Model-based reinforcement learning in differential graphical games

Rushikesh Kamalapurkar, Justin R. Klotz, Patrick Walters, and Warren E. Dixon

***Abstract*—This paper seeks to combine differential game theory with the actor-critic-identifier architecture to determine forward-in-time, approximate optimal controllers for formation tracking in multi-agent systems, where the agents have uncertain heterogeneous nonlinear dynamics. A continuous control strategy is proposed, using communication feedback from extended neighbors on a communication topology that has a spanning tree. A model-based reinforcement learning technique is developed to cooperatively control a group of agents to track a trajectory in a desired formation. Simulation results are presented to demonstrate the performance of the developed technique.**

I. Introduction

In the past few decades, reinforcement learning (RL)-based techniques have been established as primary tools for online real-time optimization [1]–[7]. RL techniques are valuable not only for optimization but also for control synthesis in complex systems such as a distributed network of cognitive agents. Combined efforts from multiple autonomous agents can yield tactical advantages including: improved munitions effects; distributed sensing, detection, and threat response; and distributed communication pipelines [8], [9]. While coordinating behaviors among autonomous agents is a challenging problem that has received mainstream focus, unique challenges arise when seeking optimal autonomous collaborative behaviors. For example, most collaborative control literature focuses on centralized approaches that require all nodes to continuously communicate with a central agent, yielding a heavy communication demand that is subject to failure due to delays, and missing information [10]. Furthermore, the central agent is required to carry enough on-board computational resources to process the data and to generate command signals. These challenges motivate the need to minimize communication for guidance, navigation and control tasks, and to distribute the computational burden among the agents.

Since all the agents in a network have independent collaborative or competitive objectives, the resulting optimization problem is a multi-objective optimization problem. Differential game theory is often used to define optimality in multi-objective optimization problems [11]–[16]. For example, a Nash equilibrium solution to a multi-objective optimization problem is said to be achieved if none of the players can benefit from a unilateral deviation from the equilibrium [17]. 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 policy. Hence, Nash equilibrium has been a widely used solution concept in differential game-based control techniques. Online real-time solutions to differential games with centralized objectives are presented in results such as [18]–[22]; however, since these results solve problems with centralized objectives (i.e., each agent minimizes or maximizes a cost function that penalizes the states of all the agents in the network), they are not applicable for a network of agents with independent decentralized objectives (i.e., each agent minimizes or maximizes a cost function that penalizes only the error states corresponding to itself).

In this paper, the objective is to obtain an online forward-in-time feedback-Nash equilibrium solution (cf. [23]–[28]) to an infinite-horizon formation tracking problem, where each agent desires to follow a mobile leader while the group maintains a desired formation. The agents try to minimize cost functions that penalize their own formation tracking errors and their own control efforts.

Various methods have been developed to solve optimal tracking problems for linear systems. In [29]–[32], optimal controllers are developed to cooperatively control agents with linear dynamics. In [33], a differential game-based approach is developed for unmanned aerial vehicles to achieve distributed Nash strategies. In [34], an optimal consensus algorithm is developed for a cooperative team of agents with linear dynamics using only partial information.

For nonlinear systems, a MPC-based approach is presented in [35]; however, no stability or convergence analysis is presented. A stable distributed MPC-based approach is presented in [36] for nonlinear discrete-time systems with known nominal dynamics. Asymptotic stability is proved without any interaction between the nodes; however, a nonlinear optimal control problem needs to be solved at every iteration to implement the controller. An optimal tracking approach for formation control is presented in [37] using single network adaptive critics where the value function is learned offline. Recently, a leader-based consensus algorithm is developed in [38] where exact model of the system dynamics is utilized, and convergence to optimality is obtained under a persistence of excitation condition.

For multi-agent problems with decentralized objectives, the desired action by an individual agent depends on the actions and the resulting trajectories of its neighbors; hence, the error

Rushikesh Kamalapurkar, is with the School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA. Email: rushikesh.kamalapurkar@okstate.edu. Justin R. Klotz, Patrick Walters, and Warren E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA. Email: {jklotz, walters8, wdixon}@ufl.edu.

This research is supported in part by National Science Foundation award numbers 1217908 and 1509516, and Office of Naval Research award numbers N00014-13-1-0151 and N00014-16-1-2091. 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.

system for each agent is a complex nonautonomous dynamical system. Nonautonomous systems, in general, have non-stationary value functions. Since non-stationary functions are difficult to approximate using parameterized function approximation schemes such as neural networks (NNs), designing optimal policies for nonautonomous systems is challenging.

Since the external influence from neighbors renders the dynamics of each agent nonautonomous, optimization in a network of agents presents challenges similar to optimal tracking problems. Using insights gained from the authors' previous work on optimal tracking problems [39], this paper develops a model-based RL technique to generate feedback-Nash equilibrium policies online, for agents in a network with cooperative or competitive objectives. In particular, the network of agents is separated into autonomous subgraphs, and the differential game is solved separately on each subgraph.

The primary contribution of this paper is the formulation and online approximate feedback-Nash equilibrium solution of an optimal network formation tracking problem. A relative control error minimization technique is introduced to facilitate the formulation of a feasible infinite-horizon total-cost differential graphical game. Dynamic programming-based feedback-Nash equilibrium solution of the differential graphical game is facilitated via the development of a set of coupled Hamilton-Jacobi (HJ) equations. The developed approximate feedback-Nash equilibrium solution is analyzed using a Lyapunov-based stability analysis to demonstrate ultimately bounded formation tracking in the presence of uncertainties.

II. Notation

Throughout the paper, $\mathbb{R}^n$ denotes $n-$dimensional Euclidean space, $\mathbb{R}_{>a}$ denotes the set of real numbers strictly greater than $a \in \mathbb{R}$, and $\mathbb{R}_{\geq a}$ denotes the set of real numbers greater than or equal to $a \in \mathbb{R}$. Unless otherwise specified, the domain of all the functions is assumed to be $\mathbb{R}_{\geq 0}$. Functions with domain $\mathbb{R}_{\geq 0}$ are defined by abuse of notation using only their image. For example, the function $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ is defined by abuse of notation as $x \in \mathbb{R}^n$. By abuse of notation, the state variables are also used to denote state trajectories. For example, the state variable x in the equation $\dot{x} = f(x) + u$ is also used as $x(t)$ to denote the state trajectory, i.e., the general solution $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ to $\dot{x} = f(x) + u$ evaluated at time t. Unless otherwise specified, all the mathematical quantities are assumed to be time-varying. Unless otherwise specified, an equation of the form $g(x) = f + h(y,t)$ is interpreted as $g(x(t)) = f(t) + h(y(t),t)$ for all $t \in \mathbb{R}_{\geq 0}$, and a definition of the form $g(x,y) \triangleq f(y) + h(x)$ for functions $g : A \times B \to C$, $f : B \to C$ and $h : A \to C$ is interpreted as $g(x,y) \triangleq f(y) + h(x), \forall (x,y) \in A \times B$. The total derivative $\frac{\partial f(x)}{\partial x}$ is denoted by ∇f and the partial derivative $\frac{\partial f(x,y)}{\partial x}$ is denoted by $\nabla_x f(x,y)$. An $n \times n$ identity matrix is denoted by I_n, $n \times m$ matrices of zeros and ones are denoted by $\mathbf{0}_{n\times m}$ and $\mathbf{1}_{n\times m}$, respectively, and $\mathbf{1}_S$ denotes the indicator function of the set S.

III. Graph Theory Preliminaries

Consider a set of N autonomous agents moving in the state space $\mathbb{R}^n$. The control objective is for the agents to maintain a desired formation with respect to a leader. The state of the leader is denoted by $x_0 \in \mathbb{R}^n$. The agents are assumed to be on a network with a fixed communication topology modeled as a static directed graph (i.e. digraph).

Each agent forms a node in the digraph. The set of all nodes excluding the leader is denoted by $\mathcal{N} = \{1, \cdots N\}$ and the leader is denoted by node 0. If node i can receive information from node j then there exists a directed edge from the j^{th} to the i^{th} node of the digraph, denoted by the ordered pair (j,i). Let E denote the set of all edges. Let there be a positive weight $a_{ij} \in \mathbb{R}$ associated with each edge (j,i). Note that $a_{ij} \neq 0$ if and only if $(j,i) \in E$. The digraph is assumed to have no repeated edges, i.e., $(i,i) \notin E, \forall i$, which implies $a_{ii} = 0, \forall i$. The neighborhood sets of node i are denoted by $\mathcal{N}_{-i}$ and $\mathcal{N}_i$, defined as $\mathcal{N}_{-i} \triangleq \{j \in \mathcal{N} \mid (j,i) \in E\}$ and $\mathcal{N}_i \triangleq \mathcal{N}_{-i} \cup \{i\}$.

To streamline the analysis, an adjacency matrix $\mathcal{A} \in \mathbb{R}^{N\times N}$ is defined as $\mathcal{A} \triangleq [a_{ij} \mid i,j \in \mathcal{N}]$, a diagonal pinning gain matrix $\mathcal{A}_0 \in \mathbb{R}^{N\times N}$ is defined as $\mathcal{A}_0 \triangleq \text{diag}([a_{10}, \cdots, a_{N0}])$, an in-degree matrix $\mathcal{D} \in \mathbb{R}^{N\times N}$ is defined as $\mathcal{D} \triangleq \text{diag}(d_i)$, where $d_i \triangleq \sum_{j\in\mathcal{N}_i} a_{ij}$, and a graph Laplacian matrix $\mathcal{L} \in \mathbb{R}^{N\times N}$ is defined as $\mathcal{L} \triangleq \mathcal{D} - \mathcal{A}$. The graph is assumed to have a spanning tree, i.e., given any node i, there exists a directed path from the leader 0 to node i. A node j is said to be an extended neighbor of node i if there exists a directed path from node j to node i. The extended neighborhood set of node i, denoted by $\mathcal{S}_{-i}$, is defined as the set of all extended neighbors of node i. Formally, $\mathcal{S}_{-i} \triangleq \{j \in \mathcal{N} \mid j \neq i \wedge \exists \kappa \leq N,\ \{j_1, \cdots j_\kappa\} \subset \mathcal{N} \mid \{(j,j_1),(j_1,j_2),\cdots,(j_\kappa,i)\} \subset 2^E\}$. Let $\mathcal{S}_i \triangleq \mathcal{S}_{-i} \cup \{i\}$, and let the edge weights be normalized such that $\sum_j a_{ij} = 1$ for all $i \in \mathcal{N}$. Note that the sub-graphs are nested in the sense that $\mathcal{S}_j \subseteq \mathcal{S}_i$ for all $j \in \mathcal{S}_i$.

IV. Problem Formulation

The state $x_i \in \mathbb{R}^n$ of each agent evolves according to the control affine dynamics

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

where $u_i \in \mathbb{R}^{m_i}$ denotes the control input, and $f_i : \mathbb{R}^n \to \mathbb{R}^n$ and $g_i : \mathbb{R}^n \to \mathbb{R}^{n\times m_i}$ are locally Lipschitz continuous functions.

Assumption 1. The dynamics of the leader are described by $\dot{x}_0 = f_0(x_0)$, where $f_0 : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitz continuous function. The function f_0, and the initial condition $x_0(t_0)$ are selected such that the trajectory $x_0(t)$ is uniformly bounded for all $t \in \mathbb{R}_{\geq t_0}$.

The control objective is for the agents to maintain a predetermined formation (with respect to an inertial reference frame) around the leader while minimizing their own cost functions. For all $i \in \mathcal{N}$, the i^{th} agent is aware of its constant desired relative position $x_{dij} \in \mathbb{R}^n$ with respect to all its neighbors $j \in \mathcal{N}_{-i}$, such that the desired formation is realized when $x_i - x_j \to x_{dij}$ for all $i,j \in \mathcal{N}$.[1] To facilitate the control design, the formation is expressed in terms of a set of constant vectors $\{x_{di0} \in \mathbb{R}^n\}_{i\in\mathcal{N}}$ where each x_{di0} denotes the constant

[1]The vectors x_{dij} are assumed to be fixed in an inertial reference frame, i.e., the final desired formation is rigid and its motion in an inertial reference frame can be described as pure translation.

final desired position of agent i with respect to the leader. The vectors $\{x_{di0}\}_{i\in\mathcal{N}}$ are unknown to the agents not connected to the leader, and the known desired inter agent relative position can be expressed in terms of $\{x_{di0}\}_{i\in\mathcal{N}}$ as $x_{dij} = x_{di0} - x_{dj0}$. The control objective is thus satisfied when $x_i \to x_{di0} + x_0$ for all $i \in \mathcal{N}$. To quantify the objective, local neighborhood tracking error signals are defined as

$$e_i = \sum_{j\in\{0\}\cup\mathcal{N}_{-i}} a_{ij}\left((x_i - x_j) - x_{dij}\right). \tag{2}$$

To facilitate the analysis, the error signals in (2) are expressed in terms of the unknown leader-relative desired positions as

$$e_i = \sum_{j\in\{0\}\cup\mathcal{N}_{-i}} a_{ij}\left((x_i - x_{di0}) - (x_j - x_{dj0})\right). \tag{3}$$

Stacking the error signals in a vector $\mathcal{E} \triangleq \left[\begin{array}{cccc} e_1^T, & e_2^T, & \cdots, & e_N^T \end{array}\right]^T \in \mathbb{R}^{nN}$ the equation in (3) can be expressed in a matrix form

$$\mathcal{E} = \left((\mathcal{L} + \mathcal{A}_0) \otimes I_n\right)\left(\mathcal{X} - \mathcal{X}_d - \mathcal{X}_0\right), \tag{4}$$

where $\mathcal{X} = \left[\begin{array}{cccc} x_1^T, & x_2^T, & \cdots, & x_N^T \end{array}\right]^T \in \mathbb{R}^{nN}$, $\mathcal{X}_d = \left[\begin{array}{cccc} x_{d10}^T, & x_{d20}^T, & \cdots, & x_{dN0}^T \end{array}\right]^T \in \mathbb{R}^{nN}$, $\mathcal{X}_0 = \left[\begin{array}{cccc} x_0^T, & x_0^T, & \cdots, & x_0^T \end{array}\right]^T \in \mathbb{R}^{nN}$, and $\otimes$ denotes the Kronecker product. Using (4), it can be concluded that provided the matrix $\left((\mathcal{L} + \mathcal{A}_0) \otimes I_n\right) \in \mathbb{R}^{nN\times nN}$ is nonsingular, $\|\mathcal{E}\| \to 0$ implies $x_i \to x_{di0} + x_0$ for all $i \in \mathcal{N}$, and hence, the satisfaction of control objective. The matrix $\left((\mathcal{L} + \mathcal{A}_0) \otimes I_n\right)$ is nonsingular provided the graph has a spanning tree with the leader at the root [40]. To facilitate the formulation of an optimization problem, the following section explores the functional dependence of the state-value functions for the network of agents.

A. Elements of the value function

The dynamics for the open-loop neighborhood tracking error are $\dot{e}_i = \sum_{j\in\{0\}\cup\mathcal{N}_{-i}} a_{ij}\Big(f_i(x_i) + g_i(x_i)u_i - f_j(x_j) - g_j(x_j)u_j\Big)$. Under the temporary assumption that each controller u_i is an error-feedback controller, i.e. $u_i(t) = \hat{u}_i(e_i(t), t)$, the error dynamics are expressed as $\dot{e}_i = \sum_{j\in\{0\}\cup\mathcal{N}_{-i}} a_{ij}\Big(f_i(x_i) + g_i(x_i)\hat{u}_i(e_i, t) - f_j(x_j) - g_j(x_j)\hat{u}_j(e_j, t)\Big)$. Thus, the error trajectory $\{e_i(t)\}_{t=t_0}^{\infty}$, where t_0 denotes the initial time, depends on $\hat{u}_j(e_j(t), t)$, $\forall j \in \mathcal{N}_i$. Similarly, the error trajectory $\{e_j(t)\}_{t=t_0}^{\infty}$ depends on $\hat{u}_k(e_k(t), t), \forall k \in \mathcal{N}_j$. Recursively, the trajectory $\{e_i(t)\}_{t=t_0}^{\infty}$ depends on $\hat{u}_j(e_j(t), t)$, and hence, on $e_j(t), \forall j \in \mathcal{S}_i$. Thus, even if the controller for each agent is restricted to use local error feedback, the resulting error trajectories are interdependent. In particular, a change in the initial condition of one agent in the extended neighborhood causes a change in the error trajectories corresponding to all the extended neighbors. Consequently, the value function corresponding to an infinite-horizon optimal control problem where each agent tries to minimize $\int_{t_0}^{\infty}\left(Q(e_i(\tau)) + R(u_i(\tau))\right)d\tau$, where $Q : \mathbb{R}^n \to \mathbb{R}$ and $R : \mathbb{R}^{m_i} \to \mathbb{R}$ are positive definite functions, is dependent on the error states of all the extended neighbors.

Since the steady-state controllers required for formation tracking are generally nonzero, quadratic total-cost optimal control problems result in infinite costs, and hence, are infeasible. In the following section, relative steady-state controllers are derived to facilitate the formulation of a feasible optimal control problem.

B. Optimal formation tracking problem

When the agents are perfectly tracking the desired trajectory in the desired formation, even though the states of all the agents are different, the time-derivatives of the states of all the agents are identical. Hence, in steady state, the control signal applied by each agent must be such that the time derivatives of the states corresponding to the set of extended neighbors are identical. In particular, the relative control signal $u_{ij} \in \mathbb{R}^{m_i}$ that will keep node i in its desired relative position with respect to node $j \in \mathcal{S}_{-i}$, i.e., $x_i = x_j + x_{dij}$, must be such that the time derivative of x_i is the same as the time derivative of x_j. Using the dynamics of the agents from (1), and substituting the desired relative positions $x_j + x_{dij}$ for the states x_i, the relative control signals u_{ij} must satisfy

$$f_i(x_j + x_{dij}) + g_i(x_j + x_{dij})u_{ij} = \dot{x}_j. \tag{5}$$

The relative steady-state control signals can be expressed in an explicit form provided the following assumption is satisfied.

Assumption 2. The matrix $g_i(x)$ is full rank for all $i \in \mathcal{N}$ and for all $x \in \mathbb{R}^n$; furthermore, the relative steady-state control signal expressed as $u_{ij} = f_{ij}(x_j) + g_{ij}(x_j)u_j$, satisfies (5) along the desired trajectory, where $f_{ij}(x_j) \triangleq g_i^+(x_j + x_{dij})\left(f_j(x_j) - f_i(x_j + x_{dij})\right) \in \mathbb{R}^{m_i}$, $g_{ij}(x_j) \triangleq g_i^+(x_j + x_{dij})g_j(x_j) \in \mathbb{R}^{m_i\times m_j}$, $g_0(x) \triangleq 0$ for all $x \in \mathbb{R}^n$, $u_{i0} \equiv 0$ for all $i \in \mathcal{N}$, and $g_i^+(x)$ denotes a pseudoinverse of the matrix $g_i(x)$ for all $x \in \mathbb{R}^n$ and for all $i \in \mathcal{N}$.

Assumption 2 places restrictions on the control-effectiveness matrices. The matrices $g_i(x)$ are full rank for a large class of systems including, but not limited to, kinematic wheels and fully actuated Euler-Lagrange systems with invertible inertia matrices. The second part of Assumption 2 requires the existence of a feedback controller that can keep the system on the desired trajectory if the system starts on the desired trajectory. This assumption depends on the systems, the network, the desired formation, and the desired trajectory; hence, insights into its satisfaction are hard to obtain in general. The satisfaction of this assumption needs to be verified on a case-by-case basis. For example, consider a kinematic wheel modeled as

$$\dot{x} = g(x)u, \quad g(x) = \begin{bmatrix} \cos(x_3) & 0 \\ \sin(x_3) & 0 \\ 0 & 1 \end{bmatrix}. \tag{6}$$

In this case, provided the formation satisfies $x_{dij}(3) = 0$, that is, the target formation is such that all the kinematic wheels have the same steering angle, the functions f_{ij} and g_{ij} can be computed as $f_{ij} = 0$, and $g_{ij} = I_2$. The relative steady-state

control is then $u_{ij} = u_j$, which satisfies $g(x_j + x_{dij})\, u_j = \dot{x}_j$, and hence, Assumption 2 holds.

To facilitate the formulation of an optimal formation tracking problem, define the control errors $\mu_i \in \mathbb{R}^{m_i}$ as

$$\mu_i \triangleq \sum_{j \in \mathcal{N}_{-i} \cup \{0\}} a_{ij} (u_i - u_{ij}). \tag{7}$$

The control errors $\{\mu_i\}$ are treated as the design variables in the remainder of this paper. Since the control errors $\{\mu_i\}$ are designed and the controllers $\{u_i\}$ are implemented in practice, it is essential to invert the relationship in (7). To facilitate the inversion, let $\mathcal{S}_i^o \triangleq \{1, \cdots, s_i\}$, where $s_i \triangleq |\mathcal{S}_i|$. Let $\lambda_i : \mathcal{S}_i^o \to \mathcal{S}_i$ be a bijective map such that $\lambda_i(1) = i$. For notational brevity, let $(\cdot)_{\mathcal{S}_i}$ denote the concatenated vector $\left[(\cdot)_{\lambda_i^1}^T, (\cdot)_{\lambda_i^2}^T, \cdots, (\cdot)_{\lambda_i^{s_i}}^T\right]^T$, let $(\cdot)_{\mathcal{S}_{-i}}$ denote the concatenated vector $\left[(\cdot)_{\lambda_i^2}^T, \cdots, (\cdot)_{\lambda_i^{s_i}}^T\right]^T$, let $\sum^i$ denote $\sum_{j \in \mathcal{N}_{-i} \cup \{0\}}$, let λ_i^j denote $\lambda_i(j)$, let $\mathcal{E}_i \triangleq \left[e_{\mathcal{S}_i}^T, x_{\lambda_i^1}^T\right]^T \in \mathbb{R}^{n(s_i+1)}$, and let $\mathcal{E}_{-i} \triangleq \left[e_{\mathcal{S}_{-i}}^T, x_{\lambda_i^1}^T\right]^T \in \mathbb{R}^{ns_i}$. Then, the control error vectors $\mu_{\mathcal{S}_i} \in \mathbb{R}^{\sum_{k \in \mathcal{S}_i} m_k}$ can be expressed as

$$\mu_{\mathcal{S}_i} = \mathscr{L}_{gi}(\mathcal{E}_i)\, u_{\mathcal{S}_i} - F_i(\mathcal{E}_i), \tag{8}$$

where the matrices $\mathscr{L}_{gi} : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}^{\sum_{k \in \mathcal{S}_i} m_k \times \sum_{k \in \mathcal{S}_i} m_k}$ are defined by

$$[\mathscr{L}_{gi}(\mathcal{E}_i)]_{kl} = \begin{cases} -a_{\lambda_i^k \lambda_i^l} g_{\lambda_i^k \lambda_i^l}\left(x_{\lambda_i^l}\right), & \forall l \neq k, \\ \sum^{\lambda_i^k} a_{\lambda_i^k j} I_{m_{\lambda_i^k}}, & \forall l = k, \end{cases}$$

where $k, l = 1, 2, \cdots, s_i$, and $F_i : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}^{\sum_{k \in \mathcal{S}_i} m_k}$ are defined as

$$F_i(\mathcal{E}_i) \triangleq \left[\sum^i a_{\lambda_i^1 j} f_{\lambda_i^1 j}^T(x_j), \cdots, \sum^{\lambda_i^{s_i}} a_{\lambda_i^{s_i} j} f_{\lambda_i^{s_i} j}^T(x_j)\right]^T.$$

Assumption 3. The matrix $\mathscr{L}_{gi}(\mathcal{E}_i(t))$ is invertible for all $t \in \mathbb{R}$ and for all $i \in \mathcal{N}$.

Assumption 3 is a controllability-like condition. Intuitively, Assumption 3 requires the control effectiveness matrices to be compatible to ensure the existence of relative control inputs that allow the agents to follow the desired trajectory in the desired formation. Assumption 3 depends on the systems, the network, the desired formation, and the desired trajectory; hence, insights into its satisfaction are hard to obtain in general. The satisfaction of this assumption needs to be verified on a case-by-case basis. For example, consider the kinematic wheel in (6). Provided the formation satisfies $x_{dij}(3) = 0$, that is, the target formation is such that all the kinematic wheels have the same steering angle, we have $g_{ij} = I_2$, and hence, the matrices $\mathscr{L}_{gi}$ are given by

$$[\mathscr{L}_{gi}(\mathcal{E}_i)]_{kl} = \begin{cases} -a_{\lambda_i^k \lambda_i^l} I_2, & \forall l \neq k, \\ \sum^{\lambda_i^k} a_{\lambda_i^k j} I_2, & \forall l = k, \end{cases}$$

It can be shown that $\mathscr{L}_{gi} = \mathcal{L}_{\mathcal{S}_i} \otimes I_2$, where $\mathcal{L}_{\mathcal{S}_i}$ denotes the Laplacian matrix corresponding to the subgraph $\mathcal{S}_i$. Hence, the graph connectivity condition ensures that the matrices $\mathscr{L}_{gi}$ are invertible, and in this specific case, Assumption 3 holds.

Using Assumption 3, the control vectors can be expressed as

$$u_{\mathcal{S}_i} = \mathscr{L}_{gi}^{-1}(\mathcal{E}_i)\, \mu_{\mathcal{S}_i} + \mathscr{L}_{gi}^{-1}(\mathcal{E}_i)\, F_i(\mathcal{E}_i). \tag{9}$$

Let $\mathscr{L}_{gi}^k$ denote the $\left(\lambda_i^{-1}(k)\right)^{\text{th}}$ block row of $\mathscr{L}_{gi}^{-1}$. Then, the controllers u_i can be implemented as

$$u_i = \mathscr{L}_{gi}^i(\mathcal{E}_i)\, \mu_{\mathcal{S}_i} + \mathscr{L}_{gi}^i(\mathcal{E}_i)\, F_i(\mathcal{E}_i), \tag{10}$$

and for any $j \in \mathcal{N}_{-i}$,

$$u_j = \mathscr{L}_{gi}^j(\mathcal{E}_i)\, \mu_{\mathcal{S}_i} + \mathscr{L}_{gi}^j(\mathcal{E}_i)\, F_i(\mathcal{E}_i). \tag{11}$$

Using (10) and (11), the error and the state dynamics for the agents can be represented as

$$\dot{e}_i = \mathscr{F}_i(\mathcal{E}_i) + \mathscr{G}_i(\mathcal{E}_i)\, \mu_{\mathcal{S}_i}, \tag{12}$$

and

$$\dot{x}_i = \mathcal{F}_i(\mathcal{E}_i) + \mathcal{G}_i(\mathcal{E}_i)\, \mu_{\mathcal{S}_i}, \tag{13}$$

where

$$\mathscr{F}_i(\mathcal{E}_i) \triangleq \sum^i a_{ij} g_i(x_i)\, \mathscr{L}_{gi}^i(\mathcal{E}_i)\, F_i(\mathcal{E}_i) - \sum^i a_{ij} f_j(x_j) - \sum^i a_{ij} g_j(x_j)\, \mathscr{L}_{gi}^j(\mathcal{E}_i)\, F_i(\mathcal{E}_i) + \sum^i a_{ij} f_i(x_i),$$

$$\mathscr{G}_i(\mathcal{E}_i) \triangleq \sum^i a_{ij} \left(g_i(x_i)\, \mathscr{L}_{gi}^i(\mathcal{E}_i) - g_j(x_j)\, \mathscr{L}_{gi}^j(\mathcal{E}_i)\right),$$

$$\mathcal{F}_i(\mathcal{E}_i) \triangleq f_i(x_i) + g_i(x_i)\, \mathscr{L}_{gi}^i(\mathcal{E}_i)\, F_i(\mathcal{E}_i),$$

and $\mathcal{G}_i(\mathcal{E}_i) \triangleq g_i(x_i)\, \mathscr{L}_{gi}^i(\mathcal{E}_i)$.

Let $h_{ei}^{\overline{\mu}_i, \overline{\mu}_{\mathcal{S}_{-i}}}(t, t_0, \mathcal{E}_{i0})$ and $h_{xi}^{\overline{\mu}_i, \overline{\mu}_{\mathcal{S}_{-i}}}(t, t_0, \mathcal{E}_{i0})$ denote the trajectories of (12) and (13), respectively, with the initial time t_0, initial condition $\mathcal{E}_i(t_0) = \mathcal{E}_{i0}$, and policies $\overline{\mu}_j : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}^{m_i}$, $j \in \mathcal{S}_i$, and let $\mathcal{H}_i \triangleq \left[(h_e)_{\mathcal{S}_i}^T, h_{x\lambda_i^1}^T\right]^T$. Define the cost functionals

$$J_i(e_i(\cdot), \mu_i(\cdot)) \triangleq \int_0^\infty r_i(e_i(\sigma), \mu_i(\sigma))\, \mathrm{d}\sigma \tag{14}$$

where $r_i : \mathbb{R}^n \times \mathbb{R}^{m_i} \to \mathbb{R}_{\geq 0}$ denote the local costs defined as $r_i(e_i, \mu_i) \triangleq Q_i(e_i) + \mu_i^T R_i \mu_i$, where $Q_i : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ are positive definite functions, and $R_i \in \mathbb{R}^{m_i \times m_i}$ are constant positive definite matrices. The objective of each agent is to minimize the cost functional in (14). To facilitate the definition of a feedback-Nash equilibrium solution, define the value functions $V_i : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}_{\geq 0}$ as

$$V_i^{\overline{\mu}_i, \overline{\mu}_{\mathcal{S}_{-i}}}(\mathcal{E}_i) \triangleq \int_t^\infty r_i\left(h_{ei}^{\overline{\mu}_i, \overline{\mu}_{\mathcal{S}_{-i}}}(\sigma, t, \mathcal{E}_i), \overline{\mu}_i\left(\mathcal{H}_i^{\overline{\mu}_i, \overline{\mu}_{\mathcal{S}_{-i}}}(\sigma, t, \mathcal{E}_i)\right)\right) \mathrm{d}\sigma, \tag{15}$$

where $V_i^{\overline{\mu}_i, \overline{\mu}_{\mathcal{S}_{-i}}}(\mathcal{E}_i)$ denotes the total cost-to-go for Agent i under the policies $\overline{\mu}_{\mathcal{S}_i}$, when the sub-graph $\mathcal{S}_i$ starts from the state $\mathcal{E}_i$. Note that the value functions in (15) are time-invariant because the dynamical systems $\left\{\dot{e}_j = \mathscr{F}_j(\mathcal{E}_i) + \mathscr{G}_j(\mathcal{E}_i)\, \mu_{\mathcal{S}_j}\right\}_{j \in \mathcal{S}_i}$ and $\dot{x}_i = \mathcal{F}_i(\mathcal{E}_i) + \mathcal{G}_i(\mathcal{E}_i)\, \mu_{\mathcal{S}_i}$ together form an autonomous dynamical system.

A graphical feedback-Nash equilibrium solution within the subgraph $\mathcal{S}_i$ is defined as the tuple of policies $\left\{\mu_j^* : \mathbb{R}^{n(s_j+1)} \to \mathbb{R}^{m_j}\right\}_{j\in\mathcal{S}_i}$ such that the value functions in (15) satisfy

$$V_j^*\left(\mathcal{E}_j\right) \triangleq V_j^{\mu_j^*,\mu_{\mathcal{S}_{-j}}^*}\left(\mathcal{E}_j\right) \leq V_j^{\overline{\mu}_j,\mu_{\mathcal{S}_{-j}}^*}\left(\mathcal{E}_j\right),$$

for all $j \in \mathcal{S}_i$, for all $\mathcal{E}_i \in \mathbb{R}^{n(s_i+1)}$ and for all admissible policies $\overline{\mu}_j$. Provided a feedback-Nash equilibrium solution exists and the value functions (15) are continuously differentiable for all $i \in \mathcal{N}$, the feedback-Nash equilibrium value functions can be characterized in terms of the following system of HJ equations:

$$\sum_{j\in\mathcal{S}_i} \nabla_{e_j} V_i^*\left(\mathcal{E}_i\right)\left(\mathscr{F}_j\left(\mathcal{E}_i\right) + \mathscr{G}_j\left(\mathcal{E}_i\right)\mu_{\mathcal{S}_j}^*\left(\mathcal{E}_i\right)\right) + \nabla_{x_i} V_i^*\left(\mathcal{E}_i\right)\left(\mathcal{F}_i\left(\mathcal{E}_i\right) + \mathcal{G}_i\left(\mathcal{E}_i\right)\mu_{\mathcal{S}_i}^*\left(\mathcal{E}_i\right)\right) + \overline{Q}_i\left(\mathcal{E}_i\right) + \mu_i^{*T}\left(\mathcal{E}_i\right) R_i \mu_i^*\left(\mathcal{E}_i\right) = 0,\ \forall \mathcal{E}_i \in \mathbb{R}^{n(s_i+1)}, \quad (16)$$

where $\overline{Q}_i : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}$ is defined as $\overline{Q}_i\left(\mathcal{E}_i\right) \triangleq Q_i\left(e_i\right)$.

Theorem 1. *Provided a feedback-Nash equilibrium solution exists and that the value functions in (15) are continuously differentiable, the system of HJ equations in (16) constitutes a necessary and sufficient condition for $\left\{\mu_j^* : \mathbb{R}^{n(s_j+1)} \to \mathbb{R}^{m_j}\right\}_{j\in\mathcal{S}_i}$ to be a feedback-Nash equilibrium solution within the subgraph $\mathcal{S}_i$.*

Proof: Consider the cost functional in (14), and assume that all the extended neighbors of the i^{th} agent follow their feedback-Nash equilibrium policies. The value function corresponding to any admissible policy $\overline{\mu}_i$ can be expressed as

$$V_i^{\overline{\mu}_i,\mu_{\mathcal{S}_{-i}}^*}\left(\left[e_i^T,\ \mathcal{E}_{-i}^T\right]^T\right) = \int_t^\infty r_i\left(h_{ei}^{\overline{\mu}_i,\mu_{\mathcal{S}_{-i}}^*}\left(\sigma,t,\mathcal{E}_i\right), \overline{\mu}_i\left(\mathcal{H}_i^{\overline{\mu}_i,\mu_{\mathcal{S}_{-i}}^*}\left(\sigma,t,\mathcal{E}_i\right)\right)\right)\mathrm{d}\sigma.$$

Treating the dependence on $\mathcal{E}_{-i}$ as explicit time dependence define

$$\overline{V}_i^{\overline{\mu}_i,\mu_{\mathcal{S}_{-i}}^*}\left(e_i,t\right) \triangleq V_i^{\overline{\mu}_i,\mu_{\mathcal{S}_{-i}}^*}\left(\left[e_i^T,\ \mathcal{E}_{-i}^T\left(t\right)\right]^T\right), \quad (17)$$

for all $e_i \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{\geq 0}$. Assuming that the optimal controller that minimizes (14) when all the extended neighbors follow their feedback-Nash equilibrium policies exists, and that the optimal value function $\overline{V}_i^* \triangleq \overline{V}_i^{\mu_i^*,\mu_{\mathcal{S}_{-i}}^*}$ exists and is continuously differentiable, optimal control theory for single objective optimization problems (cf. [41]) can be used to derive the following necessary and sufficient condition

$$\frac{\partial \overline{V}_i^*\left(e_i,t\right)}{\partial e_i}\left(\mathscr{F}_i\left(\mathcal{E}_i\right) + \mathscr{G}_i\left(\mathcal{E}_i\right)\mu_{\mathcal{S}_i}^*\left(\mathcal{E}_i\right)\right) + \frac{\partial \overline{V}_i^*\left(e_i,t\right)}{\partial t} + Q_i\left(e_i\right) + \mu_i^{*T}\left(\mathcal{E}_i\right) R_i \mu_i^*\left(\mathcal{E}_i\right) = 0. \quad (18)$$

Using (17), the partial derivative with respect to the state can be expressed as

$$\frac{\partial \overline{V}_i^*\left(e_i,t\right)}{\partial e_i} = \frac{\partial V_i^*\left(\mathcal{E}_i\right)}{\partial e_i}, \quad (19)$$

for all $e_i \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{\geq 0}$, and the partial derivative with respect to time can be expressed as

$$\frac{\partial \overline{V}_i^*\left(e_i,t\right)}{\partial t} = \frac{\partial V_i^*\left(\mathcal{E}_i\right)}{\partial x_i}\left(\mathcal{F}_i\left(\mathcal{E}_i\right) + \mathcal{G}_i\left(\mathcal{E}_i\right)\mu_{\mathcal{S}_i}^*\left(\mathcal{E}_i\right)\right) + \sum_{j\in\mathcal{S}_{-i}} \frac{\partial V_i^*\left(\mathcal{E}_i\right)}{\partial e_j}\left(\mathscr{F}_j\left(\mathcal{E}_i\right) + \mathscr{G}_j\left(\mathcal{E}_i\right)\mu_{\mathcal{S}_j}^*\left(\mathcal{E}_i\right)\right), \quad (20)$$

for all $e_i \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{\geq 0}$. Substituting (19) and (20) into (18) and repeating the process for each i, the system of HJ equations in (16) is obtained. ■

Minimizing the HJ equations using the stationary condition, the feedback-Nash equilibrium solution is expressed in the explicit form

$$\mu_i^*\left(\mathcal{E}_i\right) = -\frac{1}{2}R_i^{-1}\sum_{j\in\mathcal{S}_i}\left(\mathscr{G}_j^i\left(\mathcal{E}_i\right)\right)^T\left(\nabla_{e_j} V_i^*\left(\mathcal{E}_i\right)\right)^T - \frac{1}{2}R_i^{-1}\left(\mathcal{G}_i^i\left(\mathcal{E}_i\right)\right)^T\left(\nabla_{x_i} V_i^*\left(\mathcal{E}_i\right)\right)^T, \quad (21)$$

for all $\mathcal{E}_i \in \mathbb{R}^{n(s_i+1)}$, where $\mathscr{G}_j^i \triangleq \mathscr{G}_j \frac{\partial \mu_{\mathcal{S}_j}^*}{\partial \mu_i^*}$, and $\mathcal{G}_i^i \triangleq \mathcal{G}_i \frac{\partial \mu_{\mathcal{S}_i}^*}{\partial \mu_i^*}$. Since an analytical solution of system of HJ equations in (16) is generally infeasible to obtain, the feedback-Nash value functions and the feedback-Nash policies are approximated using parametric approximation schemes $\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)$ and $\hat{\mu}_i\left(\mathcal{E}_i, \hat{W}_{ai}\right)$, respectively, where $\hat{W}_{ci} \in \mathbb{R}^{L_i}$ and $\hat{W}_{ai} \in \mathbb{R}^{L_i}$ are parameter estimates. Substitution of the approximations $\hat{V}_i$ and $\hat{\mu}_i$ in (16) leads to a set of Bellman errors (BEs) δ_i defined as

$$\delta_i\left(\mathcal{E}_i, \hat{W}_{ci}, \left(\hat{W}_a\right)_{\mathcal{S}_i}\right) \triangleq \hat{\mu}_i^T\left(\mathcal{E}_i, \hat{W}_{ai}\right) R \hat{\mu}_i\left(\mathcal{E}_i, \hat{W}_{ai}\right) + \sum_{j\in\mathcal{S}_i} \nabla_{e_j}\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)\mathscr{G}_j\left(\mathcal{E}_j\right)\hat{\mu}_{\mathcal{S}_j}\left(\mathcal{E}_j, \left(\hat{W}_a\right)_{\mathcal{S}_j}\right) + \nabla_{x_i}\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)\left(\mathcal{F}_i\left(\mathcal{E}_i\right) + \mathcal{G}_i\left(\mathcal{E}_i\right)\hat{\mu}_{\mathcal{S}_i}\left(\mathcal{E}_i, \left(\hat{W}_a\right)_{\mathcal{S}_i}\right)\right) + \sum_{j\in\mathcal{S}_i} \nabla_{e_j}\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)\mathscr{F}_j\left(\mathcal{E}_j\right) + Q_i\left(e_i\right). \quad (22)$$

Approximation of the feedback-Nash equilibrium policies is realized by tuning the estimates $\hat{V}_i$ and $\hat{\mu}_i$ so as to minimize the BEs δ_i. However, computation of δ_i in (22) and u_{ij} in (7) requires exact model knowledge. In the following, a CL-based system identifier is developed to relax the exact model knowledge requirement and to facilitate the implementation of model-based RL via BE extrapolation (cf. [39]). In particular, the developed controllers do not require the knowledge of the system drift functions f_i.

V. System Identification

On any compact set $\chi \subset \mathbb{R}^n$ the function f_i can be represented using a NN as

$$f_i\left(x\right) = \theta_i^T\sigma_{\theta i}\left(x\right) + \epsilon_{\theta i}\left(x\right), \quad (23)$$

for all $x \in \mathbb{R}^n$, where $\theta_i \in \mathbb{R}^{P_i+1\times n}$ denote the unknown output-layer NN weights, $\sigma_{\theta i} : \mathbb{R}^n \to \mathbb{R}^{P_i+1}$ denotes a

bounded NN basis function, $\epsilon_{\theta i} : \mathbb{R}^n \to \mathbb{R}^n$ denotes the function reconstruction error, and $P_i \in \mathbb{N}$ denotes the number of NN neurons. Using the universal function approximation property of single layer NNs, provided the rows of $\sigma_{\theta i}(x)$ form a proper basis, there exist constant ideal weights θ_i and positive constants $\overline{\theta_i} \in \mathbb{R}$ and $\overline{\epsilon_{\theta i}} \in \mathbb{R}$ such that $\|\theta_i\|_F \leq \overline{\theta_i} < \infty$ and $\sup_{x\in\chi} \|\epsilon_{\theta i}(x)\| \leq \overline{\epsilon_{\theta i}}$, where $\|\cdot\|_F$ denotes the Frobenius norm, i.e., $\|\theta\|_F \triangleq \sqrt{\text{tr}(\theta^T\theta)}$.

Assumption 4. The bounds $\overline{\theta_i}$ and $\overline{\epsilon_{\theta i}}$ are known for all $i \in \mathcal{N}$.

Using an estimate $\hat{\theta}_i \in \mathbb{R}^{P_i+1\times n}$ of the weight matrix θ_i, the function f_i can be approximated by the function $\hat{f}_i : \mathbb{R}^n \times \mathbb{R}^{P_i+1\times n} \to \mathbb{R}^n$ defined by $\hat{f}_i\left(x, \hat{\theta}\right) \triangleq \hat{\theta}^T\sigma_{\theta i}(x)$. Based on (23), an estimator for online identification of the drift dynamics is developed as

$$\dot{\hat{x}}_i = \hat{\theta}_i^T\sigma_{\theta i}(x_i) + g_i(x_i)u_i + k_i\tilde{x}_i, \tag{24}$$

where $\tilde{x}_i \triangleq x_i - \hat{x}_i$, and $k_i \in \mathbb{R}$ is a positive constant learning gain. The following assumption facilitates concurrent learning (CL)-based system identification.

Assumption 5. [42], [43] A history stack containing recorded state-action pairs $\left\{x_i^k, u_i^k\right\}_{k=1}^{M_{\theta i}}$ along with numerically computed state derivatives $\left\{\dot{\bar{x}}_i^k\right\}_{k=1}^{M_{\theta i}}$ that satisfies

$$\lambda_{\min}\left(\sum_{k=1}^{M_{\theta i}} \sigma_{\theta i}^k\left(\sigma_{\theta i}^k\right)^T\right) = \underline{\sigma_{\theta i}} > 0,$$
$$\left\|\dot{\bar{x}}_i^k - \dot{x}_i^k\right\| < \overline{d_i},\ \forall k \tag{25}$$

is available a priori. In (25), $\sigma_{\theta i}^k \triangleq \sigma_{\theta i}\left(x_i^k\right)$, $\overline{d_i}, \underline{\sigma_{\theta i}} \in \mathbb{R}$ are known positive constants, and $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue.

The weight estimates $\hat{\theta}_i$ are updated using the following CL-based update law:

$$\dot{\hat{\theta}}_i = k_{\theta i}\Gamma_{\theta i}\sum_{k=1}^{M_{\theta i}} \sigma_{\theta i}^k\left(\dot{\bar{x}}_i^k - g_i^k u_i^k - \hat{\theta}_i^T\sigma_{\theta i}^k\right)^T + \Gamma_{\theta i}\sigma_{\theta i}(x_i)\tilde{x}_i^T, \tag{26}$$

where $g_i^k \triangleq g_i\left(x_i^k\right)$, $k_{\theta i} \in \mathbb{R}$ is a constant positive CL gain, and $\Gamma_{\theta i} \in \mathbb{R}^{P_i+1\times P_i+1}$ is a constant, diagonal, and positive definite adaptation gain matrix.

To facilitate the subsequent stability analysis, a candidate Lyapunov function $V_{0i} : \mathbb{R}^n \times \mathbb{R}^{P_i+1\times n} \to \mathbb{R}$ is selected as

$$V_{0i}\left(\tilde{x}_i, \tilde{\theta}_i\right) \triangleq \frac{1}{2}\tilde{x}_i^T\tilde{x}_i + \frac{1}{2}\text{tr}\left(\tilde{\theta}_i^T\Gamma_{\theta i}^{-1}\tilde{\theta}_i\right), \tag{27}$$

where $\tilde{\theta}_i \triangleq \theta_i - \hat{\theta}_i$ and $\text{tr}(\cdot)$ denotes the trace of a matrix. Using (24)-(26), the identity $\text{tr}\left(\tilde{\theta}^T\left(\sum_{j=1}^{M_{\theta i}}\sigma_{\theta i}^j\sigma_{\theta i}^j\right)\tilde{\theta}\right) = \left(\text{vec}\left(\tilde{\theta}_i\right)\right)^T\left(\left(\sum_{j=1}^{M_{\theta i}}\sigma_{\theta i}^j\sigma_{\theta i}^j\right)\otimes I_{p+1}\right)\left(\text{vec}\left(\tilde{\theta}_i\right)\right)$, and the facts that $\lambda_{\min}\left\{\left(\left(\sum_{j=1}^{M_{\theta i}}\sigma_{\theta i}^j\sigma_{\theta i}^j\right)\otimes I_{p+1}\right)\right\} = \lambda_{\min}\left\{\sum_{j=1}^{M_{\theta i}}\sigma_{\theta i}^j\sigma_{\theta i}^j\right\}$ and $\lambda_{\max}\left\{\left(\left(\sum_{j=1}^{M_{\theta i}}\sigma_{\theta i}^j\sigma_{\theta i}^j\right)\otimes I_{p+1}\right)\right\} = \lambda_{\max}\left\{\sum_{j=1}^{M_{\theta i}}\sigma_{\theta i}^j\sigma_{\theta i}^j\right\}$ (cf. [44, Theorem 4.2.12]), the following bound on the time derivative of V_{0i} is established:

$$\dot{V}_{0i} \leq -k_i\|\tilde{x}_i\|^2 - k_{\theta i}\underline{\sigma_{\theta i}}\left\|\tilde{\theta}_i\right\|_F^2 + \overline{\epsilon_{\theta i}}\|\tilde{x}_i\| + k_{\theta i}\overline{d_{\theta i}}\left\|\tilde{\theta}_i\right\|_F, \tag{28}$$

where $\overline{d_{\theta i}} \triangleq \overline{d}_i\sum_{k=1}^{M_{\theta i}}\left\|\sigma_{\theta i}^k\right\| + \sum_{k=1}^{M_{\theta i}}\left(\left\|\epsilon_{\theta i}^k\right\|\left\|\sigma_{\theta i}^k\right\|\right)$. Using (27) and (28), a Lyapunov-based stability analysis can be used to show that $\hat{\theta}_i$ converges exponentially to a neighborhood around θ_i.

VI. Approximation of the BE and the relative steady-state controller

Using the approximations $\hat{f}_i$ for the functions f_i, the BEs in (22) can be approximated as

$$\hat{\delta}_i\left(\mathcal{E}_i, \hat{W}_{ci}, \left(\hat{W}_a\right)_{\mathcal{S}_i}, \hat{\theta}_{\mathcal{S}_i}\right) \triangleq \hat{\mu}_i^T\left(\mathcal{E}_i, \hat{W}_{ai}\right)R_i\hat{\mu}_i\left(\mathcal{E}_i, \hat{W}_{ai}\right)$$
$$+ \nabla_{x_i}\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)\left(\hat{\mathcal{F}}_i\left(\mathcal{E}_i, \hat{\theta}_{\mathcal{S}_i}\right) + \mathcal{G}_i(\mathcal{E}_i)\hat{\mu}_{\mathcal{S}_i}\left(\mathcal{E}_i, \left(\hat{W}_a\right)_{\mathcal{S}_j}\right)\right)$$
$$+ \sum_{j\in\mathcal{S}_i}\nabla_{e_j}\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)\mathscr{G}_j(\mathcal{E}_j)\hat{\mu}_{\mathcal{S}_j}\left(\mathcal{E}_j, \left(\hat{W}_a\right)_{\mathcal{S}_j}\right)$$
$$+ \sum_{j\in\mathcal{S}_i}\nabla_{e_j}\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right)\hat{\mathscr{F}}_j\left(\mathcal{E}_j, \hat{\theta}_{\mathcal{S}_j}\right) + Q_i(e_i). \tag{29}$$

In (29),

$$\hat{\mathscr{F}}_i\left(\mathcal{E}_i, \hat{\theta}_{\mathcal{S}_i}\right) \triangleq \sum^i a_{ij}\left(\hat{f}_i\left(x_i, \hat{\theta}_i\right) - \hat{f}_j\left(x_j, \hat{\theta}_j\right)\right)$$
$$+\sum^i a_{ij}\left(g_i(x_i)\mathscr{L}_{gi}^i - g_j(x_j)\mathscr{L}_{gi}^j\right)\hat{F}_i\left(\mathcal{E}_i, \hat{\theta}_{\mathcal{S}_i}\right),$$
$$\hat{\mathcal{F}}_i\left(\mathcal{E}_i, \hat{\theta}_{\mathcal{S}_i}\right) \triangleq \hat{\theta}_i^T\sigma_{\theta i}(x_i) + g_i(x_i)\mathscr{L}_{gi}^i\hat{F}_i\left(\mathcal{E}_i, \hat{\theta}_{\mathcal{S}_i}\right),$$
$$\hat{F}_i\left(\mathcal{E}_i, \hat{\theta}_{\mathcal{S}_i}\right) \triangleq \begin{bmatrix} \left(\sum^{\lambda_i^1} a_{\lambda_i^1 j}\hat{f}_{\lambda_i^1 j}\left(x_{\lambda_i^1}, \hat{\theta}_{\lambda_i^1}, x_j, \hat{\theta}_j\right)\right) \\ \vdots \\ \left(\sum^{\lambda_i^{s_i}} a_{\lambda_i^{s_i} j}\hat{f}_{\lambda_i^{s_i} j}\left(x_{\lambda_i^{s_i}}, \hat{\theta}_{\lambda_i^{s_i}}, x_j, \hat{\theta}_j\right)\right) \end{bmatrix},$$
$$\hat{f}_{ij}\left(x_i, \hat{\theta}_i, x_j, \hat{\theta}_j\right) \triangleq g_i^+(x_j + x_{dij})\hat{f}_j\left(x_j, \hat{\theta}_j\right)$$
$$-g_i^+(x_j + x_{dij})\hat{f}_i\left(x_j + x_{dij}, \hat{\theta}_i\right).$$

The approximations $\hat{F}_i$, $\hat{\mathscr{F}}_i$, and $\hat{\mathcal{F}}_i$ are related to the original unknown functions as $\hat{F}_i(\mathcal{E}_i, \theta_{\mathcal{S}_i}) + B_i(\mathcal{E}_i) = F_i(\mathcal{E}_i)$, $\hat{\mathscr{F}}_i(\mathcal{E}_i, \theta_{\mathcal{S}_i}) + \mathscr{B}_i(\mathcal{E}_i) = \mathscr{F}_i(\mathcal{E}_i)$, and $\hat{\mathcal{F}}_i(\mathcal{E}_i, \theta_{\mathcal{S}_i}) + \mathcal{B}_i(\mathcal{E}_i) = \mathcal{F}_i(\mathcal{E}_i)$, where B_i, $\mathscr{B}_i$, and $\mathcal{B}_i$ are $O\left(\left(\overline{\epsilon_\theta}\right)_{\mathcal{S}_i}\right)$ terms that denote bounded function approximation errors.

Using the approximations $\hat{f}_i$, an implementable form of the controllers in (9) is expressed as

$$u_{\mathcal{S}_i} = \mathscr{L}_{gi}^{-1}(\mathcal{E}_i)\hat{\mu}_{\mathcal{S}_i}\left(\mathcal{E}_i, \left(\hat{W}_a\right)_{\mathcal{S}_i}\right) + \mathscr{L}_{gi}^{-1}\hat{F}_i(\mathcal{E}_i, \theta_{\mathcal{S}_i}). \tag{30}$$

Using (8) and (30), an unmeasurable form of the virtual controllers implemented on the systems (12) and (13) is given by

$$\mu_{\mathcal{S}_i} = \hat{\mu}_{\mathcal{S}_i}\left(\mathcal{E}_i, \left(\hat{W}_a\right)_{\mathcal{S}_i}\right) - \hat{F}_i\left(\mathcal{E}_i, \tilde{\theta}_{\mathcal{S}_i}\right) - B_i(\mathcal{E}_i). \tag{31}$$

VII. Value function approximation

On any compact set $\chi \in \mathbb{R}^{n(s_i+1)}$, the value functions can be represented as

$$V_i^*(\mathcal{E}_i) = W_i^T\sigma_i(\mathcal{E}_i) + \epsilon_i(\mathcal{E}_i),\ \forall\mathcal{E}_i \in \mathbb{R}^{n(s_i+1)}, \tag{32}$$

where $W_i \in \mathbb{R}^{L_i}$ are ideal NN weights, $\sigma_i : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}^{L_i}$ are NN basis functions and $\epsilon_i : \mathbb{R}^{n(s_i+1)} \to \mathbb{R}$ are function approximation errors. Using the universal function approximation property of single layer NNs, provided $\sigma_i(\mathcal{E}_i)$ forms a proper basis, there exist constant ideal weights W_i and positive constants $\overline{W_i} \in \mathbb{R}$ and $\overline{\epsilon_i}, \overline{\nabla \epsilon_i} \in \mathbb{R}$ such that $\|W_i\| \leq \overline{W_i} < \infty$, $\sup_{\mathcal{E}_i \in \chi} \|\epsilon_i(\mathcal{E}_i)\| \leq \overline{\epsilon_i}$, and $\sup_{\mathcal{E}_i \in \chi} \|\nabla \epsilon_i(\mathcal{E}_i)\| \leq \overline{\nabla \epsilon_i}$.

Assumption 6. The constants $\overline{\epsilon_i}$, $\overline{\nabla \epsilon_i}$, and $\overline{W_i}$ are known for all $i \in \mathcal{N}$.

Using (21) and (32), the feedback-Nash equilibrium policies are

$$\mu_i^*(\mathcal{E}_i) = -\frac{1}{2} R_i^{-1} G_{\sigma i}(\mathcal{E}_i) W_i - \frac{1}{2} R_i^{-1} G_{\epsilon i}(\mathcal{E}_i),$$

for all $\mathcal{E}_i \in \mathbb{R}^{n(s_i+1)}$, where $G_{\sigma i}(\mathcal{E}_i) \triangleq \sum_{j \in \mathcal{S}_i} \left(\mathscr{G}_j^i(\mathcal{E}_i)\right)^T \left(\nabla_{e_j} \sigma_i(\mathcal{E}_i)\right)^T + \left(\mathcal{G}_i^i(\mathcal{E}_i)\right)^T \left(\nabla_{x_i} \sigma_i(\mathcal{E}_i)\right)^T$ and $G_{\epsilon i}(\mathcal{E}_i) \triangleq \sum_{j \in \mathcal{S}_i} \left(\mathscr{G}_j^i(\mathcal{E}_i)\right)^T \left(\nabla_{e_j} \epsilon_i(\mathcal{E}_i)\right)^T + \left(\mathcal{G}_i^i(\mathcal{E}_i)\right)^T \left(\nabla_{x_i} \epsilon_i(\mathcal{E}_i)\right)^T$. The value functions and the policies are approximated using NNs as

$$\hat{V}_i\left(\mathcal{E}_i, \hat{W}_{ci}\right) \triangleq \hat{W}_{ci}^T \sigma_i(\mathcal{E}_i),$$
$$\hat{\mu}_i\left(\mathcal{E}_i, \hat{W}_{ai}\right) \triangleq -\frac{1}{2} R_i^{-1} G_{\sigma i}(\mathcal{E}_i) \hat{W}_{ai}, \tag{33}$$

where $\hat{W}_{ci}$ and $\hat{W}_{ai}$ are estimates of the ideal weights W_i, introduced in (22).

VIII. Simulation of experience via BE extrapolation

A consequence of Theorem 1 is that the BE provides an indirect measure of how close the estimates $\hat{W}_{ci}$ and $\hat{W}_{ai}$ are to the ideal weights W_i. From a reinforcement learning perspective, each evaluation of the BE along the system trajectory can be interpreted as experience gained by the critic, and each evaluation of the BE at points not yet visited can be interpreted as simulated experience. In previous results such as [4], [20], [21], [29], [45], the critic is restricted to the experience gained (in other words BEs evaluated) along the system state trajectory. The development in [20], [21], [29], [45] can be extended to employ simulated experience; however, the extension requires exact model knowledge. In results such as [4], the formulation of the BE does not allow for simulation of experience. The formulation in (29) employs the system identifier developed in Section V to facilitate approximate evaluation of the BE at off-trajectory points.

To simulate experience, a set of points $\left\{\mathcal{E}_i^k\right\}_{k=1}^{M_i}$ is selected corresponding to each agent i , and the instantaneous BE in (22) is approximated at the current state and at the selected points using (37). The approximation at the current state is denoted by $\hat{\delta}_{ti}$ and the approximation at the selected points is denoted by $\hat{\delta}_{ti}^k$, where $\hat{\delta}_{ti}$ and $\hat{\delta}_{ti}^k$ are defined as

$$\hat{\delta}_{ti}(t) \triangleq \hat{\delta}_i\left(\mathcal{E}_i(t), \hat{W}_{ci}(t), \left(\hat{W}_a(t)\right)_{\mathcal{S}_i}, \left(\hat{\theta}(t)\right)_{\mathcal{S}_i}\right),$$
$$\hat{\delta}_{ti}^k(t) \triangleq \hat{\delta}_i\left(\mathcal{E}_i^k, \hat{W}_{ci}(t), \left(\hat{W}_a(t)\right)_{\mathcal{S}_i}, \left(\hat{\theta}(t)\right)_{\mathcal{S}_i}\right).$$

Note that once $\{e_j\}_{j \in \mathcal{S}_i}$ and x_i are selected, the i^{th} agent can compute the states of all the remaining agents in the sub-graph. For notational brevity, the arguments to the functions σ_i, $\hat{\mathscr{F}}_i$, $\mathscr{G}_i$, $\mathcal{G}_i$, $\hat{\mathcal{F}}_i$, $\hat{\mu}_i$, $G_{\sigma i}$, $G_{\epsilon i}$, and ϵ_i are suppressed hereafter.

The critic uses simulated experience to update the value function weights using a least squares-based update law

$$\dot{\hat{W}}_{ci} = -\eta_{c1i} \Gamma_i \frac{\omega_i}{\rho_i} \hat{\delta}_{ti} - \frac{\eta_{c2i} \Gamma_i}{M_i} \sum_{k=1}^{M_i} \frac{\omega_i^k}{\rho_i^k} \hat{\delta}_{ti}^k,$$
$$\dot{\Gamma}_i = \left(\beta_i \Gamma_i - \eta_{c1i} \Gamma_i \frac{\omega_i \omega_i^T}{\rho_i^2} \Gamma_i\right) \mathbf{1}_{\left\{\|\Gamma_i\| \leq \overline{\Gamma}_i\right\}}, \tag{34}$$

where $\rho_i \triangleq 1 + \nu_i \omega_i^T \Gamma_i \omega_i$, $\Gamma_i \in \mathbb{R}^{L_i \times L_i}$ denotes the time-varying least-squares learning gain, $\overline{\Gamma}_i \in \mathbb{R}$ denotes the saturation constant, $\|\Gamma_i(t_0)\| \leq \overline{\Gamma}_i$, and $\eta_{c1i}, \eta_{c2i}, \beta_i, \nu_i \in \mathbb{R}$ are constant positive learning gains. In (34),

$$\omega_i \triangleq \sum_{j \in \mathcal{S}_i} \nabla_{e_j} \sigma_i \left(\hat{\mathscr{F}}_j + \mathscr{G}_j \hat{\mu}_{\mathcal{S}_j}\right) + \nabla_{x_i} \sigma_i \left(\hat{\mathcal{F}}_i + \mathcal{G}_i \hat{\mu}_{\mathcal{S}_i}\right),$$
$$\omega_i^k \triangleq \sum_{j \in \mathcal{S}_i} \nabla_{e_j} \sigma_i^k \left(\hat{\mathscr{F}}_j^k + \mathscr{G}_j^k \hat{\mu}_{\mathcal{S}_j}^k\right) + \nabla_{x_i} \sigma_i^k \left(\hat{\mathcal{F}}_i^k + \mathcal{G}_i^k \hat{\mu}_{\mathcal{S}_i}^k\right),$$

where for a function $\phi_i(\mathcal{E}_i, (\cdot))$, the notation ϕ_i^k indicates evaluation at $\mathcal{E}_i = \mathcal{E}_i^k$; i.e., $\phi_i^k \triangleq \phi_i\left(\mathcal{E}_i^k, (\cdot)\right)$. The actor updates the policy weights using the following update law derived based on the Lyapunov-based stability analysis in section IX.

$$\dot{\hat{W}}_{ai} = -\eta_{a2i} \hat{W}_{ai} + \frac{1}{4} \eta_{c1i} G_{\sigma i}^T R_i^{-1} G_{\sigma i} \hat{W}_{ai} \frac{\omega_i^T}{\rho_i} \hat{W}_{ci}$$
$$+ \frac{1}{4} \sum_{k=1}^{M_i} \frac{\eta_{c2i}}{M_i} \left(G_{\sigma i}^k\right)^T R_i^{-1} G_{\sigma i}^k \hat{W}_{ai} \frac{\left(\omega_i^k\right)^T}{\rho_i^k} \hat{W}_{ci}$$
$$- \eta_{a1i} \left(\hat{W}_{ai} - \hat{W}_{ci}\right), \tag{35}$$

where $\eta_{a1i}, \eta_{a2i} \in \mathbb{R}$ are constant positive learning gains. The following assumption facilitates simulation of experience.

Assumption 7. [43] For each $i \in \mathcal{N}$, there exists a finite set of points $\left\{\mathcal{E}_i^k\right\}_{k=1}^{M_i}$ such that

$$\underline{\rho_i} \triangleq \frac{\left(\inf_{t \in \mathbb{R}_{\geq 0}} \left(\lambda_{\min} \left\{\sum_{k=1}^{M_i} \frac{\omega_i^k(t) \left(\omega_i^k\right)^T(t)}{\rho_i^k(t)}\right\}\right)\right)}{M_i} > 0, \tag{36}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue, and $\underline{\rho_i} \in \mathbb{R}$ is a positive constant.

IX. Stability analysis

To facilitate the stability analysis, the left hand side of (16) is subtracted from (29) to express the BEs in terms of the weight estimation errors as

$$\hat{\delta}_{ti} = -\tilde{W}_{ci}^T \omega_i - W_i^T \nabla_{x_i} \sigma_i(\mathcal{E}_i) \hat{\mathcal{F}}_i\left(\mathcal{E}_i, \tilde{\theta}_{\mathcal{S}_i}\right)$$
$$+ \frac{1}{4} \tilde{W}_{ai}^T G_{\sigma i}^T R_i^{-1} G_{\sigma i} \tilde{W}_{ai} - \frac{1}{2} W_i^T G_{\sigma i}^T R_i^{-1} G_{\sigma i} \tilde{W}_{ai}$$
$$+ \frac{1}{2} W_i^T \sum_{j \in \mathcal{S}_i} \nabla_{e_j} \sigma_i(\mathcal{E}_i) \mathscr{G}_j \mathcal{R}_{\mathcal{S}_j}\left(\tilde{W}_a\right)_{\mathcal{S}_j}$$

$$- W_i^T \sum_{j\in\mathcal{S}_i} \nabla_{e_j}\sigma_i\left(\mathcal{E}_i\right)\hat{\mathscr{F}}_j\left(\mathcal{E}_j,\tilde{\theta}_{\mathcal{S}_j}\right) + \frac{1}{2}W_i^T\nabla_{x_i}\sigma_i\left(\mathcal{E}_i\right)\mathcal{G}_i\mathcal{R}_{\mathcal{S}_i}\left(\tilde{W}_a\right)_{\mathcal{S}_i} + \Delta_i, \quad (37)$$

where $\hat{(\cdot)} \triangleq (\cdot) - \hat{(\cdot)}$, $\Delta_i = O\left(\left(\overline{\epsilon}\right)_{\mathcal{S}_i}, \left(\overline{\nabla\epsilon}\right)_{\mathcal{S}_i}, \left(\overline{\epsilon_\theta}\right)_{\mathcal{S}_i}\right)$, and $\mathcal{R}_{\mathcal{S}_j} \triangleq \text{diag}\left(\left[R_{\lambda_j^1}^{-1}G_{\sigma\lambda_j^1}^T, \cdots, R_{\lambda_j^{s_j}}^{-1}G_{\sigma\lambda_j^{s_j}}^T\right]\right)$ are block diagonal matrices. Consider a set of extended neighbors $\mathcal{S}_p$ corresponding to the p^{th} agent. To analyze asymptotic properties of the agents in $\mathcal{S}_p$, consider the following candidate Lyapunov function

$$V_{Lp}\left(Z_p,t\right) \triangleq \sum_{i\in\mathcal{S}_p} V_{ti}\left(e_{\mathcal{S}_i},t\right) + \sum_{i\in\mathcal{S}_p}\frac{1}{2}\tilde{W}_{ci}^T\Gamma_i^{-1}\tilde{W}_{ci} + \sum_{i\in\mathcal{S}_p}\frac{1}{2}\tilde{W}_{ai}^T\tilde{W}_{ai} + \sum_{i\in\mathcal{S}_p} V_{0i}\left(\tilde{x}_i,\tilde{\theta}_i\right), \quad (38)$$

where $Z_p \in \mathbb{R}^{(2ns_i+2L_is_i+n(P_i+1)s_i)}$ is defined as

$$Z_p \triangleq \left[e_{\mathcal{S}_p}^T, \left(\tilde{W}_c\right)_{\mathcal{S}_p}^T, \left(\tilde{W}_a\right)_{\mathcal{S}_p}^T, \tilde{x}_{\mathcal{S}_p}^T, \text{vec}\left(\tilde{\theta}_{\mathcal{S}_p}\right)^T\right]^T,$$

$\text{vec}\left(\cdot\right)$ denotes the vectorization operator, and $V_{ti} : \mathbb{R}^{ns_i}\times\mathbb{R}\to\mathbb{R}$ is defined as

$$V_{ti}\left(e_{\mathcal{S}_i},t\right) \triangleq V_i^*\left(\left[e_{\mathcal{S}_i}^T,\ x_i^T\left(t\right)\right]^T\right), \quad (39)$$

for all $e_{\mathcal{S}_i}\in\mathbb{R}^{ns_i}$ and for all $t\in\mathbb{R}_{\geq t_0}$. Since V_{ti}^* depends on t only through uniformly bounded leader trajectories, Lemma 1 from [46] can be used to show that V_{ti} is a positive definite and decrescent function.[2] Thus, using Lemma 4.3 from [47], the following bounds on the candidate Lyapunov function in (38) are established

$$\underline{v_{lp}}\left(\|Z_p\|\right) \leq V_{Lp}\left(Z_p,t\right) \leq \overline{v_{lp}}\left(\|Z_p\|\right), \quad (40)$$

for all $Z_p\in\mathbb{R}^{(2ns_i+2L_is_i+n(P_i+1)s_i)}$ and for all $t\in\mathbb{R}_{\geq t_0}$, where $\underline{v_{lp}}, \overline{v_{lp}} : \mathbb{R}\to\mathbb{R}$ are class $\mathcal{K}$ functions.

To facilitate the stability analysis, given any compact ball $\chi_p\subset\mathbb{R}^{2ns_i+2L_is_i+n(P_i+1)s_i}$ of radius $r_p\in\mathbb{R}$ centered at the origin, a positive constant $\iota_p\in\mathbb{R}$ is defined as

$$\iota_p \triangleq \sum_{i\in\mathcal{S}_p}\overline{\left\|\sum_{j\in\mathcal{S}_i}\nabla_{e_j}V_i^*\left(\mathcal{E}_i\right)\mathscr{G}_jB_j + \nabla_{x_i}V_i^*\left(\mathcal{E}_i\right)\mathcal{G}_iB_i\right\|} + \sum_{i\in\mathcal{S}_p}\frac{1}{2}\overline{\left\|\nabla_{x_i}V_i^*\left(\mathcal{E}_i\right)\mathcal{G}_i\mathcal{R}_{\mathcal{S}_i}\epsilon_{\mathcal{S}_i} + \sum_{j\in\mathcal{S}_i}\nabla_{e_j}V_i^*\left(\mathcal{E}_i\right)\mathscr{G}_j\mathcal{R}_{\mathcal{S}_j}\epsilon_{\mathcal{S}_j}\right\|} + \sum_{i\in\mathcal{S}_p}\frac{\overline{\epsilon_{\theta i}}^2}{2k_i} + \sum_{i\in\mathcal{S}_p}\frac{3\left(k_{\theta i}\overline{d_{\theta i}} + \overline{\|A_i^\theta\|\|B_i^\theta\|}\right)^2}{4\underline{\sigma_{\theta i}}}$$

$$+ \sum_{i\in\mathcal{S}_p}\frac{3}{4\left(\eta_{a1i}+\eta_{a2i}\right)}\left(\frac{1}{2}\overline{\|A_i^{a1}\|} + \eta_{a2i}\overline{W_i} + \frac{1}{4}\left(\eta_{c1i}+\eta_{c2i}\right)\overline{\left\|W_i^T\frac{\omega_i}{\rho_i}W_i^TG_{\sigma i}^TR_i^{-1}G_{\sigma i}\right\|}\right)^2 + \sum_{i\in\mathcal{S}_p}\frac{5\left(\eta_{c1i}+\eta_{c2i}\right)^2\overline{\left\|\frac{\omega_i}{\rho_i}\Delta_i\right\|}^2}{4\eta_{c2i}\underline{\rho_i}}$$

where for any function $\varpi : \mathbb{R}^l\to\mathbb{R}$, $l\in\mathbb{N}$, the notation $\overline{\|\varpi\|}$ denotes $\sup_{y\in\chi_p\cap\mathbb{R}^l}\|\varpi\left(y\right)\|$ and A_i^θ, B_i^θ, and A_i^{a1} are uniformly bounded state-dependent terms. The following sufficient gain conditions facilitate the subsequent stability analysis.

$$\frac{\eta_{c2i}\rho_i}{5} > \sum_{j\in\mathcal{S}_p}\frac{3s_p\mathbf{1}_{j\in\mathcal{S}_i}\left(\eta_{c1i}+\eta_{c2i}\right)^2\overline{\left\|A_{ij}^{1a\theta}\right\|}^2\overline{\left\|B_{ij}^{1a\theta}\right\|}^2}{4k_{\theta j}\underline{\sigma_{\theta j}}}, \quad (41)$$

$$\frac{\left(\eta_{a1i}+\eta_{a2i}\right)}{3} > \sum_{j\in\mathcal{S}_p}\frac{5s_p\mathbf{1}_{i\in\mathcal{S}_j}\left(\eta_{c1j}+\eta_{c2j}\right)^2\overline{\left\|A_{ji}^{1ac}\right\|}^2}{16\eta_{c2j}\underline{\rho_j}} + \frac{5\eta_{a1i}^2}{4\eta_{c2i}\underline{\rho_i}} + \frac{\left(\eta_{c1i}+\eta_{c2i}\right)\overline{W_i}\left\|\frac{\omega_i}{\rho_i}\right\|\overline{\left\|G_{\sigma i}^TR_i^{-1}G_{\sigma i}\right\|}}{4}, \quad (42)$$

$$v_{lp}^{-1}\left(\iota_p\right) < \overline{v_{lp}}^{-1}\left(\underline{v_{lp}}\left(r_p\right)\right), \quad (43)$$

where $A_{ij}^{1a\theta}$, $B_{ij}^{1a\theta}$, and A_{ji}^{1ac} are uniformly bounded state-dependent terms.

Theorem 2. *Provided Assumptions 1 - 7 hold and the sufficient gain conditions in (41)-(43) are satisfied, the controller in (33) along with the actor and critic update laws in (34) and (35), and the system identifier in (24) along with the weight update laws in (26) ensure that the local neighborhood tracking errors e_i are ultimately bounded and that the policies $\hat{\mu}_i$ converge to a neighborhood around the feedback-Nash policies μ_i^* for all $i\in\mathcal{N}$.*

Proof: The time derivative of the candidate Lyapunov function in (38) is given by

$$\dot{V}_{Lp} = \sum_{i\in\mathcal{S}_p}\dot{V}_{ti}\left(e_{\mathcal{S}_i},t\right) - \frac{1}{2}\sum_{i\in\mathcal{S}_p}\tilde{W}_{ci}^T\Gamma_i^{-1}\dot{\Gamma}_i\Gamma_i^{-1}\tilde{W}_{ci} - \sum_{i\in\mathcal{S}_p}\tilde{W}_{ci}^T\Gamma_i^{-1}\dot{\hat{W}}_{ci} - \sum_{i\in\mathcal{S}_p}\tilde{W}_{ai}^T\dot{\hat{W}}_{ai} + \sum_{i\in\mathcal{S}_p}\dot{V}_{0i}\left(\tilde{x}_i,\tilde{\theta}_i\right). \quad (44)$$

Using (16), (28), (31), and (37), the update laws in (34) and (35), and the definition of V_{ti} in (39), the derivative in (44) can be bounded as[3]

$$\dot{V}_{Lp} \leq \sum_{i\in\mathcal{S}_p}\left(-\frac{\eta_{c2i}\underline{\rho_i}}{5}\left\|\tilde{W}_{ci}\right\|^2 - \frac{\left(\eta_{a1i}+\eta_{a2i}\right)}{3}\left\|\tilde{W}_{ai}\right\|^2\right) + \sum_{i\in\mathcal{S}_p}\left(-\underline{q_i}\left(\|e_i\|\right) - \frac{k_i}{2}\left\|\tilde{x}_i\right\|^2 - \frac{k_{\theta i}\underline{\sigma_{\theta i}}}{3}\left\|\tilde{\theta}_i\right\|_F^2\right) + \iota_p.$$

[2] Since the graph has a spanning tree, the mapping between the errors and the states is invertible. Hence, the state of an agent can be expressed as $x_i = h_i\left(e_{\mathcal{S}_i},x_0\right)$ for some function h_i. Thus, the value function can be expressed as $V_i^*\left(e_{\mathcal{S}_i},x_0\right) = V_i^*\left(e_{\mathcal{S}_i},h\left(e_{\mathcal{S}_i},x_0\right)\right)$. Then, V_{ti}^* can be alternatively defined as $V_{ti}\left(e_{\mathcal{S}_i},t\right) \triangleq V_i^*\left(\begin{bmatrix}e_{\mathcal{S}_i}\\ x_0\left(t\right)\end{bmatrix}\right)$. Since x_0 is a uniformly bounded function of t by assumption, Lemma 1 from [46] can be used to conclude that V_{ti} is a positive definite and decrescent function.

[3] For a detailed derivation of the bound, see [48].

Let $v_{lp} : \mathbb{R} \to \mathbb{R}$ be a class $\mathcal{K}$ function such that

$$v_{lp}(\|Z_p\|) \leq \frac{1}{2}\sum_{i\in\mathcal{S}_p} \underline{q_i}(\|e_i\|) + \frac{1}{2}\sum_{i\in\mathcal{S}_p} \frac{\eta_{c2i}\underline{\rho_i}}{5}\left\|\tilde{W}_{ci}\right\|^2 + \frac{1}{2}\sum_{i\in\mathcal{S}_p} \frac{(\eta_{a1i}+\eta_{a2i})}{3}\left\|\tilde{W}_{ai}\right\|^2 + \frac{1}{2}\sum_{i\in\mathcal{S}_p} \frac{k_i}{2}\|\tilde{x}_i\|^2 + \frac{1}{2}\sum_{i\in\mathcal{S}_p} \frac{k_{\theta i}\underline{\sigma_{\theta i}}}{3}\left\|\tilde{\theta}_i\right\|_F^2, \quad (45)$$

where $\underline{q_i} : \mathbb{R} \to \mathbb{R}$ are class $\mathcal{K}$ functions such that $\underline{q_i}(\|e\|) \leq Q_i(e)$, $\forall e \in \mathbb{R}^n$, $\forall i \in \mathcal{N}$. Then, the Lyapunov derivative can be bounded as

$$\dot{V}_{Lp} \leq -v_{lp}(\|Z_p\|) \quad (46)$$

for all Z_p such that $Z_p \in \chi_p$ and $\|Z_p\| \geq v_{lp}^{-1}(\iota_p)$. Using the bounds in (40), the sufficient conditions in (41)-(43), and the inequality in (46), Theorem 4.18 in [47] can be invoked to conclude that every trajectory $Z_p(t)$ satisfying $\|Z_p(t_0)\| \leq \overline{v_{lp}}^{-1}\left(\underline{v_{lp}}(r_p)\right)$, is bounded for all $t \in \mathbb{R}_{\geq t_0}$ and satisfies

$$\limsup_{t\to\infty} \|Z_p(t)\| \leq \underline{v_{lp}}^{-1}\left(\overline{v_{lp}}\left(v_{lp}^{-1}(\iota_p)\right)\right).$$

Since the choice of the subgraph $\mathcal{S}_p$ was arbitrary, the neighborhood tracking errors e_i are ultimately bounded for all $i \in \mathcal{N}$. Furthermore, the weight estimates $\hat{W}_{ai}$ converge to a neighborhood of the ideal weights W_i; hence, invoking Theorem 1, the policies $\hat{\mu}_i$ converge to a neighborhood of the feedback-Nash equilibrium policies μ_i^* for all $i \in \mathcal{N}$. ■

X. Simulations

This section provides a simulation example to demonstrate the applicability of the developed technique. The agents are assumed to have the communication topology as shown in Figure 1 with unit pinning gains and edge weights. The motion of the agents is described by identical nonlinear one-dimensional dynamics of the form (1) where $f_i(x_i) = \theta_{i1}x_i + \theta_{i2}x_i^2$, and $g_i(x_i) = (\cos(2x_{i1}) + 2)$ for all $i = 1, \cdots, 5$. The ideal values of the unknown parameters are selected to be $\theta_{i1} = 0$, 0, 0.1, 0.5, and 0.2, and $\theta_{i2} = 1$, 0.5, 1, 1, and 1, for $i = 1, \cdots, 5$, respectively. The agents start at $x_i = 2$ for all i, and their final desired locations with respect to each other are given by $x_{d12} = 0.5$, $x_{d21} = -0.5$, $x_{d43} = -0.5$, and $x_{d53} = -0.5$. The leader traverses an exponentially decaying trajectory $x_0(t) = e^{-0.1t}$. The desired positions of agents 1 and 3 with respect to the leader are $x_{d10} = 0.75$ and $x_{d30} = 1$, respectively.[4]

For each agent i, five values of e_i, three values of x_i, and three values of errors corresponding to all the extended neighbors are selected for BE extrapolation, resulting in 5×3^{s_i} total values of $\mathcal{E}_i$. All agents estimate the unknown drift parameters using history stacks containing thirty points recorded online using a singular value maximizing algorithm (cf. [49]), and compute the required state derivatives using a fifth order Savitzky-Golay smoothing filter (cf. [50]). Figures 2 - 4

[4]The optimal control problem parameters, basis functions, and adaptation gains for all the agents and the plots for weight estimates corresponding to agents 1-5 are available in [48]

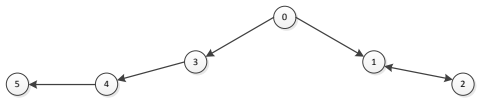

Fig. 1. Communication topology: A network containing five agents.

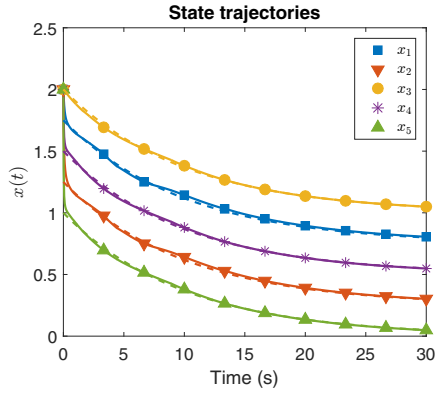

Fig. 2. State trajectories for the five agents for the one-dimensional example. The dotted lines show the desired state trajectories.

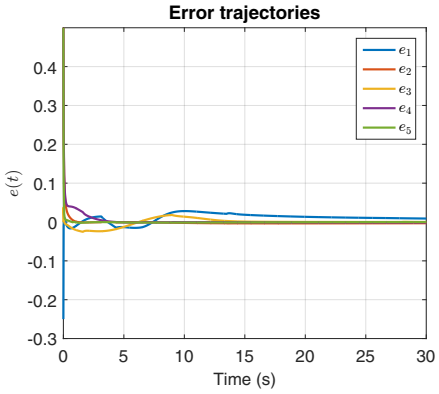

Fig. 3. Tracking error trajectories for the agents for the one-dimensional example.

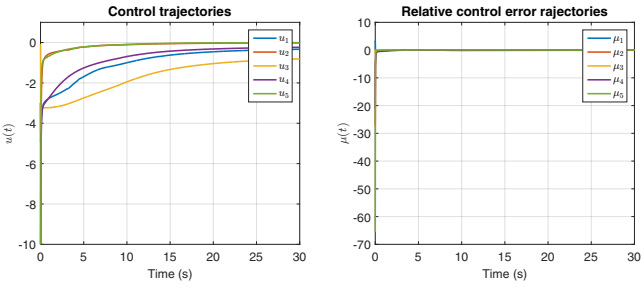

Fig. 4. Trajectories of the control input and the relative control error for all agents for the one-dimensional example.

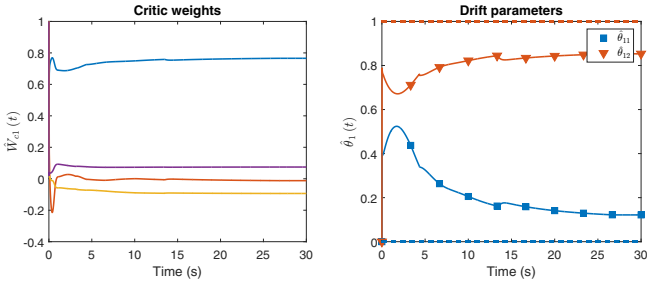

Fig. 5. Value function weights and drift dynamics parameters estimates for Agent 1 for the one-dimensional example. The dotted lines in the drift parameter plot are the ideal values of the drift parameters.

show the tracking error, the state trajectories compared with the desired trajectories, and the control inputs for all the agents demonstrating convergence to the desired formation and the desired trajectory. Note that Agents 2, 4, and 5 do not have a communication link to the leader, nor do they know their desired relative position with respect to the leader. The convergence to the desired formation is achieved via cooperative control based on decentralized objectives. Figure 5 shows the evolution and convergence of the value function weights and the parameters estimates for the drift dynamics for Agent 1. The errors between the ideal drift parameters and their respective estimates are large, however, as demonstrated by Figure 3, the resulting dynamics are sufficiently close to the actual dynamics for the developed technique to generate stabilizing policies. It is unclear whether the value function and the policy weights converge to their ideal values. Since an alternative method to solve this problem is not available to the best of the author's knowledge, a comparison between value function and policy weight estimates and their corresponding ideal values is infeasible.

XI. Concluding Remarks

A simulation-based actor-critic-identifier architecture is developed to obtain feedback-Nash equilibrium solutions to a class of differential graphical games. It is established that in a cooperative game based on minimization of the local neighborhood tracking errors, the value function corresponding to an agent depends on information obtained from all their extended neighbors. A set of coupled HJ equations are developed that serve as necessary and sufficient conditions for feedback-Nash equilibrium, and closed-form expressions for the feedback-Nash equilibrium policies are developed based on the HJ equations. The fact that the developed technique requires each agent to communicate with all of its extended neighbors motivates the search for a decentralized method to generate feedback-Nash equilibrium policies.

References

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

[2] D. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA: Athena Scientific, 2007, vol. 2.

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

[4] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[5] V. Konda and J. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2004.

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

[7] A. Heydari and S. Balakrishnan, "Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 145–157, 2013.

[8] W. Ren and R. W. Beard, *Distributed Consensus in Multi-Vehicle Cooperative Control*. New York: Springer-Verlag, 2008.

[9] E. Semsar-Kazerooni and K. Khorasani, *Team Cooperation in a Network of Multi-Vehicle Unmanned Systems: Synthesis of Consensus Algorithms*. Springer New York, 2013.

[10] R. Murray, "Recent research in cooperative control of multivehicle systems," *J. Dyn. Syst. Meas. Control*, vol. 129, pp. 571–583, 2007.

[11] M. Tidball and E. Altman, "Approximations in dynamic zero-sum games, i," *SIAM J. Control Optim.*, vol. 34, no. 1, pp. 311–328, Jan. 1996.

[12] M. Tidball, O. Pourtallier, and E. Altman, "Approximations in dynamic zero-sum games, ii," *SIAM J. Control Optim.*, vol. 35, no. 6, pp. 2101–2117, 1997.

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

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

[15] E. Altman, O. Pourtallier, A. Haurie, and F. Moresino, "Approximating nash equilibria in nonzero-sum games," *Int. Game Theory Rev.*, vol. 2, no. 2–3, pp. 155–172, 2000.

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

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

[18] K. G. Vamvoudakis and F. L. Lewis, "Online neural network solution of nonlinear two-player zero-sum games using synchronous policy iteration," in *Proc. IEEE Conf. Decis. Control*, 2010.

[19] D. Vrabie and F. L. 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.

[20] 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.

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

[22] X. Lin and C. G. Cassandras, "An optimal control approach to the multi-agent persistent monitoring problem in two-dimensional spaces," *IEEE Trans. Autom. Control*, vol. 60, no. 6, pp. 1659–1664, June 2015.

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

[24] A. Starr and C.-Y. Ho, "Nonzero-sum differential games," *J. Optim. Theory App.*, vol. 3, no. 3, pp. 184–206, 1969.

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

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

[27] A. Bressan and F. S. Priuli, "Infinite horizon noncooperative differential games," *J. Differ. Equ.*, vol. 227, no. 1, pp. 230 – 257, 2006.

[28] A. Bressan, "Noncooperative differential games," *Milan J. Math.*, vol. 79, no. 2, pp. 357–427, Dec. 2011.

[29] 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.

[30] J. Wang and M. Xin, "Integrated optimal formation control of multiple unmanned aerial vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 5, pp. 1731–1744, 2013.

[31] H. Zhang, T. Feng, G. H. Yang, and H. Liang, "Distributed cooperative optimal control for multiagent systems on directed graphs: An inverse optimal approach," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1315–1326, July 2015.

[32] S. Ghosh and J. W. Lee, "Optimal distributed finite-time consensus on unknown undirected graphs," *IEEE Trans. Control Netw. Syst.*, vol. 2, no. 4, pp. 323–334, December 2015.

[33] W. Lin, "Distributed uav formation control using differential game approach," *Aerosp. Sci. Technol.*, vol. 35, pp. 54–62, 2014.

[34] 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.

[35] D. H. Shim, H. J. Kim, and S. Sastry, "Decentralized nonlinear model predictive control of multiple flying robots," in *Proc. IEEE Conf. Decis. Control*, vol. 4, 2003, pp. 3621–3626.

[36] L. Magni and R. Scattolini, "Stabilizing decentralized model predictive control of nonlinear systems," *Automatica*, vol. 42, no. 7, pp. 1231 – 1236, 2006.

[37] A. Heydari and S. N. Balakrishnan, "An optimal tracking approach to formation control of nonlinear multi-agent systems," in *Proc. AIAA Guid. Navig. Control Conf.*, 2012.

[38] H. Zhang, J. Zhang, G. H. Yang, and Y. Luo, "Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 1, pp. 152–163, February 2015.

[39] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, to appear.

[40] 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.

[41] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press, 2012.

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

[43] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, 2016.

[44] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[45] 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, Jun. 2013, pp. 1322–1327.

[46] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.

[47] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[48] R. Kamalapurkar, "Model-based reinforcement learning for online approximate optimal control," Ph.D. dissertation, University of Florida, 2014.

[49] 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.

[50] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

Rushikesh Kamalapurkar received his M.S. and his Ph.D. degree in 2011 and 2014, respectively, from the Mechanical and Aerospace Engineering Department at the University of Florida. After working for a year as a postdoctoral research fellow with Dr. Warren E. Dixon, he was selected as the 2015-16 MAE postdoctoral teaching fellow. In 2016 he joined the School of Mechanical and Aerospace Engineering at the Oklahoma State University as an Assistant professor. His primary research interest has been intelligent, learning-based control of uncertain nonlinear dynamical systems. His work has been recognized by the 2015 University of Florida Department of Mechanical and Aerospace Engineering Best Dissertation Award, and the 2014 University of Florida Department of Mechanical and Aerospace Engineering Outstanding Graduate Research Award.

Justin R. Klotz received the Ph.D. degree in mechanical engineering from the University of Florida, Gainesville, FL, USA, in 2015, where he was awarded the Science, Mathematics and Research for Transformation (SMART) Scholarship, sponsored by the Department of Defense. His research interests include the development of Lyapunov-based techniques for reinforcement learning-based control, switching control methods, delay-affected control, and trust-based cooperative control.

Patrick Walters received the Ph.D. degree in mechanical engineering from the University of Florida, Gainesville, FL, USA, in 2015. His research interests include reinforcement learning-based feedback control, approximate dynamic programming, and robust control of uncertain nonlinear systems with a focus on the application of underwater vehicles.

Prof. Warren E. Dixon received his Ph.D. in 2000 from the Department of Electrical and Computer Engineering from Clemson University. He was selected as a Eugene P. Wigner Fellow at Oak Ridge National Laboratory (ORNL). In 2004, he joined the University of Florida in the Mechanical and Aerospace Engineering Department. His main research interest has been the development and application of Lyapunov-based control techniques for uncertain nonlinear systems. He has published 3 books, over a dozen chapters, and approximately 125 journal and 230 conference papers. His work has been recognized by the 2015 & 2009 American Automatic Control Council (AACC) O. Hugo Schuck (Best Paper) Award, the 2013 Fred Ellersick Award for Best Overall MILCOM Paper, a 2012-2013 University of Florida College of Engineering Doctoral Dissertation Mentoring Award, the 2011 American Society of Mechanical Engineers (ASME) Dynamics Systems and Control Division Outstanding Young Investigator Award, the 2006 IEEE Robotics and Automation Society (RAS) Early Academic Career Award, an NSF CAREER Award, the 2004 Department of Energy Outstanding Mentor Award, and the 2001 ORNL Early Career Award for Engineering Achievement. He is a Fellow of ASME and IEEE and is an IEEE Control Systems Society (CSS) Distinguished Lecturer. He has served as the Director of Operations for the Executive Committee of the IEEE CSS Board of Governors and as a member of the U.S. Air Force Science Advisory Board. He is currently or formerly an associate editor for ASME Journal of Journal of Dynamic Systems, Measurement and Control, Automatica, IEEE Control Systems Magazine, IEEE Transactions on Systems Man and Cybernetics: Part B Cybernetics, and the International Journal of Robust and Nonlinear Control.