# Model-based reinforcement learning for on-line feedback-Nash equilibrium solution of $N$-player nonzero-sum differential games

Rushikesh Kamalapurkar, Justin Klotz, and Warren E. Dixon

***Abstract*—This paper presents a concurrent learning-based actor-critic-identifier architecture to obtain an approximate feedback-Nash equilibrium solution to a deterministic, continuous-time, and infinite-horizon $N$-player nonzero-sum differential game on-line, without requiring persistence of excitation (PE), for non-linear control-affine systems. Convergence of the developed control policies to neighborhoods of the feedback-Nash equilibrium policies is established under a sufficient rank condition. Simulation results are presented to demonstrate the performance of the developed technique.**

## I. Introduction

Various 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. In general, an optimal solution in the usual sense does not exist; hence, alternative criteria for optimality are sought.

Differential game theory provides solution concepts for multi-player, multi-objective optimization problems [1]–[3]. 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 outcome by changing their policy if all the other players abide by the Nash equilibrium policies [4]. The Nash equilibrium provides a secure set of strategies, in the sense that none of the players have an incentive to diverge from their equilibrium policy. 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 are characterized by a coupled system of Hamilton-Jacobi (HJ) equations [5]–[7].

If the system dynamics are non-linear and uncertain, obtaining an analytical solution of the coupled HJ equations is generally infeasible; hence, dynamic programming-based approximate solutions are sought [8]–[14]. In [13], an integral reinforcement learning algorithm is presented to solve nonzero-sum differential games in linear systems without the knowledge of the drift matrix. In [14], 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 on-line for non-linear control-affine systems with known dynamics. In [15], a policy iteration-based method is used to solve a two-player zero-sum game on-line, without the knowledge of drift dynamics, for non-linear control-affine systems.

The methods in [14] and [15] solve the differential game on-line 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. A restrictive persistence of excitation (PE) condition is required to ensure boundedness and convergence of the value function weights. An ad-hoc exploration signal is added to the control signal during the learning phase to satisfy the PE condition along the system trajectories [16]–[18].

Based on the ideas in recent concurrent learning-based results in adaptive control such as [19] and [20] which show that 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 [14] and [15] to relax the PE condition. In this paper, a concurrent learning-based actor-critic architecture (cf. [21]) is used to obtain an approximate feedback-Nash equilibrium solution to an infinite-horizon $N$-player nonzero-sum differential game on-line, without requiring PE, for a non-linear control-affine system.

The solutions to the coupled HJ equations and the corresponding feedback-Nash equilibrium policies are approx-

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

This research is supported in part by NSF award numbers 0901491, 1161260, 1217908, ONR grant number N00014-13-1-0151, and a contract with the AFRL Mathematical Modeling and Optimization Institute. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agency.

imated using parametric universal function approximators. Using the known system dynamics, the Bellman errors are evaluated at a set of preselected 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 preselected points. It is shown that under a sufficient rank condition, uniformly ultimately bounded (UUB) convergence of 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, including parameter identification without an added excitation signal.

## II. Problem formulation and exact solution

Consider a class of control-affine multi-input systems

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

where $x \in \mathbb{R}^n$ is the state and $\hat{u}_i \in \mathbb{R}^{m_i}$ are the control inputs (i.e. the players). In (1), the functions $g_i : \mathbb{R}^n \to \mathbb{R}^{n\times m_i}$ are known, uniformly bounded, and locally Lipschitz, the function $f : \mathbb{R}^n \to \mathbb{R}^n$ is known and $f\left(0\right) = 0$. Let $U \triangleq \{\{u_i : \mathbb{R}^n \to \mathbb{R}^{m_i}, i = 1,..,N\} \mid$ The tuple $\{u_i,..,u_N\}$ is admissible w.r.t. (1)} be the set of admissible tuples of feedback policies. Let $V_i^{\{u_i,..,u_N\}} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ denote the value function of the $i^{th}$ player w.r.t. the tuple of feedback policies $\{u_1,..,u_N\} \in U$, defined as

$$V_i^{\{u_1,..,u_N\}}\left(x_o\right) = \int\limits_{t_o}^{\infty} r_i\left(x\left(\tau\right), u_i\left(x\left(\tau\right)\right),..,u_N\left(x\left(\tau\right)\right)\right)d\tau, \tag{2}$$

where $x\left(\tau\right)$ for $\tau \in \mathbb{R}_{\geq 0}$ denotes the trajectory of (1) obtained using the feedback policies $\hat{u}_i\left(\tau\right) = u_i\left(x\left(\tau\right)\right)$ and the initial condition $x\left(t_o\right) = x_o$. In (2), $r_i : \mathbb{R}^n \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_N} \to \mathbb{R}_{\geq 0}$ denote the instantaneous costs defined as $r_i\left(x, u_i,..,u_N\right) \triangleq x^TQ_ix + \sum_{j=1}^{N} u_j^T R_{ij} u_j$, where $Q_i \in \mathbb{R}^{n\times n}$ and $R_{ij} \in \mathbb{R}^{m\times m}$ are positive definite matrices. The control objective is to find an approximate feedback-Nash equilibrium solution to the infinite-horizon regulation differential game on-line, i.e., to find a tuple $\{u_1^*,..,u_N^*\} \in U$ such that for all $i \in \{1,..,N\}$, for all $x_o \in \mathbb{R}^n$, the corresponding value functions satisfy

$$V_i^*\left(x_o\right) \triangleq V_i^{\{u_1^*,u_2^*,..,u_i^*,..,u_N^*\}}\left(x_o\right) \leq V_i^{\{u_1^*,u_2^*,..,u_i,..,u_N^*\}}\left(x_o\right)$$

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

The exact closed-loop feedback-Nash equilibrium solution $\{u_i^*,..,u_N^*\}$ can be expressed in terms of the value functions as [3], [6], [14]

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

assuming that the solutions $\{V_1^*,..,V_N^*\}$ to the coupled Hamilton-Jacobi (HJ) equations

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

exist and are continuously differentiable. In (4), $G_j \triangleq g_jR_{jj}^{-1}g_j^T$ and $G_{ij} \triangleq g_jR_{jj}^{-1}R_{ij}R_{jj}^{-1}g_j^T$. In (3) and (4), and in the rest of the paper, the dependence of $u_i^*$, $V_i^*$, $g_i$, and $f$ on the state is omitted for notational brevity. The HJ equations in (4) are in the so-called closed-loop form; they can also be expressed in an open-loop form as

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

## III. Approximate solution

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

$$\delta_i = x^TQ_ix + \sum_{j=1}^{N}\hat{u}_j^TR_{ij}\hat{u}_j + \nabla_x\hat{V}_if + \nabla_x\hat{V}_i\sum_{j=1}^{N}g_j\hat{u}_j, \tag{6}$$

and the approximate solution is recursively improved to drive the BEs to zero.

### A. Value function approximation

Using the universal approximation property of NNs, the value functions can be represented as

$$V_i^*\left(x\right) = W_i^T\sigma_i\left(x\right) + \epsilon_i\left(x\right), \tag{7}$$

where $W_i \in \mathbb{R}^{p_{Wi}}$ denote constant vectors of unknown NN weights, $\sigma_i : \mathbb{R}^n \to \mathbb{R}^{p_{Wi}}$ denote the known NN activation functions, $p_{Wi} \in \mathbb{N}$ denote the number of hidden layer neurons, and $\epsilon_i : \mathbb{R}^n \to \mathbb{R}$ denote the unknown function reconstruction errors. The universal function approximation property guarantees that over any compact domain $\mathcal{C} \subset \mathbb{R}^n$, for all constant $\overline{\epsilon}_i, \overline{\epsilon}_i' > 0$, there exists a set of weights and basis functions such that $\|W_i\| \leq \overline{W}$, $\sup_{x\in\mathcal{C}}\|\sigma_i\left(x\right)\| \leq \overline{\sigma}_i$, $\sup_{x\in\mathcal{C}}\|\sigma_i'\left(x\right)\| \leq \overline{\sigma}_i'$, $\sup_{x\in\mathcal{C}}\|\epsilon_i\left(x\right)\| \leq \overline{\epsilon}_i$ and $\sup_{x\in\mathcal{C}}\|\epsilon_i'\left(x\right)\| \leq \overline{\epsilon}_i'$, where $\overline{W}_i, \overline{\sigma}_i, \overline{\sigma}_i', \overline{\epsilon}_i, \overline{\epsilon}_i' \in \mathbb{R}$ are positive constants. Based on (3) and (7), the feedback-Nash equilibrium solutions are

$$u_i^*\left(x\right) = -\frac{1}{2}R_{ii}^{-1}g_i^T\left(x\right)\left(\sigma_i'^T\left(x\right)W_i + \epsilon_i'^T\left(x\right)\right). \tag{8}$$

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

$$\hat{V}_i \triangleq \hat{W}_{ci}^T\sigma_i, \quad \hat{u}_i \triangleq -\frac{1}{2}R_{ii}^{-1}g_i^T\sigma_i'^T\hat{W}_{ai}, \tag{9}$$

where $\hat{W}_{ci} \in \mathbb{R}^{p_{W_i}}$, i.e., the value function weights, and $\hat{W}_{ai} \in \mathbb{R}^{p_{W_i}}$, i.e., the policy weights, are the estimates of the ideal weights $W_i$. In (9), and in the rest of the paper, the dependence of $\sigma_i$, $\sigma_i'$, $\epsilon_i$, and $\epsilon_i'$ on the state is omitted for notational brevity. 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 approximation of the BEs that is affine in the value function weights, enabling least squares-based adaptation. Based on (9), measurable approximations to the BEs in (6) are developed as

$$\hat{\delta}_i = \omega_i^T \hat{W}_{ci} + x^T Q_i x + \sum_{j=1}^{N} \frac{1}{4} \hat{W}_{aj}^T \sigma_j' G_{ij} \sigma_j'^T \hat{W}_{aj}, \quad (10)$$

where $\omega_i \triangleq \sigma_i' f - \frac{1}{2}\sum_{j=1}^{N} \sigma_i' G_j \sigma_j'^T \hat{W}_{aj}$. The following assumption is required for convergence of the concurrent learning-based value function weight estimates.

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

$$\underline{c}_{xi} \triangleq \frac{\left(\inf_{t\in\mathbb{R}_{\geq 0}} \left(\lambda_{\min}\left\{\sum_{k=1}^{M_{xi}} \frac{\omega_i^k \omega_i^{kT}}{\rho_i^k}\right\}\right)\right)}{M_{xi}} > 0, \quad (11)$$

where $\lambda_{\text{min}}$ denotes the minimum eigenvalue, and $\underline{c}_{xi} \in \mathbb{R}$ are positive constants. In (11), $\omega_i^k \triangleq \sigma_i'^{ik} f^{ik} - \frac{1}{2}\sum_{j=1}^{N} \sigma_i'^{ik} G_j^{ik} \left(\sigma_j'^{ik}\right)^T \hat{W}_{aj}$ and $\rho_i^k \triangleq 1 + \nu_i \left(\omega_i^k\right)^T \Gamma_i \omega_i^k$, where the superscripts $ik$ indicate that the corresponding functions are evaluated at $x = x_{ik}$.

Similar to the PE condition, the condition in (11) depends on the system trajectories. Hence, it is unclear how a set of points that satisfies (11) can be selected a priori. However, as demonstrated in the numerical experiment in Section V, the condition in (11) can be heuristically satisfied by collecting redundant data, i.e., by selecting $M_{xi} \gg p_{Wi}$. Furthermore, unlike the traditional PE condition, the minimum eigenvalue in (11) can be computed on-line. Hence, to satisfy the condition in (11), new data points can be selected if the minimum eigenvalues falls below $\underline{c}_{xi} + d_{xi}$ for some constant positive threshold $d_{xi}$.

The concurrent learning-based least-squares update laws for the value function weights are designed as

$$\dot{\hat{W}}_{ci} = -\eta_{c1i}\Gamma_i \frac{\omega_i}{\rho_i}\hat{\delta}_i - \frac{\eta_{c2i}\Gamma_i}{M_{xi}} \sum_{k=1}^{M_{xi}} \frac{\omega_i^k}{\rho_i^k}\hat{\delta}_i^k,$$

$$\dot{\Gamma}_i = \left(\beta_i \Gamma_i - \eta_{c1i}\Gamma_i \frac{\omega_i \omega_i^T}{\rho_i^2}\Gamma_i\right) \mathbf{1}_{\left\{\|\Gamma_i\| \leq \overline{\Gamma}_i\right\}}, \ \|\Gamma_i(t_0)\| \leq \overline{\Gamma}_i, \quad (12)$$

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

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

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

$$\dot{\hat{W}}_{ai} = -\eta_{a1i}\left(\hat{W}_{ai} - \hat{W}_{ci}\right) - \eta_{a2i}\hat{W}_{ai} + \frac{1}{4}\sum_{j=1}^{N} \eta_{c1i}\sigma_j' G_{ij} \sigma_j'^T \hat{W}_{aj}^T \frac{\omega_i^T}{\rho_i}\hat{W}_{ci}^T + \frac{1}{4}\sum_{k=1}^{M_{xi}}\sum_{j=1}^{N} \frac{\eta_{c2i}}{M_{xi}} \sigma_j'^{ik} G_{ij}^{ik} \left(\sigma_j'^{ik}\right)^T \hat{W}_{aj}^T \frac{\left(\omega_i^k\right)^T}{\rho_i^k}\hat{W}_{ci}^T, \quad (13)$$

where $\eta_{a1i}, \eta_{a2i} \in \mathbb{R}$ are positive constant adaptation gains and $G_{\sigma i} \triangleq \sigma_i' g_i R_{ii}^{-1} g_i^T \sigma_i'^T \in \mathbb{R}^{p_{W_i} \times p_{W_i}}$. The forgetting factors $\beta_i$ along with the saturation in the update laws for the least squares gain matrices in (12) ensure (cf. [22]) that the least squares gain matrices $\Gamma_i$ and their inverses are positive definite and bounded for all $i \in \{1,..,N\}$ as

$$\underline{\Gamma}_i \leq \|\Gamma_i\| \leq \overline{\Gamma}_i, \forall t \in \mathbb{R}_{\geq 0}, \quad (14)$$

where $\underline{\Gamma}_i \in \mathbb{R}$ are positive constants, and the normalized regressors are bounded as

$$\left\|\frac{\omega_i}{\rho_i}\right\| \leq \frac{1}{2\sqrt{\nu_i \underline{\Gamma}_i}}, \forall t \in \mathbb{R}_{\geq 0}.$$

## IV. Stability analysis

Subtracting (4) from (10), the approximate BEs can be expressed in an unmeasurable form as

$$\hat{\delta}_i = -\omega_i^T \tilde{W}_{ci} + \frac{1}{4}\sum_{j=1}^{N} \tilde{W}_{aj}^T \sigma_j' G_{ij} \sigma_j'^T \tilde{W}_{aj} - \frac{1}{2}\sum_{j=1}^{N} \left(W_i^T \sigma_i' G_j - W_j^T \sigma_j' G_{ij}\right) \sigma_j'^T \tilde{W}_{aj} - \epsilon_i' f + \Delta_i, \quad (15)$$

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

$$\hat{\delta}_i^k = -\omega_i^{kT} \tilde{W}_{ci} + \frac{1}{4}\sum_{j=1}^{N} \tilde{W}_{aj}^T \sigma_j'^{ik} G_{ij}^{ik} \left(\sigma_j'^{ik}\right)^T \tilde{W}_{aj} + \Delta_i^k - \frac{1}{2}\sum_{j=1}^{N} \left(W_i^T \sigma_i'^{ik} G_j^{ik} - W_j^T \sigma_j'^{ik} G_{ij}^{ik}\right) \left(\sigma_j'^{ik}\right)^T \tilde{W}_{aj}, \quad (16)$$

where the constants $\Delta_i^k \in \mathbb{R}$ are defined as $\Delta_i^k \triangleq -\epsilon_i'^{ik} f^{ik} + \Delta_i^{ik}$. To facilitate the stability analysis, a candidate Lyapunov function is defined as

$$V_L = \sum_{i=1}^{N} V_i^* + \frac{1}{2}\sum_{i=1}^{N} \tilde{W}_{ci}^T \Gamma_i^{-1} \tilde{W}_{ci} + \frac{1}{2}\sum_{i=1}^{N} \tilde{W}_{ai}^T \tilde{W}_{ai} \quad (17)$$

Since $V_i^*$ are positive definite, the bound in (14) and Lemma 4.3 in [23] can be used to bound the candidate Lyapunov function as

$$\underline{v}\left(\|Z\|\right) \leq V_L\left(Z,t\right) \leq \overline{v}\left(\|Z\|\right), \quad (18)$$

where $Z = \left[x^T, \tilde{W}_{c1}^T, .., \tilde{W}_{cN}^T, \tilde{W}_{a1}^T, .., \tilde{W}_{aN}^T\right]^T \in \mathbb{R}^{2n+2N\sum_i p_{W_i}}$ and $\underline{v}, \overline{v} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions. For any compact set $\mathcal{Z} \subset \mathbb{R}^{2n+2N\sum_i p_{W_i}}$, define

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

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

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

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

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

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

$$v_l \triangleq \frac{1}{2}\min\left(\frac{\underline{q_i}}{2}, \frac{\eta_{c2i}\underline{c}_{xi}}{4}, \frac{2\eta_{a1i}+\eta_{a2i}}{8}\right)$$

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

$$\overline{Z} \triangleq \underline{v}^{-1}\left(\overline{v}\left(\max\left(\left\|Z\left(t_0\right)\right\|, \sqrt{\frac{\iota}{v_l}}\right)\right)\right) \quad (19)$$

where $\underline{q_i}$ denote the minimum eigenvalues of $Q_i$ and the suprema exist since $\frac{\omega_i}{\rho_i}$ are uniformly bounded for all $Z$, and the functions $G_i$, $G_{ij}$, $\sigma_i'$, and $\epsilon_i'$ are continuous. In (19), $L_f \in \mathbb{R}_{\geq 0}$ denotes a Lipschitz constant such that $\left\|f\left(\varpi\right)\right\| \leq L_f\left\|\varpi\right\|$ for all $\varpi \in \mathcal{Z}\cap\mathbb{R}^n$. The sufficient conditions for UUB convergence are derived based on the subsequent stability analysis as

$$\underline{q_i} > 2\iota_{5i},$$
$$\eta_{c2i}\underline{c}_{xi} > 2\iota_{5i} + \iota_2\zeta N + \eta_{a1i},$$
$$2\eta_{a1i} + \eta_{a2i} > 4\iota_8 + \frac{2\iota_2 N}{\zeta}, \quad (20)$$

where $\zeta \in \mathbb{R}$ is a known positive adjustable constant.

**Algorithm 1** Gain Selection

First iteration:

Given $z \in \mathbb{R}_{\geq 0}$ such that $\left\|Z\left(t_0\right)\right\| < z$, let $\mathcal{Z}_1 \triangleq \left\{\xi \in \mathbb{R}^{2n+2N\sum_i\{p_{W_i}\}_1} \mid \|\xi\| \leq \underline{v}^{-1}\left(\overline{v}\left(z\right)\right)\right\}$. Using $\mathcal{Z}_1$, compute the bounds in (19) and select the gains according to (20). If $\left\{\sqrt{\frac{\iota}{v_l}}\right\}_1 \leq z$, set $\mathcal{Z} = \mathcal{Z}_1$ and terminate.

Second iteration:

If $z < \left\{\sqrt{\frac{\iota}{v_l}}\right\}_1$, let $\mathcal{Z}_2 \triangleq \left\{\xi \in \mathbb{R}^{2n+2N\sum_i\{p_{W_i}\}_1} \mid \|\xi\| \leq \underline{v}^{-1}\left(\overline{v}\left(\left\{\sqrt{\frac{\iota}{v_l}}\right\}_1\right)\right)\right\}$. Using $\mathcal{Z}_2$, compute the bounds in (19) and select the gains according to (20). If $\left\{\sqrt{\frac{\iota}{v_l}}\right\}_2 \leq \left\{\sqrt{\frac{\iota}{v_l}}\right\}_1$, set $\mathcal{Z} = \mathcal{Z}_2$ and terminate.

Third iteration:

If $\left\{\sqrt{\frac{\iota}{v_l}}\right\}_2 > \left\{\sqrt{\frac{\iota}{v_l}}\right\}_1$, increase the number of NN neurons to $\{p_{Wi}\}_3$ to ensure $\{L_f\}_2\{\overline{\epsilon}_i'\}_3 \leq \{L_f\}_2\{\overline{\epsilon}_i'\}_2, \forall i = 1, .., N$. These adjustments ensure $\{\iota\}_3 \leq \{\iota\}_2$. Set $\mathcal{Z} = \left\{\xi \in \mathbb{R}^{2n+2N\sum_i\{p_{W_i}\}_3} \mid \|\xi\| \leq \underline{v}^{-1}\left(\overline{v}\left(\left\{\sqrt{\frac{\iota}{v_l}}\right\}_2\right)\right)\right\}$ and terminate.

Since the NN function approximation error and the Lipschitz constant $L_f$ depend on the compact set that contains the state trajectories, the compact set needs to be established before the gains can be selected to satisfy the sufficient conditions in (20). 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 $\varpi$ denotes the value of $\varpi$ computed in the $i^{th}$ iteration. Since the constants $\iota$ and $v_l$ depend on $L_f$ only through the product $L_f\overline{\epsilon}_i'$, Algorithm 1 ensures that

$$\sqrt{\frac{\iota}{v_l}} \leq \frac{1}{2}\text{diam}\left(\mathcal{Z}\right), \quad (21)$$

where diam$(\mathcal{Z}) \triangleq \sup\left\{\|y_1 - y_2\| \mid y_1, y_2 \in \mathcal{Z}\right\}$.

**Theorem 1.** *Provided Assumption 1 holds and the control gains satisfy the sufficient conditions in (20), where the constants in (19) are computed based on a compact set $\mathcal{Z}$ selected using Algorithm 1, the controllers in (9) along with the adaptive update laws in (12) and (13) ensure that the state $x$, the value function weight estimation errors $\tilde{W}_{ci}$ and the policy weight estimation errors $\tilde{W}_{ai}$ are UUB, resulting in convergence of the policies $\hat{u}_i$ to neighborhoods of the feedback-Nash equilibrium policies $u_i^*$.*

*Proof:* The derivative of the candidate Lyapunov function in (17) along the trajectories of (1), (12), and (13) is given by

$$\dot{V}_L = \sum_{i=1}^{N}\left(\nabla_x V_i^*\left(f + \sum_{j=1}^{N} g_j\hat{u}_j\right)\right)$$

$$
\begin{aligned}
&+\sum_{i=1}^{N}\tilde{W}_{ci}^{T}\left(\frac{\eta_{c1i}\omega_i}{\rho_i}\hat{\delta}_i+\frac{\eta_{c2i}}{M_{xi}}\sum_{i=1}^{M_{xi}}\frac{\omega_i^k}{\rho_i^k}\hat{\delta}_i^k\right)\\
&-\frac{1}{2}\sum_{i=1}^{N}\tilde{W}_{ci}^{T}\left(\beta_i\Gamma_i^{-1}-\eta_{c1i}\frac{\omega_i\omega_i^T}{\rho_i^2}\right)\tilde{W}_{ci}\\
&-\sum_{i=1}^{N}\tilde{W}_{ai}^{T}\Bigg(-\eta_{a1i}\left(\hat{W}_{ai}^{T}-\hat{W}_{ci}^{T}\right)-\eta_{a2i}\hat{W}_{ai}^{T}\\
&+\frac{1}{4}\sum_{j=1}^{N}\eta_{c1i}\hat{W}_{ci}^{T}\frac{\omega_i}{\rho_i}\hat{W}_{aj}^{T}\sigma_j'G_{ij}\sigma_j'^T\\
&+\frac{1}{4}\sum_{k=1}^{M_{xi}}\sum_{j=1}^{N}\frac{\eta_{c2i}}{M_{xi}}\hat{W}_{ci}^{T}\frac{\omega_i^k}{\rho_i^k}\hat{W}_{aj}^{T}\sigma_j'^{ik}G_{ij}^{ik}\left(\sigma_j'^{ik}\right)^T\Bigg). \qquad (22)
\end{aligned}
$$

Substituting the unmeasurable forms of the BEs from (15) and (16) into (22) and using the triangle inequality, completion of squares, Cauchy-Schwarz inequality, Young's inequality, and provided the sufficient conditions in (20) hold, the Lyapunov derivative in (22) can be bounded as

$$
\begin{aligned}
\dot{V} \leq &-\sum_{i=1}^{N}\frac{q_i}{2}\left\|x\right\|^2-\sum_{i=1}^{N}\frac{\eta_{c2i}\underline{c}_{xi}}{4}\left\|\tilde{W}_{ci}\right\|^2\\
&-\sum_{i=1}^{N}\left(\frac{2\eta_{a1i}+\eta_{a2i}}{8}\right)\left\|\tilde{W}_{ai}\right\|^2+\iota,\\
&< -v_l\left\|Z\right\|^2,\quad \forall\left\|Z\right\|>\sqrt{\frac{\iota}{v_l}}. \qquad (23)
\end{aligned}
$$

Using (18), (21), and (23), Theorem 4.18 in [23] can be invoked to conclude that $\limsup_{t\to\infty}\left\|Z\left(t\right)\right\| \leq \underline{v}^{-1}\left(\overline{v}\left(\sqrt{\frac{\iota}{v_l}}\right)\right)$. Furthermore, the system trajectories are bounded as $\left\|Z\left(t\right)\right\|\leq\overline{Z}$ for all $t\in\mathbb{R}_{\geq0}$.

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

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

for all $i=1,..,N$, where $\overline{g_i}\triangleq\sup_x\left\|g_i\left(x\right)\right\|$. 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 (17) is a switched system. The switching happens when the least squares regression matrices $\Gamma_i$ reach their saturation bound. Similar to least squares-based adaptive control (cf. [22]), (17) can be shown to be a common Lyapunov function for the regression matrix saturation. Since (17) is a common Lyapunov function, (18), (21), and (23) establish UUB convergence of the switched system.

## V. Simulation

### *A. Problem setup*

To illustrate the performance of the developed approach, the concurrent learning-based adaptive technique is applied to the 2-player non-linear control-affine system described by (1), where $x\in\mathbb{R}^2$, $u_1,u_2\in\mathbb{R}$, and [14]

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

$$
g_1=\left[\begin{array}{c}0\\ \cos\left(2x_1\right)+2\end{array}\right],\ g_2=\left[\begin{array}{c}0\\ \sin\left(4x_1^2\right)+2\end{array}\right]. \qquad (24)
$$

The value function has the structure shown in (2) with the weights $Q_1=2Q_2=2I_2$ and $R_{11}=R_{12}=2R_{21}=2R_{22}=2I_2$, where $I_2$ is a $2\times2$ identity matrix. The concurrent learning-based scheme given in Section III-A is implemented to provide an approximate on-line feedback-Nash equilibrium solution to the given nonzero-sum two-player game.

### *B. Analytical solution*

The control affine system in (24) is selected for this simulation because it is constructed using the converse HJ approach in [24], such that the analytical feedback-Nash equilibrium solution of the nonzero-sum game is $W_1=\left(0.5,\ 0,\ 1\right)$ and $W_2=\left(0.25,\ 0,\ 0.5\right)$. 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*

Based on the structure of the feedback-Nash equilibrium value functions, the basis function for value function approximation is selected as $\sigma=[x_1^2,x_1x_2,x_2^2]^T$, and the adaptive learning parameters and initial conditions are shown for both players in Table I. Twenty-five points lying on a $5\times5$ grid around the origin are selected for the concurrent learning-based update laws in (12) and (13).

| | Player 1 | Player 2 |
|---|---|---|
| $\nu$ | 0.005 | 0.005 |
| $\eta_{c1}$ | 1 | 1 |
| $\eta_{c2}$ | 1.5 | 1 |
| $\eta_{a1}$ | 10 | 10 |
| $\eta_{a2}$ | 0.1 | 0.1 |
| $\beta$ | 3 | 3 |
| $\overline{\Gamma}$ | 10,000 | 10,000 |

(a) Gains

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

(b) Initial Conditions

Table I: ADP parameters

### *D. Simulation results*

Figure 1 shows rapid convergence of the actor and critic weight estimates (represented by solid lines) to the known feedback-Nash equilibrium values (represented by dashed lines) for both players. Figure 2 shows regulation of the state trajectories to the origin and the convergence of the approximate control trajectories to the feedback-Nash equilibrium control trajectories.

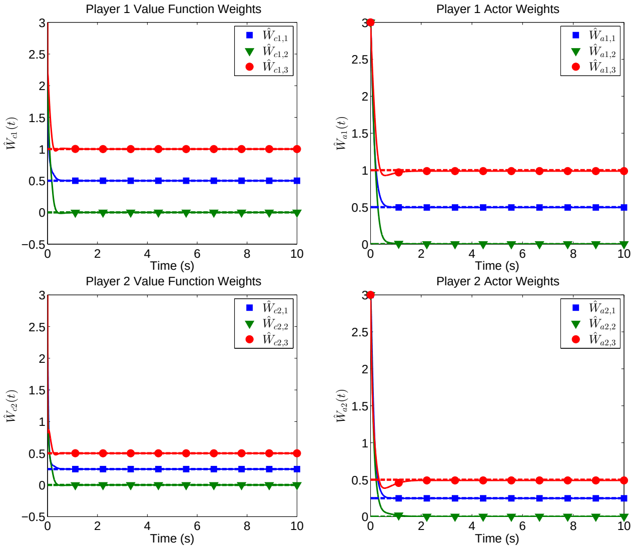

Figure 1: Convergence of policy and value function weights

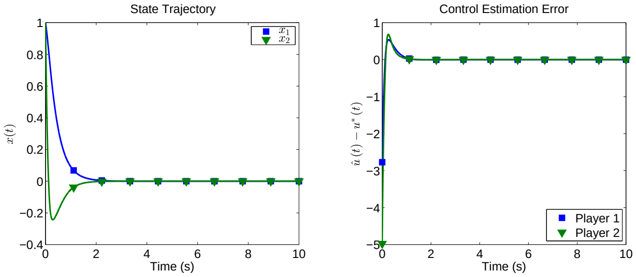

Figure 2: Regulation of the system states and the error between the approximate control and the feedback-Nash equilibrium control

## VI. Conclusion

A concurrent learning-based adaptive approach is developed to determine the feedback-Nash equilibrium solution to an $N$-player nonzero-sum game on-line. The solutions to the associated coupled HJ equations and the corresponding feedback-Nash equilibrium policies are approximated using parametric universal function approximators. Based on the system dynamics, the Bellman errors are evaluated at a set of preselected 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 the preselected points.

Unlike traditional approaches that require a PE condition for convergence, UUB convergence of the value function and policy weights to their true values, and hence, UUB convergence of the policies to the feedback-Nash equilibrium policies, is established under 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 regressor matrix. While this condition can be heuristically satisfied by choosing enough points, and can be easily verified on-line, it can not, 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 investigating the aforementioned challenges.

## 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] J. Nash, "Non-cooperative games," *Annals of Math.*, vol. 2, pp. 286–295, 1951.

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

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

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] D. Vrabie and F. Lewis, "Integral reinforcement learning for online computation of feedback nash strategies of nonzero-sum differential games," in *Proc. IEEE Conf. Decis. Control*, 2010, pp. 3066–3071.

[14] K. Vamvoudakis and F. Lewis, "Multi-player non-zero-sum games: Online adaptive learning solution of coupled hamilton-jacobi equations," *Automatica*, vol. 47, pp. 1556–1569, 2011.

[15] 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.

[16] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Proc. IEEE Conf. Decis. Control*, Dec. 2009, pp. 3598 –3605.

[17] D. Vrabie, M. Abu-Khalaf, F. Lewis, and Y. Wang, "Continuous-time ADP for linear systems with partially unknown dynamics," in *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinf. Learn.*, 2007, pp. 247–253.

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

[19] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 4, pp. 280–301, 2013.

[20] G. V. Chowdhary and E. N. Johnson, "Theory and flight-test validation of a concurrent-learning adaptive controller," *J. Guid. Contr. Dynam.*, vol. 34, no. 2, pp. 592–607, March 2011.

[21] 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.

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

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

[24] 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.