sampled data controller is developed based on policy iteration (PI), where the feedback control operation of the actor occurs at faster time scale than the learning process of the critic. The method in [31] was extended by designing a hybrid model-based online algorithm called synchronous PI which involved synchronous continuous-time adaptation of both actor and critic neural networks. Bhasin et al. [32] developed a continuous actor-critic-identifier (ACI) technique to solve the infinite horizon single player optimal control problem by using a robust dynamic neural network (DNN) to identify the dynamics and a critic NN to approximate the value function. This technique removes the requirement of complete knowledge of the system drift dynamics through the use of an indirect adaptive control technique.

Most of the previous continuous-time RL algorithms that provide an online approximate optimal solution assume that the dynamical system is affected by a single control strategy. Previous research has also investigated the generalization of RL controllers to differential game problems [31], [33]–[39]. Techniques utilizing Q-learning algorithms have been developed for a zero-sum game in [40]. An ADP procedure that provides a solution to the HJI equation associated with the two-player zero-sum nonlinear differential game is introduced in [33]. The ADP algorithm involves two iterative cost functions finding the upper and lower performance indices as sequences that converge to the saddle point solution of the game. The AC structure required for learning the saddle point solution is composed of four action networks and two critic networks. The iterative ADP solution in [34] considers solving zero-sum differential games under the condition that the saddle point does not exist, and a mixed optimal performance index function is obtained under a deterministic mixed optimal control scheme when the saddle point does not exist. Another ADP iteration technique is presented in [35], in which the nonlinear quadratic zero-sum game is transformed into an equivalent sequence of linear quadratic zero-sum games to approximate an optimal saddle point solution. In [36], an integral RL method is used to determine an online solution to the two player nonzero-sum game for a linear system without complete knowledge of the dynamics. The synchronous PI method in [31] was then further generalized to solve the two-player zero-sum game problem in [38] and a multi-player nonzero-sum game in [39] and [41] for nonlinear continuous-time systems with known dynamics. Furthermore, [42] presents a policy iteration method for an infinite horizon two-player zero-sum Nash game with unknown nonlinear continuous-time dynamics. The proposed work expands upon the applicability of [31], [38], [39], [41] by removing the assumption that the drift dynamics are known, and advances the theory in [32], [42] to solve the more general multi-player nonzero-sum differential game where the objective is to minimize a set of coupled cost functions. The single player game in [32] and two-player zero-sum game in [42] are special cases of the multi-player nonzero-sum game presented in this paper.

This paper aims to solve a N-player nonzero-sum infinite horizon differential game subject to continuous-time uncertain nonlinear dynamics. The main contribution of this work is deriving an approximate solution to a N-player nonzero-sum game with a continuous controller using an ACI technique. Previous research has focused on scalar nonlinear systems or implemented iterative/hybrid techniques that required complete knowledge of the drift dynamics. The developed technique uses N-actor and N-critic neural network structures to approximate the optimal control laws and the optimal value function set, respectively. The main traits of this online algorithm involve the use of ADP techniques and adaptive theory to determine the Nash equilibrium solution of the game in manner that does not require full knowledge of the system dynamics and approximately solves the underlying set of coupled HJB equations of the game problem. For an equivalent nonlinear system, previous research makes use of offline procedures or requires full knowledge of the system dynamics to determine the Nash equilibrium. A Lyapunov-based stability analysis shows that UUB tracking for the closed-loop system is guaranteed for the proposed ACI architecture and a convergence analysis demonstrates that the approximate control policies converge to a neighborhood of the optimal solutions.

II. N-Player Differential Game for Nonlinear Systems

Consider the N-player nonlinear, time-invariant, affine in the input dynamic system given by

$$\dot{x} = f(x) + \sum_{j=1}^{N} g_j(x) u_j \tag{2}$$

where $x \in \mathbb{R}^n$ is the state vector, $u_j \in \mathbb{R}^{m_j}$ are the control inputs, and $f : \mathbb{R}^n \to \mathbb{R}^n$, and $g_j : \mathbb{R}^n \to \mathbb{R}^{n \times m_j}$ are the drift and input matrices, respectively. Assume that $g_1, \ldots, g_N$, and f are second order differentiable, and that $f(0) = 0$ and $g_j(0) = 0$ so that $x = 0$ is an equilibrium point for (2). The infinite-horizon scalar cost functional J_i associated with each player can be defined as

$$J_i = \int_t^{\infty} r_i(x, u_1, u_2, \cdots u_N)\, ds, \tag{3}$$

where $i \in \{1, \cdots, N\}$, t is the initial time and $r_i : \mathbb{R}^{n+\sum_{j=1}^{N} m_j} \to \mathbb{R}$ is the local cost for the state and control, defined as

$$r_i(x, u_1, \cdots, u_N) = Q_i(x) + \sum_{j=1}^{N} u_j^T R_{ij} u_j, \tag{4}$$

where $Q_i : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable and positive definite functions and $R_{ij} \in \mathbb{R}^{m_j \times m_j}$ and $R_{ii} \in \mathbb{R}^{m_i \times m_i}$ are positive definite symmetric matrices.

The objective of the N-player game is to find a set of admissible feedback policies $(u_1^*, u_2^*, \ldots, u_N^*)$ such that the value function $V_i : \mathbb{R}^n \to \mathbb{R}$ given in (3)

$$V_i(x; u_1, \cdots u_N) = \int_t^{\infty} \left(Q_i(x) + \sum_{j=1}^{N} u_j^T R_{ij} u_j, \right) ds, \tag{5}$$

is minimized, where $V_i(x; u_1, \cdots u_N)$ denotes the value of state x under feedback policies $(u_1, \cdots, u_N)$. This paper will focus on the Nash equilibrium solution for the N-player game, in which the following N inequalities are satisfied for all $u_i^* \in \mathcal{U}_i, i \in N$:

$$\begin{aligned} V_1^* &\triangleq V_1(x; u_1^*, u_2^*, \cdots u_N^*) \leq V_1(x; u_1, u_2^*, \cdots u_N^*) \\ V_2^* &\triangleq V_2(x; u_1^*, u_2^*, \cdots u_N^*) \leq V_2(x; u_1^*, u_2, \cdots u_N^*) \\ &\cdots \\ V_N^* &\triangleq V_N(x; u_1^*, u_2^*, \cdots u_N^*) \leq V_N(x; u_1^*, u_2^*, \cdots u_N), \end{aligned} \tag{6}$$

where $\mathcal{U}_i$ denotes the set of admissible policies for the i^{th} player. The Nash equilibrium outcome of the N-player game is given by the N-tuple of quantities $\{V_1^*, V_2^*, \cdots, V_N^*\}$. The value functions can be alternately presented by a differential equivalent given by the following nonlinear Lyapunov equation [39]

$$\begin{aligned} 0 &= r(x, u_1, \ldots, u_N) + \nabla V_i^* \left(f(x) + \sum_{j=1}^{N} g_j(x) u_j \right), \\ V_i^*(0) &= 0, \end{aligned} \tag{7}$$

where $\nabla V_i^* \triangleq \frac{\partial V_i^*}{\partial x} \in \mathbb{R}^{1 \times n}$. Assuming the value functional is continuously differentiable, Bellman's principle of optimality can be used to derive the following optimality condition

$$\begin{aligned} 0 &= \min_{u_i} \left[\nabla V_i^* \left(f + \sum_{j=1}^{N} g_j u_j \right) + r \right], \\ V_i^*(0) &= 0,\ i \in N \end{aligned} \tag{8}$$

which is the N-coupled set of nonlinear PDEs called the HJB equation. Suitable non-negative definite solutions to (7) can be used to evaluate the infinite integral in (5) along the system trajectories. A closed-form expression of the optimal feedback control policies are given by

$$u_i^*(x) = -\frac{1}{2} R_{ii}^{-1} g_i^T(x) \left(\nabla V_i^*(x) \right)^T. \tag{9}$$

The closed-form expression for the optimal control policies in (9), obviates the need to search for a set of feedback policies that minimize the value function; however, the solution V_i^* to the HJB equation given in (8) is required. The HJB equation in (8), can be rewritten by substituting for the local cost in (4) and the optimal control policy in (9), respectively, as

$$\begin{aligned} 0 &= Q_i + \nabla V_i^* f - \frac{1}{2} \nabla V_i^* \sum_{j=1}^{N} g_j R_{jj}^{-1} g_j^T \left(\nabla V_i^* \right)^T \\ &+ \frac{1}{4} \sum_{j=1}^{N} \nabla V_i^* g_j R_{jj}^{-T} R_{ij} R_{jj}^{-1} g_j^T \left(\nabla V_i^* \right)^T, \\ V_i^*(0) &= 0. \end{aligned} \tag{10}$$

Since the HJB equation may not have an analytical solution in general, an approximate solution is sought. Although nonzero-sum games contain non-cooperative components, the solution to each player's coupled HJB equation in (10) requires knowledge of all the other player's strategies in (9). The underlying assumption of rational opponents [43] is characteristic of differential game theory problems and it implies that the players share information, yet they agree to adhere to the equilibrium policy determined from the Nash game.

III. Hamilton-Jacobi-Bellman Approximation via Actor-Critic-Identifier

This paper uses an actor-critic-identifier (ACI) [42], [44] approximation architecture to solve for (10). The ACI architecture eliminates the need for exact model knowledge and utilizes a DNN to robustly identify the system, a critic NN to approximate the value function and an actor NN to find a control policy which minimizes the value functions. The following development focuses on the solution to a two player nonzero-sum game. The approach can easily be extended to the N-player game presented in Section 2. This section introduces the actor-critic-identifier architecture, and subsequent sections give details of the design for the two player nonzero-sum game solution.

The Hamiltonian $H_i \in \mathbb{R}$ of the system in (2) can be defined as

$$H_i = r_{u_i} + \nabla V_i F_u, \tag{11}$$

where $\nabla V_i \triangleq \frac{\partial V_i}{\partial x} \in \mathbb{R}^{1\times n}$ denotes the Jacobian of the value function V_i and

$$F_u\left(x, u_1, \ldots, u_N\right) \triangleq f\left(x\right) + \sum_{j=1}^{N} g_j\left(x\right) u_j \in \mathbb{R}^n$$

denotes the system dynamics. The optimal policies in (9) and the associated value functions V_i^* satisfy the HJB equation with the corresponding Hamiltonian as

$$H_i\left(x, \nabla V_i^*, u_1^*, \ldots, u_N^*\right) = r_{u_i^*} + \nabla V_i^* F_{u^*} = 0. \tag{12}$$

Replacing the optimal Jacobian ∇V_i^* and optimal control policies u_i^* by estimates $\nabla \hat{V}_i$ and $\hat{u}_i$, respectively, yields the approximate Hamiltonian

$$H_i\left(x, \nabla \hat{V}_i, \hat{u}_1, \ldots, \hat{u}_N\right) = r_{\hat{u}_i} + \nabla \hat{V}_i F_{\hat{u}}. \tag{13}$$

The approximate Hamiltonian in (13) is dependent on complete knowledge of the system. To overcome this limitation, an online system identifier replaces the system dynamics which modifies the approximate Hamiltonian in (13) as

$$H_i\left(x, \hat{x}, \nabla \hat{V}_i, \hat{u}_1, \ldots, \hat{u}_N\right) = r_{\hat{u}_i} + \nabla \hat{V}_i \hat{F}_{\hat{u}}, \tag{14}$$

where $\hat{F}_{\hat{u}}$ is an approximation of the system dynamics $F_{\hat{u}}$. The difference between the optimal and approximate Hamiltonians equations in (12) and (14) yields the Bellman residual errors $\delta_{hjb_i} \in \mathbb{R}$ defined as

$$\begin{aligned}\delta_{hjb_i} \triangleq\ & H_i\left(x, \hat{x}, \nabla \hat{V}_i, \hat{u}_1, \ldots, \hat{u}_N\right) \\ & - H_i\left(x, \nabla V_i^*, u_1^*, \ldots, u_N^*\right).\end{aligned} \tag{15}$$

However since $H_i = 0 \quad \forall i \in N$, the Bellman residual error can be defined in a measurable form as $\delta_{hjb_i} = H_i\left(x, \hat{x}, \nabla \hat{V}_i, \hat{u}_1, \ldots, \hat{u}_N\right)$. The objective is to update both $\hat{u}_i$ (actors) and $\hat{V}_i$ (critics) simultaneously, based on the minimization of the Bellman residual errors δ_{hjb_i}. All together, the actors $\hat{u}_i$, the critics $\hat{V}_i$, and the identifiers $\hat{F}_{\hat{u}}$, constitute the ACI architecture. To facilitate the subsequent analysis the following properties are given.

Property 1. Given a continuous function $h : \mathbb{S} \to \mathbb{R}^n$, where $\mathbb{S}$ is a compact set, there exist ideal weights W,V such that the function can be represented by a NN as $h\left(x\right) = W^T \sigma\left(V^T x\right) + \varepsilon\left(x\right)$, where $\sigma\left(\cdot\right)$ is a nonlinear activation function and $\varepsilon\left(x\right)$ is the function reconstruction error.

Property 2. The NN activation function $\sigma\left(\cdot\right)$ and its time derivative $\sigma'\left(\cdot\right)$ with respect to its argument is bounded.

Property 3. The ideal NN weight matrices are bounded by known positive constants [45], i.e., $||W|| \leq \bar{W}$ and $||V|| \leq \bar{V}$.

Property 4. The NN function reconstruction errors and their derivatives are bounded [45], i.e., $||\varepsilon|| \leq \bar{\varepsilon}$ and $||\varepsilon'|| \leq \bar{\varepsilon}'$.

IV. System Identifier

Consider the two-player case for the dynamics given in (2) as

$$\dot{x} = f\left(x\right) + g_1\left(x\right) u_1 + g_2\left(x\right) u_2, \tag{16}$$

$$x\left(0\right) = x_0,$$

where $u_1, u_2 \in \mathbb{R}^n$ are the control inputs, and the state $x \in \mathbb{R}^n$ is assumed to be measurable. The following assumptions about the system will be utilized in the subsequent development.

Assumption 1. The input matrices g_1 and g_2 are known and bounded i.e. $\|g_1\| \leq \bar{g}_1$ and $\|g_2\| \leq \bar{g}_2$ where $\bar{g}_1$ and $\bar{g}_2$ are known positive constants.

Assumption 2. The control inputs u_1 and u_2 are bounded i.e. $u_1, u_2 \in \mathcal{L}_\infty$.

Based on Property 1, the nonlinear system in (16) can be represented using a multi-layer NN as

$$\begin{aligned}\dot{x} \quad = F_u\left(x, u_1, u_2\right) = \quad & W_f^T \sigma_f\left(V_f^T x\right) + \varepsilon_f\left(x\right) \\ & + g_1\left(x\right) u_1 + g_2\left(x\right) u_2,\end{aligned} \tag{17}$$

where $W_f \in \mathbb{R}^{N_f+1\times n}$, $V_f \in \mathbb{R}^{n\times N_f}$ are unknown ideal NN weight matrices with N_f representing the neurons in the output layers. The activation function is given by $\sigma_f = \sigma\left(V_f^T x\right) \in \mathbb{R}^{N_f+1}$, and $\varepsilon_f : \mathbb{R}^n \to \mathbb{R}^n$ is the function reconstruction error in approximating the function f. The proposed multi-layer dynamic neural network (MLDNN) used to identify the system in (16) is

$$\begin{aligned}\dot{\hat{x}} &= \hat{F}_u\left(x, \hat{x}, u_1, u_2\right) \\ &= \hat{W}_f^T \hat{\sigma}_f + g_1\left(x\right) u_1 + g_2\left(x\right) u_2 + \mu,\end{aligned} \tag{18}$$

where $\hat{x} \in \mathbb{R}^n$ is the state of the MLDNN, $\hat{W}_f \in \mathbb{R}^{N_f+1\times n}$, $\hat{V}_f \in \mathbb{R}^{n\times N_f}$ are the estimates of the ideal weights of the NNs, $\hat{\sigma}_f = \hat{\sigma}\left(\hat{V}_f^T \hat{x}\right) \in \mathbb{R}^{N_f+1}$ are the NN activation functions, and $\mu \in \mathbb{R}^n$ denotes the RISE feedback term defined as

$$\mu \triangleq k\left(\tilde{x}\left(t\right) - \tilde{x}\left(0\right)\right) + \nu, \tag{19}$$

where the measurable identification error $\tilde{x} \in \mathbb{R}^n$ is defined as

$$\tilde{x} \triangleq x - \hat{x}, \tag{20}$$

and $\nu \in \mathbb{R}^n$ is a generalized Filippov solution to the differential equation

$$\dot{\nu} = \left(k\alpha + \gamma\right)\tilde{x} + \beta_1 sgn\left(\tilde{x}\right); \qquad \nu\left(0\right) = 0,$$

where k, α, γ $\beta \in \mathbb{R}$ are positive constant gains, and $sgn\left(\cdot\right)$ denotes a vector signum function. The identification error dynamics are developed by taking the time derivative of (20) and substituting for (17) and (18) as

$$\begin{aligned}\dot{\tilde{x}} &= \tilde{F}_u\left(x, \hat{x}, u_1, u_2\right) \\ &= W_f^T \sigma_f - \hat{W}_f^T \hat{\sigma}_f + \varepsilon_f\left(x\right) - \mu,\end{aligned} \tag{21}$$

where $\tilde{F}_u = F_u - \hat{F}_u$. An auxiliary identification error is defined as

$$r \triangleq \dot{\tilde{x}} + \alpha\tilde{x}. \tag{22}$$

Taking the time derivative of (22) and using (21) yields

$$\begin{aligned}\dot{r} = {} & W_f^T \sigma_f' V_f^T \dot{x} - \dot{\hat{W}}_f^T \hat{\sigma}_f - \hat{W}_f^T \hat{\sigma}_f' \dot{\hat{V}}_f^T \hat{x} - \hat{W}_f^T \hat{\sigma}_f' \hat{V}_f^T \dot{\hat{x}} \\ & + \dot{\varepsilon}_f(x) - kr - \gamma\tilde{x} - \beta_1 sgn(\tilde{x}) + \alpha\dot{\tilde{x}},\end{aligned} \tag{23}$$

where $\hat{\sigma}'_f = d\sigma\left(V^T\hat{x}\right)/d\left(V^T\hat{x}\right)\big|_{V^T=\hat{V}^T} \in \mathbb{R}^{(N_f+1)\times N_f}$. The weight update laws for the DNN in (18) are developed based on the subsequent stability analysis as

$$\dot{\hat{W}}_f = proj(\Gamma_{wf}\hat{\sigma}'_f\hat{V}_f^T\dot{\hat{x}}\tilde{x}^T),\ \dot{\hat{V}}_f = proj(\Gamma_{vf}\dot{\hat{x}}\tilde{x}^T\hat{W}_f^T\hat{\sigma}'_f), \tag{24}$$

where $proj(\cdot)$ is a smooth projection operator [46], [47], and $\Gamma_{wf} \in \mathbb{R}^{N_f+1\times N_f+1}$, $\Gamma_{vf} \in \mathbb{R}^{n\times n}$ are positive constant adaptation gain matrices. Adding and subtracting $\frac{1}{2}W_f^T\hat{\sigma}'_f\hat{V}_f^T\dot{\hat{x}} + \frac{1}{2}\hat{W}_f^T\hat{\sigma}'_fV_f^T\dot{\hat{x}}$, and grouping similar terms, the expression in (23) can be rewritten as

$$\dot{r} = \tilde{N} + N_{B1} + \hat{N}_{B2} - kr - \gamma\tilde{x} - \beta_1 sgn(\tilde{x}), \tag{25}$$

where the auxiliary signals, $\tilde{N}$, N_{B1}, and $\hat{N}_{B2} \in \mathbb{R}^n$ in (25) are defined as

$$\begin{aligned}\tilde{N} \triangleq\ & \alpha\dot{\tilde{x}} - \dot{\hat{W}}_f^T\hat{\sigma}_f - \hat{W}_f^T\hat{\sigma}'_f\dot{\hat{V}}_f^T\hat{x} \\ & + \frac{1}{2}W_f^T\hat{\sigma}'_f\hat{V}_f^T\dot{\tilde{x}} + \frac{1}{2}\hat{W}_f^T\hat{\sigma}'_fV_f^T\dot{\tilde{x}},\end{aligned} \tag{26}$$

$$\begin{aligned}N_{B1} \triangleq\ & W_f^T\sigma'_fV_f^T\dot{x} - \frac{1}{2}W_f^T\hat{\sigma}'_f\hat{V}_f^T\dot{x} \\ & - \frac{1}{2}\hat{W}_f^T\hat{\sigma}'_fV_f^T\dot{x} + \dot{\varepsilon}_f(x),\end{aligned} \tag{27}$$

$$\hat{N}_{B2} \triangleq \frac{1}{2}\tilde{W}_f^T\hat{\sigma}'_f\hat{V}_f^T\dot{\hat{x}} + \frac{1}{2}\hat{W}_f^T\hat{\sigma}'_f\tilde{V}_f^T\dot{\hat{x}}. \tag{28}$$

To facilitate the subsequent stability analysis, an auxiliary term $N_{B2} \in \mathbb{R}^n$ is defined by replacing $\dot{\hat{x}}$ in $\hat{N}_{B2}$ by $\dot{x}$, and $\tilde{N}_{B2} \triangleq \hat{N}_{B2} - N_{B2}$. The terms N_{B1} and N_{B2} are grouped as $N_B \triangleq N_{B1} + N_{B2}$. Using Properties 1-4, Assumption 1, (22), (24), (27) and (28) the following inequalities can be obtained

$$\left\|\tilde{N}\right\| \leq \rho_1(\|z\|)\|z\|,\ \|N_{B1}\| \leq \zeta_1,\ \|N_{B2}\| \leq \zeta_2, \tag{29}$$

$$\left\|\dot{N}_B\right\| \leq \zeta_3 + \zeta_4\rho_2(\|z\|)\|z\|, \tag{30}$$

$$\left\|\dot{\tilde{x}}^T\tilde{N}_{B2}\right\| \leq \zeta_5\|\tilde{x}\|^2 + \zeta_6\|r\|^2, \tag{31}$$

where $z \triangleq \left[\tilde{x}^T\ r^T\right]^T \in \mathbb{R}^{2n}$, and $\rho_1, \rho_2 : \mathbb{R} \to \mathbb{R}$ are positive, globally invertible, non-decreasing functions, and $\zeta_i \in \mathbb{R}$, $i = 1, ..., 6$ are computable positive constants. To facilitate the subsequent stability analysis, let $\mathcal{D} \subset \mathbb{R}^{2n+2}$ be a domain containing $y = 0$, where $y \in \mathbb{R}^{2n+2}$ is defined as

$$y \triangleq \left[\tilde{x}^T\ \ r^T\ \ \sqrt{P}\ \ \sqrt{Q_f}\right]^T, \tag{32}$$

where the auxiliary signal $P \in \mathbb{R}$ is the generalized Filippov solution to the differential equation [48]

$$\begin{aligned}\dot{P} =\ & -r^T(N_{B1} - \beta_1 sgn(\tilde{x})) - \dot{\tilde{x}}^T N_{B2} \\ & + \beta_2\rho_2(\|z\|)\|z\|\|\tilde{x}\|, \\ P(0) =\ & \beta_1\sum_{i=1}^{n}|\tilde{x}_i(0)| - \tilde{x}^T(0)N_B(0),\end{aligned} \tag{33}$$

where $\beta_1, \beta_2 \in \mathbb{R}$ are chosen according to the sufficient conditions[1]

$$\beta_1 > \max(\zeta_1 + \zeta_2,\ \zeta_1 + \frac{\zeta_3}{\alpha}), \qquad \beta_2 > \zeta_4, \tag{34}$$

[1]The derivation of the sufficient conditions in (34) is provided in the Appendix.

such that $P(t) \geq 0$ for all $t \in [0, \infty)$. The auxiliary function $Q_f : \mathbb{R}^{n(2N_f+1)} \to \mathbb{R}$ in (32) is defined as $Q_f \triangleq \frac{1}{4}\alpha\left[tr(\tilde{W}_f^T\Gamma_{wf}^{-1}\tilde{W}_f) + tr(\tilde{V}_f^T\Gamma_{vf}^{-1}\tilde{V}_f)\right]$, where $tr(\cdot)$ denotes the trace of a matrix.

Theorem 1. *For the system in (16), the identifier developed in (18) along with its weight update laws in (24) ensures asymptotic identification of the state and its derivative, in the sense that*

$$\lim_{t\to\infty}\|\tilde{x}(t)\| = 0 \qquad \textit{and} \qquad \lim_{t\to\infty}\left\|\dot{\tilde{x}}(t)\right\| = 0,$$

provided Assumptions 1 and 2 hold, and the control gains k and γ are chosen sufficiently large based on the initial conditions of the states,[2] and satisfy the following sufficient conditions

$$\alpha\gamma > \zeta_5, \qquad k > \zeta_6, \tag{35}$$

where ζ_5 and ζ_6 are introduced in (31), and β_1, β_2 introduced in (33), are chosen according to the sufficient conditions in (34).

Proof: To facilitate the subsequent development, let the gains k and γ_f be split as $k \triangleq k_1 + k_2$, $\gamma_f \triangleq \gamma_1 + \gamma_2$ and let $\lambda \triangleq min\{\alpha\gamma_1 - \zeta_5,\ k_1 - \zeta_6\}$, $\rho(\|z\|)^2 \triangleq \rho_1(\|z\|)^2 + \rho_2(\|z\|)^2$, and $\eta \triangleq min\{k_2, \frac{\alpha\gamma_2}{\beta_2^2}\}$. Let $\mathcal{D} \triangleq \left\{y(t) \in \mathbb{R}^{2n+2} \mid \|y\| \leq \rho^{-1}\left(2\sqrt{\lambda\eta}\right)\right\}$ and let $V_I : \mathcal{D} \to \mathbb{R}$ be a positive definite function defined as

$$V_I \triangleq \frac{1}{2}r^Tr + \frac{1}{2}\gamma_f\tilde{x}^T\tilde{x} + P + Q_f, \tag{36}$$

which satisfies the following inequalities:

$$U_1(y) \leq V_I(y) \leq U_2(y), \tag{37}$$

where U_1, $U_2 : \mathbb{R}^{2n+2} \to \mathbb{R}$ are continuous positive definite functions defined as

$$U_1(y) \triangleq \frac{1}{2}min(1, \gamma_f)\|y\|^2 \qquad U_2(y) \triangleq max(1, \gamma_f)\|y\|^2.$$

Let $\dot{y} = h(y, t)$ represent the closed-loop differential equations in (21), (24), (25), and (33), where $h : \mathbb{R}^{2n+2} \times [0, \infty) \to \mathbb{R}^{2n+2}$ denotes the right-hand side of the closed-loop error signals. Using Filippov's theory of differential inclusion [49], the existence of solutions can be established for $\dot{y} \in K[h](y, t)$, where $K[h] \triangleq \bigcap_{\delta>0}\bigcap_{\mu M=0}\overline{co}h(B(y, \delta)\backslash M, t)$, where $\bigcap_{\mu M=0}$ denotes the intersection over all sets M of Lebesgue measure zero, $\overline{co}$ denotes convex closure, and $B(y, \delta) = \left\{w \in R^{4n+2} \mid \|y - w\| < \delta\right\}$. The generalized time derivative of (36) exists almost everywhere (a.e.), and $\dot{V}_I(y) \in^{a.e.} \dot{\tilde{V}}_I(y)$ where

$$\dot{\tilde{V}}_I = \bigcap_{\xi\in\partial V_I(y)}\xi^TK\left[\dot{r}^T\ \ \dot{\tilde{x}}^T\ \ \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\ \ \frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}\right]^T,$$

where ∂V_I is the generalized gradient of V_I [50]. Since $V_I : \mathcal{D} \to \mathbb{R}$ is continuously differentiable $\dot{\tilde{V}}_I$ can be simplified as

$$\dot{\tilde{V}}_I = \nabla V_I^TK\left[\dot{r}^T\ \ \dot{\tilde{x}}^T\ \ \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\ \ \frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}\right]^T$$

[2]See subsequent stability analysis.

$$= \left[r^T \ \gamma_f \tilde{x}^T \ 2P^{\frac{1}{2}} \ 2Q^{\frac{1}{2}}\right] K \left[\dot{r}^T \ \dot{\tilde{x}}^T \ \frac{1}{2}P^{-\frac{1}{2}}\dot{P} \ \frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}\right]^T.$$

Using the calculus for $K[\cdot]$ from [51], and substituting the dynamics from (25) and (33), yields

$$\begin{aligned}\dot{\tilde{V}}_I \subset\ & r^T(\tilde{N} + N_{B1} + \hat{N}_{B2} - kr - \beta_1 K[sgn(\tilde{x})] - \gamma_f \tilde{x}) \\ & + \gamma_f \tilde{x}^T(r - \alpha\tilde{x}) - r^T(N_{B1} - \beta_1 K[sgn(\tilde{x})]) \\ & - \dot{\tilde{x}}^T N_{B2} + \beta_2 \rho_2(\|z\|)\|z\|\|\tilde{x}\| \\ & - \frac{1}{2}\alpha\left[tr(\tilde{W}_f^T \Gamma_{wf}^{-1}\dot{\hat{W}}_f) + tr(\tilde{V}_f^T \Gamma_{vf}^{-1}\dot{\hat{V}}_f)\right], \end{aligned} \tag{38}$$

where $K[sgn(\tilde{x})] = SGN(\tilde{x})$ [51], such that $SGN(\tilde{x}_i) = 1$ if $\tilde{x}_i > 0$, $[-1, 1]$ if $\tilde{x}_i = 0$, and -1 if $\tilde{x}_i < 0$. Substituting (24), canceling common terms, and rearranging the expression yields

$$\begin{aligned}\dot{\tilde{V}}_I \overset{a.e.}{\leq}\ & -\alpha\gamma_f \tilde{x}^T\tilde{x} - kr^T r + r^T\tilde{N} + \frac{1}{2}\alpha\tilde{x}^T\tilde{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}} \\ & + \frac{1}{2}\alpha\tilde{x}^T\hat{W}_f^T\hat{\sigma}_f'\tilde{V}_f^T\dot{\hat{x}} + \dot{\tilde{x}}^T(\hat{N}_{B2} - N_{B2}) \\ & + \beta_2\rho_2(\|z\|)\|z\|\|\tilde{x}\| - \frac{1}{2}\alpha tr(\tilde{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}}\tilde{x}^T) \\ & - \frac{1}{2}\alpha tr(\tilde{V}_f^T\dot{\hat{x}}\tilde{x}^T\hat{W}_f^T\hat{\sigma}_f'). \end{aligned} \tag{39}$$

The set inclusion in (38) reduces to the scalar inequality in (39) because the RHS of (38) is set valued only on the Lebesgue negligible set of times $\{t \mid \tilde{x}(t) = 0\}$. Substituting for $k \triangleq k_1 + k_2$ and $\gamma_f \triangleq \gamma_1 + \gamma_2$, using (24), (29), and (31), and completing the squares, the expression in (39) can be upper bounded as

$$\begin{aligned}\dot{\tilde{V}}_I \overset{a.e.}{\leq}\ & -(\alpha\gamma_1 - \zeta_5)\|\tilde{x}\|^2 - (k_1 - \zeta_6)\|r\|^2 \\ & + \frac{\rho_1(\|z\|)^2}{4k_2}\|z\|^2 + \frac{\beta_2^2\rho_2(\|z\|)^2}{4\alpha\gamma_2}\|z\|^2. \end{aligned} \tag{40}$$

Provided the sufficient conditions in (35) are satisfied, the expression in (40) can be rewritten as

$$\dot{\tilde{V}}_I \overset{a.e.}{\leq} -\lambda\|z\|^2 + \frac{\rho(\|z\|)^2}{4\eta}\|z\|^2 \overset{a.e.}{\leq} -U(y),\ \forall y \in \mathcal{D}. \tag{41}$$

In (41), $U(y) = c\|z\|^2$ is a continuous, positive semi-definite function defined on $\mathcal{D}$, where c is a positive constant.

The inequalities in (37) and (41) can be used to show that $V_I \in \mathcal{L}_\infty$; hence, $\tilde{x}, r \in \mathcal{L}_\infty$. Using (22), standard linear analysis can be used to show that $\dot{\tilde{x}} \in \mathcal{L}_\infty$, and since $\dot{x} \in \mathcal{L}_\infty$, $\dot{\hat{x}} \in \mathcal{L}_\infty$. Since $\hat{W}_f \in \mathcal{L}_\infty$ from the use of projection in (24), $\hat{\sigma}_f \in \mathcal{L}_\infty$ from Property 2, and $u \in \mathcal{L}_\infty$ from Assumption 2, (18) can be used to conclude that $\mu \in \mathcal{L}_\infty$. Using the above bounds and the fact that $\hat{\sigma}_f', \dot{\varepsilon}_f \in \mathcal{L}_\infty$, it can be shown from (23) that $\dot{r} \in \mathcal{L}_\infty$. Let $\mathcal{S} \subset \mathcal{D}$ denote a set defined as

$$\mathcal{S} \triangleq \left\{ y \subset \mathcal{D} \mid U_2(y) < \frac{1}{2}\left(\rho^{-1}\left(2\sqrt{\lambda\eta}\right)\right)^2 \right\}. \tag{42}$$

From (41), [52, Corollary 1] can be invoked to show that $c\|z(t)\|^2 \to 0$ as $t \to \infty$, $\forall y(0) \in \mathcal{S}$. Using the definition of z the following result can be shown

$$\|\tilde{x}(t)\|, \left\|\dot{\tilde{x}}(t)\right\|, \|r(t)\| \to 0 \quad \text{as} \quad t \to \infty, \quad \forall y(0) \in \mathcal{S}.$$

Note that the region of attraction in (42) can be made arbitrarily large to include any initial conditions by increasing the control gain η. ■

V. Actor-critic design

Using Property 1 and (9), the optimal value function and the optimal controls can be represented by NNs as

$$\begin{aligned} V_1^*(x) &= W_1^T\phi_1(x) + \varepsilon_1(x); \\ u_1^*(x) &= -\frac{1}{2}R_{11}^{-1}g_1^T(x)\left(\phi_1'^T(x)W_1 + \varepsilon_1'(x)^T\right), \\ V_2^*(x) &= W_2^T\phi_2(x) + \varepsilon_2(x); \\ u_2^*(x) &= -\frac{1}{2}R_{22}^{-1}g_2^T(x)\left(\phi_2'^T(x)W_2 + \varepsilon_2'(x)^T\right), \end{aligned} \tag{43}$$

where $W_1, W_2 \in \mathbb{R}^N$ are unknown ideal NN weights, N is the number of neurons, $\phi_i = [\phi_{i1} \ \ \phi_{i2} \ldots \phi_{iN}]^T : \mathbb{R}^n \to \mathbb{R}^N$ are smooth NN activation functions, such that $\phi_{ij}(0) = 0$ and $\phi_{ij}'(0) = 0 \ \ j = 1 \ldots N$ and $i = 1, 2$, and $\varepsilon_1, \varepsilon_2 : \mathbb{R}^n \to \mathbb{R}$ are the function reconstruction errors.

Assumption 3. The NN activation functions $\{\phi_{ij} : j = 1 \ldots N,\ i = 1, 2\}$ are chosen such that as $N \to \infty$, ϕ provides a complete independent basis for V_1^* and V_2^*.

Using Assumption 3 and Weierstrass higher-order approximation Theorem, both V_i^* and ∇V_i^* can be uniformly approximated by NNs in (43), i.e. as $N \to \infty$, the approximation errors $\varepsilon_i, \varepsilon_i' \to 0$ for $i = 1, 2$, respectively. The critic $\hat{V}$ and the actor $\hat{u}$ approximate the optimal value function and the optimal controls in (43), and are given as

$$\begin{aligned} \hat{V}_1(x) &= \hat{W}_{1c}^T\phi_1(x),\ \hat{u}_1(x) = -\frac{1}{2}R_{11}^{-1}g_1^T(x)\phi_1'^T(x)\hat{W}_{1a} \\ \hat{V}_2(x) &= \hat{W}_{2c}^T\phi_2(x),\ \hat{u}_2(x) = -\frac{1}{2}R_{22}^{-1}g_2^T(x)\phi_2'^T(x)\hat{W}_{2a}, \end{aligned} \tag{44}$$

where $\hat{W}_{1c}, \hat{W}_{2c} \in \mathbb{R}^N$ and $\hat{W}_{1a}, \hat{W}_{2a} \in \mathbb{R}^N$ are estimates of the ideal weights of the critic and actor NNs, respectively. The weight estimation errors for the critic and actor are defined as $\tilde{W}_{ic} \triangleq W_i - \hat{W}_{ic}$ and $\tilde{W}_{ia} \triangleq W_i - \hat{W}_{ia}$ for $i = 1, 2$, respectively. The actor and critic NN weights are both updated based on minimizing the Bellman error δ_{hjb} in (14), which can be rewritten by substituting $\hat{V}_1$ and $\hat{V}_2$ from (44) as

$$\begin{aligned} \delta_{hjb_1} &= \hat{W}_{1c}^T\phi_1'\hat{F}_{\hat{u}} + r_1(x, \hat{u}_1, \hat{u}_2) = \hat{W}_{1c}^T\omega_1 + r_1(x, \hat{u}_1, \hat{u}_2), \\ \delta_{hjb_2} &= \hat{W}_{2c}^T\phi_2'\hat{F}_{\hat{u}} + r_2(x, \hat{u}_1, \hat{u}_2) = \hat{W}_{1c}^T\omega_1 + r_1(x, \hat{u}_1, \hat{u}_2), \end{aligned} \tag{45}$$

where $\omega_i(x, \hat{u}, t) \triangleq \phi_i'\hat{F}_{\hat{u}} \in \mathbb{R}^N$ for $i = 1, 2$, is the critic NN regressor vector.

A. Least squares update for the critic

Consider the integral squared Bellman error E_c

$$E_c(\hat{W}_{1c}, \hat{W}_{2c}, t) = \int_0^t \left(\delta_{hjb_1}^2(\tau) + \delta_{hjb_2}^2(\tau)\right) d\tau. \tag{46}$$

The LS update law for the critic $\hat{W}_{1c}$ is generated by minimizing the total prediction error in (46)

$$\frac{\partial E_c}{\partial \hat{W}_{1c}} = 2\int_0^t \delta_{hjb_1}(\tau)\frac{\partial \delta_{hjb_1}(\tau)}{\partial \hat{W}_{1c}(\tau)}\, d\tau = 0$$

$$= \hat{W}_{1c}^T \int_0^t \omega_1(\tau)\omega_1(\tau)^T \, d\tau + \int_0^t \omega_1(\tau)^T r_1(\tau) \, d\tau = 0$$

$$\hat{W}_{1c} = -\left(\int_0^t \omega_1(\tau)\omega_1(\tau)^T \, d\tau \right)^{-1} \int_0^t \omega_1(\tau) r_1(\tau) \, d\tau,$$

which gives the LS estimate of the critic weights, provided $\left(\int_0^t \omega_1(\tau)\omega_1(\tau)^T \, d\tau\right)^{-1}$ exists. Likewise, the LS update law for the critic $\hat{W}_{2c}$ is generated by

$$\hat{W}_{2c} = -\left(\int_0^t \omega_2(\tau)\omega_2(\tau)^T \, d\tau \right)^{-1} \int_0^t \omega_2(\tau) r_2(\tau) \, d\tau.$$

The recursive formulation of the normalized LS algorithm [53] gives the update laws for the two critic weights as

$$\dot{\hat{W}}_{1c} = -\eta_{1c}\Gamma_{1c} \frac{\omega_1}{1+\nu_1 \omega_1^T \Gamma_{1c} \omega_1} \delta_{hjb_1},$$
$$\dot{\hat{W}}_{2c} = -\eta_{2c}\Gamma_{2c} \frac{\omega_2}{1+\nu_2 \omega_2^T \Gamma_{2c} \omega_2} \delta_{hjb_2}, \tag{47}$$

where $\nu_1, \nu_2, \eta_{1c}, \eta_{2c} \in \mathbb{R}$ are constant positive gains and $\Gamma_{ic} \triangleq \left(\int_0^t \omega(\tau)\omega(\tau)^T \, d\tau\right)^{-1} \in \mathbb{R}^{N\times N}$ for $i = 1, 2$, are symmetric estimation gain matrices generated by

$$\dot{\Gamma}_{1c} = -\eta_{1c}\left(-\lambda_1 \Gamma_{1c} + \Gamma_{1c} \frac{\omega_1 {\omega_1}^T}{1+\nu_1 \omega_1^T \Gamma_{1c} \omega_1} \Gamma_{1c} \right),$$
$$\dot{\Gamma}_{2c} = -\eta_{2c}\left(-\lambda_2 \Gamma_{2c} + \Gamma_{2c} \frac{\omega_2 {\omega_2}^T}{1+\nu_2 \omega_2^T \Gamma_{2c} \omega_2} \Gamma_{2c} \right), \tag{48}$$

where $\lambda_1, \lambda_2 \in (0, 1)$ are forgetting factors. The use of forgetting factors ensures that Γ_{1c} and Γ_{2c} are positive-definite for all time and prevents arbitrarily small values in some directions, making adaptation in those directions very slow (also called the covariance wind-up problem) [54], [55]. Thus, the covariance matrices $(\Gamma_{1c}, \Gamma_{2c})$ can be bounded as

$$\varphi_{11} I \leq \Gamma_{1c} \leq \varphi_{01} I, \; \varphi_{12} I \leq \Gamma_{2c} \leq \varphi_{02} I. \tag{49}$$

B. Gradient update for the actor

The actor update, like the critic update in Section V-A, is based on the minimization of the Bellman error δ_{hjb}. However, unlike the critic weights, the actor weights appear nonlinearly in δ_{hjb}, making it problematic to develop a LS update law. Hence, a gradient update law is developed for the actor which minimizes the squared Bellman error $E_a \triangleq \delta_{hjb_1}^2 + \delta_{hjb_2}^2$, whose gradients are given as

$$\frac{\partial E_a}{\partial \hat{W}_{1a}} = (\hat{W}_{1a} - \hat{W}_{1c})^T \phi_1' G_1 {\phi_1'}^T \delta_{hjb_1} + (\hat{W}_{1a}^T \phi_1' G_{21} - \hat{W}_{2c}^T \phi_2' G_1) {\phi_1'}^T \delta_{hjb_2};$$
$$\frac{\partial E_a}{\partial \hat{W}_{2a}} = (\hat{W}_{2a}^T \phi_2' G_{12} - \hat{W}_{1c}^T \phi_1' G_2) {\phi_2'}^T \delta_{hjb_1} + (\hat{W}_{2a} - \hat{W}_{2c})^T \phi_2' G_2 {\phi_2'}^T \delta_{hjb_2}, \tag{50}$$

where $G_i \triangleq g_i R_{ii}^{-1} g_i \in \mathbb{R}^{n\times n}$ and $G_{ji} \triangleq g_i R_{ii}^{-1} R_{ji} R_{ii}^{-1} g_i \in \mathbb{R}^{n\times n}$, for $i = 1, 2$ and $j = 1, 2$, are symmetric matrices. Using (50), the actor NNs are updated as

$$\dot{\hat{W}}_{1a} = proj\left\{ -\frac{\Gamma_{11a}}{\sqrt{1+\omega_1^T\omega_1}} \frac{\partial E_a}{\partial \hat{W}_{1a}} - \Gamma_{12a}(\hat{W}_{1a} - \hat{W}_{1c}) \right\}$$
$$\dot{\hat{W}}_{2a} = proj\left\{ -\frac{\Gamma_{21a}}{\sqrt{1+\omega_2^T\omega_2}} \frac{\partial E_a}{\partial \hat{W}_{2a}} - \Gamma_{22a}(\hat{W}_{2a} - \hat{W}_{2c}) \right\}, \tag{51}$$

where $\Gamma_{11a}, \Gamma_{12a}, \Gamma_{21a}, \Gamma_{22a} \in \mathbb{R}$ are positive adaptation gains, and $proj\{\cdot\}$ is a projection operator used to bound the weight estimates[3] [46], [47]. The first term in (51) is normalized and the last term is added as feedback for stability (based on the subsequent stability analysis).

VI. Stability Analysis

The dynamics of the critic weight estimation errors $\tilde{W}_{1c}$ and $\tilde{W}_{2c}$ can be developed using (11)-(14), (45) and (47), as

$$\dot{\tilde{W}}_{1c} = \eta_{1c}\Gamma_{1c} \frac{\omega_1}{1+\nu_1\omega_1^T\Gamma_{1c}\omega_1} \Big[-\tilde{W}_{1c}^T \omega_1 - W_1^T \phi_1' \tilde{F}_{\hat{u}} - {u_1^*}^T R_{11} u_1^* - \varepsilon_{1v}' F_{u^*} + \hat{u}_1^T R_{11} \hat{u}_1 + W_1^T \phi_1' \left(g_1(\hat{u}_1 - u_1^*) + g_2(\hat{u}_2 - u_2^*) \right) - {u_2^*}^T R_{12} u_2^* + \hat{u}_2^T R_{12} \hat{u}_2 \Big];$$

$$\dot{\tilde{W}}_{2c} = \eta_{2c}\Gamma_{2c} \frac{\omega_2}{1+\nu_2\omega_2^T\Gamma_{2c}\omega_2} \Big[-\tilde{W}_{2c}^T \omega_2 - W_2^T \phi_2' \tilde{F}_{\hat{u}} - {u_2^*}^T R_{22} u_2^* - \varepsilon_{2v}' F_{u^*} + \hat{u}_2^T R_{22} \hat{u}_2 W_2^T \phi_2' \left(g_1(\hat{u}_1 - u_1^*) + g_2(\hat{u}_2 - u_2^*) \right) - {u_1^*}^T R_{21} u_1^* + \hat{u}_1^T R_{21} \hat{u}_1 \Big]. \tag{52}$$

Substituting for (u_1^*, u_2^*) and $(\hat{u}_1, \hat{u}_2)$ from (43) and (44), respectively, in (52) yields

$$\dot{\tilde{W}}_{1c} = -\eta_{1c}\Gamma_{1c}\psi_1\psi_1^T\tilde{W}_{1c} + \eta_{1c}\Gamma_{1c} \frac{\omega_1}{1+\nu_1\omega_1^T\Gamma_{1c}\omega_1} \Big[-W_1^T\phi_1'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_{2a}^T\phi_2' G_{12} {\phi_2'}^T \tilde{W}_{2a} - \frac{1}{4}\varepsilon_2' G_{12} {\varepsilon_2'}^T + \frac{1}{2}\left(\tilde{W}_{2a}\phi_2' + {\varepsilon_2'}^T\right)\left(G_2 {\phi_1'}^T W_1 - G_{12}{\phi_2'}^T W_2\right) + \frac{1}{4}\tilde{W}_{1a}^T\phi_1' G_1 {\phi_1'}^T \tilde{W}_{1a} - \frac{1}{4}\varepsilon_1' G_1 {\varepsilon_1'}^T - \varepsilon_1' F_{u^*} \Big];$$

$$\dot{\tilde{W}}_{2c} = -\eta_{2c}\Gamma_{2c}\psi_2\psi_2^T\tilde{W}_{2c} + \eta_{2c}\Gamma_{2c} \frac{\omega_2}{1+\nu_2\omega_2^T\Gamma_{2c}\omega_2} \Big[-W_2^T\phi_2'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_{1a}^T\phi_1' G_{21} {\phi_1'}^T \tilde{W}_{1a} - \frac{1}{4}\varepsilon_1' G_{21} {\varepsilon_1'}^T + \frac{1}{2}\left(\tilde{W}_{1a}\phi_1' + {\varepsilon_1'}^T\right)\left(G_1 {\phi_2'}^T W_2 - G_{21}{\phi_1'}^T W_1\right) + \frac{1}{4}\tilde{W}_{2a}^T\phi_2' G_2 {\phi_2'}^T \tilde{W}_{2a} - \frac{1}{4}\varepsilon_2' G_2 {\varepsilon_2'}^T - \varepsilon_2' F_{u^*} \Big]; \tag{53}$$

[3]Instead of the projection algorithm, $\sigma-$modification-like terms $-\Gamma_{13a}\hat{W}_{1a}$ and $-\Gamma_{13a}\hat{W}_{1a}$ can be added to the update laws to ensure that the weights $\hat{W}_{1a}$ and $\hat{W}_{2a}$ remain bounded, resulting in additional gain conditions.

where $\psi_i(t) \triangleq \frac{\omega_i(t)}{\sqrt{1+\nu_i\omega_i(t)^T\Gamma_{ic}(t)\omega_i(t)}} \in \mathbb{R}^N$ are the normalized critic regressor vectors for $i = 1, 2$, respectively, bounded as

$$\|\psi_1\| \leq \frac{1}{\sqrt{\nu_1\varphi_{11}}},\ \|\psi_2\| \leq \frac{1}{\sqrt{\nu_2\varphi_{12}}}, \tag{54}$$

where φ_{11} and φ_{12} are introduced in (49). The error systems in (53) can be represented as the following perturbed systems

$$\dot{\tilde{W}}_{1c} = \Omega_1 + \Lambda_{01}\Delta_1, \qquad \dot{\tilde{W}}_{2c} = \Omega_2 + \Lambda_{02}\Delta_2, \tag{55}$$

where $\Omega_i(\tilde{W}_{ic}, t) \triangleq -\eta_{ic}\Gamma_{ic}\psi_i\psi_i^T\tilde{W}_{ic} \in \mathbb{R}^N \quad i = 1, 2$, denotes the nominal system, $\Lambda_{0i} \triangleq \frac{\eta_{ic}\Gamma_{ic}\omega_i}{1+\nu_i\omega_i^T\Gamma_{ic}\omega_i}$ denotes the perturbation gain, and the perturbations $\Delta_i \in \mathbb{R}^N$ are denoted as

$$\begin{aligned}\Delta_i \triangleq \Bigg[&-W_i^T\phi_i'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_{ia}^T\phi_i'G_i\phi_i'^T\tilde{W}_{ia} - \varepsilon_i'F_{u^*} \\ &+\frac{1}{4}\tilde{W}_{ka}^T\phi_k'G_{ik}\phi_k'^T\tilde{W}_{ka} - \frac{1}{4}\varepsilon_k'G_{ik}\varepsilon_k'^T - \frac{1}{4}\varepsilon_i'G_i\varepsilon_i'^T \\ &+\frac{1}{2}\left(\tilde{W}_{ka}\phi_k' + \varepsilon_k'^T\right)\left(G_k\phi_i'^TW_i - G_{ik}\phi_k'^TW_k\right)\Bigg],\end{aligned}$$

where $i = 1, 2$ and $k = 3 - i$. Using Theorem 2.5.1 in [53], it can be shown that the nominal systems

$$\dot{\tilde{W}}_{1c} = -\eta_{1c}\Gamma_{1c}\psi_1\psi_1^T\tilde{W}_{1c},\ \dot{\tilde{W}}_{2c} = -\eta_{2c}\Gamma_{2c}\psi_2\psi_2^T\tilde{W}_{2c}, \tag{56}$$

are exponentially stable if the bounded signals $(\psi_1(t), \psi_2(t))$ are uniformly persistently exciting (u-PE) as [56]

$$\mu_{i2}I \geq \int_{t_0}^{t_0+\delta_i} \psi_i(\tau)\psi_i(\tau)^T d\tau \geq \mu_{i1}I \qquad \forall t_0 \geq 0,\ i = 1, 2,$$

where $\mu_{i1}, \mu_{i2}, \delta_i \in \mathbb{R}$ are positive constants independent of the initial conditions. Since Ω_i is continuously differentiable in $\tilde{W}_{ic}$ and the Jacobian $\frac{\partial\Omega_i}{\partial\tilde{W}_{ic}} = -\eta_{ic}\Gamma_{ic}\psi_i\psi_i^T$ is bounded for the exponentially stable system (56) for $i = 1, 2$, the converse Lyapunov Theorem 4.14 in [57] can be used to show that there exists a function $V_c : \mathbb{R}^N \times [0,\ \infty) \to \mathbb{R}$, which satisfies the following inequalities

$$\begin{aligned}c_{11}\left\|\tilde{W}_{1c}\right\|^2 + c_{12}\left\|\tilde{W}_{2c}\right\|^2 &\leq V_c(\tilde{W}_{1c}, \tilde{W}_{2c}, t) \\ V_c(\tilde{W}_{1c}, \tilde{W}_{2c}, t) &\leq c_{21}\left\|\tilde{W}_{1c}\right\|^2 + c_{22}\left\|\tilde{W}_{2c}\right\|^2 \\ -c_{31}\left\|\tilde{W}_{1c}\right\|^2 - c_{32}\left\|\tilde{W}_{2c}\right\|^2 &\geq \frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial\tilde{W}_{1c}}\Omega_1(\tilde{W}_{1c}, t) \\ &\quad + \frac{\partial V_c}{\partial\tilde{W}_{2c}}\Omega_2(\tilde{W}_{2c}, t) \\ \left\|\frac{\partial V_c}{\partial\tilde{W}_{1c}}\right\| \leq c_{41}\left\|\tilde{W}_{1c}\right\| &\qquad \left\|\frac{\partial V_c}{\partial\tilde{W}_{2c}}\right\| \leq c_{42}\left\|\tilde{W}_{2c}\right\|,\end{aligned} \tag{57}$$

for some positive constants $c_{1i},\ c_{2i},\ c_{3i},\ c_{4i} \in \mathbb{R}$ for $i = 1, 2$. Using Properties 1-4, Assumption 1, the projection bounds in (51), the fact that $F_{u^*} \in \mathcal{L}_\infty$ (using 17), and provided the conditions of Theorem 1 hold (required to prove that $\tilde{F}_{\hat{u}} \in \mathcal{L}_\infty$), the following bounds are developed to facilitate the subsequent stability proof

$$\begin{aligned}\kappa_1 &\geq \left\|\tilde{W}_{1a}\right\|; & \kappa_2 &\geq \left\|\tilde{W}_{2a}\right\|, \\ \kappa_3 &\geq \left\|\phi_1'G_1\phi_1'^T\right\|; & \kappa_4 &\geq \left\|\phi_2'G_2\phi_2'^T\right\|, \\ \kappa_5 &\geq \|\Delta_1\|; & \kappa_6 &\geq \|\Delta_2\|, \\ \kappa_7 &\geq \frac{1}{4}\|G_1 - G_{21}\|\|\nabla V_1^*\|^2 + \frac{1}{4}\|G_2 - G_{12}\|\|\nabla V_2^*\|^2 \\ &\quad + \frac{1}{2}\left\|\nabla V_1^*(G_2 + G_1)\nabla V_2^{*T}\right\|, \\ \kappa_8 &\geq \Bigg\|-\frac{1}{2}(\nabla V_1^* - \nabla V_2^*)\left(G_1\phi_1'^TW_{1a} - G_2\phi_2'^TW_{2a}\right) \\ &\quad + \frac{1}{2}(\nabla V_1^* - \nabla V_2^*)\left(G_1\phi_1'^T\tilde{W}_{1a} - G_2\phi_2'^T\tilde{W}_{2a}\right)\Bigg\|, \\ \kappa_9 &\geq \left\|\phi_1'G_{21}\phi_1'^T\right\|; & \kappa_{10} &\geq \left\|\phi_2'G_1\phi_1'^T\right\|, \\ \kappa_{11} &\geq \left\|\phi_1'G_2\phi_2'^T\right\|; & \kappa_{12} &\geq \left\|\phi_2'G_{12}\phi_2'^T\right\|,\end{aligned} \tag{58}$$

where $\kappa_j \in \mathbb{R}$ for $j = 1, \ldots, 12$ are computable positive constants.

Theorem 2. *If Assumptions 1-3 hold, the regressors ψ_i for $i = 1, 2$ are u-PE, and provided (34), (35), and the following sufficient gain conditions are satisfied*

$$\begin{aligned}c_{31} &> \Gamma_{11a}\kappa_1\kappa_3 + \Gamma_{21a}\kappa_2\kappa_{11}, \\ c_{32} &> \Gamma_{21a}\kappa_2\kappa_4 + \Gamma_{11a}\kappa_1\kappa_{10},\end{aligned}$$

where $\Gamma_{11a}, \Gamma_{21a}, c_{31}, c_{32}, \kappa_1, \kappa_2, \kappa_3$, and κ_4 are introduced in (51), (57), and (58), then the controller in (44), the actor-critic weight update laws in (47)-(48) and (51), and the identifier in (18) and (24), guarantee that the state of the system $x(t)$, and the actor-critic weight estimation errors $\left(\tilde{W}_{1a}(t), \tilde{W}_{2a}(t)\right)$ and $\left(\tilde{W}_{1c}(t), \tilde{W}_{2c}(t)\right)$ are UUB.

Proof: To investigate the stability of (16) with control inputs $\hat{u}_1$ and $\hat{u}_2$, and the perturbed system (55), consider $V_L : \mathbb{S}\times\mathbb{R}^N\times\mathbb{R}^N\times[0,\infty) \to \mathbb{R}$ as the continuously differentiable, positive-definite Lyapunov function candidate, given as

$$\begin{aligned}V_L \triangleq\ &V_1^*(x) + V_2^*(x) + V_c(\tilde{W}_{1c}, \tilde{W}_{2c}, t) \\ &+ \frac{1}{2}\tilde{W}_{1a}^T\tilde{W}_{1a} + \frac{1}{2}\tilde{W}_{2a}^T\tilde{W}_{2a},\end{aligned}$$

where V_i^* for $i = 1, 2$ (the optimal value function for (16), is the Lyapunov function for (16), and V_c is the Lyapunov function for the exponentially stable system in (56). Since (V_1^*, V_2^*) are continuously differentiable and positive-definite from (5), from Lemma 4.3 in [57], there exist class $\mathcal{K}$ functions α_1 and α_2 defined on $[0,\ r]$, where $B_r \subset \mathcal{X}$, such that

$$\alpha_1(\|x\|) \leq V_1^*(x) + V_2^*(x) \leq \alpha_2(\|x\|),\ \forall x \in B_r. \tag{59}$$

Using (57) and (59), V_L can be bounded as

$$\begin{aligned}V_L &\geq \alpha_1(\|x\|) + c_{11}\left\|\tilde{W}_{1c}\right\|^2 + c_{12}\left\|\tilde{W}_{2c}\right\|^2 \\ &\quad + \frac{1}{2}\left(\left\|\tilde{W}_{1a}\right\|^2 + \left\|\tilde{W}_{2a}\right\|^2\right) \\ V_L &\leq \alpha_2(\|x\|) + c_{21}\left\|\tilde{W}_{1c}\right\|^2 + c_{22}\left\|\tilde{W}_{2c}\right\|^2\end{aligned}$$

$$+\frac{1}{2}\left(\left\|\tilde{W}_{1a}\right\|^2+\left\|\tilde{W}_{2a}\right\|^2\right).$$

which can be written as $\alpha_3(\|w\|) \leq V_L(x,\tilde{W}_{1c},\tilde{W}_{2c},\tilde{W}_{1a},\tilde{W}_{2a},t) \leq \alpha_4(\|w\|)$, $\forall w \in B_s$, where $w \triangleq [x^T\ \tilde{W}_{1c}^T\ \tilde{W}_{2c}^T\ \tilde{W}_{1a}^T\ \tilde{W}_{2a}^T]^T \in \mathbb{R}^{n+4N}$, α_3 and α_4 are class $\mathcal{K}$ functions defined on $[0,\ s]$, where $B_s \subset \mathbb{S}\times\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N$ is a ball of radius s centered at the origin. Taking the time derivative of $V_L(\cdot)$ yields

$$\begin{aligned}\dot{V}_L &= \left(\nabla V_1^* + \nabla V_2^*\right)\left(f+g_1\hat{u}_1+g_2\hat{u}_2\right)\\ &+\frac{\partial V_c}{\partial t}+\frac{\partial V_c}{\partial \tilde{W}_{1c}}\Omega_1+\frac{\partial V_c}{\partial \tilde{W}_{1c}}\Lambda_{01}\Delta_1+\frac{\partial V_c}{\partial \tilde{W}_{2c}}\Omega_2\\ &+\frac{\partial V_c}{\partial \tilde{W}_{2c}}\Lambda_{02}\Delta_2-\tilde{W}_{1a}^T\dot{\tilde{W}}_{1a}-\tilde{W}_{2a}^T\dot{\tilde{W}}_{2a},\end{aligned} \tag{60}$$

where the time derivatives of $V_i^*$for $i=1,2$, are taken along the trajectories of the system (16) with control inputs $(\hat{u}_1,\hat{u}_2)$ and the time derivative of V_c is taken along the along the trajectories of the perturbed system (55). Using (12), $\nabla V_i^* f = -\nabla V_i^*\left(g_1u_1^*+g_2u_2^*\right)-Q_i(x)-\sum_{j=1}^{2}u_j^{*T}R_{ij}u_j^*$ for $i=1,2$. Substituting for the $\nabla V_i^* f$ terms in (60), using the fact that $\nabla V_i^* g_i = -2u_i^{*^T}R_{ii}$ from (9), and using (51) and (57), (60) can be upper bounded as

$$\begin{aligned}\dot{V}_L &\leq -Q-u_1^{*^T}\left(R_{11}+R_{21}\right)u_1^*-u_2^{*^T}\left(R_{22}+R_{12}\right)u_2^*\\ &+2u_1^{*^T}R_{11}(u_1^*-\hat{u}_1)+2u_2^{*^T}R_{22}(u_2^*-\hat{u}_2)\\ &+\nabla V_1^* g_2\left(\hat{u}_2-u_2^*\right)+\nabla V_2^* g_1\left(\hat{u}_1-u_1^*\right)\\ &+c_{41}\Lambda_{01}\left\|\tilde{W}_{1c}\right\|\|\Delta_1\|-c_{31}\left\|\tilde{W}_{1c}\right\|^2\\ &+c_{42}\Lambda_{02}\left\|\tilde{W}_{2c}\right\|\|\Delta_2\|-c_{32}\left\|\tilde{W}_{2c}\right\|^2\\ &+\tilde{W}_{1a}^T\left[\frac{\Gamma_{11a}}{\sqrt{1+\omega_1^T\omega_1}}\frac{\partial E_a}{\partial \hat{W}_{1a}}+\Gamma_{12a}(\hat{W}_{1a}-\hat{W}_{1c})\right]\\ &+\tilde{W}_{2a}^T\left[\frac{\Gamma_{21a}}{\sqrt{1+\omega_2^T\omega_2}}\frac{\partial E_a}{\partial \hat{W}_{2a}}+\Gamma_{22a}(\hat{W}_{2a}-\hat{W}_{2c})\right],\end{aligned} \tag{61}$$

where $Q \triangleq Q_1+Q_2$. Substituting for u_i^*, $\hat{u}_i$, δ_{hjb_i}, and Δ_i for $i=1,2$ using (9), (44), (52), and (55), respectively, and using (49) and (54) in (61), yields

$$\begin{aligned}\dot{V}_L &\leq \frac{1}{4}\|G_1-G_{21}\|\|\nabla V_1^*\|^2+\frac{1}{4}\|G_2-G_{12}\|\|\nabla V_2^*\|^2\\ &+\frac{1}{2}\left\|\nabla V_1^*\left(G_1+G_2\right)\nabla V_2^{*T}\right\|-Q\\ &-\frac{1}{2}\left(\nabla V_1^*-\nabla V_2^*\right)\left(G_1\phi_1'^T W_{1a}-G_2\phi_2'^T W_{2a}\right)\\ &+\frac{1}{2}\left(\nabla V_1^*-\nabla V_2^*\right)\left(G_1\phi_1'^T \tilde{W}_{1a}-G_2\phi_2'^T \tilde{W}_{2a}\right)\\ &+c_{41}\frac{\eta_{1c}\varphi_{01}}{2\sqrt{\nu_1\varphi_{11}}}\|\Delta_1\|\left\|\tilde{W}_{1c}\right\|-c_{31}\left\|\tilde{W}_{1c}\right\|^2\\ &+c_{42}\frac{\eta_{2c}\varphi_{02}}{2\sqrt{\nu_2\varphi_{12}}}\|\Delta_2\|\left\|\tilde{W}_{2c}\right\|-c_{32}\left\|\tilde{W}_{2c}\right\|^2\\ &-\Gamma_{12a}\left\|\tilde{W}_{1a}\right\|^2-\Gamma_{22a}\left\|\tilde{W}_{2a}\right\|^2\\ &+\Gamma_{12a}\left\|\tilde{W}_{1a}\right\|\left\|\tilde{W}_{1c}\right\|+\Gamma_{22a}\left\|\tilde{W}_{2a}\right\|\left\|\tilde{W}_{2c}\right\|\\ &+\frac{\Gamma_{11a}}{\sqrt{1+\omega_1^T\omega_1}}\tilde{W}_{1a}^T\Big(\\ &\left(\tilde{W}_{1c}-\tilde{W}_{1a}\right)^T\phi_1'G_1\phi_1'^T\left(-\tilde{W}_{1c}^T\omega_1+\Delta_1\right)\\ &+\left(\tilde{W}_{1a}^T\phi_1'G_{21}-\tilde{W}_{2c}^T\phi_2'G_2\right)\phi_1'^T\left(-\tilde{W}_{2c}^T\omega_2+\Delta_2\right)\\ &+\left(W_1^T\phi_1'G_{21}-W_2^T\phi_2'G_1\right)\phi_1'^T\left(-\tilde{W}_{2c}^T\omega_2+\Delta_2\right)\Big)\\ &+\frac{\Gamma_{21a}}{\sqrt{1+\omega_2^T\omega_2}}\tilde{W}_{2a}^T\Big(\\ &\left(\tilde{W}_{2c}-\tilde{W}_{2a}\right)^T\phi_2'G_2\phi_2'^T\left(-\tilde{W}_{2c}^T\omega_2+\Delta_2\right)\\ &+\left(\tilde{W}_{2a}^T\phi_2'G_{12}-\tilde{W}_{1c}^T\phi_1'G_2\right)\phi_2'^T\left(-\tilde{W}_{1c}^T\omega_1+\Delta_1\right)\\ &+\left(W_2^T\phi_2'G_{12}-W_1^T\phi_1'G_2\right)\phi_2'^T\left(-\tilde{W}_{1c}^T\omega_1+\Delta_1\right)\Big).\end{aligned} \tag{62}$$

Using the bounds developed in (58), (62) can be further upper bounded as

$$\begin{aligned}\dot{V}_L &\leq -Q-\left(c_{31}-\Gamma_{11a}\kappa_1\kappa_3-\Gamma_{21a}\kappa_2\kappa_{11}\right)\left\|\tilde{W}_{1c}\right\|^2\\ &-\left(c_{32}-\Gamma_{21a}\kappa_2\kappa_4-\Gamma_{11a}\kappa_1\kappa_{10}\right)\left\|\tilde{W}_{2c}\right\|^2+\Phi_2\left\|\tilde{W}_{2c}\right\|\\ &-\Gamma_{12a}\left\|\tilde{W}_{1a}\right\|^2-\Gamma_{22a}\left\|\tilde{W}_{2a}\right\|^2+\Phi_1\left\|\tilde{W}_{1c}\right\|\\ &+\Gamma_{11a}\kappa_1\left(\kappa_1\left(\kappa_3\kappa_5+\kappa_6\kappa_9\right)+\kappa_6\left(\bar{W}_1\kappa_9+\bar{W}_2\kappa_{10}\right)\right)\\ &+\Gamma_{21a}\kappa_2\left(\kappa_2\left(\kappa_4\kappa_6+\kappa_5\kappa_{12}\right)+\kappa_5\left(\bar{W}_1\kappa_{11}+\bar{W}_2\kappa_{12}\right)\right)\\ &+\kappa_7+\kappa_8,\end{aligned}$$

where

$$\begin{aligned}\Phi_1 &\triangleq \Big(\frac{c_{41}\eta_{1c}\varphi_{01}}{2\sqrt{\nu_1\varphi_{11}}}\kappa_5+\Gamma_{11a}\left(\kappa_1\kappa_3\left(\kappa_1+\kappa_5\right)\right)+\Gamma_{12a}\kappa_1\\ &\quad+\Gamma_{21a}\kappa_2\left(\kappa_{11}\left(\kappa_5+\bar{W}_1\right)+\kappa_{12}\left(\kappa_2+\bar{W}_2\right)\right)\Big),\\ \Phi_2 &\triangleq \Big(\frac{c_{42}\eta_{2c}\varphi_{02}}{2\sqrt{\nu_2\varphi_{12}}}\kappa_6+\Gamma_{21a}\left(\kappa_2\kappa_4\left(\kappa_2+\kappa_6\right)\right)+\Gamma_{22a}\kappa_2\\ &\quad+\Gamma_{11a}\kappa_1\left(\kappa_9\left(\kappa_1+\bar{W}_1\right)+\kappa_{10}\left(\kappa_6+\bar{W}_2\right)\right)\Big).\end{aligned}$$

Provided $c_{31}>\Gamma_{11a}\kappa_1\kappa_3+\Gamma_{21a}\kappa_2\kappa_{11}$ and $c_{32}>\Gamma_{21a}\kappa_2\kappa_4+\Gamma_{11a}\kappa_1\kappa_{10}$, completing the square yields

$$\begin{aligned}\dot{V}_L &\leq -Q-\Gamma_{22a}\left\|\tilde{W}_{2a}\right\|^2-\Gamma_{12a}\left\|\tilde{W}_{1a}\right\|^2\\ &-(1-\theta_1)(c_{31}-\Gamma_{11a}\kappa_1\kappa_3-\Gamma_{21a}\kappa_2\kappa_{11})\left\|\tilde{W}_{1c}\right\|^2\\ &-(1-\theta_2)(c_{32}-\Gamma_{21a}\kappa_2\kappa_4-\Gamma_{11a}\kappa_1\kappa_{10})\left\|\tilde{W}_{2c}\right\|^2\\ &+\Gamma_{11a}\kappa_1\left(\kappa_1\left(\kappa_3\kappa_5+\kappa_6\kappa_9\right)+\kappa_6\left(\bar{W}_1\kappa_9+\bar{W}_2\kappa_{10}\right)\right)\\ &+\Gamma_{21a}\kappa_2\left(\kappa_2\left(\kappa_4\kappa_6+\kappa_5\kappa_{12}\right)+\kappa_5\left(\bar{W}_1\kappa_{11}+\bar{W}_2\kappa_{12}\right)\right)\\ &+\frac{\Phi_1^2}{4\theta_1(c_{31}-\Gamma_{11a}\kappa_1\kappa_3-\Gamma_{21a}\kappa_2\kappa_{11})}+\kappa_7\\ &+\frac{\Phi_2^2}{4\theta_2(c_{32}-\Gamma_{21a}\kappa_2\kappa_4-\Gamma_{11a}\kappa_1\kappa_{10})}+\kappa_8,\end{aligned} \tag{63}$$

where $\theta_1,\theta_2\in(0,1)$. Since Q is positive definite, according to Lemma 4.3 in [57], there exist class $\mathcal{K}$ functions α_5 and α_6 such that

$$\alpha_5(\|w\|)\leq F\left(\|w\|\right)\leq\alpha_6(\|w\|)\qquad\forall w\in B_s, \tag{64}$$

where

$$F\left(\|w\|\right) = Q + \Gamma_{22a}\left\|\tilde{W}_{2a}\right\|^2 + \Gamma_{12a}\left\|\tilde{W}_{1a}\right\|^2 + (1-\theta_1)(c_{31} - \Gamma_{11a}\kappa_1\kappa_3 - \Gamma_{21a}\kappa_2\kappa_{11})\left\|\tilde{W}_{1c}\right\|^2 + (1-\theta_2)(c_{32} - \Gamma_{21a}\kappa_2\kappa_4 - \Gamma_{11a}\kappa_1\kappa_{10})\left\|\tilde{W}_{2c}\right\|^2$$

Using (64), the expression in (63) can be further upper bounded as $\dot{V}_L \leq -\alpha_5(\|w\|) + \Upsilon$, where

$$\Upsilon = \Gamma_{11a}\kappa_1\left(\kappa_1\left(\kappa_3\kappa_5 + \kappa_6\kappa_9\right) + \kappa_6\left(\bar{W}_1\kappa_9 + \bar{W}_2\kappa_{10}\right)\right) + \Gamma_{21a}\kappa_2\left(\kappa_2\left(\kappa_4\kappa_6 + \kappa_5\kappa_{12}\right) + \kappa_5\left(\bar{W}_1\kappa_{11} + \bar{W}_2\kappa_{12}\right)\right) + \frac{\Phi_1^2}{4\theta_1(c_{31} - \Gamma_{11a}\kappa_1\kappa_3 - \Gamma_{21a}\kappa_2\kappa_{11})} + \kappa_7 + \frac{\Phi_2^2}{4\theta_2(c_{32} - \Gamma_{21a}\kappa_2\kappa_4 - \Gamma_{11a}\kappa_1\kappa_{10})} + \kappa_8,$$

which proves that $\dot{V}_L$ is negative whenever w lies outside the compact set $\Omega_w \triangleq \left\{w : \|w\| \leq \alpha_5^{-1}\left(\Upsilon\right)\right\}$, and hence, $\|w(t)\|$ is UUB, according to Theorem 4.18 in [57]. ■

Remark 3. Since the actor, critic and identifier are continuously updated, the developed RL algorithm can be compared to fully optimistic PI in machine learning literature [11], where policy evaluation and policy improvement are done after every state transition, unlike traditional PI, where policy improvement is done after convergence of the policy evaluation step. Convergence behavior of optimistic PI is not fully understood, and by considering an adaptive control framework, this result investigates the convergence and stability behavior of fully optimistic PI in continuous-time.

Remark 4. The PE requirement in Theorem 2 is equivalent to the exploration paradigm in RL which ensures sufficient sampling of the state space and convergence to the optimal policy [15].

VII. Nash Solution

The subsequent theorem demonstrates that the actor NN approximations converge to the approximate coupled Hamiltonians in (10). It can also be shown that the approximate controllers in (44) approximate the optimal solutions to the 2-player Nash game for the dynamic system given in (16).

Corollary 5. *Provided the assumptions and sufficient gain constraints in Theorem 2 hold, then the actor NNs $\hat{W}_{1a}$ and $\hat{W}_{2a}$ converge to the approximate coupled HJB solution, in the sense that the Hamiltonians in (13) are UUB.*

Proof: Substituting the approximate control laws in (44), in the approximate Hamiltonians $\{H_1, H_2\}$ in (13), yields

$$\begin{aligned} H_1 &= r_{\hat{u}_1} + \nabla\hat{V}_1 F_{\hat{u}} \\ &= Q_1\left(x\right) + \nabla\hat{V}_1 f\left(x\right) + \frac{1}{4}\nabla\hat{V}_1 G_1 \nabla\hat{V}_1^T \\ &\quad + \frac{1}{4}\nabla\hat{V}_2 G_{12}\nabla\hat{V}_2^T - \frac{1}{2}\nabla\hat{V}_1\left(G_1\nabla\hat{V}_1^T + G_2\nabla\hat{V}_2^T\right), \end{aligned}$$

and

$$\begin{aligned} H_2 &= r_{\hat{u}_2} + \nabla\hat{V}_2 F_{\hat{u}} \\ &= Q_2\left(x\right) + \nabla\hat{V}_2 f\left(x\right) + \frac{1}{4}\nabla\hat{V}_2 G_2 \nabla\hat{V}_2^T \\ &\quad + \frac{1}{4}\nabla\hat{V}_1 G_{21}\nabla\hat{V}_1^T - \frac{1}{2}\nabla\hat{V}_2\left(G_1\nabla\hat{V}_1^T + G_2\nabla\hat{V}_2^T\right). \end{aligned}$$

After adding and subtracting

$$\nabla V_i f = -\nabla V_i\left(g_1 u_1^* + g_2 u_2^*\right) - Q_i(x) - \sum_{j=1}^{2} u_j^{*T} R_{ij} u_j^*,$$

for $i = 1, 2$ and performing basic algebraic operations, and using $\nabla\tilde{V}_i^T \triangleq \nabla V_i^T - \nabla\hat{V}_i^T = \phi_i'^T\tilde{W}_{ia} + \varepsilon_i'^T$, the Hamiltonians can be rewritten as

$$\begin{aligned} H_1 &= -\nabla\tilde{V}_1 f\left(x\right) - \frac{1}{4}\nabla\tilde{V}_1 G_1\nabla\tilde{V}_1^T - \frac{1}{2}\nabla V_2 G_{12}\nabla\tilde{V}_2^T \\ &\quad - \frac{1}{2}\nabla\tilde{V}_1 G_2\nabla\tilde{V}_2^T + \frac{1}{2}\nabla V_1 G_1\nabla\tilde{V}_1^T + \frac{1}{4}\nabla\tilde{V}_2 G_{12}\nabla\tilde{V}_2^T \\ &\quad + \frac{1}{2}\nabla V_1 G_2\nabla\tilde{V}_2^T + \frac{1}{2}\nabla\tilde{V}_1 G_2\nabla V_2^T, \end{aligned} \tag{65}$$

and

$$\begin{aligned} H_2 &= -\nabla\tilde{V}_2 f\left(x\right) - \frac{1}{4}\nabla\tilde{V}_2 G_2\nabla\tilde{V}_2^T - \frac{1}{2}\nabla V_1 G_{21}\nabla\tilde{V}_1^T \\ &\quad - \frac{1}{2}\nabla\tilde{V}_2 G_1\nabla\tilde{V}_1^T + \frac{1}{2}\nabla V_2 G_2\nabla\tilde{V}_2^T + \frac{1}{4}\nabla\tilde{V}_1 G_{21}\nabla\tilde{V}_1^T \\ &\quad + \frac{1}{2}\nabla V_2 G_1\nabla\tilde{V}_1^T + \frac{1}{2}\nabla\tilde{V}_2 G_1\nabla V_1^T, \end{aligned} \tag{66}$$

If the assumptions and sufficient gain constraints in Theorem 2 hold, then the right side of (65) and (66) can be upper bounded by a function that is UUB, i.e., $\|H_i\| \leq \Theta_i\left(\tilde{W}_{1a}, \tilde{W}_{2a}, t\right)$ for $i = 1, 2$. Thus the approximate HJBs are also UUB. ■

Corollary 6. *Provided the assumptions and sufficient gain constraints in Theorem 2 hold, the approximate control laws in (44) converge to the approximate Nash solution of the game.*

Proof: Consider the control errors $(\tilde{u}_1, \tilde{u}_2)$ between the optimal control laws in (9) and the approximate control laws in (44) given as $\tilde{u}_1 \triangleq u_1^* - \hat{u}_1$, $\tilde{u}_2 \triangleq u_2^* - \hat{u}_2$. Substituting for the optimal control laws in (9) and the approximate control laws in (44) and using $\tilde{W}_{ia} = W_i - \hat{W}_{ia}$ for $i = 1, 2$, yields

$$\begin{aligned} \tilde{u}_1 &= -\frac{1}{2}R_{11}^{-1} g_1^T(x)\phi_1'(x)\left(\tilde{W}_{1a} + \varepsilon_1'(x)^T\right) \\ \tilde{u}_2 &= -\frac{1}{2}R_{22}^{-1} g_2^T(x)\phi_2'(x)\left(\tilde{W}_{2a} + \varepsilon_2'(x)^T\right). \end{aligned} \tag{67}$$

Using Properties 1-4, (67) can be upper bounded as

$$\begin{aligned} \|\tilde{u}_1\| &\leq \frac{1}{2}\lambda_{\min}\left(R_{11}^{-1}\right)\bar{g}_1\left\|\phi_1'\right\|\left(\left\|\tilde{W}_{1a}\right\| + \bar{\varepsilon'}_1\right) \\ \|\tilde{u}_2\| &\leq \frac{1}{2}\lambda_{\min}\left(R_{22}^{-1}\right)\bar{g}_2\left\|\phi_2'\right\|\left(\left\|\tilde{W}_{2a}\right\| + \bar{\varepsilon'}_2\right). \end{aligned}$$

Given that the assumptions and sufficient gain constraints in Theorem 2 hold, then all terms to the right of the inequality can be bounded by a function that is UUB, therefore the control errors $(\tilde{u}_1, \tilde{u}_2)$ are UUB and the approximate control laws $(\hat{u}_1, \hat{u}_2)$ give the approximate Nash equilibrium solution. ■

VIII. Simulations

A. Two-player game with a known Nash equilibrium solution

The following two player non-zero sum game considered in [31], [38], [39], [58] has a known analytical solution, and hence is utilized in this paper to demonstrate the performance of the developed technique. The system dynamics are given by $\dot{x} = f(x) + g_1(x) u_1 + g_2(x) u_2$, where

$$f(x) = \begin{bmatrix} (x_2 - 2x_1) \\ \left(-\frac{1}{2}x_1 - x_2 + \frac{1}{4}x_2 \left(\cos(2x_1) + 2\right)^2 \right) \\ \left(+\frac{1}{4}x_2 \left(\sin\left(4x_1^2\right) + 2\right)^2 \right) \end{bmatrix},$$

$$g_1(x) = \begin{bmatrix} 0 & \cos(2x_1) + 2 \end{bmatrix}^T, \quad (68)$$

$$g_2(x) = \begin{bmatrix} 0 & \sin\left(4x_1^2\right) + 2 \end{bmatrix}^T. \quad (69)$$

The objective is to design u_1 and u_2 to minimize the cost functionals in (3), where the local cost is given by $r_i = x^T Q_i x + u_i^T R_{ii} u_i + u_j^T R_{ij} u_j$, $i = 1, 2$, $j = 3 - i$, where $R_{11} = 2R_{22} = 2$, $R_{12} = 2R_{21} = 2$, $Q_1 = 2Q_2 = \mathbb{I}_{2\times 2}$. The known analytical solutions for the optimal value functions of player 1 and player 2 are given as $V_1^*(x) = \frac{1}{2}x_1^2 + x_2^2$, $V_2^*(x) = \frac{1}{4}x_1^2 + \frac{1}{2}x_2^2$, and the corresponding optimal control inputs are given as $u_1^* = -(cos(2x_1) + 2) x_2$, $u_2^* = -\frac{1}{2}\left(sin\left(4x_1^2\right) + 2\right) x_2$.

To implement the developed technique, the activation function for critic NNs are selected as $\phi_i = \begin{bmatrix} x_1^2 & x_1 x_2 & x_2^2 \end{bmatrix}^T$, $i = 1, 2$, while the activation function for the identifier DNN is selected as a symmetric sigmoid with 5 neurons in the hidden layer. The identifier gains are selected as $k = 300$, $\alpha = 200$, $\gamma_f = 5$, $\beta_1 = 0.2$, $\Gamma_{wf} = 0.1\mathbb{I}_{6\times 6}$, $\Gamma_{vf} = 0.1\mathbb{I}_{2\times 2}$, and the gains of the actor-critic learning laws are selected as $\Gamma_{11a} = \Gamma_{12a} = 10$, $\Gamma_{21a} = \Gamma_{22a} = 20$, $\eta_{1c} = 50$, $\eta_{2c} = 10$, $\nu_1 = \nu_2 = 0.001$, $\lambda_1 = \lambda_2 = 0.03$. The covariance matrix is initialized to $\Gamma(0) = 5000\mathbb{I}_{3\times 3}$, the NN weights for state derivative estimator are randomly initialized with values between $[-1, 1]$, the weights for the actor and the critic are initialized to $[3, 3, 3]^T$, the state estimates are initialized to zero, and the states are initialized to $x(0) = [3, -1]$. Similar to results such as [37]–[39], [41], [59], a small amplitude exploratory signal (noise) is added to the control to excite the states for the first 6 seconds of the simulation, as seen from the evolution of states and control in Figure 1. The identifier approximates the system dynamics, and the state derivative estimation error is shown in Figure 1. The time histories of the critic NN weights and the actors NN weights are given in Figure 2, where solid lines denote the weight estimates and dotted lines denote the true values of the weights. Persistence of excitation ensures that the weights converge to their known ideal values in less than 5 seconds of simulation. The use of two separate neural networks facilitates the design of least squares-based update laws in (47). The least squares-based update laws result is a performance benefit over single NN-based results such as [41], where the convergence of weights is obtained after about 250 seconds of simulation.

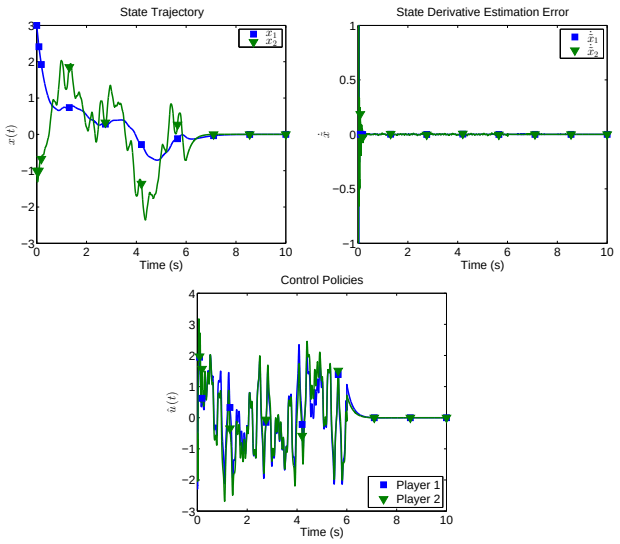

Figure 1. The evolution of the system states, state derivative estimates and control signals for the two-player nonzero-sum game, with persistently excited input for the first 6 seconds.

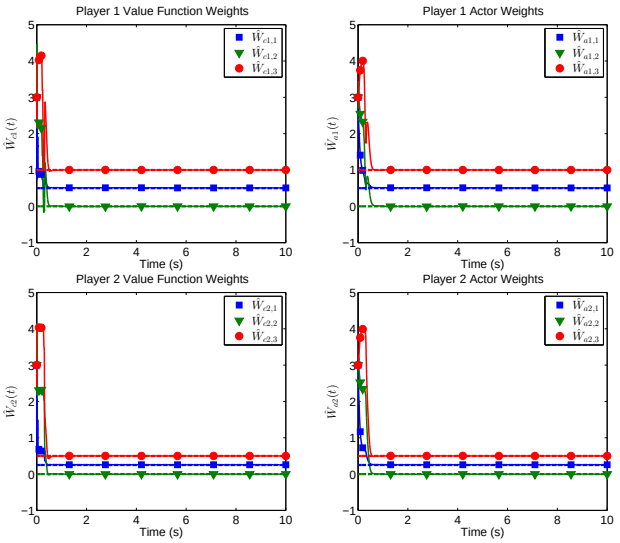

Figure 2. Convergence of actor and critic weights for player 1 and player 2 in the nonzero-sum game.

B. Three player game

To demonstrate the performance of the developed technique in the multi-player case, the two player simulation is augmented with another actor. The resulting dynamics are given by $\dot{x} = f(x) + g_1(x) u_1 + g_2(x) u_2 + g_3(x) u_3$, where

$$f(x) = \begin{bmatrix} (x_2 - 2x_1) \\ \left(-\frac{1}{2}x_1 - x_2 + \frac{1}{4}x_2 \left(\cos(2x_1) + 2\right)^2 \right) \\ \left(+\frac{1}{4}x_2 \left(\sin\left(4x_1^2\right) + 2\right)^2 \right) \\ \left(+\frac{1}{4}x_2 \left(cos\left(4x_1^2\right) + 2\right)^2 \right) \end{bmatrix},$$

$$g_3(x) = \begin{bmatrix} 0 & \cos\left(4x_1^2\right) + 2 \end{bmatrix}^T, \quad (70)$$

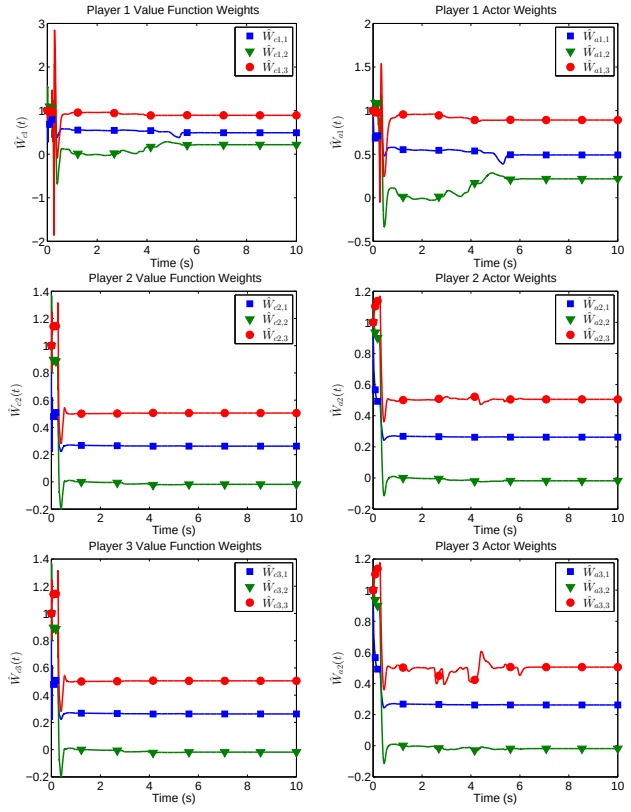

Figure 3. Convergence of actor and critic weights for the three-player nonzero-sum game

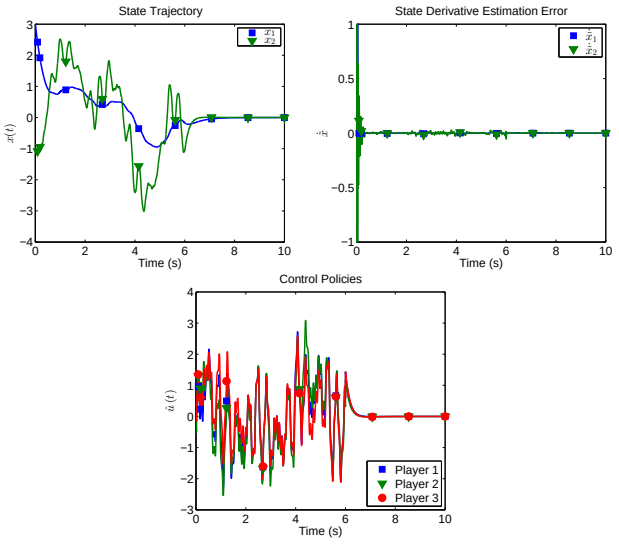

Figure 4. The evolution of the system states, state derivative estimates and control signals for the three-player nonzero-sum game, with persistently excited input for the first 6 seconds.

and g_1 and g_2 are the same as (69). Figure 1 demonstrates the convergence of the actor and the critic weights. Since the Nash equilibrium solution is unknown for the dynamics in (70), the obtained weights are not compared against their true values. Figure 2 demonstrates the regulation of the system states and the state derivative estimation error to the origin, and the boundedness of the control signals.

Remark 7. An implementation issue in using the developed algorithm as well as results such as [37]–[39], [41], [59] is to ensure PE of the critic regressor vector. Unlike linear systems, where PE of the regressor translates to the sufficient richness of the external input, no verifiable method exists to ensure PE in nonlinear systems. In this simulation, a small amplitude exploratory signal consisting of a sum of sines and cosines of varying frequencies is added to the control to ensure PE qualitatively, and convergence of critic weights to their optimal values is achieved. The exploratory signal $n(t)$, designed using trial and error, is present in the first 6 seconds of the simulation and is given by

$$n(t) = \sin(5\pi t) + \sin(et) + \sin^5(t) + \cos^5(20t) + \sin^2(-1.2t)\cos(0.5t).$$

IX. Conclusion

A generalized solution for a N-player nonzero-sum differential game is developed by utilizing an HJB approximation by an ACI architecture. The ACI architecture implements the actor and critic approximation simultaneously and in real-time. The use of a robust DNN-based identifier circumvents the need for complete model knowledge, yielding an identifier which is proven to be asymptotically convergent. A gradient-based weight update law is used for the critic NN to approximate the value function. Using the identifier and the critic, an approximation to the optimal control laws is developed which stabilizes the closed-loop system and approaches the optimal solutions to the N-player nonzero-sum game.

While this result provides an approach for approximating solutions to nonzero-sum differential games, it relies on limiting assumptions such as: the existence and uniqueness of a set Nash solutions for the nonzero-sum game, knowledge of the upper bound of the input matrices of the unknown dynamics, and persistence of excitation for convergence of learning parameters. Future research will focus on relaxing the aforementioned assumptions to broaden the applicability of ADP techniques for nonzero-sum differential games.

References

[1] R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*, ser. Dover Books on Mathematics. Dover Publications, 1999.

[2] S. Tijs, *Introduction to Game Theory*. Hindustan Book Agency, 2003.

[3] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory: Second Edition*, ser. Classics in Applied Mathematics. SIAM, 1999.

[4] M. Abu-Khalaf and F. 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, 2005.

[5] A. Starr and Ho, "Further properties of nonzero-sum differential games," *J. Optim. Theory App.*, vol. 4, pp. 207–219, 1969.

[6] J. Engwerda and A. Weeren, "The open-loop nash equilibrium in lq-games revisited," *Center for Economic Research*, 1995.

[7] J. Case, "Toward a theory of many player differential games," *SIAM J. Control*, vol. 7, pp. 179–197, 1969.

[8] A. Friedman, *Differential games*. Wiley, 1971.

[9] T. Basar and P. Bernhard, *Hinfinity- optimal control and related minimax design problems: A dynamic game approach*. Birkhï¿œuser, 1995.

[10] P. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, D. A. White and D. A. Sofge, Eds. New York: Van Nostrand Reinhold, 1992.

[11] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[12] D. V. Prokhorov and I. Wunsch, D. C., "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, pp. 997–1007, 1997.

[13] 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. Part B Cybern.*, vol. 38, pp. 943–949, 2008.

[14] ——, "Model-free q-learning designs for linear discrete-time zero-sum games with application to H_∞ control," *Automatica*, vol. 43, pp. 473–481, 2007.

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

[16] B. Widrow, N. Gupta, and S. Maitra, "Punish/reward: Learning with a critic in adaptive threshold systems," *IEEE Trans. Syst. Man Cybern.*, vol. 3, no. 5, pp. 455–465, 1973.

[17] A. Barto, R. Sutton, and C. Anderson, "Neuron-like adaptive elements that can solve difficult learning control problems," *IEEE Trans. Syst. Man Cybern.*, vol. 13, no. 5, pp. 834–846, 1983.

[18] R. Sutton, A. Barto, and R. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Contr. Syst. Mag.*, vol. 12, no. 2, pp. 19–22, 1992.

[19] J. Campos and F. Lewis, "Adaptive critic neural network for feedforward compensation," in *Proc. Am. Control Conf.*, vol. 4, 1999.

[20] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Adaptive critic designs for discrete-time zero-sum games with application to h-[infinity] control," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 37, pp. 240–247, 2007.

[21] S. Balakrishnan, "Adaptive-critic-based neural networks for aircraft optimal control," *J. Guid. Contr. Dynam.*, vol. 19, no. 4, pp. 893–898, 1996.

[22] G. Lendaris, L. Schultz, and T. Shannon, "Adaptive critic design for intelligent steering and speed control of a 2-axle vehicle," in *Int. Joint Conf. Neural Netw.*, 2000, pp. 73–78.

[23] S. Ferrari and R. Stengel, "An adaptive critic global controller," in *Proc. Am. Control Conf.*, vol. 4, 2002, pp. 2665–2670.

[24] D. Han and S. Balakrishnan, "State-constrained agile missile control with adaptive-critic-based neural networks," *IEEE Trans. Control Syst. Technol.*, vol. 10, no. 4, pp. 481–489, 2002.

[25] P. He and S. Jagannathan, "Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 37, no. 2, pp. 425–436, 2007.

[26] L. Baird, "Advantage updating," Wright Lab, Wright-Patterson Air Force Base, OH, Tech. Rep., 1993.

[27] K. Doya, "Reinforcement learning in continuous time and space," *Neural Comput.*, vol. 12, no. 1, pp. 219–245, 2000.

[28] J. Murray, C. Cox, G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 32, no. 2, pp. 140–153, 2002.

[29] R. Beard, G. Saridis, and J. Wen, "Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation," *Automatica*, vol. 33, pp. 2159–2178, 1997.

[30] 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, 2009.

[31] K. Vamvoudakis and F. Lewis, "Online synchronous policy iteration method for optimal control," in *Recent Advances in Intelligent Control Systems*, W. Yu, Ed. Springer, 2009, pp. 357–374.

[32] S. Bhasin, N. Sharma, P. Patre, and W. E. Dixon, "Asymptotic tracking by a reinforcement learning-based adaptive critic controller," *J. Control Theory Appl.*, vol. 9, no. 3, pp. 400–409, 2011.

[33] Q. Wei and H. Zhang, "A new approach to solve a class of continuous-time nonlinear quadratic zero-sum game using adp," in *IEEE Int. Conf. Netw. Sens. Control*, 2008, pp. 507–512.

[34] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, pp. 207–214, 2010.

[35] X. Zhang, H. Zhang, Y. Luo, and M. Dong, "Iteration algorithm for solving the optimal strategies of a class of nonaffine nonlinear quadratic zero-sum games," in *Proc. IEEE Conf. Decis. Control*, May 2010, pp. 1359–1364.

[36] A. Mellouk, Ed., *Advances in Reinforcement Learning*. InTech, 2011.

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

[38] ——, "Online neural network solution of nonlinear two-player zero-sum games using synchronous policy iteration," in *Proc. IEEE Conf. Decis. Control*, 2010.

[39] ——, "Multi-player non-zero-sum games: Online adaptive learning solution of coupled hamilton-jacobi equations," *Automatica*, vol. 47, pp. 1556–1569, 2011.

[40] M. Littman, "Value-function reinforcement learning in markov games," *Cogn. Syst. Res.*, vol. 2, no. 1, pp. 55–66, 2001.

[41] H. Zhang, L. Cui, and Y. Luo, "Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network adp," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 206–216, 2013.

[42] M. Johnson, S. Bhasin, and W. E. Dixon, "Nonlinear two-player zero-sum game approximate solution using a policy iteration algorithm," in *Proc. IEEE Conf. Decis. Control*, 2011, pp. 142–147.

[43] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed., ser. Modern Birkhäuser Classics. Boston: Birkhäuser, 2008.

[44] S. Bhasin, "Reinforcement learning and optimal control methods for uncertain nonlinear systems," Ph.D. dissertation, University of Florida, 2011.

[45] F. L. Lewis, R. Selmic, and J. Campos, *Neuro-Fuzzy Control of Industrial Systems with Actuator Nonlinearities*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2002.

[46] W. E. Dixon, A. Behal, D. M. Dawson, and S. Nagarkatti, *Nonlinear Control of Engineering Systems: A Lyapunov-Based Approach*. Birkhauser: Boston, 2003.

[47] M. Krstic, P. V. Kokotovic, and I. Kanellakopoulos, *Nonlinear and Adaptive Control Design*. John Wiley & Sons, 1995.

[48] P. M. Patre, W. MacKunis, K. Kaiser, and W. E. Dixon, "Asymptotic tracking for uncertain dynamic systems via a multilayer neural network feedforward and RISE feedback control structure," *IEEE Trans. Automat. Control*, vol. 53, no. 9, pp. 2180–2185, 2008.

[49] A. F. Filippov, *Differential Equations with Discontinuous Right-hand Sides*. Kluwer Academic Publishers, 1988.

[50] F. H. Clarke, *Optimization and nonsmooth analysis*. SIAM, 1990.

[51] B. Paden and S. Sastry, "A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators," *IEEE Trans. Circuits Syst.*, vol. 34 no. 1, pp. 73–82, 1987.

[52] N. Fischer, R. Kamalapurkar, and W. E. Dixon, "LaSalle-Yoshizawa corollaries for nonsmooth systems," *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2333–2338, 2013.

[53] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*. Upper Saddle River, NJ: Prentice-Hall, 1989.

[54] R. M. Johnstone, C. R. Johnson, R. R. Bitmead, and B. D. O. Anderson, "Exponential convergence of recursive least squares with exponential forgetting factor," in *Proc. IEEE Conf. Decis. Control*, vol. 21, 1982, pp. 994–997.

[55] P. Ioannou and J. Sun, *Robust Adaptive Control*. Prentice Hall, 1996.

[56] A. Loría and E. Panteley, "Uniform exponential stability of linear time-varying systems: revisited," *Syst. Control Lett.*, vol. 47, no. 1, pp. 13 – 24, 2002.

[57] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.

[58] V. Nevistic and J. A. Primbs, "Constrained nonlinear optimal control: a converse HJB approach," California Institute of Technology, Pasadena, CA 91125, Tech. Rep. CIT-CDS 96-021, 1996.

[59] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. L. Lewis, and W. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 89–92, 2013.