Concurrent Learning-Based Network Synchronization

J. Klotz, R. Kamalapurkar, and W. E. Dixon

***Abstract*—A data-driven concurrent learning-based control law is developed for the synchronization of a leader-follower network of agents with uncertain nonlinear dynamics wherein only a subset of the follower agents is connected to the leader. The development is facilitated by the use of online data-driven adaptive update policies to approximately learn a distributed control law which satisfies a given performance metric without the need for persistence of excitation (PE). A neighbor-decoupled control structure is introduced which provides greater flexibility in the consideration of individual neighbors during synchronization and makes the control of each agent a differential game.**

I. Introduction

A network of cooperating agents can provide a more reliable and capable strategy of accomplishing objectives; e.g., multi-agent systems such as teams of robotic systems and distributed sensors benefit from the ability to interact with or sense the environment in a collaborative manner simultaneously. Synchronization of a distributed leader-follower network is achieved by cooperatively driving the states of follower agents (nodes) to the states of neighboring follower agents and the leader using local interactions in the network. Specifying desired network behavior by assigning a distributed performance metric (e.g. quadratic cost function) to each agent in the network allows for the development of distributed controllers which accomplish the desired global network goals [1]. However, because the performance of an agent depends upon the behavior of network neighbors, an agent's method for attaining desirable performance is inherently coupled with that of its neighbors, making it difficult to ascertain minimizing control policies. To address this challenge, the interaction of the networked agents is modeled with graphical differential game theory, wherein each agent acts as a player in the network-wide game. This approach provides coupled Hamilton-Jacobi (HJ) equations for which the solution results in the distributed control laws which minimize the given cost functions.

In general, the analytical solution of the coupled HJ equations for agents with nonlinear dynamics is infeasible. To this end, adaptive dynamic programming (ADP) techniques have been employed to approximate solutions for the coupled HJ equations [2]–[6]. The results in [7] and [8], among others, demonstrate that recurrent learning-based ADP techniques can be employed to approximately determine the solutions of the coupled HJ equations. The works in [7] and [8], similar to the work in [9]–[11], guarantee approximate solutions to the HJ equations and uniformly ultimately bounded (UUB) stability of the networked systems under the assumption of persistence of excitation (PE), which is a condition that asserts there is sufficient volatility in the dynamical system for all time. PE, which in general is not verifiable online, helps provide data richness for adaptive algorithms and is often provided by injecting ad hoc exploratory signals to the agents' control signals. However, augmenting a control signal with an exciting signal affects the system states in an (often unknown) manner and can alter the intended controller performance. Moreover, the effects of including the exciting signal in the control input is generally not considered in stability analyses, leaving the controller examination incomplete.

A concurrent learning-based (cf. [12], [13]) approach can avoid the PE assumption by using recorded data, in addition to instantaneously available data, to guarantee data richness in adaptive algorithms. As recently shown in [14], online data-driven concurrent learning-based techniques are capable of approximately solving the optimal policy of a single dynamical system using an assumption weaker than the PE condition; to guarantee data richness, the concurrent learning-based approach in [14] assumes there exists a set of preselected state space points which, when evaluated with the update regressor, provide full rank to the enacted data-driven adaptive update laws. Thus, full rank of the adaptive update laws is garnered by evaluating the update regressor over an advantageous set of points instead of only along the trajectory of the dynamical system.

This research presents a data-driven concurrent learning-based ADP method of synchronizing a leader-follower network of agents which have uncertain nonlinear dynamics wherein only a subset of the follower agents is connected to the leader. The control methodology is structured as an infinite horizon nonzero-sum graphical game wherein each agent computes a control policy online which minimizes a given performance metric. Motivated by the work in [2], [15], [16], the control of each agent is treated as a subgame in which there is a player assigned to each neighbor that minimizes a performance metric respectively associated with its assigned neighbor. This neighbor-decoupled strategy allows

J. Klotz, R. Kamalapurkar, and W. E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville FL 32611-6250 USA. Email: {jklotz, rkamalapurkar, wdixon}@ufl.edu.

This research is supported in part by NSF award numbers 0901491, 1161260, 1217908, ONR grant number N00014-13-1-0151, and 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.

for greater flexibility in how each neighbor is considered during the formulation of a synchronizing control policy.

II. Problem Definition

A. Graph Theory Preliminaries

Consider an undirected graph, $\mathcal{G} = \{\mathcal{V}, E\}$, which describes the communication topology of a set of N networked agents $\mathcal{V} = \{1, 2, \ldots, N\}$ with communication links $E \subseteq \mathcal{V} \times \mathcal{V}$. If agents $i \in \mathcal{V}$ and $j \in \mathcal{V}$ communicate with each other, the pairs (i, j), (j, i) are defined such that $(i, j), (j, i) \in E$. It is assumed that the network contains no self-loops, i.e., $(i, i) \notin E$. The neighborhood of an agent i is the set of all networked agents which communicate with i and is defined as $\mathcal{N} \triangleq \{j \in \mathcal{V} \mid (i, j) \in E\}$. Let $a_{ij} \in \mathbb{R}$ be a constant link weight such that $a_{ij} = a_{ji} > 0$ if $(i, j) \in E$ and $a_{ij} = a_{ji} = 0$ otherwise. The graph adjacency matrix $A \in \mathbb{R}^{N \times N}$ is defined as $A \triangleq [a_{ij} \mid i, j = 1, \ldots, N]$.

Let the graph $\bar{\mathcal{G}} = \left\{\bar{\mathcal{V}}, \bar{E}\right\}$ represent $\mathcal{G}$ augmented by a network leader, denoted by $\{0\}$, such that $\bar{\mathcal{V}} = \mathcal{V} \cup \{0\}$. The link set $\bar{E}$ is constructed such that $E \subset \bar{E}$ and $(i, 0) \in \bar{E}$ if agent i communicates with the leader. Using $\bar{E}$, the neighborhood $\bar{\mathcal{N}}_i$ is defined as $\bar{\mathcal{N}}_i \triangleq \left\{j \in \bar{\mathcal{V}} \mid (i, j) \in \bar{E}\right\}$. A pinning gain matrix $A_0 \in \mathbb{R}^{N \times N}$ is defined as $A_0 \triangleq \operatorname{diag}(a_{i0}) \mid i = 1, \ldots, N$, where the weight $a_{i0} \in \mathbb{R}$ is defined such that $a_{i0} > 0$ if $(i, 0) \in \bar{E}$ and $a_{i0} = 0$ otherwise.

B. Network Objectives

Let each agent $i \in \mathcal{V}$ have dynamics of the form

$$\dot{x}_i = f_i(x_i) + g_i u_i, \tag{1}$$

where $x_i \in \mathcal{S}$ is the measurable state, $f_i : \mathcal{S} \to \mathbb{R}^n$ represents uncertain continuous drift dynamics, $g_i \in \mathbb{R}^{n \times m}$ is a known, constant control-effectiveness matrix, $u_i \in \mathbb{R}^m$ is the control input, and the set $\mathcal{S} \subset \mathbb{R}^n$ is the state space.

Agents in the network seek to cooperatively synchronize in state towards the leader state, which is selected to be the origin ($x_o = 0 \in \mathcal{S}$) for simplicity of exposition. The synchronizing control input for each agent must satisfy a desired behavior specified by a quadratic cost function.

C. General Approach

This approach considers a class of nonlinear systems specified by the following assumption.

Assumption 1. The continuous drift dynamics, f_i, $i = 1, \ldots, N$, satisfies $f_i(0) = 0$ and contains unknown constants that are linear-in-the-parameters (LP).

To quantify an agent's progress towards synchronization, a neighborhood tracking error $e_i \in S$ is designed for agent $i \in \mathcal{V}$ as [17]

$$e_i \triangleq \sum_{j \in \bar{\mathcal{N}}_i} a_{ij} (x_i - x_j). \tag{2}$$

To facilitate the development of a control policy and to allow for greater flexibility in how individual neighbors are treated during synchronization, a partial neighborhood tracking error $e_{ij} \in S$ is defined as $e_{ij} \triangleq a_{ij}(x_i - x_j)$ and the control policy u_i in (1) is structured as an $|\mathcal{N}_i|$ player game such that

$$u_i = \sum_{j \in \bar{\mathcal{N}}_i} u_{ij}, \tag{3}$$

where $u_{ij} \in \mathbb{R}^m$ is a player and $|\cdot|$ denotes set cardinality for a set argument. Note that, as opposed to results such as [7], [8], [17] wherein the control signal evaluates the summation in (2) wholly, the controller in (3) is structured such that the relative state difference between agents is considered individually. To this end, each player minimizes the cost

$$J_{ij} \triangleq \frac{1}{2} \int_0^\infty \mathcal{L}_{ij}(e_{ij}, u_{ij})\, dt, \tag{4}$$

where $\mathcal{L}_{ij}(e_{ij}, u_{ij}) \triangleq e_{ij}^T Q_{ij} e_{ij} + u_{ij}^T R_i u_{ij}$ and $Q_{ij} \in \mathbb{R}^{n \times n}$, $R_i \in \mathbb{R}^{m \times m}$ are user-specified positive definite, constant, symmetric weighting matrices which allow for customization of the desired performance.

III. System Identification

To approximate the drift dynamics for use in the subsequent control development, each agent i approximates its own drift dynamics, f_i, using a concurrent learning-based update scheme.

Let $f_i(x_i) = Y_i(x_i)\theta_i^*$ represent the LP form of the drift dynamics of agent i, where $Y_i : \mathbb{R}^n \to \mathbb{R}^{n \times p_i}$ is a known regression matrix and $\theta_i^* \in \mathbb{R}^{p_i}$ represents a constant vector of unknown parameters. The function $\hat{f}_i : \mathbb{R}^n \times \mathbb{R}^{p_i} \to \mathbb{R}^n$ is used as an estimate of the unknown function f_i and is defined as $\hat{f}_i\left(x_i, \hat{\theta}_i\right) \triangleq Y_i(x_i)\hat{\theta}_i$, where $\hat{\theta}_i \in \mathbb{R}^{p_i}$ is an estimate of the unknown vector θ_i^*.

The drift dynamics are estimated by constructing the identifier

$$\dot{\hat{x}}_i = \hat{f}_i + g_i u_i + k_{\tilde{x}i}\tilde{x}_i, \tag{5}$$

where $\tilde{x}_i$ is the state estimation error defined as $\tilde{x}_i \triangleq x_i - \hat{x}_i$ and $k_{\tilde{x}i} \in \mathbb{R}^{n \times n}$ is a constant positive definite diagonal gain matrix. The identification error dynamics can be obtained using (1) and (5) as

$$\dot{\tilde{x}}_i = Y_i\tilde{\theta}_i - k_{\tilde{x}i}\tilde{x}_i, \tag{6}$$

where $\tilde{\theta}_i \triangleq \theta_i^* - \hat{\theta}_i$.

A. Estimate update policy

A concurrent learning-based approach is used to update the estimate $\hat{\theta}_i$. The following assumption specifies the observability condition used to guarantee parameter identification.

Assumption 2. [12], [13] A finite set of time instances $\{t_l \mid l = 1, \ldots, L_1\}$ exists for each agent i such that

$$\operatorname{rank}\left(\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l\right) = p_i, \tag{7}$$

where $Y_i^l \triangleq Y(x_i(t_l))$.

A data stack is maintained by recording the states $\left\{x_i^l \triangleq x_i\left(t_l\right) \mid l = 1, \ldots, L_1\right\}$ and corresponding control values $\left\{u_i^l \triangleq u_i\left(t_l\right) \mid l = 1, \ldots, L_1\right\}$; the stack updating is governed by using a singular value maximizing algorithm (cf. [13]).

The PE condition assumes enough excitation is present in the estimated system for all time and is used prevalently to guarantee parameter convergence in estimation and adaptive control; however, the condition in (7) only assumes that the system states are exciting over a finite period of time. Furthermore, satisfaction of the condition in (7) can be verified online. Note that the satisfaction of the condition in (7) can be more easily accomplished by gathering more data than necessary, i.e., $L_1 \gg p_i$.

The update law for $\hat{\theta}_i$ is designed as

$$\dot{\hat{\theta}}_i = \Gamma_{\hat{\theta}i} Y_i^T \tilde{x}_i + \Gamma_{\hat{\theta}i} k_{\hat{\theta}i} \sum_{l=1}^{L_1} \left(Y_i^l\right)^T \left(\dot{x}_i^l - g_i u_i^l - Y_i^l \hat{\theta}_i\right), \tag{8}$$

where $\Gamma_{\hat{\theta}i} \in \mathbb{R}^{p_i \times p_i}$ is a constant positive definite gain matrix and $k_{\hat{\theta}i} \in \mathbb{R}_{>0}$ is a constant concurrent learning gain. Note that (8) depends on the unknown past state derivative $\dot{x}_i^l = \dot{x}_i\left(t_l\right)$; because the value $\dot{x}_i^l$ is a past value, it can be numerically calculated using preceding and proceeding recorded state information using numerical smoothing techniques. This online method of computing the state derivative facilitates facilitates the online implementation of (8).

Using (1) and the definitions of $\hat{f}_i$ and $\tilde{\theta}_i$, the update law in (8) can also be expressed as

$$\dot{\hat{\theta}}_i = \Gamma_{\hat{\theta}i} Y_i^T \tilde{x}_i + \Gamma_{\hat{\theta}i} k_{\hat{\theta}i} \left(\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l\right) \tilde{\theta}_i. \tag{9}$$

B. Estimate convergence

Let $V_{I,i} : \mathbb{R}^{n+p_i} \to \mathbb{R}_{\geq 0}$ be a positive definite continuously differentiable Lyapunov function defined as

$$V_{I,i} \triangleq \frac{1}{2} \tilde{x}_i^T \tilde{x}_i + \frac{1}{2} \tilde{\theta}_i^T \Gamma_{\hat{\theta}i}^{-1} \tilde{\theta}_i. \tag{10}$$

The positive definite Lyapunov function in (10) satisfies the inequalities

$$\underline{V}_{I,i} \left\|z_i\right\|^2 \leq V_{I,i} \leq \bar{V}_{I,i} \left\|z_i\right\|^2, \tag{11}$$

where $\underline{V}_{I,i} \triangleq \frac{1}{2} \min\left(1, \lambda_{\min}\left(\Gamma_{\hat{\theta}i}\right)\right)$ and $\bar{V}_{I,i} \triangleq \frac{1}{2} \max\left(1, \lambda_{\max}\left(\Gamma_{\hat{\theta}i}\right)\right)$ are positive known constants, $z_i \triangleq \left[\tilde{x}_i^T, \tilde{\theta}_i^T\right]^T$, and $\lambda_{\min}\left(\cdot\right)$ and $\lambda_{\max}\left(\cdot\right)$ denote the minimum and maximum eigenvalues, respectively. Using (6) and (9), the derivative of (10) can be expressed as

$$\dot{V}_{I,i} = -\tilde{x}_i^T k_{\tilde{x}i} \tilde{x}_i - \tilde{\theta}_i^T k_{\hat{\theta}i} \left(\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l\right) \tilde{\theta}_i. \tag{12}$$

Provided the identifier condition in Assumption 2 is satisfied, the matrix $\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l$ is positive definite. Using this fact and the inequalities in (11), (12) can be bounded as

$$\dot{V}_{I,i} \leq -c_{I,i} \left\|z_i\right\|^2 \leq -\frac{c_{I,i}}{\bar{V}_{I,i}} V_{I,i}, \tag{13}$$

where $c_{I,i} \triangleq \min\left(\lambda_{\min}\left(k_{\tilde{x}i}\right), k_{\hat{\theta}i} \underline{Y}_i\right)$, where $\underline{Y}_i \triangleq \lambda_{\min}\left(\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l\right) > 0$. The inequalities in (11) and (13) can be used to show that $\left\|\tilde{x}_i\left(t\right)\right\|, \left\|\tilde{\theta}_i\left(t\right)\right\| \to 0$ exponentially fast. Furthermore, $Y_i\left(t\right) \in \mathcal{L}_\infty$ if $x_i\left(t\right) \in \mathcal{L}_\infty$.[1] Thus, it follows from (6) that $\left\|\dot{\tilde{x}}_i\left(t\right)\right\| \to 0$ exponentially fast. In the presence of state derivative estimation errors, the parameter estimation error $\tilde{\theta}_i$ can be shown to be uniformly ultimately bounded (UUB), where the magnitude of the ultimate bound depends upon the derivative estimate error [13].

Because the data record $\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l$ is updated during estimation, the dynamical system represented by (6) and (9) is a switched system. Let $T_L \in \mathbb{R}_{>0}$ be the elapsed time required to collect sufficient data to satisfy the condition in (7). Assuming that the system states are sufficiently exciting over the time interval $[0, T]$ and the recorded data $\sum_{l=1}^{L_1} \left(Y_i^l\right)^T Y_i^l$ are collected using a singular value maximizing algorithm, (13) ensures that (10) is a common Lyapunov function for the switched system of (6) and (9) (cf. [13]).

IV. Controller Development

Throughout the control development, let the subscript j denote a neighbor of agent i such that $j \in \bar{\mathcal{N}}_i$.

A. Desired control policy

Using the principle of optimality, the aforementioned cost function can be minimized by minimizing the value function $V_{ij} : \mathcal{S} \to \mathbb{R}_{\geq 0}$, which is defined as

$$V_{ij} \triangleq \frac{1}{2} \int_t^\infty \mathcal{L}_{ij}\left(e_{ij}\left(\tau\right), u_{ij}\left(\tau\right)\right) d\tau, \tag{14}$$

where t_o denotes the initial time. The optimal value function $V_{ij}^* : \mathcal{S} \to \mathbb{R}_{\geq 0}$, minimized by the control contribution of player u_{ij} is

$$V_{ij}^* = \min_{\substack{u_{ij}:\mathcal{S} \to \mathbb{R}^m \\ u_{ij} \in U_{ij}}} \frac{1}{2} \int_t^\infty \mathcal{L}_{ij}\left(e_{ij}\left(\tau\right), u_{ij}\left(\tau\right)\right) d\tau, \tag{15}$$

where U_{ij} is the set of admissible control policies for player u_{ij} [18].

Because the minimization of each value function V_{ij} is inherently coupled with the minimization of other value functions through mutual dependence on neighbor's states, an individual value function depends on all tracking errors in the undirected network. The coupled Hamilton-Jacobi equation may be constructed as

$$\begin{aligned} H_{ij}^* = \frac{1}{2} \mathcal{L}_{ij}^* + \sum_{(k,l) \in \bar{E}} \frac{\partial V_{ij}^*}{\partial e_{kl}} a_{kl} \Big(& f_k + g_k u_k^* \\ & - \mathbf{1}_{\{l \neq 0\}} \left(f_l + g_l u_l^*\right) \Big) = 0, \end{aligned} \tag{16}$$

where $\mathcal{L}_{ij}^* \triangleq \mathcal{L}_{ij}\left(e_{ij}, u_{ij}^*\right)$, u_{ij}^* is the admissible minimizer in (15), $\dot{x}_i^* \triangleq f_i\left(x_i\right) + g_i u_i^*$ if $i \in \mathcal{V}$ and $\dot{x}_0^* = 0$, and $u_i^* =$

[1] It is shown in the following stability analysis that $x_i\left(t\right) \in \mathcal{L}_\infty \, \forall i \in \mathcal{V}$.

$\sum_{j\in\bar{\mathcal{N}}_i} u^*_{ij}$. The closed-form solution of u^*_{ij} is obtained from (16) by solving $\frac{\partial H^*_{ij}\left(e_{ij},u^*_{ij}\right)}{u^*_{ij}} = 0$. Provided the continuously differentiable solution for V^*_{ij} in (16) exists, the stabilizing desired control policy may be expressed as[2]

$$u^*_{ij} = -R_i^{-1} g_i^T \left(\sum_{k\in\bar{\mathcal{N}}_i} a_{ik} \left(\frac{\partial V^*_{ij}}{\partial e_{ik}} \right)^T - \sum_{k\in\mathcal{N}_i} a_{ki} \left(\frac{\partial V^*_{ij}}{\partial e_{ki}} \right)^T \right). \tag{17}$$

B. Desired policy approximation

In general, the expression in (17) cannot be solved analytically. For implementation purposes, an approximation protocol for the desired policy for each agent is detailed in this section. An ADP-based approach is used to develop a two-tier architecture for approximately learning both the desired value function and policy online simultaneously. Since this approach uses neural networks for learning, the following assumption is necessary for function approximation.

Assumption 3. The set $\mathcal{S}$ is compact.

Note that the set $\mathcal{S}$ is compact if all initial conditions $x_i(t_0)$ are bounded (see [8, Remark 1] for details). For notational brevity, let the vector $\Xi \in \mathbb{R}^{n|\bar{E}|}$ be a composite column containing each error signal e_{ij}, $(i,j) \in \bar{E}$.

Using the universal function approximation property of NNs, a NN containing w_{ij} neurons can be used to equivalently represent the desired value function V^*_{ij} as

$$V^*_{ij}(\Xi) = W^T_{ij} \sigma_{ij}(\Xi) + \epsilon_{ij}(\Xi), \tag{18}$$

with arbitrarily small ϵ_{ij}, where $W_{ij} \in \mathbb{R}^{w_{ij}}$ is an unknown ideal NN weight matrix bounded above by a known constant $\bar{W}_{ij} \in \mathbb{R}_{>0}$ as $\|W_{ij}\| \leq \bar{W}_{ij}$, $\sigma_{ij} : \mathcal{S} \to \mathbb{R}^{w_{ij}}$ is a selected nonlinear, continuously differentiable, bounded activation function such that $\sigma_{ij}(0) = 0$ and $\frac{\partial \sigma_{ij}}{\partial \Xi}(0) = 0$, and $\epsilon_{ij} : \mathcal{S} \to \mathbb{R}$ is an unknown function reconstruction error. The reconstruction error satisfies $\sup_{\varrho\in\mathcal{S}} |\epsilon_{ij}(\varrho)| \leq \bar{\epsilon}_{ij}$ and $\sup_{\varrho\in\mathcal{S}} \left| \frac{\partial \epsilon_{ij}}{\partial \Xi} |_{\varrho} \right| \leq \bar{\epsilon}'_{ij}$, where $\bar{\epsilon}_{ij}$, $\bar{\epsilon}'_{ij} \in \mathbb{R}_{>0}$ are constant upper bounds.

Let $\sigma'_{ijkl} \triangleq \frac{\partial \sigma_{ij}}{\partial e_{kl}}$ and $\epsilon'_{ijkl} \triangleq \frac{\partial \epsilon_{ij}}{\partial e_{kl}}$. Using (18), u^*_{ij} may be alternatively expressed as

$$u^*_{ij} = -R_i^{-1} g_i^T \left(\left(\sum_{k\in\bar{\mathcal{N}}_i} a_{ik} \sigma'^T_{ijik} - \sum_{k\in\mathcal{N}_i} a_{ki} \sigma'^T_{ijki} \right) W_{ij} + \sum_{k\in\bar{\mathcal{N}}_i} a_{ik} \epsilon'^T_{ijik} - \sum_{k\in\mathcal{N}_i} a_{ki} \epsilon'^T_{ijki} \right). \tag{19}$$

Because the function reconstruction error ϵ_{ij} is unknown, the desired value function V^*_{ij} and control policy u^*_{ij} are approximated as

$$\hat{V}_{ij} = \hat{W}^T_{cij} \sigma_{ij},$$

$$u_{ij} = -R_i^{-1} g_i^T \left(\sum_{k\in\bar{\mathcal{N}}_i} a_{ik} \sigma'^T_{ijik} - \sum_{k\in\mathcal{N}_i} a_{ki} \sigma'^T_{ijki} \right) \hat{W}_{aij}, \tag{20}$$

where $\hat{W}_{cij}(t)$, $\hat{W}_{aij}(t) \in \mathbb{R}^{w_{ij}}$ are the "critic" and "actor" estimates, respectively, of the ideal NN weight W_{ij}.

The Bellman error (BE) for an agent i with respect to a neighbor j, $\delta_{ij}(\cdot) \in \mathbb{R}$, constitutes a metric for "closeness" to optimality and is defined as the difference between the HJ and the approximation of the HJ. The BE can be represented in unmeasurable form as $\delta_{ij} \triangleq \hat{H}_{ij} - H^*_{ij}$; however, because $H^*_{ij} = 0$, the BE is represented in a measurable form as

$$\delta_{ij} = \hat{H}_{ij}, \tag{21}$$

where $\hat{H}_{ij}$ is defined as

$$\hat{H}_{ij} \triangleq \frac{1}{2} e^T_{ij} Q_{ij} e_{ij} + \frac{1}{2} u^T_{ij} R_i u_{ij} + \hat{W}^T_{cij} \chi_{ij}, \tag{22}$$

and the regressor vector $\chi_{ij} \in \mathbb{R}^{w_{ij}}$ is defined as

$$\chi_{ij} \triangleq \sum_{(k,l)\in\bar{E}} a_{kl} \frac{\partial \sigma_{ij}}{\partial e_{kl}} \left(\hat{f}_k + g_k u_k - \mathbf{1}_{\{l\neq 0\}} \left(\hat{f}_l + g_l u_l \right) \right). \tag{23}$$

C. NN update policy

The solution of the policy in (19) may be approximated by designing NN weight update policies $\hat{W}_{cij}$ and $\hat{W}_{aij}$ which use the BE as feedback to minimize $\int_0^\infty (\delta_{ij}(\tau))\, d\tau$. Online NN learning, i.e., the minimization of $\int_0^\infty (\delta_{ij}(\tau))\, d\tau$, is often guaranteed by augmenting a control signal with an exploratory signal to ensure PE, i.e., volatility of the update regressor for all time (cf. [6], [11], [19]). As NN learning is being performed, the resulting NN weights are used in the controller to drive the states to accomplish an objective (e.g., regulation to the origin); this seemingly counter-intuitive notion of having both volatility and control over the states provides motivation to take advantage of a concurrent learning-based approach which assumes state excitation over only a finite interval of time.

Contrary to PE-based approaches and similar to [14], this result ensures data richness by using estimated system dynamics to approximately evaluate the BE at any desired point in the state space. Let $s_{ij} \triangleq \left\{ \varrho_l \in \mathcal{S} \mid l = 1, \ldots, \left|\bar{\mathcal{V}}\right| \right\}$ be a set of sampling state space points and let $S_{ij} \triangleq \left\{ s^{c_l}_{ij} \mid c_l = 1, \ldots, L_2 \right\}$ be a collection of these sets. Similar to (7), the following condition provides data richness for use of the BE evaluated at a prespecified set of state space points.

Assumption 4. For each $(i,j) \in \bar{E}$, there exists a set of state space points S_{ij} such that $\forall t \in \mathbb{R}_{\geq 0}$,

$$\mu_{ij} \triangleq \frac{1}{L_2} \inf_{t\in\mathbb{R}_{\geq 0}} \lambda_{\min} \left(\sum_{c_l=1}^{L_2} \frac{\chi^{c_l}_{ij} \left(\chi^{c_l}_{ij}\right)^T}{\gamma^{c_l}_{ij}} \right) > 0, \tag{24}$$

[2] To verify that the control policy u^*_{ij} is indeed stabilizing, a Lyapunov stability analysis may be performed using the Lyapunov equation $V_{L*} = \sum_{i\in\mathcal{V}} \sum_{j\in\bar{\mathcal{N}}_i} V^*_{ij}$ and the property $\sum_{(k,l)\in\bar{E}} \frac{\partial V^*_{ij}}{\partial e_{kl}} a_{kl} \left(f_k + g_k u_k - \mathbf{1}_{l\neq 0} (f_l + g_l u_l) \right) = -\frac{1}{2} \mathcal{L}^*_{ij}$.

where $(\cdot)_{ij}^{c_l}$ denotes evaluation at the sample state-space points s_{ij}, $\gamma_{ij} \triangleq 1 + \lambda_{ij} \left(\chi_{ij}\right)^T \Gamma_{ij} \chi_{ij} \in \mathbb{R}$ provides normalization, $\lambda_{ij} \in \mathbb{R}$ is a positive constant normalization gain, $\Gamma_{ij} \in \mathbb{R}^{w_{ij} \times w_{ij}}$ is a positive definite least-squares gain matrix, and $\chi_{ij}^{c_l} \triangleq \chi_{ij}\left(s_{ij}^{c_l}, \omega_a, \hat{\theta}_i, \hat{\theta}_j\right)$, where $\omega_a \triangleq \left\{\hat{W}_{aij} \mid j \in \bar{\mathcal{V}}\right\}$. In general, it is not feasible to guarantee (24) a priori. However, similar to the condition in (7), this approach benefits from the capability of accommodating this restriction by storing more information than theoretically necessary; i.e., the condition in (24) may be satisfied by collecting data such that $L_2 \gg w_{ij}$.

Note that, contrary to the PE assumption, the preceding assumption can be verified online in part by computing the rank of the record matrix $\sum_{c_l=1}^{L_2} \frac{\chi_{ij}^{c_l}\left(\chi_{ij}^{c_l}\right)^T}{\gamma_{ij}^{c_l}}$.

The BE may be evaluated approximately at a sampling point $s_{ij}^{c_l}$ as

$$\delta_{ij}^{c_l} \triangleq \frac{1}{2}\left(e_{ij}^{c_l}\right)^T Q_{ij} e_{ij}^{c_l} + \frac{1}{2}\left(u_{ij}^{c_l}\right)^T R_i u_{ij}^{c_l} + \hat{W}_{cij}^T \chi_{ij}^{c_l},$$

where $u_{ij}^{c_l} \triangleq -R_i^{-1} g_i^T \left(\sum_{k \in \bar{\mathcal{N}}_i} a_{ik} \left(\sigma_{ijik}^{\prime c_l}\right)^T - \sum_{k \in \mathcal{N}_i} a_{ki} \left(\sigma_{ijki}^{\prime c_l}\right)^T\right) \hat{W}_{aij}$. A least-squares based concurrent learning update law is designed to update the critic NN weight estimate using the BE as [14]

$$\dot{\hat{W}}_{cij} = -\phi_{c1ij} \Gamma_{ij} \frac{\chi_{ij}}{\gamma_{ij}} \delta_{ij} - \frac{\phi_{c2ij}}{L_2} \Gamma_{ij} \sum_{c_l=1}^{L_2} \frac{\chi_{ij}^{c_l}}{\gamma_{ij}^{c_l}} \delta_{ij}^{c_l}, \qquad (25)$$

$$\dot{\Gamma}_{ij} = \left(\beta_{ij} \Gamma_{ij} - \phi_{c1ij} \Gamma_{ij} \frac{\chi_{ij} \chi_{ij}^T}{\gamma_{ij}^2} \Gamma_{ij}\right) \mathbf{1}_{\left\{\left\|\Gamma_{ij}\right\| \leq \bar{\Gamma}_{ij}\right\}}, \qquad (26)$$

where $\phi_{c1ij}, \phi_{c2ij} \in \mathbb{R}_{>0}$ are constant adaptation gains, the initial condition of Γ_{ij} is positive definite and bounded such that $\left\|\Gamma_{ij}\left(t_0\right)\right\| \leq \bar{\Gamma}_{ij}$, $\beta_{ij} \in \mathbb{R}_{>0}$ is a constant forgetting factor, $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function, and $\bar{\Gamma}_{ij} \in \mathbb{R}_{>0}$ is a saturation constant. The construction of the update law in (26) ensures that Γ_{ij} satisfies the bounds

$$\underline{\Gamma}_{ij} \leq \left\|\Gamma_{ij}(t)\right\| \leq \bar{\Gamma}_{ij}, \ \forall t \in \mathbb{R}_{\geq 0}, \qquad (27)$$

where $\underline{\Gamma}_{ij} \in \mathbb{R}_{>0}$ is constant, and that Γ_{ij} remains positive definite [20].

The actor NN weight estimate is updated to follow the critic NN weight as

$$\begin{aligned} \dot{\hat{W}}_{aij} = & -\phi_{a1ij}\left(\hat{W}_{aij} - \hat{W}_{cij}\right) - \phi_{a2ij} \hat{W}_{aij} \\ & + \left(\frac{\phi_{c1ij} G_{\sigma ij}^T \hat{W}_{aij} \chi_{ij}^T}{2 \gamma_{ij}}\right. \\ & \left. + \sum_{c_l=1}^{L_2} \frac{\phi_{c2ij}\left(G_{\sigma ij}^{c_l}\right)^T \hat{W}_{aij}\left(\chi_{ij}^{c_l}\right)^T}{2 L_2 \gamma_{ij}^{c_l}}\right) \hat{W}_{cij}, \qquad (28) \end{aligned}$$

where $\phi_{a1ij}, \phi_{a2ij} \in \mathbb{R}_{>0}$ are constant adaptation gains and $G_{\sigma ij} \triangleq \left(\sum_{k \in \bar{\mathcal{N}}_i} a_{ik} \sigma_{ijik}^{\prime} - \sum_{k \in \mathcal{N}_i} a_{ki} \sigma_{ijki}^{\prime}\right) g_i R_i^{-1} g_i^T \left(\sum_{k \in \bar{\mathcal{N}}_i} a_{ik} \sigma_{ijik}^{\prime T} - \sum_{k \in \mathcal{N}_i} a_{ki} \sigma_{ijki}^{\prime T}\right) \in \mathbb{R}^{w_{ij}}$.

Note that, using this approach, the recorded data points need not be collected from the state trajectory only; given the estimated dynamics, full rank of the update law in (25) is guaranteed by preselecting a set of points which satisfy the condition in (24). Thus, sufficient exploration of the state space is not necessary to accommodate NN training.

Since the network objective is regulation of all agents' states, an intuitive selection of data points is a set of uniformly distributed bounded points about the origin.

V. Stability Analysis

The following theorem describes the stability of the networked systems and describes the performance of the ADP-based control scheme.

Theorem 1. *For each agent $i \in \mathcal{V}$ and neighbor $j \in \bar{\mathcal{N}}_i$, using the identifier in (8) and the controller in (20) along with the adaptive update laws in (25)-(28), e_{ij}, $\tilde{W}_{aij}$, $\tilde{W}_{cij}$ are UUB, resulting in approximate network synchronization and NN weight estimation, provided Assumptions 1-4 hold, the gains $k_{\hat{\theta}_i}, \phi_{a1ij}, \phi_{a2ij}, \phi_{c2ij}$ are selected sufficiently large, and the concurrent learning sample points are selected to obtain a sufficiently large μ_{ij}.*

Proof: Due to paper length restrictions, a sketch of the proof is outlined. Let a positive definite, continuously differentiable candidate Lyapunov function $V_L : \mathbb{R}^{2Nn + \sum_{i \in \mathcal{V}}\left(\sum_{j \in \bar{\mathcal{N}}_i} 2w_{ij} + p_i\right)} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be defined as

$$\begin{aligned} V_L \triangleq \sum_{i \in \mathcal{V}} \Bigg(& \sum_{j \in \bar{\mathcal{N}}_i} \left(V_{ij}^* + \frac{1}{2} \tilde{W}_{cij}^T \Gamma_{ij}^{-1} \tilde{W}_{cij} + \frac{1}{2} \tilde{W}_{aij}^T \tilde{W}_{aij}\right) \\ & + V_{I,i}\Bigg). \qquad (29) \end{aligned}$$

Using the inequalities in (11) and (27), the positive definite property of V_{ij}^* and [21, Lemma 4.3], V_L may be bounded as

$$\underline{v}_L\left(\|Z\|^2\right) \leq V_L(Z, t) \leq \bar{v}_L\left(\|Z\|^2\right) \quad \forall t \in \mathbb{R}_{\geq 0} \qquad (30)$$

for all $Z \in \mathbb{R}^{2Nn + \sum_{i \in \mathcal{V}}\left(\sum_{j \in \bar{\mathcal{N}}_i} 2w_{ij} + p_i\right)}$, where $\underline{v}_L, \bar{v}_L : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions and

$$Z \triangleq \left[x_1^T, \ldots, x_N^T, \tilde{\theta}_1^T, \ldots, \tilde{\theta}_N^T, \tilde{x}_1^T, \ldots, \tilde{x}_N^T, Z_1^T, \ldots, Z_N^T\right]^T,$$

where $Z_i \triangleq \left[\tilde{W}_{ai\zeta_i}^T, \ldots, \tilde{W}_{ai\left|\bar{\mathcal{N}}_i\right|}^T, \tilde{W}_{ci\zeta_i}^T, \ldots, \tilde{W}_{ci\left|\bar{\mathcal{N}}_i\right|}^T\right]^T$ and $\zeta_i \in \bar{\mathcal{N}}_i$. ■

After using the Cauchy-Schwarz and triangle inequalities and completing the squares, the derivative of (29) is upper-bounded as

$$\dot{V}_L \leq -\frac{1}{2} v_L \|Z\|^2 \quad \forall \|Z\| > \sqrt{\frac{2\iota}{v_L}}, \qquad (31)$$

where $v_L \in \mathbb{R}$ is positive provided $k_{\hat{\theta}_i}, \phi_{a1ij}, \phi_{a2ij}, \phi_{c2ij}, \mu_{ij}$ are sufficiently large and $\iota \in \mathbb{R}$ is a positive bounded constant which is dependent on the agents' dynamics and neural network function reconstruction. With (30) and (31), Theorem 4.18 in [21] is invoked to conclude that $Z(t)$ is UUB. Thus, because e_i is UUB $\forall i \in \mathcal{V}$ and $\bar{\mathcal{G}}$ is connected, approximate synchronization is achieved [22]. Furthermore, because each $\tilde{W}_{aij}$ is UUB, each $\left\|u_{ij}^{*} - u_{ij}\right\|$ is UUB.

Remark 1. The ultimate bound can be made arbitrarily small by increasing the gains ϕ_{a1ij}, ϕ_{a2ij}, ϕ_{c2ij} and the parameter μ_{ij} (by selecting more state space evaluation points).

VI. Conclusion

A concurrent learning-based controller is designed to synchronize a communication-centralized leader-follower network of agents with uncertain nonlinear dynamics. The developed controller introduces a control structure which treats the control of each agent as a multi-player nonzero-sum game. The novel per-neighbor control structure allows for greater flexibility in the consideration of individual neighbors during synchronization. The stability of all networked agents and approximate learning of the desired control policies are guaranteed by the use of data-driven concurrent learning-based ADP and a Lyapunov-based stability analysis. Future work may consider completely unknown drift dynamics and communication topologies described by directed graphs.

References

[1] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 38, no. 2, pp. 156–172, 2008.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598 – 1611, 2012.

[8] R. Kamalapurkar, H. T. Dinh, P. Walters, and W. E. Dixon, "Approximate optimal cooperative decentralized control for consensus in a topological network of agents with uncertain nonlinear dynamics," in *Proc. Am. Control Conf.*, Washington, DC, June 2013, pp. 1322–1327.

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

[10] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, ser. IEEE Press Series on Computational Intelligence. Wiley and IEEE Press, 2012, ch. An Actor-Critic-Identifier Architecture for Adaptive Approximate Optimal Control, pp. 258–278.

[11] T. Dierks and S. Jagannathan, "Optimal control of affine nonlinear continuous-time systems," in *Proc. Am. Control Conf.*, 2010, pp. 1568–1573.

[12] 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.

[13] 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.

[14] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Concurrent learning-based approximate optimal regulation," in *Proc. IEEE Conf. Decis. Control*, Florence, IT, Dec. 2013, pp. 6256–6261.

[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] E. Semsar-Kazerooni and K. Khorasani, "Optimal consensus algorithms for cooperative team of agents subject to partial information," *Automatica*, vol. 44, no. 11, pp. 2766 – 2777, 2008.

[17] S. Khoo and L. Xie, "Robust finite-time consensus tracking algorithm for multirobot systems," *IEEE/ASME Trans. Mechatron.*, vol. 14, no. 2, pp. 219–228, 2009.

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

[19] 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.

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

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

[22] G. Chen and F. L. Lewis, "Distributed adaptive tracking control for synchronization of unknown networked Lagrangian systems," *IEEE Trans. Syst. Man Cybern.*, vol. 41, no. 3, pp. 805–816, 2011.