Concurrent Learning-based Approximate Feedback-Nash Equilibrium Solution of N-player Nonzero-sum Differential Games

Rushikesh Kamalapurkar Justin R. Klotz Warren E. Dixon

***Abstract*—This paper presents a concurrent learning-based actor-critic-identifier architecture to obtain an approximate feedback-Nash equilibrium solution to an infinite horizon N-player nonzero-sum differential game. The solution is obtained online for a nonlinear control-affine system with uncertain linearly parameterized drift dynamics. It is shown that under a condition milder than persistence of excitation (PE), uniformly ultimately bounded convergence of the developed control policies to the feedback-Nash equilibrium policies can be established. Simulation results are presented to demonstrate the performance of the developed technique without an added excitation signal.**

***Index Terms*—Nonlinear system, optimal adaptive control, dynamic programming, data driven control.**

I. Introduction

CLASSICAL optimal control problems are formulated in Bernoulli form as the need to find a single control input that minimizes a single cost functional under boundary constraints and dynamical constraints imposed by the system[1−2]. A multitude of relevant control problems can be modeled as multi-input systems, where each input is computed by a player, and each player attempts to influence the system state to minimize its own cost function. In this case, the optimization problem for each player is coupled with the optimization problem for other players, and hence, in general, an optimal solution in the usual sense does not exist, motivating the formulation of alternative optimality criteria.

Differential game theory provides solution concepts for many multi-player, multi-objective optimization problems[3−5]. For example, a set of policies is called a Nash equilibrium solution to a multi-objective optimization problem if none of the players can improve their outcomes by changing their policies while all the other players abide by the Nash equilibrium policies[6]. Thus, Nash equilibrium solutions provide a secure set of strategies, in the sense that none of the players have an incentive to diverge from their equilibrium policies. Hence, Nash equilibrium has been a widely used solution concept in differential game-based control techniques.

In general, Nash equilibria are not unique. For a closed-loop differential game (i.e., the control is a function of the state and time) with perfect information (i.e., all the players know the complete state history), there can be infinitely many Nash equilibria. If the policies are constrained to be feedback policies, the resulting equilibria are called (sub)game perfect Nash equilibria or feedback-Nash equilibria. The value functions corresponding to feedback-Nash equilibria satisfy a coupled system of Hamilton-Jacobi (HJ) equations[7−10].

If the system dynamics are nonlinear and uncertain, an analytical solution of the coupled HJ equations is generally infeasible; and hence, dynamic programming-based approximate solutions are sought[11−18]. In [16], an integral reinforcement learning algorithm is presented to solve nonzero-sum differential games in linear systems without the knowledge of the drift matrix. In [17], a dynamic programming-based technique is developed to find an approximate feedback-Nash equilibrium solution to an infinite horizon N-player nonzero-sum differential game online for nonlinear control-affine systems with known dynamics. In [19], a policy iteration-based method is used to solve a two-player zero-sum game online for nonlinear control-affine systems without the knowledge of drift dynamics.

The methods in [17] and [19] solve the differential game online using a parametric function approximator such as a neural network (NN) to approximate the value functions. Since the approximate value functions do not satisfy the coupled HJ equations, a set of residual errors (the so-called Bellman errors (BEs)) is computed along the state trajectories and is used to update the estimates of the unknown parameters in the function approximator using least-squares or gradient-based techniques. Similar to adaptive control, a restrictive persistence of excitation (PE) condition is required to ensure boundedness and convergence of the value function weights. Similar to reinforcement learning, an ad-hoc exploration signal is added to the control signal during the learning phase to satisfy the PE condition along the system trajectories[20−22].

It is unclear how to analytically determine an exploration signal that ensures PE for nonlinear systems; and hence, the exploration signal is typically computed via a simulation-based trial and error approach. Furthermore, the existing online approximate optimal control techniques such as [16 − 17, 19, 23] do not consider the ad-hoc signal in the Lyapunov-based analysis. Hence, the stability of the overall closed-loop implementation is not established. These stability concerns, along with concerns that the added probing signal can result in increased control effort and oscillatory transients, provide motivation for the subsequent development.

Based on the ideas in recent concurrent learning-based adap-

Manuscript received September 14, 2013; accepted January 17, 2014. This work was supported by National Science Foundation Award (1161260, 1217908), Office of Naval Research (N00014-13-1-0151), and a contract with the Air Force Research Laboratory Mathematical Modeling and Optimization Institute. Recommended by Associate Editor Zhongsheng Hou

Citation: Rushikesh Kamalapurkar, Justin R. Klotz, Warren E. Dixon. Concurrent learning-based approximate feedback-Nash equilibrium solution of N-player nonzero-sum differential games. *IEEE/CAA Journal of Automatica Sinica*, 2014, **1**(3): 239−247

Rushikesh Kamalapurkar is with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville 32608, USA (e-mail: rkamalapurkar@ufl.edu).

Justin R. Klotz is with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville 32608, USA (e-mail: jklotz@ufl.edu).

Warren E. Dixon is with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville 32608, USA (e-mail: wdixon@ufl.edu).

tive control results such as [24] and [25] that show a concurrent learning-based adaptive update law can exploit recorded data to augment the adaptive update laws to establish parameter convergence under conditions milder than PE, this paper extends the work in [17] and [19] to relax the PE condition. In this paper, a concurrent learning-based actor-critic-identifier architecture[23] is used to obtain an approximate feedback-Nash equilibrium solution to an infinite horizon N-player nonzero-sum differential game online, without requiring PE, for a nonlinear control-affine system with uncertain linearly parameterized drift dynamics.

A system identifier is used to estimate the unknown parameters in the drift dynamics. The solutions to the coupled HJ equations and the corresponding feedback-Nash equilibrium policies are approximated using parametric universal function approximators. Based on estimates of the unknown drift parameters, estimates for the BEs are evaluated at a set of pre-selected points in the state-space. The value function and the policy weights are updated using a concurrent learning-based least-squares approach to minimize the instantaneous BEs and the BEs evaluated at pre-selected points. Simultaneously, the unknown parameters in the drift dynamics are updated using a history stack of recorded data via a concurrent learning-based gradient descent approach. It is shown that under a condition milder than PE, uniformly ultimately bounded (UUB) convergence of the unknown drift parameters, the value function weights and the policy weights to their true values can be established. Simulation results are presented to demonstrate the performance of the developed technique without an added excitation signal.

II. Problem formulation and exact solution

Consider a class of control-affine multi-input systems

$$\dot{x} = f(x) + \sum_{i=1}^{N} g_i(x)\,\hat{u}_i, \tag{1}$$

where $x \in \mathbf{R}^n$ is the state and $\hat{u}_i \in \mathbf{R}^{m_i}$ are the control inputs (i.e., the players). In (1), the unknown function $f : \mathbf{R}^n \to \mathbf{R}^n$ is linearly parameterizable, the function $g_i : \mathbf{R}^n \to \mathbf{R}^{n\times m_i}$ is known and uniformly bounded, f and g_i are locally Lipschitz, and $f(0) = 0$. Let

$$U := \{\{u_i : \mathbf{R}^n \to \mathbf{R}^{m_i}, i = 1, \cdots, N\}, \text{ such that the tuple } \{u_i, \cdots, u_N\} \text{ is admissible w.r.t. (1)}\}$$

be the set of all admissible tuples of feedback policies. Let $V_i^{\{u_i,\cdots,u_N\}} : \mathbf{R}^n \to \mathbf{R}_{\geq 0}$ denote the value function of the ith player w.r.t. the tuple of feedback policies $\{u_1, \cdots, u_N\} \in U$, defined as

$$V_i^{\{u_1,\cdots,u_N\}}(x) = \int_t^{\infty} r_i\left(\phi(\tau, x), u_i\left(\phi(\tau, x)\right), \cdots, u_N\left(\phi(\tau, x)\right)\right) \mathrm{d}\tau, \tag{2}$$

where $\phi(\tau, x)$ for $\tau \in [t, \infty)$ denotes the trajectory of (1) obtained using the feedback policies $\hat{u}_i(\tau) = u_i(\phi(\tau, x))$ and the initial condition $\phi(t, x) = x$. In (2), $r_i : \mathbf{R}^n \times \mathbf{R}^{m_1} \times \cdots \times \mathbf{R}^{m_N} \to \mathbf{R}_{\geq 0}$ denotes the instantaneous cost defined as $r_i(x, u_i, \cdots, u_N) := x^{\mathrm{T}} Q_i x + \sum_{j=1}^{N} u_j^{\mathrm{T}} R_{ij} u_j$, where $Q_i \in \mathbf{R}^{n\times n}$ is a positive definite matrix. The control objective is to find an approximate feedback-Nash equilibrium solution to the infinite horizon regulation differential game online, i.e., to find a tuple $\{u_1^*, \cdots, u_N^*\} \in U$ such that for all $i \in \{1, \cdots, N\}$, for all $x \in \mathbf{R}^n$, the corresponding value functions satisfy

$$V_i^*(x) := V_i^{\{u_1^*, u_2^*, \cdots, u_i^*, \cdots, u_N^*\}}(x) \leq V_i^{\{u_1^*, u_2^*, \cdots, u_i, \cdots, u_N^*\}}(x)$$

for all u_i such that $\{u_1^*, u_2^*, \cdots, u_i, \cdots, u_N^*\} \in U$.

The exact closed-loop feedback-Nash equilibrium solution $\{u_i^*, \cdots, u_N^*\}$ can be expressed in terms of the value functions as[5, 8−9, 17]

$$u_i^* = -\frac{1}{2} R_{ii}^{-1} g_i^{\mathrm{T}} \left(\nabla_x V_i^*\right)^{\mathrm{T}}, \tag{3}$$

where $\nabla_x := \frac{\partial}{\partial x}$ and the value functions $\{V_1^*, \cdots, V_N^*\}$ are the solutions to the coupled HJ equations

$$x^{\mathrm{T}} Q_i x + \sum_{j=1}^{N} \frac{1}{4} \nabla_x V_j^* G_{ij} \left(\nabla_x V_j^*\right)^{\mathrm{T}} + \nabla_x V_i^* f - \frac{1}{2} \nabla_x V_i^* \sum_{j=1}^{N} G_j \left(\nabla_x V_j^*\right)^{\mathrm{T}} = 0. \tag{4}$$

In (4), $G_j := g_j R_{jj}^{-1} g_j^{\mathrm{T}}$ and $G_{ij} := g_j R_{jj}^{-1} R_{ij} R_{jj}^{-1} g_j^{\mathrm{T}}$. The HJ equations in (4) are in the so-called closed-loop form; they can be expressed in an open-loop form as

$$x^{\mathrm{T}} Q_i x + \sum_{j=1}^{N} u_j^{*\mathrm{T}} R_{ij} u_j^* + \nabla_x V_i^* f + \nabla_x V_i^* \sum_{j=1}^{N} g_j u_j^* = 0. \tag{5}$$

III. Approximate solution

Computation of an analytical solution to the coupled nonlinear HJ equations in (4) is, in general, infeasible. Hence, an approximate solution $\left\{\hat{V}_1, \cdots, \hat{V}_N\right\}$ is sought. Based on $\left\{\hat{V}_1, \cdots, \hat{V}_N\right\}$, an approximation $\{\hat{u}_i, \cdots, \hat{u}_N\}$ to the closed-loop feedback-Nash equilibrium solution is computed. Since the approximate solution, in general, does not satisfy the HJ equations, a set of residual errors (the so-called Bellman errors (BEs)) is computed as

$$\delta_i = x^{\mathrm{T}} Q_i x + \sum_{j=1}^{N} \hat{u}_j^{\mathrm{T}} R_{ij} \hat{u}_j + \nabla_x \hat{V}_i f + \nabla_x \hat{V}_i \sum_{j=1}^{N} g_j \hat{u}_j, \tag{6}$$

and the approximate solution is recursively improved to drive the BEs to zero. The computation of the BEs in (6) requires knowledge of the drift dynamics f. To eliminate this requirement, a concurrent learning-based system identifier is developed in the following section.

A. System Identification

Let $f(x) = Y(x)\,\theta$ be the linear parametrization of the drift dynamics, where $Y : \mathbf{R}^n \to \mathbf{R}^{n\times p_\theta}$ denotes the locally Lipschitz regression matrix, and $\theta \in \mathbf{R}^{p_\theta}$ denotes the vector of constant, unknown drift parameters. The system identifier is designed as

$$\dot{\hat{x}} = Y(x)\,\hat{\theta} + \sum_{i=1}^{N} g_i(x)\,\hat{u}_i + k_x \tilde{x}, \tag{7}$$

where the measurable state estimation error $\tilde{x}$ is defined as $\tilde{x} := x - \hat{x}$, $k_x \in \mathbf{R}^{n\times n}$ is a positive definite, constant diagonal observer gain matrix, and $\hat{\theta} \in \mathbf{R}^{p_\theta}$ denotes the

vector of estimates of the unknown drift parameters. In traditional adaptive systems, the estimates are updated to minimize the instantaneous state estimation error, and convergence of parameter estimates to their true values can be established under a restrictive PE condition. In this result, a concurrent learning-based data-driven approach is developed to relax the PE condition to a weaker, verifiable rank condition as follows.

Assumption 1[24−25]**.** A history stack $\mathcal{H}_{id}$ containing state-action tuples $\left\{\left(x_j, \hat{u}_{1_j}, \cdots, \hat{u}_{N_j},\right) \mid j = 1, \cdots, M_\theta\right\}$ recorded along the trajectories of (1) is available a priori, such that

$$\text{rank}\left(\sum_{j=1}^{M_\theta} Y_j^{\mathrm{T}} Y_j\right) = p_\theta, \tag{8}$$

where $Y_j = Y(x_j)$, and p_θ denotes the number of unknown parameters in the drift dynamics.

To facilitate the concurrent learning-based parameter update, numerical methods are used to compute the state derivative $\dot{x}_j$ corresponding to $(x_j, \hat{u}_j)$. The update law for the drift parameter estimates is designed as

$$\dot{\hat{\theta}} = \Gamma_\theta Y^{\mathrm{T}} \tilde{x} + \Gamma_\theta k_\theta \sum_{j=1}^{M_\theta} Y_j^{\mathrm{T}} \left(\dot{x}_j - \sum_{i=1}^{N} g_{i_j} \hat{u}_{i_j} - Y_j \hat{\theta}\right), \tag{9}$$

where $g_{i_j} := g_i(x_j)$, $\Gamma_\theta \in \mathbf{R}^{p\times p}$ is a constant positive definite adaptation gain matrix, and $k_\theta \in \mathbf{R}$ is a constant positive concurrent learning gain. The update law in (9) requires the unmeasurable state derivative $\dot{x}_j$. Since the state derivative at a past recorded point on the state trajectory is required, past and future recorded values of the state can be used along with accurate noncausal smoothing techniques to obtain good estimates of $\dot{x}_j$. In the presence of derivative estimation errors, the parameter estimation errors can be shown to be UUB, where the size of the ultimate bound depends on the error in the derivative estimate[24].

To incorporate new information, the history stack is updated with new data. Thus, the resulting closed-loop system is a switched system. To ensure the stability of the switched system, the history stack is updated using a singular value maximizing algorithm[24]. Using (1), the state derivative can be expressed as

$$\dot{x}_j - \sum_{i=1}^{N} g_{i_j} \hat{u}_{i_j} = Y_j \theta,$$

and hence, the update law in (9) can be expressed in the advantageous form

$$\dot{\tilde{\theta}} = -\Gamma_\theta Y^{\mathrm{T}} \tilde{x} - \Gamma_\theta k_\theta \left(\sum_{j=1}^{M_\theta} Y_j^{\mathrm{T}} Y_j\right) \tilde{\theta}, \tag{10}$$

where $\tilde{\theta} := \theta - \hat{\theta}$ denotes the drift parameter estimation error. The closed-loop dynamics of the state estimation error are given by

$$\dot{\tilde{x}} = Y\tilde{\theta} - k_x \tilde{x}. \tag{11}$$

B. Value Function Approximation

The value functions, i.e., the solutions to the HJ equations in (4), are continuously differentiable functions of the state. Using the universal approximation property of NNs, the value functions can be represented as

$$V_i^*(x) = W_i^{\mathrm{T}} \sigma_i(x) + \epsilon_i(x), \tag{12}$$

where $W_i \in \mathbf{R}^{p_{W_i}}$ denotes the constant vector of unknown NN weights, $\sigma_i : \mathbf{R}^n \to \mathbf{R}^{p_{W_i}}$ denotes the known NN activation function, $p_{Wi} \in \mathbf{N}$ denotes the number of hidden layer neurons, and $\epsilon_i : \mathbf{R}^n \to \mathbf{R}$ denotes the unknown function reconstruction error. The universal function approximation property guarantees that over any compact domain $\mathcal{C} \subset \mathbf{R}^n$, for all constants $\overline{\epsilon}_i, \overline{\epsilon}_i' > 0$, there exist a set of weights and basis functions such that $\|W_i\| \leq \overline{W}$, $\sup_{x\in\mathcal{C}} \|\sigma_i(x)\| \leq \overline{\sigma}_i$, $\sup_{x\in\mathcal{C}} \|\sigma_i'(x)\| \leq \overline{\sigma}_i'$, $\sup_{x\in\mathcal{C}} \|\epsilon_i(x)\| \leq \overline{\epsilon}_i$ and $\sup_{x\in\mathcal{C}} \|\epsilon_i'(x)\| \leq \overline{\epsilon}_i'$, where $\overline{W}_i, \overline{\sigma}_i, \overline{\sigma}_i', \overline{\epsilon}_i, \overline{\epsilon}_i' \in \mathbf{R}$ are positive constants. Based on (3) and (12), the feedback-Nash equilibrium solutions are given by

$$u_i^*(x) = -\frac{1}{2} R_{ii}^{-1} g_i^{\mathrm{T}}(x) \left(\sigma_i'^{\mathrm{T}}(x) W_i + \epsilon_i'^{\mathrm{T}}(x)\right). \tag{13}$$

The NN-based approximations to the value functions and the controllers are defined as

$$\hat{V}_i := \hat{W}_{ci}^{\mathrm{T}} \sigma_i, \quad \hat{u}_i := -\frac{1}{2} R_{ii}^{-1} g_i^{\mathrm{T}} \sigma_i'^{\mathrm{T}} \hat{W}_{ai}, \tag{14}$$

where $\hat{W}_{ci} \in \mathbf{R}^{p_{W_i}}$, i.e., the value function weights, and $\hat{W}_{ai} \in \mathbf{R}^{p_{W_i}}$, i.e., the policy weights, are the estimates of the ideal weights W_i. The use of two different sets of estimates to approximate the same set of ideal weights is motivated by the subsequent stability analysis and the fact that it facilitates an approximate formulation of the BEs which is affine in the value function weights, enabling least squares-based adaptation. Based on (14), measurable approximations to the BEs in (6) are developed as

$$\hat{\delta}_i = \omega_i^{\mathrm{T}} \hat{W}_{ci} + x^{\mathrm{T}} Q_i x + \sum_{j=1}^{N} \frac{1}{4} \hat{W}_{aj}^{\mathrm{T}} \sigma_j' G_{ij} \sigma_j'^{\mathrm{T}} \hat{W}_{aj}, \tag{15}$$

where $\omega_i := \sigma_i' Y \hat{\theta} - \frac{1}{2} \sum_{j=1}^{N} \sigma_i' G_j \sigma_j'^{\mathrm{T}} \hat{W}_{aj}$. The following assumption, which is generally weaker than the PE assumption, is required for convergence of the concurrent learning-based value function weight estimates.

Assumption 2. For each $i \in \{1, \cdots, N\}$, there exist a finite set of M_{xi} points $\{x_{ij} \in \mathbf{R}^n \mid j = 1, \cdots, M_{xi}\}$ such that for all $t \in \mathbf{R}_{\geq 0}$,

$$\text{rank}\left(\sum_{k=1}^{M_{xi}} \frac{\omega_i^k(t) \left(\omega_i^k\right)^{\mathrm{T}}(t)}{\rho_i^k(t)}\right) = p_{W_i},$$
$$\underline{c}_{xi} := \frac{\left(\inf\limits_{t\in\mathbf{R}_{\geq 0}} \left(\lambda_{\min}\left\{\sum\limits_{k=1}^{M_{xi}} \frac{\omega_i^k(t)\left(\omega_i^k\right)^{\mathrm{T}}(t)}{\rho_i^k(t)}\right\}\right)\right)}{M_{xi}} > 0, \tag{16}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue, and $\underline{c}_{xi} \in \mathbf{R}$ is a positive constant. In (16), $\omega_i^k(t) := \sigma_i'^{ik} Y^{ik} \hat{\theta}(t) - \frac{1}{2} \sum_{j=1}^{N} \sigma_i'^{ik} G_j^{ik} \left(\sigma_j'^{ik}\right)^{\mathrm{T}} \hat{W}_{aj}(t)$, where the superscript ik indicates that the term is evaluated at $x = x_{ik}$, and $\rho_i^k := 1 + \nu_i \left(\omega_i^k\right)^{\mathrm{T}} \Gamma_i \omega_i^k$, where $\nu_i \in \mathbf{R}_{>0}$ is the normalization gain and $\Gamma_i \in \mathbf{R}^{P_{W_i}\times P_{W_i}}$ is the adaptation gain matrix.

The concurrent learning-based least-squares update law for the value function weights is designed as

$$\dot{\hat{W}}_{ci} = -\eta_{c1i}\Gamma_i\frac{\omega_i}{\rho_i}\hat{\delta}_i - \frac{\eta_{c2i}\Gamma_i}{M_{xi}}\sum_{k=1}^{M_{xi}}\frac{\omega_i^k}{\rho_i^k}\hat{\delta}_i^k,$$
$$\dot{\Gamma}_i = \left(\beta_i\Gamma_i - \eta_{c1i}\Gamma_i\frac{\omega_i\omega_i^{\mathrm{T}}}{\rho_i^2}\Gamma_i\right)\mathbf{1}_{\{\|\Gamma_i\|\leq\overline{\Gamma}_i\}},$$
$$\|\Gamma_i(t_0)\| \leq \overline{\Gamma}_i, \tag{17}$$

where $\rho_i := 1+\nu_i\omega_i^{\mathrm{T}}\Gamma_i\omega_i$, $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function, $\overline{\Gamma}_i > 0 \in \mathbf{R}$ is the saturation constant, $\beta_i \in \mathbf{R}$ is the constant positive forgetting factor, $\eta_{c1i}, \eta_{c2i} \in \mathbf{R}$ are constant positive adaptation gains, and the approximate BE $\hat{\delta}_i^k$ is defined as

$$\hat{\delta}_i^k := \left(\omega_i^k\right)^{\mathrm{T}}\hat{W}_{ci} + x_{ik}^{\mathrm{T}}Q_ix_{ik} + \sum_{j=1}^{N}\frac{\hat{W}_{aj}^{\mathrm{T}}\sigma_j'^{ik}G_{ij}^{ik}\left(\sigma_j'^{ik}\right)^{\mathrm{T}}\hat{W}_{aj}}{4}.$$

The policy weight update laws are designed based on the subsequent stability analysis as

$$\dot{\hat{W}}_{ai} = -\eta_{a1i}\left(\hat{W}_{ai} - \hat{W}_{ci}\right) - \eta_{a2i}\hat{W}_{ai} + \frac{1}{4}\sum_{j=1}^{N}\eta_{c1i}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}\hat{W}_{aj}^{\mathrm{T}}\frac{\omega_i^{\mathrm{T}}}{\rho_i}\hat{W}_{ci}^{\mathrm{T}} + \frac{1}{4}\sum_{k=1}^{M_{xi}}\sum_{j=1}^{N}\frac{\eta_{c2i}}{M_{xi}}\sigma_j'^{ik}G_{ij}^{ik}\left(\sigma_j'^{ik}\right)^{\mathrm{T}}\hat{W}_{aj}^{\mathrm{T}}\frac{\left(\omega_i^k\right)^{\mathrm{T}}}{\rho_i^k}\hat{W}_{ci}^{\mathrm{T}}, \tag{18}$$

where $\eta_{a1i}, \eta_{a2i} \in \mathbf{R}$ are positive constant adaptation gains. The forgetting factor β_i along with the saturation in the update law for the least-squares gain matrix in (17) ensure that the least-squares gain matrix Γ_i and its inverse is positive definite and bounded for all $i \in \{1, \cdots, N\}$ as[26]

$$\underline{\Gamma}_i \leq \|\Gamma_i(t)\| \leq \overline{\Gamma}_i, \quad \forall t \in \mathbf{R}_{\geq 0}, \tag{19}$$

where $\underline{\Gamma}_i \in \mathbf{R}$ is a positive constant, and the normalized regressor is bounded as

$$\left\|\frac{\omega_i}{\rho_i}\right\| \leq \frac{1}{2\sqrt{\nu_i\underline{\Gamma}_i}}.$$

IV. Stability analysis

Subtracting (4) from (15), the approximate BE can be expressed in an unmeasurable form as

$$\hat{\delta}_i = \omega_i^{\mathrm{T}}\hat{W}_{ci} + x^{\mathrm{T}}Q_ix + \sum_{j=1}^{N}\frac{1}{4}\hat{W}_{aj}^{\mathrm{T}}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}\hat{W}_{aj} - \left(x^{\mathrm{T}}Q_ix + \sum_{j=1}^{N}u_j^{*\mathrm{T}}R_{ij}u_j^* + \nabla_xV_i^*f + \nabla_xV_i^*\sum_{j=1}^{N}g_ju_j^*\right).$$

Substituting for V^* and u^* from (12) and (13) and using $f = Y\theta$, the approximate BE can be expressed as

$$\hat{\delta}_i = \omega_i^{\mathrm{T}}\hat{W}_{ci} + \sum_{j=1}^{N}\frac{1}{4}\hat{W}_{aj}^{\mathrm{T}}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}\hat{W}_{aj} - W_i^{\mathrm{T}}\sigma_i'Y\theta - \epsilon_i'Y\theta - \sum_{j=1}^{N}\frac{1}{4}\left(W_j^{\mathrm{T}}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}W_j + 2\epsilon_j'G_{ij}\sigma_j'^{\mathrm{T}}W_j + \epsilon_j'G_{ij}\epsilon_j'^{\mathrm{T}}\right) + \frac{1}{2}\sum_{j=1}^{N}\left(W_i^{\mathrm{T}}\sigma_i'G_j\sigma_j'^{\mathrm{T}}W_j + \epsilon_i'G_j\sigma_j'^{\mathrm{T}}W_j + W_i^{\mathrm{T}}\sigma_i'G_j\epsilon_j'^{\mathrm{T}}\right) + \frac{1}{2}\sum_{j=1}^{N}\epsilon_i'G_j\epsilon_j'^{\mathrm{T}}.$$

Adding and subtracting $\frac{1}{4}\hat{W}_{aj}^{\mathrm{T}}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}W_j + \omega_i^{\mathrm{T}}W_i$ yields

$$\hat{\delta}_i = -\omega_i^{\mathrm{T}}\tilde{W}_{ci} + \frac{1}{4}\sum_{j=1}^{N}\tilde{W}_{aj}^{\mathrm{T}}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}\tilde{W}_{aj} - W_i^{\mathrm{T}}\sigma_i'Y\tilde{\theta} - \frac{1}{2}\sum_{j=1}^{N}\left(W_i^{\mathrm{T}}\sigma_i'G_j - W_j^{\mathrm{T}}\sigma_j'G_{ij}\right)\sigma_j'^{\mathrm{T}}\tilde{W}_{aj} - \epsilon_i'Y\theta + \Delta_i, \tag{20}$$

where $\Delta_i := \frac{1}{2}\sum_{j=1}^{N}\left(W_i^{\mathrm{T}}\sigma_i'G_j - W_j^{\mathrm{T}}\sigma_j'G_{ij}\right)\epsilon_j'^{\mathrm{T}} + \frac{1}{2}\sum_{j=1}^{N}W_j^{\mathrm{T}}\sigma_j'G_j\epsilon_i'^{\mathrm{T}} + \frac{1}{2}\sum_{j=1}^{N}\epsilon_i'G_j\epsilon_j'^{\mathrm{T}} - \sum_{j=1}^{N}\frac{1}{4}\epsilon_j'G_{ij}\epsilon_j'^{\mathrm{T}}$. Similarly, the approximate BE evaluated at the selected points can be expressed in an unmeasurable form as

$$\hat{\delta}_i^k = -\omega_i^{k\mathrm{T}}\tilde{W}_{ci} + \frac{1}{4}\sum_{j=1}^{N}\tilde{W}_{aj}^{\mathrm{T}}\sigma_j'^{ik}G_{ij}^{ik}\left(\sigma_j'^{ik}\right)^{\mathrm{T}}\tilde{W}_{aj} + \Delta_i^k - \frac{1}{2}\sum_{j=1}^{N}\left(W_i^{\mathrm{T}}\sigma_i'^{ik}G_j^{ik} - W_j^{\mathrm{T}}\sigma_j'^{ik}G_{ij}^{ik}\right)\left(\sigma_j'^{ik}\right)^{\mathrm{T}}\tilde{W}_{aj} - W_i^{\mathrm{T}}\sigma_i'^{ik}Y^{ik}\tilde{\theta}, \tag{21}$$

where the constant $\Delta_i^k \in \mathbf{R}$ is defined as $\Delta_i^k := -\epsilon_i'^{ik}Y^{ik}\theta + \Delta_i^{ik}$. To facilitate the stability analysis, a candidate Lyapunov function is defined as

$$V_L = \sum_{i=1}^{N}V_i^* + \frac{1}{2}\sum_{i=1}^{N}\tilde{W}_{ci}^{\mathrm{T}}\Gamma_i^{-1}\tilde{W}_{ci} + \frac{1}{2}\sum_{i=1}^{N}\tilde{W}_{ai}^{\mathrm{T}}\tilde{W}_{ai} + \frac{1}{2}\tilde{x}^{\mathrm{T}}\tilde{x} + \frac{1}{2}\tilde{\theta}^{\mathrm{T}}\Gamma_\theta^{-1}\tilde{\theta}. \tag{22}$$

Since V_i^* are positive definite, the bound in (19) and Lemma 4.3 in [27] can be used to bound the candidate Lyapunov function as

$$\underline{v}(\|Z\|) \leq V_L(Z,t) \leq \overline{v}(\|Z\|), \tag{23}$$

where $Z = \left[x^{\mathrm{T}}, \tilde{W}_{c1}^{\mathrm{T}}, \cdots, \tilde{W}_{cN}^{\mathrm{T}}, \tilde{W}_{a1}^{\mathrm{T}}, \cdots, \tilde{W}_{aN}^{\mathrm{T}}, \tilde{x}, \tilde{\theta}\right]^{\mathrm{T}} \in \mathbf{R}^{2n+2N\sum_i p_{W_i}+p_\theta}$ and $\underline{v}, \overline{v} : \mathbf{R}_{\geq 0} \to \mathbf{R}_{\geq 0}$ are class $\mathcal{K}$

functions. For any compact set $\mathcal{Z} \subset \mathbf{R}^{2n+2N\sum_i p_{W_i}+p_\theta}$, define

$$\iota_1 := \max_{i,j}\left(\sup_{Z\in\mathcal{Z}}\left\|\frac{1}{2}W_i^{\mathrm{T}}\sigma_i'G_j\sigma_j'^{\mathrm{T}}+\frac{1}{2}\epsilon_i'G_j\sigma_j'^{\mathrm{T}}\right\|\right),$$

$$\iota_2 := \max_{i,j}\Bigg(\sup_{Z\in\mathcal{Z}}\Bigg\|\frac{\eta_{c1i}\omega_i}{4\rho_i}\left(3W_j\sigma_j'G_{ij}-2W_i^{\mathrm{T}}\sigma_i'G_j\right)\sigma_j'^{\mathrm{T}}+ \sum_{k=1}^{M_{xi}}\frac{\eta_{c2i}\omega_i^k}{4M_{xi}\rho_i^k}\left(3W_j^{\mathrm{T}}\sigma_j'^{ik}G_{ij}^{ik}-2W_i^{\mathrm{T}}\sigma_i'^{ik}G_j^{ik}\right)\left(\sigma_j'^{ik}\right)^{\mathrm{T}}\Bigg\|\Bigg),$$

$$\iota_3 := \max_{i,j}\Bigg(\sup_{Z\in\mathcal{Z}}\Bigg\|\frac{1}{2}\sum_{i,j=1}^{N}\left(W_i^{\mathrm{T}}\sigma_i'+\epsilon_i'\right)G_j\epsilon_j'^{\mathrm{T}}- \frac{1}{4}\sum_{i,j=1}^{N}\left(2W_j^{\mathrm{T}}\sigma_j'+\epsilon_j'\right)G_{ij}\epsilon_j'^{\mathrm{T}}\Bigg\|\Bigg),$$

$$\iota_4 := \max_{i,j}\left(\sup_{Z\in\mathcal{Z}}\left\|\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}\right\|\right),\ \iota_{5i} := \frac{\eta_{c1i}L_Y\overline{\epsilon}_i'\overline{\theta}}{4\sqrt{\nu_i\underline{\Gamma}_i}},$$

$$\iota_{6i} := \frac{\eta_{c1i}L_Y\overline{W}_i\overline{\sigma}_i'}{4\sqrt{\nu_i\underline{\Gamma}_i}},\ \iota_{7i} := \frac{\eta_{c2i}\max_k\left\|\sigma_i'^{ik}Y^{ik}\right\|\overline{W}_i}{4\sqrt{\nu_i\underline{\Gamma}_i}},$$

$$\iota_8 := \sum_{i=1}^{N}\frac{\left(\eta_{c1i}+\eta_{c2i}\right)\overline{W}_i\iota_4}{8\sqrt{\nu_i\underline{\Gamma}_i}},\ \iota_{9i} := \left(\iota_1 N+\left(\eta_{a2i}+\iota_8\right)\overline{W}_i\right),$$

$$\iota_{10i} := \frac{\eta_{c1i}\sup_{Z\in\mathcal{Z}}\|\Delta_i\|+\eta_{c2i}\max_k\left\|\Delta_i^k\right\|}{2\sqrt{\nu_i\underline{\Gamma}_i}},$$

$$v_l := \sqrt{\frac{1}{2}\min\left(\frac{\underline{q_i}}{2},\frac{\eta_{c2i}\underline{c}_{xi}}{4},\underline{k_x},\frac{2\eta_{a1i}+\eta_{a2i}}{8},\frac{k_\theta\underline{y}}{2}\right)},$$

$$\iota := \sqrt{\sum_{i=1}^{N}\left(\frac{2\iota_{9i}^2}{2\eta_{a1i}+\eta_{a2i}}+\frac{\iota_{10i}^2}{\eta_{c2i}\underline{c}_{xi}}\right)+\iota_3}, \tag{24}$$

where $\underline{y}$ denotes the minimum eigenvalue of $\sum_{j=1}^{M_\theta}Y_j^TY_j$, $\underline{q_i}$ denotes the minimum eigenvalue of Q_i, $\underline{k_x}$ denotes the minimum eigenvalue of k_x, and the suprema exist since $\frac{\omega_i}{\rho_i}$ is uniformly bounded for all Z, and the functions G_i, G_{ij}, σ_i, and ϵ_i' are continuous. In (24), $L_Y \in \mathbf{R}_{\geq 0}$ denotes the Lipschitz constant such that $\|Y(\varpi)\| \leq L_Y\|\varpi\|$ for all $\varpi \in \mathcal{Z}\cap\mathbf{R}^n$. The sufficient conditions for UUB convergence are derived based on the subsequent stability analysis as

$$\begin{aligned}&\underline{q_i} > 2\iota_{5i},\\ &\eta_{c2i}\underline{c}_{xi} > 2\iota_{5i}+2\zeta_1\iota_{7i}+\iota_2\zeta_2 N+\eta_{a1i}+2\zeta_3\iota_{6i}\overline{Z},\\ &2\eta_{a1i}+\eta_{a2i} > 4\iota_8+\frac{2\iota_2 N}{\zeta_2},\\ &k_\theta\underline{y} > \frac{2\iota_{7i}}{\zeta_1}+2\frac{\iota_{6i}}{\zeta_3}\overline{Z},\end{aligned} \tag{25}$$

where $\overline{Z} := \underline{v}^{-1}\left(\overline{v}\left(\max\left(\|Z(t_0)\|,\frac{\iota}{v_l}\right)\right)\right)$ and $\zeta_1,\zeta_2,\zeta_3 \in \mathbf{R}$ are known positive adjustable constants.

Since the NN function approximation error and the Lipschitz constant L_Y depend on the compact set that contains the state trajectories, the compact set needs to be established before the gains can be selected using (25). Based on the subsequent stability analysis, an algorithm is developed to compute the required compact set (denoted by $\mathcal{Z}$) based on the initial conditions. In Algorithm 1, the notation $\{\varpi\}_i$ for any parameter ϖ denotes the value of ϖ computed in the i-th iteration. Since the constants ι and v_l depend on L_Y only through the products $L_Y\overline{\epsilon}_i'$ and $L_Y\zeta_3$, Algorithm 1 ensures that

$$\frac{\iota}{v_l} \leq \frac{1}{2}\mathrm{diam}(\mathcal{Z}), \tag{26}$$

where $\mathrm{diam}(\mathcal{Z})$ denotes the diameter of set $\mathcal{Z}$.

Algorithm 1. Gain Selection

First iteration:

Given $z \in \mathbf{R}_{\geq 0}$ such that $\|Z(t_0)\| < z$, let $\mathcal{Z}_1 := \left\{\xi \in \mathbf{R}^{2n+2N\sum_i\{p_{W_i}\}_1+p_\theta} \mid \|\xi\| \leq \underline{v}^{-1}(\overline{v}(z))\right\}$. Using $\mathcal{Z}_1$, compute the bounds in (24) and select the gains according to (25). If $\left\{\frac{\iota}{v_l}\right\}_1 \leq z$, set $\mathcal{Z} = \mathcal{Z}_1$ and terminate.

Second iteration:

If $z < \left\{\frac{\iota}{v_l}\right\}_1$, let $\mathcal{Z}_2 := \left\{\xi \in \mathbf{R}^{2n+2N\sum_i\{p_{W_i}\}_1+p_\theta} \mid \|\xi\| \leq \underline{v}^{-1}\left(\overline{v}\left(\left\{\frac{\iota}{v_l}\right\}_1\right)\right)\right\}$.

Using $\mathcal{Z}_2$, compute the bounds in (24) and select the gains according to (25). If $\left\{\frac{\iota}{v_l}\right\}_2 \leq \left\{\frac{\iota}{v_l}\right\}_1$, set $\mathcal{Z} = \mathcal{Z}_2$ and terminate.

Third iteration:

If $\left\{\frac{\iota}{v_l}\right\}_2 > \left\{\frac{\iota}{v_l}\right\}_1$, increase the number of NN neurons to $\{p_{W_i}\}_3$ to ensure $\{L_Y\}_2\{\overline{\epsilon}_i'\}_3 \leq \{L_Y\}_2\{\overline{\epsilon}_i'\}_2$, $\forall i = 1,\cdots,N$, decrease constant ζ_3 to ensure $\{L_Y\}_2\ \{\zeta_3\}_3 \leq \{L_Y\}_2\ \{\zeta_3\}_2$, and increase gain k_θ to satisfy the gain conditions in (25). These adjustments ensure $\{\iota\}_3 \leq \{\iota\}_2$. Set $\mathcal{Z} = \left\{\xi \in \mathbf{R}^{2n+2N\sum_i\{p_{W_i}\}_3+p_\theta} \mid \|\xi\| \leq \underline{v}^{-1}\left(\overline{v}\left(\left\{\frac{\iota}{v_l}\right\}_2\right)\right)\right\}$ and terminate.

Theorem 1. Provided Assumptions $1\sim 2$ hold and the control gains satisfy the sufficient conditions in (25), where the constants in (24) are computed based on the compact set $\mathcal{Z}$ selected using Algorithm 1, the system identifier in (7) along with the adaptive update law in (9), and the controllers in (14) along with the adaptive update laws in (17) and (18) ensure that the state x, the state estimation error $\tilde{x}$, the value function weight estimation errors $\tilde{W}_{ci}$ and the policy weight estimation errors $\tilde{W}_{ai}$ are UUB, resulting in UUB convergence of the policies $\hat{u}_i$ to the feedback-Nash equilibrium policies u_i^*.

Proof. The derivative of the candidate Lyapunov function in (22) along the trajectories of (1), (10), (11), (17), and (18) is given by

$$\begin{aligned}\dot{V}_L = &\sum_{i=1}^{N}\left(\nabla_x V_i^*\left(f+\sum_{j=1}^{N}g_ju_j\right)\right)+\tilde{x}^{\mathrm{T}}\left(Y\tilde{\theta}-k_x\tilde{x}\right)+\\ &\sum_{i=1}^{N}\tilde{W}_{ci}^{\mathrm{T}}\left(\frac{\eta_{c1i}\omega_i}{\rho_i}\hat{\delta}_i+\frac{\eta_{c2i}}{M_{xi}}\sum_{i=1}^{M_{xi}}\frac{\omega_i^k}{\rho_i^k}\hat{\delta}_i^k\right)-\\ &\frac{1}{2}\sum_{i=1}^{N}\tilde{W}_{ci}^{\mathrm{T}}\left(\beta_i\Gamma_i^{-1}-\eta_{c1i}\frac{\omega_i\omega_i^{\mathrm{T}}}{\rho_i^2}\right)\tilde{W}_{ci}+\\ &\tilde{\theta}^{\mathrm{T}}\left(-Y^{\mathrm{T}}\tilde{x}-k_\theta\left(\sum_{j=1}^{M_\theta}Y_j^{\mathrm{T}}Y_j\right)\tilde{\theta}\right)-\end{aligned}$$

$$\sum_{i=1}^{N}\tilde{W}_{ai}^{\mathrm{T}}\Bigg(-\eta_{a1i}\left(\hat{W}_{ai}^{\mathrm{T}}-\hat{W}_{ci}^{\mathrm{T}}\right)-\eta_{a2i}\hat{W}_{ai}^{\mathrm{T}}+\frac{1}{4}\sum_{j=1}^{N}\eta_{c1i}\hat{W}_{ci}^{\mathrm{T}}\frac{\omega_i}{\rho_i}\hat{W}_{aj}^{\mathrm{T}}\sigma_j'G_{ij}\sigma_j'^{\mathrm{T}}+\frac{1}{4}\sum_{k=1}^{M_{xi}}\sum_{j=1}^{N}\frac{\eta_{c2i}}{M_{xi}}\hat{W}_{ci}^{\mathrm{T}}\frac{\omega_i^k}{\rho_i^k}\hat{W}_{aj}^{\mathrm{T}}\sigma_j'^{ik}G_{ij}^{ik}\left(\sigma_j'^{ik}\right)^{\mathrm{T}}\Bigg). \quad (27)$$

Substituting the unmeasurable forms of the BEs from (20) and (21) into (27), and using the triangle inequality, the Cauchy-Schwarz inequality and Young′s inequality, the Lyapunov derivative in (27) can be bounded as

$$\dot{V}\leq-\sum_{i=1}^{N}\frac{\underline{q_i}}{2}\left\|x\right\|^2-\sum_{i=1}^{N}\frac{\eta_{c2i}\underline{c_{xi}}}{2}\left\|\tilde{W}_{ci}\right\|^2-\underline{k_x}\left\|\tilde{x}\right\|^2-\frac{k_\theta\underline{y}}{2}\left\|\tilde{\theta}\right\|^2-\sum_{i=1}^{N}\left(\frac{2\eta_{a1i}+\eta_{a2i}}{4}\right)\left\|\tilde{W}_{ai}\right\|^2+\sum_{i=1}^{N}\iota_{9i}\left\|\tilde{W}_{ai}\right\|+\sum_{i=1}^{N}\iota_{10i}\left\|\tilde{W}_{ci}\right\|-\sum_{i=1}^{N}\left(\frac{q_i}{2}-\iota_{5i}\right)\left\|x\right\|^2-\sum_{i=1}^{N}\left(\frac{\eta_{c2i}\underline{c_{xi}}}{2}-\iota_{5i}-\zeta_1\iota_{7i}-\frac{1}{2}\iota_2\zeta_2N-\frac{1}{2}\eta_{a1i}-\zeta_3\iota_{6i}\left\|x\right\|\right)\left\|\tilde{W}_{ci}\right\|^2+\sum_{i=1}^{N}\left(\frac{k_\theta\underline{y}}{2}-\frac{\iota_{7i}}{\zeta_1}-\frac{\iota_{6i}}{\zeta_3}\left\|x\right\|\right)\left\|\tilde{\theta}_i\right\|^2+\sum_{i=1}^{N}\left(\frac{2\eta_{a1i}+\eta_{a2i}}{4}-\iota_8-\frac{\iota_2N}{2\zeta_2}\right)\left\|\tilde{W}_{ai}\right\|^2+\iota_3. \quad (28)$$

Provided the sufficient conditions in (25) hold and the conditions

$$\frac{\eta_{c2i}\underline{c_{xi}}}{2}>\iota_{5i}+\zeta_1\iota_{7i}+\frac{1}{2}\iota_2\zeta_2N+\frac{1}{2}\eta_{a1i}+\zeta_3\iota_{6i}\left\|x\right\|,\quad\frac{k_\theta\underline{y}}{2}>\frac{\iota_{7i}}{\zeta_1}+\frac{\iota_{6i}}{\zeta_3}\left\|x\right\| \quad (29)$$

hold for all $t\in\mathbf{R}_{\geq0}$. Completing the squares in (28), the bound on the Lyapunov derivative can be expressed as

$$\dot{V}\leq-\sum_{i=1}^{N}\frac{\underline{q_i}}{2}\left\|x\right\|^2-\sum_{i=1}^{N}\frac{\eta_{c2i}\underline{c_{xi}}}{4}\left\|\tilde{W}_{ci}\right\|^2-\underline{k_x}\left\|\tilde{x}\right\|^2-\sum_{i=1}^{N}\left(\frac{2\eta_{a1i}+\eta_{a2i}}{8}\right)\left\|\tilde{W}_{ai}\right\|^2-\frac{k_\theta\underline{y}}{2}\left\|\tilde{\theta}\right\|^2+\iota\leq-\left(v_l\left\|Z\right\|\right)^2,\quad\forall\left\|Z\right\|>\frac{\iota}{v_l},\ Z\in\mathcal{Z}. \quad (30)$$

Using (23), (26), and (30), Theorem 4.18 in [27] can be invoked to conclude that $\lim_{t\to\infty}\sup\left\|Z\left(t\right)\right\|\leq\underline{v}^{-1}\left(\overline{v}\left(\frac{\iota}{v_l}\right)\right)$. Furthermore, the system trajectories are bounded as $\left\|Z\left(t\right)\right\|\leq\overline{Z}$ for all $t\in\mathbf{R}_{\geq0}$. Hence, the conditions in (25) are sufficient for the conditions in (29) to hold for all $t\in\mathbf{R}_{\geq0}$.

The error between the feedback-Nash equilibrium policy and the approximate policy can be expressed as

$$\left\|u_i^*-\hat{u}_i\right\|\leq\frac{1}{2}\left\|R_{ii}\right\|\overline{g_i}\overline{\sigma_i'}\left(\left\|\tilde{W}_{ai}\right\|+\overline{\epsilon_i'}\right),$$

for all $i=1,\cdots,N$, where $\overline{g_i}:=\sup_x\left\|g_i\left(x\right)\right\|$. Since the weights $\tilde{W}_{ai}$ are UUB, UUB convergence of the approximate policies to the feedback-Nash equilibrium policies is obtained. □

Remark 1. The closed-loop system analyzed using the candidate Lyapunov function in (22) is a switched system. The switching happens when the history stack is updated and when the least-squares regression matrices Γ_i reach their saturation bound. Similar to least squares-based adaptive control[26], (22) can be shown to be a common Lyapunov function for the regression matrix saturation, and the use of a singular value maximizing algorithm to update the history stack ensures that (22) is a common Lyapunov function for the history stack updates[24]. Since (22) is a common Lyapunov function, (23), (26) and (30) establish UUB convergence of the switched system.

V. Simulation

A. Problem Setup

To portray the performance of the developed approach, the concurrent learning-based adaptive technique is applied to the nonlinear control-affine system[17]

$$\dot{x}=f\left(x\right)+g_1\left(x\right)u_1+g_2\left(x\right)u_2, \quad (31)$$

where $x\in\mathbf{R}^2$, $u_1,u_2\in\mathbf{R}$, and

$$f=\begin{bmatrix}x_2-2x_1\\\left(-\frac{1}{2}x_1-x_2+\frac{1}{4}x_2\left(\cos\left(2x_1\right)+2\right)^2\right)\\+\frac{1}{4}x_2\left(\sin\left(4x_1^2\right)+2\right)^2\end{bmatrix},\quad g_1=\begin{bmatrix}0\\\cos\left(2x_1\right)+2\end{bmatrix},\ g_2=\begin{bmatrix}0\\\sin\left(4x_1^2\right)+2\end{bmatrix}.$$

The value function has the structure shown in (2) with the weights $Q_1=2Q_2=2I_2$ and $R_{11}=R_{12}=2R_{21}=2R_{22}=2$, where I_2 is a 2×2 identity matrix. The system identification protocol given in Section III-A and the concurrent learning-based scheme given in Section III-B are implemented simultaneously to provide an approximate online feedback-Nash equilibrium solution to the given nonzero-sum two-player game.

B. Analytical Solution

The control affine system in (31) is selected for this simulation because it is constructed using the converse HJ approach[28] such that the analytical feedback-Nash equilibrium solution of the nonzero-sum game is

$$V_1^*=\begin{bmatrix}0.5\\0\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}x_1^2\\x_1x_2\\x_2^2\end{bmatrix},\quad V_2^*=\begin{bmatrix}0.25\\0\\0.5\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}x_1^2\\x_1x_2\\x_2^2\end{bmatrix},$$

and the feedback-Nash equilibrium control policies for Player 1 and Player 2 are

$$u_1^*=-\frac{1}{2}R_{11}^{-1}g_1^{\mathrm{T}}\begin{bmatrix}2x_1&0\\x_2&x_1\\0&2x_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}0.5\\0\\1\end{bmatrix},\quad u_2^*=-\frac{1}{2}R_{22}^{-1}g_2^{\mathrm{T}}\begin{bmatrix}2x_1&0\\x_2&x_1\\0&2x_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}0.25\\0\\0.5\end{bmatrix}.$$

Since the analytical solution is available, the performance of the developed method can be evaluated by comparing the obtained approximate solution against the analytical solution.

C. Simulation Parameters

The dynamics are linearly parameterized as $f = Y(x)\theta$, where

$$Y(x) = \begin{bmatrix} x_2 & 0 \\ x_1 & 0 \\ 0 & x_1 \\ 0 & x_2 \\ 0 & x_2\left(\cos\left(2x_1\right)+2\right)^2 \\ 0 & x_2\left(\sin\left(4x_1^2\right)+2\right)^2 \end{bmatrix}^{\mathrm{T}}$$

is known and the constant vector of parameters $\theta = \left[1, -2, -\frac{1}{2}, -1, \frac{1}{4}, -\frac{1}{4}\right]^{\mathrm{T}}$ is assumed to be unknown. The initial guess for θ is selected as $\hat{\theta}(t_0) = 0.5 \cdot [1,1,1,1,1,1]^{\mathrm{T}}$. The system identification gains are chosen as $k_x = 5$, $\Gamma_\theta = \mathrm{diag}\{20, 20, 100, 100, 60, 60\}$, $k_\theta = 1.5$. A history stack of 30 points is selected using a singular value maximizing algorithm[24] for the concurrent learning-based update law in (9), and the state derivatives are estimated using a fifth order Savitzky-Golay filter[29]. Based on the structure of the feedback-Nash equilibrium value functions, the basis function for value function approximation is selected as $\sigma = [x_1^2, x_1x_2, x_2^2]^{\mathrm{T}}$, and the adaptive learning parameters and initial conditions are shown for both players in Tables I and II. Twenty-five points lying on a 5×5 grid around the origin are selected for the concurrent learning-based update laws in (17) and (18).

TABLE I
SIMAULATION PARMAMERERS

	Player 1	Player 2
ν	0.005	0.005
η_{c1}	1	1
η_{c2}	1.5	1
η_{a1}	10	10
η_{a2}	0.1	0.1
β	3	3
$\bar{\Gamma}$	10 000	10 000

D. Simulation Results

Figs. $1 \sim 4$ show the rapid convergence of the actor and critic weights to the approximate feedback-Nash equilibrium values for both players, resulting in the value functions and control policies

$$\hat{V}_1 = \begin{bmatrix} 0.5021 \\ -0.0159 \\ 0.9942 \end{bmatrix}^{\mathrm{T}} \sigma, \quad \hat{V}_2 = \begin{bmatrix} 0.2510 \\ -0.0074 \\ 0.4968 \end{bmatrix}^{\mathrm{T}} \sigma,$$

$$\hat{u}_1 = -\frac{1}{2} R_{11}^{-1} g_1^{\mathrm{T}} \begin{bmatrix} 2x_1 & 0 \\ x_2 & x_1 \\ 0 & 2x_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0.4970 \\ -0.0137 \\ 0.9810 \end{bmatrix},$$

$$\hat{u}_2 = -\frac{1}{2} R_{22}^{-1} g_2^{\mathrm{T}} \begin{bmatrix} 2x_1 & 0 \\ x_2 & x_1 \\ 0 & 2x_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0.2485 \\ -0.0055 \\ 0.4872 \end{bmatrix}.$$

Fig. 5 demonstrates that (without the injection of a PE signal) the system identification parameters also approximately converge to the correct values. The state and control signal trajectories are displayed in Figs. 6 and 7.

TABLE II
SIMULATION INITIAL CONDITIONS

	Player 1	Player 2
$\hat{W}_c(t_0)$	$[3,3,3]^{\mathrm{T}}$	$[3,3,3]^{\mathrm{T}}$
$\hat{W}_a(t_0)$	$[3,3,3]^{\mathrm{T}}$	$[3,3,3]^{\mathrm{T}}$
$\Gamma(t_0)$	$100I_3$	$100I_3$
$x(t_0)$	$[1,1]^{\mathrm{T}}$	$[1,1]^{\mathrm{T}}$
$\hat{x}(t_0)$	$[0,0]^{\mathrm{T}}$	$[0,0]^{\mathrm{T}}$

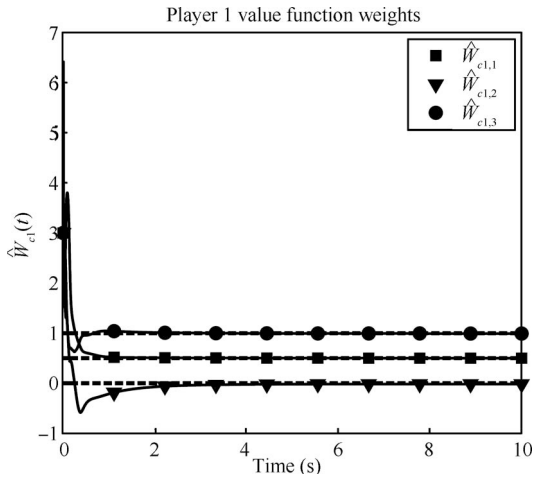

Fig. 1. Player 1 critic weights convergence.

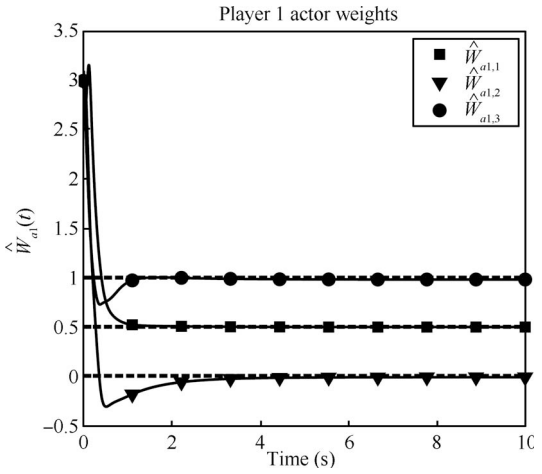

Fig. 2. Player 1 actor weights convergence.

VI. Conclusion

A concurrent learning-based adaptive approach is developed to determine the feedback-Nash equilibrium solution to an N-player nonzero-sum game online. The solutions to the associated coupled HJ equations and the corresponding feedback-Nash equilibrium policies are approximated using parametric universal function approximators. Based on estimates of the unknown drift parameters, estimates for the BEs are evaluated at a set of preselected points in the state-space. The value

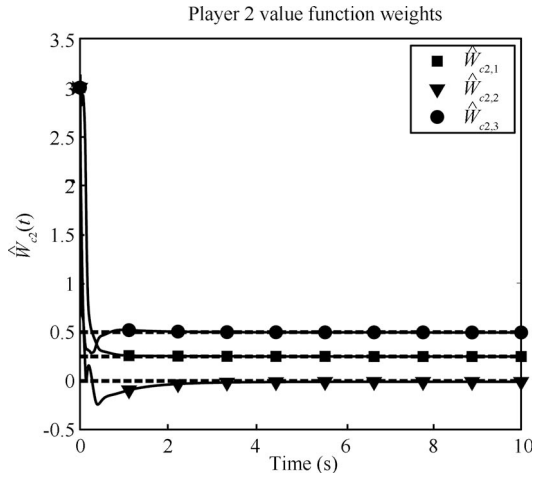

Fig. 3. Player 2 critic weights convergence.

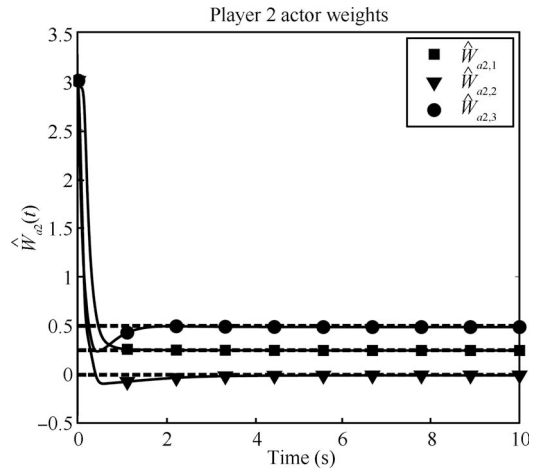

Fig. 4. Player 2 actor weights convergence.

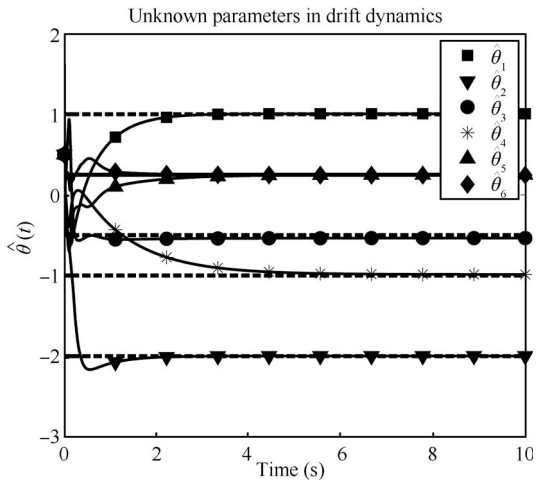

Fig. 5. System identification parameters convergence.

function and the policy weights are updated using a concurrent learning-based least-squares approach to minimize the instantaneous BEs and the BEs evaluated at the preselected points. Simultaneously, the unknown parameters in the drift dynamics are updated using a history stack of recorded data via a concurrent learning-based gradient descent approach.

Unlike traditional approaches that require a restrictive PE condition for convergence, UUB convergence of the drift parameters, the value function and policy weights to their true values, and hence, UUB convergence of the policies to the feedback-Nash equilibrium policies, are established under weaker rank conditions using a Lyapunov-based analysis. Simulations are performed to demonstrate the performance of the developed technique.

The developed result relies on a sufficient condition on the minimum eigenvalue of a time-varying regression matrix. While this condition can be heuristically satisfied by choosing enough points, and can be easily verified online, it cannot, in general, be guaranteed a priori. Furthermore, finding a sufficiently good basis for value function approximation is, in general, nontrivial and can be achieved only through prior knowledge or trial and error. Future research will focus on extending the applicability of the developed technique by relieving the aforementioned shortcomings.

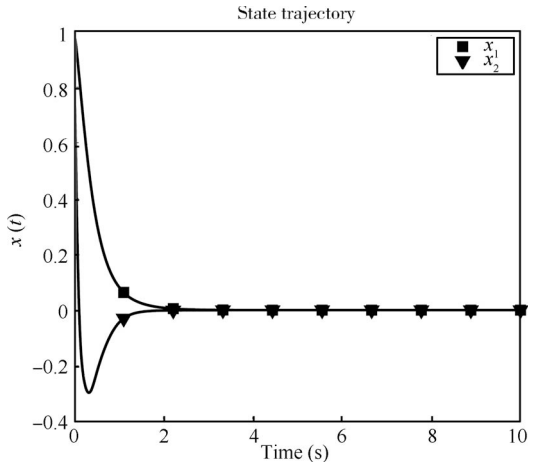

Fig. 6. State trajectory convergence to the origin.

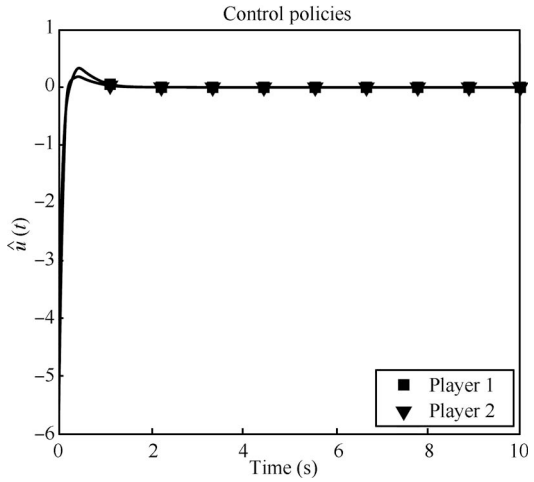

Fig. 7. Control policies of Players 1 and 2.

REFERENCES

[1] Kirk D E. *Optimal Control Theory: An Introduction*. New York: Dover Publications, 2004.

[2] Lewis F L, Vrabie D, Syrmos V L. *Optimal Control (Third edition)*. New York: John Wiley & Sons, 2012.

[3] Isaacs R. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. New York: Dover Publications, 1999.

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

[5] Basar T, Olsder G. *Dynamic Noncooperative Game Theory* (Second edition). Philadelphia, PA: SIAM, 1999.

[6] Nash J. Non-cooperative games. *The Annals of Mathematics*, 1951, **54**(2): 286−295

[7] Case J H. Toward a theory of many player differential games. *SIAM Journal on Control*, 1969, **7**(2): 179−197

[8] Starr A W, Ho C Y. Nonzero-sum differential games. *Journal of Optimization Theory and Applications*, 1969, **3**(3): 184−206

[9] Starr A, Ho C Y. Further properties of nonzero-sum differential games. *Journal of Optimization Theory and Applications*, 1969, **3**(4): 207−219

[10] Friedman A. *Differential Games*. New York: John Wiley and Sons, 1971.

[11] Littman M. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2001, **2**(1): 55−66

[12] Wei Q L, Zhang H G. A new approach to solve a class of continuous-time nonlinear quadratic zero-sum game using ADP. In: Proceedings of the IEEE International Conference on Networking Sensing and Control. Sanya, China: IEEE, 2008. 507−512

[13] Vamvoudakis K, Lewis F. Online synchronous policy iteration method for optimal control. In: Proceedings of the Recent Advances in Intelligent Control Systems. London: Springer, 2009. 357−374

[14] Zhang H G, Wei Q L, Liu D R. An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games. *Automatica*, 2010, **47**: 207−214

[15] Zhang X, Zhang H G, Luo Y H, Dong M. Iteration algorithm for solving the optimal strategies of a class of nonaffine nonlinear quadratic zero-sum games. In: Proceedings of the IEEE Conference Decision and Control. Xuzhou, China: IEEE, 2010. 1359−1364

[16] Vrabie D, Lewis F. Integral reinforcement learning for online computation of feedback Nash strategies of nonzero-sum differential games. In: Proceedings of the 49th IEEE Conference Decision and Control. Atlanta, GA: IEEE, 2010. 3066−3071

[17] Vamvoudakis K G, Lewis F L. Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*, 2011, **47**(8): 1556−1569

[18] Zhang H, Liu D, Luo Y, Wang D. *Adaptive Dynamic Programming for Control-Algorithms and Stability (Communications and Control Engineering)*. London: Springer-Verlag, 2013

[19] Johnson M, Bhasin S, Dixon W E. Nonlinear two-player zero-sum game approximate solution using a policy iteration algorithm. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference. Orlando, FL, USA: IEEE, 2011. 142−147

[20] Mehta P, Meyn S. Q-learning and pontryagin′s minimum principle. In: Proceedings of the IEEE Conference on Decision and Control. Shanghai, China: IEEE, 2009. 3598−3605

[21] Vrabie D, Abu-Khalaf M, Lewis F L, Wang Y Y. Continuous-time ADP for linear systems with partially unknown dynamics. In: Proceedings of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning. Honolulu, HI: IEEE, 2007. 247−253

[22] Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998. 4

[23] Bhasin S, Kamalapurkar R, Johnson M, Vamvoudakis K, Lewis F L, Dixon W. A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, 2013, **49**(1): 89−92

[24] Chowdhary G, Yucelen T, Mühlegg M, Johnson E N. Concurrent learning adaptive control of linear systems with exponentially convergent bounds. *International Journal of Adaptive Control and Signal Processing*, 2012, **27**(4): 280−301

[25] Chowdhary G V, Johnson E N. Theory and flight-test validation of a concurrent-learning adaptive controller. *Journal of Guidance, Control, and Dynamics*, 2011, **34**(2): 592−607

[26] Ioannou P, Sun J. *Robust Adaptive Control*. New Jersey: Prentice Hall, 1996. 198

[27] Khalil H K. *Nonlinear Systems (Third edition)*. New Jersey: Prentice Hall, 2002. 172

[28] Nevistić V, Primbs J A. Constrained nonlinear optimal control: a converse HJB approach. California Institute of Technology, Pasadena, CA 91125, Techical Report CIT-CDS 96-021, 1996. 5

[29] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 1964, **36**(8): 1627−1639

Rushikesh Kamalapurkar Received his bachelor degree in mechanical engineering from Visvesvaraya National Institute of Technology, Nagpur, India. He worked for two years as a design engineer at Larsen and Toubro Ltd., Mumbai, India. He is currently pursuing the Ph. D. degree in the Department of Mechanical and Aerospace Engineering in University of Florida under the supervision of Dr. Warren E. Dixon. His research interest covers learning based control, dynamic programming, optimal control, reinforcement learning and adaptive control for uncertain nonlinear dynamical systems. Corresponding author of this paper.

Justin R. Klotz Received the B. S. and M. S. degrees in mechanical engineering from the University of Florida in 2011 and 2013, respectively. He is currently working towards the Ph. D. degree with a concentration on control theory at the University of Florida under the SMART Scholarship. He was a research assistant for the Air Force Research Laboratory at Eglin Air Force Base for the summers of 2012 and 2013. His research interest covers cooperative network control and optimal control of uncertain nonlinear systems.

Warren E. Dixon Received his Ph. D. degree in 2000 from the Department of Electrical and Computer Engineering, Clemson University. After completing his doctoral studies he was selected as an Eugene P. Wigner Fellow at Oak Ridge National Laboratory (ORNL). In 2004, Dr. Dixon joined the University of Florida in the Department of Mechanical and Aerospace Engineering. His research interest covers the development and application of Lyapunov-based control techniques for uncertain nonlinear systems. He has published 3 books, an edited collection, 9 chapters, and over 250 referred journal and conference papers. His work has been recognized by the 2013 Fred Ellersick Award for Best Overall MILCOM Paper, 2012-2013 University of Florida College of Engineering Doctoral Dissertation Mentoring Award, 2011 American Society of Mechanical Engineers (ASME) Dynamics Systems and Control Division Outstanding Young Investigator Award, 2009 American Automatic Control Council (AACC) O. Hugo Schuck (Best Paper) Award, 2006 IEEE Robotics and Automation Society (RAS) Early Academic Career Award, an NSF CAREER Award (2006−2011), 2004 DOE Outstanding Mentor Award, and the 2001 ORNL Early Career Award for Engineering Achievement. He is an IEEE Control Systems Society (CSS) Distinguished Lecturer. He currently serves as the director of operations for the Executive Committee of the IEEE CSS Board of Governors. He is currently or was formerly an associate editor for *ASME Journal of Dynamic Systems, Measurement and Control*, *Automatica*, *IEEE Transactions on Systems Man and Cybernetics: Part B Cybernetics*, and the *International Journal of Robust and Nonlinear Control*.