Adaptive Optimal Control for Stochastic Multiplayer Differential Games Using On-Policy and Off-Policy Reinforcement Learning

Mushuang Liu, *Student Member, IEEE*, Yan Wan, *Senior Member, IEEE*, Frank L. Lewis, *Life Fellow, IEEE*, and Victor G. Lopez, *Student Member, IEEE*

***Abstract*— Control-theoretic differential games have been used to solve optimal control problems in multiplayer systems. Most existing studies on differential games either assume deterministic dynamics or dynamics corrupted with additive noise. In realistic environments, multidimensional environmental uncertainties often modulate system dynamics in a more complicated fashion. In this article, we study stochastic multiplayer differential games, where the players' dynamics are modulated by randomly time-varying parameters. We first formulate two differential games for systems of general uncertain linear dynamics, including the two-player zero-sum and multiplayer nonzero-sum games. We then show that optimal control policies, which constitute the Nash equilibrium solutions, can be derived from the corresponding Hamiltonian functions. Stability is proven using the Lyapunov type of analysis. In order to solve the stochastic differential games online, we integrate reinforcement learning (RL) and an effective uncertainty sampling method called the multivariate probabilistic collocation method (MPCM). Two learning algorithms, including the on-policy integral RL (IRL) and off-policy IRL, are designed for the formulated games, respectively. We show that the proposed learning algorithms can effectively find the Nash equilibrium solutions for the stochastic multiplayer differential games.**

***Index Terms*— Integral reinforcement learning (IRL), multiplayer systems, neural networks (NNs), systems with randomly time-varying parameters, uncertainty quantification.**

I. Introduction

GAME theory has been widely used in multiplayer systems to obtain decisions that optimize individual payoffs [1]–[6]. In the traditional game theory, a player finds the best strategy to minimize a static and immediate cost [1]–[3]. Recently, differential games were combined with control theory to study dynamical systems that involve the evolution of the players' payoff functions [5]–[7]. Widely used differential games include the two-player zero-sum game, which provides solutions for pursuit–evasion games and H_∞ design for disturbance attenuation [7], and the multiplayer nonzero-sum game, which finds applications in, e.g., the control of transportation networks and the cooperative control of multiple robots with individual goals [6]. Most existing studies on differential games assume deterministic dynamics. In reality, multidimensional uncertainties, such as uncertain player intentions and environmental impacts, often modulate system dynamics in a complicated fashion. As such, in this article, we formulate and study practical Nash solutions for new stochastic two-player zero-sum and multiplayer nonzero-sum games, where the system dynamics are modulated by multidimensional time-varying random parameters.

For deterministic differential games, the Nash equilibrium solutions rely on solving the Hamilton–Jacobi–Bellman (HJB) equations for nonlinear systems or the game algebraic Riccati equation (GARE) for linear systems. However, solving HJB or GARE equations analytically is difficult or even impossible [6]. Moreover, this method requires the full knowledge of system dynamics, and only provides an offline solution. As such, the reinforcement learning (RL) method has been developed to solve the differential games online [8]–[12]. We also explore RL to develop online solutions in this article for the new games with dynamics modulated by uncertainties.

The RL method was developed based on the idea that successful decisions should be remembered as a reinforcement signal, such that they are more likely to be used in future decisions [13]–[19]. RL has been used to find the Nash equilibrium solutions online for multiplayer differential games. In particular, for the two-player zero-sum game, [8] presented an adaptive dynamic programming (ADP)-based learning algorithm and used integral RL (IRL) to find the optimal policies online. However, the developed method uses a two-loop iteration algorithm to update the policies of the two players in sequence, which can be time-consuming. To deal with this problem, [20] developed a single-loop iteration algorithm that updates the two players' control policies simultaneously. In addition, to deal with the systems with unknown dynamics, [21] developed a model-free IRL for the two-player zero-sum differential game using Q-learning. For the multiplayer nonzero-sum game, [11] developed an ADP algorithm that finds the Nash equilibrium online using IRL and partial information of the system dynamics. To deal with the systems of totally unknown dynamics, [12] established an off-policy IRL to solve the nonlinear continuous-time

Manuscript received February 12, 2019; revised October 19, 2019; accepted January 11, 2020. Date of publication February 27, 2020; date of current version December 1, 2020. This work was supported in part by the Office of Naval Research (ONR) under Grant N00014-18-1-2221 and in part by the NSF under Grant 1714519 and Grant 1839804. *(Corresponding author: Yan Wan.)*

Mushuang Liu, Yan Wan, and Victor G. Lopez are with the Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: mushuang.liu@mavs.uta.edu; yan.wan@uta.edu; victor.lopezmejia@mavs.uta.edu).

Frank L. Lewis is with the UTA Research Institute, The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: lewis@uta.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2969215

multiplayer nonzero-sum games at the cost of additional computation. The off-policy method solves the value function and optimal control policies simultaneously using both critic and actor neural networks (NNs) and does not require knowledge of the system dynamics. All these aforementioned studies assume time-invariant and deterministic system dynamics.

To address uncertainties in differential games operating in realistic environments, practical uncertainty evaluation methods are needed to evaluate expected costs [22]–[26]. The most widely used simulation-based uncertainty evaluation methods are the Monte Carlo (MC) method and its variants, including the Markov chain MC and sequential MC [27]–[29]. However, the MC-based methods require a large number of simulations to estimate the expected cost function accurately, which makes them unrealistic for online algorithms. To improve the computational efficiency, other uncertainty sampling methods, including the Latin hypercube sampling [30], importance sampling [31], multilevel MC [32], and greedy and adaptive sampling [33], [34] have also been developed. However, none of them can estimate expected system outputs accurately with a computational load. To deal with this challenge, [35] and [36] developed effective uncertainty evaluation methods, named multivariate probabilistic collocation method (MPCM) and its variant (MPCM-OFFD), which accurately estimate the expected output mean of a system mapping by smartly selecting a small set of samples according to the uncertainties' statistics (e.g., probability density functions). References [37] and [38] further integrated the MPCM with the discrete-time RL to solve optimal control problems online for uncertain systems. Here, in this article, we study the integrated MPCM and IRL to effectively solve stochastic multiplayer differential games online.

This article, for the first time in the literature to the best of our knowledge, analyzes multiplayer differential games for systems modulated by general randomly time-varying parameters and develops effective online learning methods to solve such stochastic games. The main contributions of this article are fourfold: 1) the formulation of two-player zero-sum and multiplayer nonzero-sum games for systems modulated by time-varying random parameters, which captures stochastic environmental impacts and random player intentions [39]–[41]; 2) the analysis of the formulated differential game properties, including the stability and the Nash equilibrium; 3) a novel policy iteration algorithm that integrates IRL and an effective uncertainty sampling method, i.e., MPCM, to provide an effective online solution for these stochastic games; and 4) the integration of off-policy IRL and the MPCM to solve these stochastic games online without knowing the system dynamics.

The rest of this article is organized as follows. Section II formulates the stochastic two-player zero-sum and multiplayer nonzero-sum games and presents the preliminaries to facilitate the analysis in this article. Sections III and IV study the properties and online solutions of these two stochastic games. Section V presents the simulation studies that demonstrate performances of the proposed solutions, and Section VI concludes this article.

II. Problem Formulation and Preliminaries

In this section, we first formulate two stochastic multiplayer games with general linear uncertain dynamics, including the two-player zero-sum and multiplayer nonzero-sum games. Preliminaries are then introduced to facilitate the analysis in this article.

A. Problem Formulation

Game 1 (Stochastic Two-Player Zero-Sum Game): Consider a generic two-player linear system with a randomly time-varying vector $\mathbf{a}(t)$ of dimension m

$$\dot{\mathbf{x}} = A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d} \tag{1}$$

where $\mathbf{x} = \mathbf{x}(t) \in \mathbb{R}^n$ is the system state vector, $\mathbf{u} = \mathbf{u}(t) \in \mathbb{R}^p$ is the control input, and $\mathbf{d} = \mathbf{d}(t) \in \mathbb{R}^q$ is the adversarial control input. $\mathbf{A}(\mathbf{a})$, $\mathbf{B}$, and $\mathbf{C}$ are the drift dynamics, input dynamics, and adversarial input dynamics, respectively. Each element of $\mathbf{a}(t)$, $a_p(t)$ $(p = 1, 2, \ldots, m)$, changes independently over time with pdf $f_{A_p}(a_p(t))$, and the sample functions of $a_p(t)$ are well-behaved so that the sample equations for (1) are ordinary differential equations [42], [43].

This stochastic game formulation has a wide range of potential applications, e.g., the pursuit–evasion games and H_∞ design for disturbance attenuation in real environments modulated by uncertain parameters. One specific example is the aircraft dynamics described as $\dot{\mathbf{v}}(t) = -K\mathbf{v}(t) + \mathbf{F}_u(t) + \mathbf{F}_d(t)$. Here, $\mathbf{v}$ is the velocity, $\mathbf{F}_u(t)$ is the controlled thrust force, $\mathbf{F}_d(t)$ is the disturbance force, and K is the air resistance coefficient. The air resistance coefficient, related to air density and air humidity, is a randomly time-varying parameter affected by uncertain weather conditions. The statistics (e.g., pdfs) of such weather conditions can be obtained from probabilistic weather forecasts.

The expected cost to optimize is

$$\begin{aligned} J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}) &= E\left[\int_0^\infty r(\mathbf{x}, \mathbf{u}, \mathbf{d})dt\right] \\ &= E\left[\int_0^\infty (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2)dt\right] \end{aligned} \tag{2}$$

where $\mathbf{Q}$ and $\mathbf{R}$ are positive semidefinite and positive definite matrices, respectively, and γ is the amount of attenuation from the disturbance input to the defined performance.

The value function $V(\mathbf{x}(t))$ corresponding to the performance index is defined as

$$V(\mathbf{x}(t)) = E\left[\int_t^\infty (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2)d\tau\right]. \tag{3}$$

Define the two-player zero-sum differential game as

$$V^*(\mathbf{x}(0)) = \min_{\mathbf{u}} \max_{\mathbf{d}} J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}) \tag{4}$$

where $V^*(\mathbf{x}(0))$ is the optimal value function. In the two-player zero-sum game, one player $\mathbf{u}$ seeks to minimize the value function, and the other $\mathbf{d}$ seeks to maximize it.

Game 2 (Stochastic Multiplayer Nonzero-Sum Game): Consider a generic N-player linear system with a time-varying

uncertain vector $\mathbf{a}(t)$ of dimension m. The system dynamics is

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}\mathbf{u}_j \tag{5}$$

where $\mathbf{x} = \mathbf{x}(t) \in \mathbb{R}^n$ is the system state vector, $\mathbf{u}_j = \mathbf{u}_j(t) \in \mathbb{R}^p$ is the control input of player j, and $\mathbf{A}(\mathbf{a})$ and $\mathbf{B}$ are the drift dynamics and input dynamics, respectively. Each element of $\mathbf{a}(t)$, $a_p(t)$ $(p = 1, 2, \ldots, m)$, changes independently over time with pdf $f_{A_p}(a_p(t))$, and the sample functions of $a_p(t)$ are well-behaved so that the sample equations (5) are ordinary differential equations [42], [43].

The expected cost to optimize for player i is

$$\begin{aligned} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}) &= E\left[\int_0^\infty r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i})dt\right] \\ &= E\left[\int_0^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right)dt\right] \end{aligned} \tag{6}$$

where $\mathbf{u}_{-i}$ is the supplementary set of $\mathbf{u}_i$: $\mathbf{u}_{-i} = \{\mathbf{u}_j, j \in (1, 2, \ldots, i-1, i+1, \ldots, N)\}$. $\mathbf{Q}_i$ and $\mathbf{R}_{ij}$ $(i \neq j)$ are positive semidefinite matrices, and $\mathbf{R}_{ii}$ is positive definite.

The value function for player i is defined as

$$V_i(\mathbf{x}(t)) = E\left[\int_t^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right)d\tau\right]. \tag{7}$$

Define the multiplayer differential game as

$$V_i^*(\mathbf{x}(0)) = \min_{\mathbf{u}_i} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}) \tag{8}$$

where $V_i^*(\mathbf{x}(0))$ is the optimal value function for player i. In the multiplayer game, each player tries to minimize its cost by choosing its control policy $\mathbf{u}_i$ based on the full-state information of the system.

B. Preliminaries

Definition 1 [42]: The equilibrium solution of a system is said to be stable in the mean (norm) if for any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$, such that for any initial condition satisfying $\|\mathbf{x}_0\| < \delta(\epsilon)$

$$E\{\|\mathbf{x}(t)\|\} < \epsilon$$

for all $t \geq t_0$.

It is assumed that the system described in (1) is stabilizable in the mean, that is, there exist control policies $\mathbf{u} = -K_u\mathbf{x}$ and $\mathbf{d} = -K_d\mathbf{x}$, such that the closed-loop system $\dot{\mathbf{x}} = (\mathbf{A}(\mathbf{a}) - \mathbf{B}K_u - \mathbf{C}K_d)\mathbf{x}$ is stable in the mean.

Definition 2 [42]: The equilibrium solution is said to be asymptotically stable in the mean (norm) if it is stable in the mean and moreover, there exists a $\delta(t_0) > 0$, such that for any initial condition satisfying $\|x_0\| < \delta(t_0)$

$$\lim_{t\to\infty} E\{\|\mathbf{x}(t)\|\} \to 0.$$

Definition 3 [44]: The system (1) is said to have L_2-gain less than or equal to γ if the following disturbance attenuation condition is satisfied for all $T \geq 0$ and all $\mathbf{d} \in L_2[0, \infty)$ with $\mathbf{x}(0) = \mathbf{0}$, where $\mathbf{0}$ is a zero matrix with proper dimensions

$$\frac{\int_0^T \|\mathbf{z}(\tau)\|^2 d\tau}{\int_0^T \|\mathbf{d}(\tau)\|^2 d\tau} \leq \gamma^2$$

where $\|\mathbf{z}(t)\|^2 = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}$, $\mathbf{d}(t)$ is the disturbance input, and γ is the amount of attenuation.

It is assumed that γ in (2) satisfies $\gamma > \gamma^*$, where γ^* is the smallest γ that satisfies the disturbance attenuation condition for all possible $\mathbf{A}(\mathbf{a})$, to make sure that the system is always stabilizable.

Definition 4 [5]: Policies $\{\mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_N^*\}$ are said to constitute a Nash equilibrium solution for the N-player game if the following equation is satisfied for $\forall \mathbf{u}_i \forall i \in N$:

$$J_i^*(\mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_i^*, \ldots, \mathbf{u}_N^*) \leq J_i(\mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_N^*).$$

The N-tuple $\{J_1^*, J_2^*, \ldots, J_N^*\}$ is known as a Nash equilibrium value set of the N-player game.

Lemma 1 [42, Th. II]: Consider a system $\dot{\mathbf{x}} = f(\mathbf{x}(t), \mathbf{a}(t))$, where $\mathbf{a}(t)$ is a vector of time-varying random parameters. If there exists a Lyapunov function $\tilde{V}(\mathbf{x}(t))$ defined over the state space and satisfies the conditions listed as follows (1–4), then the equilibrium solution of the system is asymptotically stable in the mean.

1) $\tilde{V}(\mathbf{0}) = 0$.
2) $\tilde{V}(\mathbf{x}(t))$ is continuous with both $\mathbf{x}$ and t, and the first partial derivatives in these variables exist.
3) $\tilde{V}(\mathbf{x}(t)) \geq b\|\mathbf{x}\|$ for some $b > 0$.
4) $E\left[\dot{\tilde{V}}(\mathbf{x}(t))\right]$ is a negative definite.

III. Stochastic Two-Player Zero-Sum Game

In this section, we study the properties and optimal solutions of the stochastic two-player zero-sum game. Section III-A studies the stability and the Nash equilibrium of the proposed game, and Section III-B develops both on-policy and off-policy IRL solutions to solve the game online.

A. Stability and Nash Equilibrium

With the value function defined in (3), the following Bellman equation can be derived by taking derivative of $V(\mathbf{x}(t))$ with respect to time t

$$r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E\left[\frac{\partial V^{\mathrm{T}}}{\partial \mathbf{x}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})\right] = 0. \tag{9}$$

with the Hamiltonian function

$$\begin{aligned} H\left(\mathbf{x}, \mathbf{u}, \mathbf{d}, \frac{\partial V}{\partial \mathbf{x}}\right) &= r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E\left[\frac{\partial V^{\mathrm{T}}}{\partial \mathbf{x}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})\right]. \end{aligned} \tag{10}$$

The optimal control policies $\mathbf{u}^*$ and $\mathbf{d}^*$ can be derived by employing the stationary conditions in the Hamiltonian function [5, p. 447]

$$\begin{aligned} \frac{\partial H}{\partial \mathbf{u}} = 0 \to \mathbf{u}^* &= -\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V^*}{\partial \mathbf{x}} \\ \frac{\partial H}{\partial \mathbf{d}} = 0 \to \mathbf{d}^* &= \frac{1}{2\gamma^2}\mathbf{C}^{\mathrm{T}}\frac{\partial V^*}{\partial \mathbf{x}}. \end{aligned} \tag{11}$$

Substituting (11) into the Bellman equation (9), the following HJB equation is obtained

$$H\left(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X^*\right) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + E\left[V_X^{*\mathrm{T}} A(\mathbf{a})\mathbf{x} - \frac{1}{4}V_X^{*\mathrm{T}}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}V_X^* + \frac{1}{4\gamma^2}V_X^{*\mathrm{T}}\mathbf{C}\mathbf{C}^{\mathrm{T}}V_X^*\right] = 0, \quad V(\mathbf{0}) = 0 \quad (12)$$

where $V_X^* = \partial V^*/\partial \mathbf{x}$.

Note that the HJB equation (12) contains the randomly time-varying vector, $\mathbf{a}(t)$. Compared with the HJB equation defined in deterministic systems, (12) is harder to solve, as it involves the evaluation of uncertainty, which can be computationally expensive. In Section III-B, we introduce an effective uncertainty evaluation method and show its integration with learning methods to solve the HJB equation (12) online.

Lemma 2: For any admissible control and disturbance policies $\mathbf{u}$ and $\mathbf{d}$, let $V \geq 0$ be the corresponding solution to the Bellman equation (10), and then, the following equation holds [5, Lemma 10.2-1]:

$$H(\mathbf{x}, \mathbf{u}, \mathbf{d}, V_X) = H(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X) + (\mathbf{u} - \mathbf{u}^*)^T\mathbf{R}(\mathbf{u} - \mathbf{u}^*) - \gamma^2\|\mathbf{d} - \mathbf{d}^*\|^2$$

where $\mathbf{u}^*$ and $\mathbf{d}^*$ are described in (11), and $V_X = \partial V/\partial \mathbf{x}$.

Proof: See Appendix A. □

Theorem 1: Let $V(\mathbf{x}(t)) > 0$ be a smooth function satisfying the HJB equation described in (12), and then, the following statements hold.

1) The system (1) is asymptotically stable in the mean with the policies $\mathbf{u}^*$ and $\mathbf{d}^*$ described in (11).
2) The solution (i.e., policies $\mathbf{u}^*$ and $\mathbf{d}^*$) derived in (11) provides a saddle point solution to the game, and the system is in Nash equilibrium with this solution.

Proof: See Appendix B. □

B. Approximate Solutions Using On-Policy and Off-Policy IRL and the MPCM

Solving the HJB equation (12) analytically is extremely difficult or even impossible [44]. Here, we integrate IRL and the MPCM to provide effective online algorithms to approximate the solution of the HJB equation.

The IRL Bellman equation can be written as

$$V(\mathbf{x}(t)) = E\left[\int_t^{t+T} r(\mathbf{x}(\tau), \mathbf{u}(\tau), \mathbf{d}(\tau))d\tau + V(\mathbf{x}(t+T))\right] \quad (13)$$

where T is the time interval.

It is assumed that there exists a weight $\mathbf{W}$, such that the value function is approximated as

$$V(\mathbf{x}) = \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}) \quad (14)$$

where $\phi(\mathbf{x})$ is the polynomial basis function vector.

1) On-Policy IRL: With the value function approximation (VFA), one can find the optimal solution from the policy iteration (PI) algorithm by iteratively conducting two steps: policy evaluation that evaluates the value function $V(\mathbf{x})$ using (13) and policy improvement [5] that finds the optimal solution based on the current approximated value function using (11). For systems with uncertain system dynamics, the policy evaluation step involves uncertainty evaluation, which is typically solved by the MC method, too slow to be used for online solutions.

Here, we utilize an effective uncertainty evaluation method, called the MPCM [35]. To map to the MPCM framework, we denote the generic function whose expectation to be evaluated as $G(a_1, a_2, \ldots, a_m)$, which is modulated by m uncertain parameters, i.e., $a_1, a_2, \ldots, a_m$, with the degree of each parameter up to $2n_p - 1$, where $p = 1, 2, \ldots, m$. The MPCM accurately evaluates the output mean of G by conducting the following three steps: 1) selecting a limited number of sample points according to the Gaussian quadrature rules and the pdfs of the uncertain parameters, i.e., $f_{A_p}(a_p(t))$; 2) evaluating the system outputs at selected sample points; and 3) finding the output mean of G from a reduced-order mapping G'. The properties of the MPCM are briefly described in the following lemma. For the detailed MPCM design procedure, please refer to [35].

Lemma 3 [35, Th. 2]: Consider a system mapping modulated by m independent uncertain parameters

$$G(a_1, a_2, \ldots, a_m) = \sum_{q_1=0}^{2n_1-1}\sum_{q_2=0}^{2n_2-1}\cdots\sum_{q_m=0}^{2n_m-1}\psi_{q_1,q_2,\ldots,q_m}\prod_{p=1}^{m}a_p^{q_p} \quad (15)$$

where a_p is an uncertain parameter with the degree up to $2n_p - 1$, $p = 1, 2, \ldots, m$. n_p is a positive integer, and $\psi_{q_1,q_2,\ldots,q_m} \in \mathbb{R}$ are the coefficients. Each uncertain parameter a_p follows an independent pdf $f_{A_p}(a_p)$. The MPCM approximates $G(a_1, a_2, \ldots, a_m)$ with the following low-order mapping:

$$G'(a_1, a_2, \ldots, a_m) = \sum_{q_1=0}^{n_1-1}\sum_{q_2=0}^{n_2-1}\cdots\sum_{q_m=0}^{n_m-1}\Omega_{q_1,q_2,\ldots,q_m}\prod_{p=1}^{m}a_p^{q_p}$$

with $E[G(a_1, a_2, \ldots, a_m)] = E[G'(a_1, a_2, \ldots, a_m)]$, where $\Omega_{q_1,q_2,\ldots,q_m} \in \mathbb{R}$ are coefficients.

As shown in Lemma 3, the MPCM reduces the number of simulations from $2^m\prod_{p=1}^m n_p$ to $\prod_{p=1}^m n_p$. Despite the significant reduction of computation by 2^m, the MPCM accurately predicts the output mean [35]. Here, we integrate the MPCM into IRL to provide effective online learning-based solutions for differential games of systems with randomly time-varying parameters.

Define a system mapping subject to uncertain parameters $\mathbf{a}(t)$, $G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a}) = \int_t^{t+T} r(\mathbf{x}(\tau), \mathbf{u}(\tau), \mathbf{d}(\tau))d\tau + V(\mathbf{x}(t+T))$. Given the current system state $\mathbf{x}(t)$ and admissible control and disturbance policies $\mathbf{u}(t)$ and $\mathbf{d}(t)$, the value function described in (13) can be approximated by the mean output of the system mapping $G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})$ (i.e., $V(\mathbf{x}) = E\left[G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})\right]$), using the MPCM. In particular, we select a set of samples based on the pdfs of uncertain parameters, $f_{A_p}(a_p)$, and run simulations at these

samples to estimate $E\left[G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})\right]$. Under the assumption that each uncertain parameter a_p has a degree up to $2n_p-1$, $G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})$ has the following form:

$$G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a}) = \sum_{q_1=0}^{2n_1-1}\sum_{q_2=0}^{2n_2-1}\cdots\sum_{q_m=0}^{2n_m-1}\psi_{q_1,q_2,\ldots,q_m}(\mathbf{x},\mathbf{u},\mathbf{d})\prod_{p=1}^{m}a_p^{q_p}. \quad (16)$$

With this mapping, the value function can be estimated from the mean output of a reduced-order mapping according to Lemma 3 as

$$V(\mathbf{x}(t)) = E[G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})] = E\left[G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})\right] \quad (17)$$

where $G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})$ is the reduced-order mapping derived from the MPCM procedure [35]

$$G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a}) = \sum_{q_1=0}^{n_1-1}\sum_{q_2=0}^{n_2-1}\cdots\sum_{q_m=0}^{n_m-1}\Omega_{q_1,q_2,\ldots,q_m}(\mathbf{x},\mathbf{u},\mathbf{d})\prod_{p=1}^{m}a_p^{q_p}. \quad (18)$$

The PI algorithm that integrates IRL and the MPCM for the two-player zero-sum game with uncertain system dynamics is summarized in Algorithm 1.

Theorem 2: Consider the stochastic two-player zero-sum game shown in (1)–(4). The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume the following: 1) VFA in (14) holds; 2) the relation between the value function $V(\mathbf{x}(t))$ and the uncertain parameters $\mathbf{a}(t)$ can be approximated by a polynomial system mapping (19) with the form of (15); and 3) Algorithm 1 converges. Then, the policies $\mathbf{u}$ and $\mathbf{d}$ derived from Algorithm 1 are optimal policies.

Proof: See Appendix C. □

2) Off-Policy IRL: Algorithm 1 learns the optimal solution online with knowledge of the system dynamics (i.e., matrix $\mathbf{B}$ and $\mathbf{C}$). In addition, the on-policy learning algorithm requires both control and disturbance policies to be adjustable to learn the optimal solution.

In this section, we provide an off-policy IRL algorithm and use three NNs, including critic NN, actor NN, and disturbance NN, to learn the optimal solution online without requiring to know the system dynamics, i.e., matrix $\mathbf{B}$ and $\mathbf{C}$, or manipulating the disturbance policies.

To this end, we introduce auxiliary variables $\mathbf{u}^{(s)}$ and $\mathbf{d}^{(s)}$, and hence, the system dynamics described in (1) is further written as

$$\dot{\mathbf{x}} = A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u}^{(s)} + \mathbf{C}\mathbf{d}^{(s)} + \mathbf{B}\left(\mathbf{u}-\mathbf{u}^{(s)}\right) + \mathbf{C}\left(\mathbf{d}-\mathbf{d}^{(s)}\right) \quad (21)$$

where $\mathbf{u}$ and $\mathbf{d}$ are behavior policies applied to the system to generate data for learning, and $\mathbf{u}^{(s)}$ and $\mathbf{d}^{(s)}$ are the desired policies to be updated.

Algorithm 1 Policy Iteration Algorithm for Two-Player Zero-Sum Game With Uncertain System Dynamics

1: Initialize the players with initial state $\mathbf{x}(0)$ and admissible control and disturbance policies $\mathbf{u}(0)$ and $\mathbf{d}(0)$.

2: Apply the MPCM procedure [35, Section II] to select a set of samples for the uncertain vector $\mathbf{a}(t)$.

3: For each iteration s, find the value of

$$\int_t^{t+T} r\left(\mathbf{x}(\tau),\mathbf{u}^{(s)}(\tau),\mathbf{d}^{(s)}(\tau)\right)d\tau + \mathbf{W}^{(s+1)^{\mathrm{T}}}\phi(\mathbf{x}(t+T))$$

at each MPCM sample.

4: Find the value function $V^{(s)}(\mathbf{x}(t))$ using the MPCM [35], which is the mean output of the mapping $G_{V^{(s)}}(\cdot)$ subject to uncertain parameters $\mathbf{a}(t)$,

$$G_{V^{(s)}}\left(\mathbf{x},\mathbf{u}^{(s)},\mathbf{d}^{(s)},\mathbf{a}\right) = \mathbf{W}^{(s+1)^{\mathrm{T}}}\phi(\mathbf{x}(t+T)) + \int_t^{t+T} r\left(\mathbf{x}(\tau),\mathbf{u}^{(s)}(\tau),\mathbf{d}^{(s)}(\tau)\right)d\tau. \quad (19)$$

5: Update the value function weight vector $\mathbf{W}^{(s)}$ according to the estimated $V^{(s)}(\mathbf{x}(t))$.

$$\mathbf{W}^{(s)^{\mathrm{T}}}\phi(\mathbf{x}(t)) = V^{(s)}(\mathbf{x}(t)).$$

6: Update the control and disturbance policies $\mathbf{u}^{(s+1)}$ and $\mathbf{d}^{(s+1)}$ as

$$\mathbf{u}^{(s+1)} = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V^{(s)}}{\partial \mathbf{x}}$$
$$\mathbf{d}^{(s+1)} = \frac{1}{2\gamma^2}\mathbf{C}^{\mathrm{T}}\frac{\partial V^{(s)}}{\partial \mathbf{x}}. \quad (20)$$

7: Repeat procedures 3−6.

Differentiating the value function $V^{(s)}(\mathbf{x}(t))$ of the system (21), one has

$$\begin{aligned}\dot{V}^{(s)}(\mathbf{x}(t)) &= E\left[V_X^{(s)^{\mathrm{T}}}\left(A(\mathbf{a})\mathbf{x}+\mathbf{B}\mathbf{u}^{(s)}+\mathbf{C}\mathbf{d}^{(s)}\right)\right] \\ &\quad + V_X^{(s)^{\mathrm{T}}}\left(\mathbf{B}\left(\mathbf{u}-\mathbf{u}^{(s)}\right)+\mathbf{C}\left(\mathbf{d}-\mathbf{d}^{(s)}\right)\right) \\ &= -\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{(s)^{\mathrm{T}}}\mathbf{R}\mathbf{u}^{(s)}-\gamma^2\|\mathbf{d}^{(s)}\|^2\right) \\ &\quad - 2\mathbf{u}^{(s+1)^{\mathrm{T}}}\mathbf{R}\left(\mathbf{u}-\mathbf{u}^{(s)}\right) \\ &\quad + 2\gamma^2\mathbf{d}^{(s+1)^{\mathrm{T}}}\left(\mathbf{d}-\mathbf{d}^{(s)}\right). \end{aligned} \quad (22)$$

The second equality is obtained by combining the Hamiltonian functions (10) and (20).

Integrating both sides in (22), one has

$$\begin{aligned} V^{(s)}(\mathbf{x}(t+T)) - V^{(s)}(\mathbf{x}(t)) &= E\left[\int_t^{t+T}\left(-\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}-\mathbf{u}^{(s)^{\mathrm{T}}}\mathbf{R}\mathbf{u}^{(s)}+\gamma^2\|\mathbf{d}^{(s)}\|^2\right)d\tau\right] \\ &\quad + \int_t^{t+T}\left(-2\mathbf{u}^{(s+1)^{\mathrm{T}}}\mathbf{R}\left(\mathbf{u}-\mathbf{u}^{(s)}\right)\right. \\ &\quad \left. + 2\gamma^2\mathbf{d}^{(s+1)^{\mathrm{T}}}\left(\mathbf{d}-\mathbf{d}^{(s)}\right)\right)d\tau. \end{aligned} \quad (23)$$

Algorithm 2 Off-Policy IRL for Two-Player Zero-Sum Game With Uncertain System Dynamics

1: Initialize the players with initial state $\mathbf{x}(0)$ and admissible control and disturbance policies $\mathbf{u}(0)$ and $\mathbf{d}(0)$.
2: Apply the MPCM procedure [35, Section II] to select a set of samples for the uncertain vector $\mathbf{a}(t)$.
3: For each iteration s, find the value of

$$\int_t^{t+T} \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{(s)^{\mathrm{T}}}\mathbf{R}\mathbf{u}^{(s)} - \gamma^2\|\mathbf{d}^{(s)}\|^2\right) d\tau + V^{(s)}(\mathbf{x}(t+T)) \tag{25}$$

at each MPCM sample.
4: Find the mean output of mapping $G^o_{V^{(s)}}(\cdot)$ subject to uncertain parameters $\mathbf{a}(t)$ using the MPCM [35],

$$G^o_{V^{(s)}}\left(\mathbf{x}, \mathbf{u}^{(s)}, \mathbf{d}^{(s)}, \mathbf{a}\right) = V^{(s)}(\mathbf{x}(t+T)) + \int_t^{t+T} \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{(s)^{\mathrm{T}}}\mathbf{R}\mathbf{u}^{(s)} - \gamma^2\|\mathbf{d}^{(s)}\|^2\right) d\tau. \tag{26}$$

5: Solve the following equation for $V^{(s)}(\mathbf{x})$, $\mathbf{u}^{(s+1)}$, and $\mathbf{d}^{(s+1)}$ simultaneously.

$$E\left[G^o_{V^{(s)}}\left(\mathbf{x}, \mathbf{u}^{(s)}, \mathbf{d}^{(s)}, \mathbf{a}\right)\right] = V^{(s)}(\mathbf{x}(t)) - \int_t^{t+T} \left(2\mathbf{u}^{(s+1)^{\mathrm{T}}}\mathbf{R}\left(\mathbf{u} - \mathbf{u}^{(s)}\right) - 2\gamma^2\mathbf{d}^{(s+1)^{\mathrm{T}}}\left(\mathbf{d} - \mathbf{d}^{(s)}\right)\right) d\tau. \tag{27}$$

6: Repeat procedures 3–5.

Note that for any fixed admissible control and disturbance behavior policies $\mathbf{u}$ and $\mathbf{d}$, (23) can be solved for the value function $V^{(s)}$ and the optimal control and disturbance polices $\mathbf{u}^{(s+1)}$ and $\mathbf{d}^{(s+1)}$ simultaneously using the following NNs:

$$\begin{aligned} V^{(s)}(\mathbf{x}) &= \mathbf{W}^{(s)^{\mathrm{T}}}\phi(\mathbf{x}) \\ \mathbf{u}^{(s+1)}(\mathbf{x}) &= \mathbf{W}_{\mathbf{u}}^{(s+1)^{\mathrm{T}}}\phi_{\mathbf{u}}(\mathbf{x}) \\ \mathbf{d}^{(s+1)}(\mathbf{x}) &= \mathbf{W}_{\mathbf{d}}^{(s+1)^{\mathrm{T}}}\phi_{\mathbf{d}}(\mathbf{x}). \end{aligned} \tag{24}$$

The off-policy IRL and the MPCM is described in Algorithm 2.

Theorem 3: Consider the stochastic two-player zero-sum game shown in (1)–(4). The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume the following: 1) VFA in (24) holds; 2) the relation between the value function $V(\mathbf{x}(t))$ and the uncertain parameters $\mathbf{a}(t)$ can be approximated by a polynomial system mapping (26) with the form of (15); and 3) Algorithm 2 converges. Then, the policies derived from off-policy IRL described in Algorithm 2 are optimal policies.

Proof: See Appendix D. □

Remark 1: In both algorithms, the disturbance needs to be measurable. For the off-policy algorithm, the disturbance policy is not required to be adjustable. In particular, in the off-policy algorithm, the control and disturbance policies $\mathbf{u}$ and $\mathbf{d}$ that are applied to the system can be different from the control and disturbance policies $\mathbf{u}^{(s)}$ and $\mathbf{d}^{(s)}$ that are evaluated and updated. As such, in contrast to the on-policy IRL, the applied disturbance input $\mathbf{d}$ in the off-policy IRL can be the actual external disturbance that is not adjustable.

Remark 2: Note that the admissible control and disturbance policies initialize the first step in Algorithm 2. They refer to control and disturbance policies that can make the system stable. In the off-policy IRL, the exact system dynamics $\mathbf{B}$ and $\mathbf{C}$ are unknown. However, the ranges of parameters in the system dynamics are often available due to the system's physical properties to obtain an estimated range of admissible control policies to initialize the off-policy IRL algorithm. It is also, often, of practice to first try a PID controller for an unknown system, which gives a range of admissible control policies for the initialization step.

Remark 3: Algorithms 1 and 2 integrate IRL and MPCM, for the first time in the literature, to solve the stochastic two-player zero-sum game. The uncertainty evaluation in such stochastic optimal control problems is typically solved by MC method and its variants, which is time-consuming to use for online solutions. The proposed algorithms find the optimal solutions accurately with computational efficiency, as indicated in Lemma 3 and Theorems 2 and 3. The potential applications of the two algorithms include the pursuit–evasion games and H_∞ design for disturbance attenuation in real environments modulated by uncertain parameters.

IV. Multiplayer Nonzero-Sum Game

This section studies the stochastic N-player nonzero-sum game. Each player aims to find its optimal control policy to minimize its own cost function. The properties and optimal solution of this game are analyzed in Section IV-A, and online learning algorithms are provided in Section IV-B.

A. Stability and Global Nash Equilibrium

Consider the value function described in (7), and the differential Bellman equation can be found by taking derivative of $V_i(\mathbf{x}(t))$ with respect to time t

$$r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}) + E\left[\frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}}\left(\mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j\right)\right] = 0. \tag{28}$$

The Hamiltonian function is

$$H_i\left(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}}\right) = r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}) + E\left[\frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}}\left(\mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j\right)\right]. \tag{29}$$

The optimal control policy $\mathbf{u}_i^*$ is derived by employing the stationary condition in the Hamiltonian function

$$\frac{\partial H_i}{\partial \mathbf{u}_i} = 0 \rightarrow \mathbf{u}_i^* = -\frac{1}{2}\mathbf{R}_{ii}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_i^*}{\partial \mathbf{x}}. \tag{30}$$

Substituting (30) into the Bellman equation (28), the following HJB equation is obtained:

$$\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + E\left[\frac{1}{4}\sum_{j=1}^{N}\frac{\partial V_j^{*\mathrm{T}}}{\partial \mathbf{x}}\mathbf{B}\mathbf{R}_{jj}^{-\mathrm{T}}\mathbf{R}_{ij}\mathbf{R}_{jj}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_j^*}{\partial \mathbf{x}} + \frac{\partial V_i^{*\mathrm{T}}}{\partial \mathbf{x}}\left(A(\mathbf{a})\mathbf{x} - \frac{1}{2}\sum_{j=1}^{N}\mathbf{B}\mathbf{R}_{jj}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_j^*}{\partial \mathbf{x}}\right)\right] = 0. \tag{31}$$

Lemma 4: For any admissible control policy $\mathbf{u}_i$, let $V_i \geq 0$ be the corresponding solution to the Bellman equation (28), and then, the following equation holds:

$$H_i\left(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \frac{\partial V_i}{\partial \mathbf{x}}^{\mathrm{T}}\right) = H_i\left(\mathbf{x}, \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \frac{\partial V_i}{\partial \mathbf{x}}^{\mathrm{T}}\right) + \sum_{j=1}^{N}\left(\mathbf{u}_j - \mathbf{u}_j^*\right)^{\mathrm{T}}\mathbf{R}_{ij}\left(\mathbf{u}_j - \mathbf{u}_j^*\right) + \frac{\partial V_i}{\partial \mathbf{x}}^{\mathrm{T}}\sum_{j=1}^{N}\mathbf{B}\left(\mathbf{u}_j - \mathbf{u}_j^*\right) + 2\sum_{j=1}^{N}\left(\mathbf{u}_j^*\right)^{\mathrm{T}}\mathbf{R}_{ij}\left(\mathbf{u}_j - \mathbf{u}_j^*\right).$$

Proof: See Appendix E. □

Theorem 4: Let V_i be a smooth function satisfying the HJB equation (31), and then, the following statements hold.

1) The system (5) is asymptotically stable in the mean with the control policy $\mathbf{u}_i^*$ described in (30).
2) The control policies $[\mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_N^*]$ derived in (30) are global Nash equilibrium policies.

Proof: See Appendix F. □

B. Approximate Solutions Using On-Policy and Off-Policy IRL and MPCM

The IRL Bellman equation for each player is given as [6]

$$V_i(\mathbf{x}(t)) = E\left[\int_t^{t+T} r_i(\mathbf{x}(\tau), \mathbf{u}_i(\tau), \mathbf{u}_{-i}(\tau))d\tau + V_i(\mathbf{x}(t+T))\right] \tag{32}$$

where T is the time interval.

Assume there exists a weight $\mathbf{W}_i$ for each player i, such that the value function $V_i(\mathbf{x})$ can be approximated as

$$V_i(\mathbf{x}) = \mathbf{W}_i^{\mathrm{T}}\phi_i(\mathbf{x}) \tag{33}$$

where $\phi_i(\mathbf{x})$ is the polynomial basis function vector for player i. Then, based on this VFA, the optimal control policy for each player can be learned iteratively from the online learning algorithms by integrating IRL and the MPCM.

1) On-policy IRL: Define a system mapping $G_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a}) = \int_t^{t+T} r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i})d\tau + V_i(\mathbf{x}(t+T))$. Then, given any admissible control policies $\mathbf{u}_i$ and $\mathbf{u}_{-i}$, the value function described in (32) can be approximated by the expected output of $G_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})$, i.e., $V_i(\mathbf{x}) = E\left[G_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})\right]$ using the MPCM

$$V_i(\mathbf{x}(t)) = E\left[G_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})\right] = E\left[G'_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})\right] \tag{34}$$

Algorithm 3 Policy Iteration for Multiplayer Nonzero-Sum Game With Uncertain System Dynamics

1: Initialize each player with initial state $\mathbf{x}(0)$ and admissible control policy $\mathbf{u}_i(0)$, $i = 1, 2, \ldots, N$.
2: Apply the MPCM procedure [35, Section II] to select a set of samples for the uncertain vector $\mathbf{a}(t)$.
3: For each iteration s, find the value of

$$\int_t^{t+T} r_i\left(\mathbf{x}(\tau), \mathbf{u}_i^{(s)}(\tau), \mathbf{u}_{-i}^{(s)}(\tau)\right)d\tau + \mathbf{W}_i^{(s+1)^{\mathrm{T}}}\phi_i(\mathbf{x}(t+T)) \tag{35}$$

at each MPCM sample.
4: Find the value function $V_i^{(s)}(\mathbf{x}(t))$ using the MPCM [35], which is the mean output of the mapping $G_{V_i^{(s)}}(\cdot)$ subject to uncertain parameters $\mathbf{a}(t)$,

$$G_{V_i^{(s)}}\left(\mathbf{x}, \mathbf{u}_i^{(s)}, \mathbf{u}_{-i}^{(s)}, \mathbf{a}\right) = \int_t^{t+T} r_i\left(\mathbf{x}(\tau), \mathbf{u}_i^{(s)}(\tau), \mathbf{u}_{-i}^{(s)}(\tau)\right)d\tau + \mathbf{W}_i^{(s+1)^{\mathrm{T}}}\phi_i(\mathbf{x}(t+T)). \tag{36}$$

5: Update the value function weight vector $\mathbf{W}_i^{(s)}$ according to the estimated $V_i^{(s)}(\mathbf{x}(t))$.

$$\mathbf{W}_i^{(s)^{\mathrm{T}}}\phi_i(\mathbf{x}(t)) = V_i^{(s)}(\mathbf{x}(t)).$$

6: Update the control policy $\mathbf{u}_i$ using

$$\mathbf{u}_i^{(s+1)} = -\frac{1}{2}\mathbf{R}_{ii}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_i^{(s)}}{\partial \mathbf{x}}. \tag{37}$$

7: Repeat procedures 3−6.

where $G'_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})$ is the reduced-order mapping derived from the MPCM procedure [35]. The detailed algorithm is described in Algorithm 3.

Theorem 5: Consider the stochastic multiplayer nonzero-sum game shown in (5)–(8). The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume the following: 1) VFA in (33) holds; 2) the relation between the value function $V_i(\mathbf{x}(t))$ and the uncertain parameters $\mathbf{a}(t)$ can be approximated by a polynomial system mapping (36) with the form of (15); and 3) Algorithm 3 converges. Then, the solution found from Algorithm 3 is the optimal control solution.

Proof: See Appendix G. □

2) Off-Policy IRL: We introduce auxiliary variable $\mathbf{u}_j^{(s)}$ for the player j, $(j = 1, 2, \ldots, N)$ and rewrite the system dynamics described in (5) as

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j^{(s)} + \sum_{j=1}^{N}\mathbf{B}\left(\mathbf{u}_j - \mathbf{u}_j^{(s)}\right) \tag{38}$$

where $\mathbf{u}_j$ is the behavior policy applied to the system to generate the data for learning, and $\mathbf{u}_j^{(s)}$ is the desired policy to be updated for the player j.

Differentiating the value function $V_i^{(s)}(\mathbf{x}(t))$ for the system (38), one has

$$\begin{aligned}
&\dot{V}_i^{(s)}(\mathbf{x}(t)) \\
&= E\left[\frac{\partial V_i^{(s)^{\mathrm{T}}}}{\partial \mathbf{x}}\left(A(\mathbf{a})\mathbf{x}+\sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j^{(s)}+\sum_{j=1}^{N}\mathbf{B}\left(\mathbf{u}_j-\mathbf{u}_j^{(s)}\right)\right)\right] \\
&= -\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x}-\sum_{j=1}^{N}\mathbf{u}_j^{(s)^{\mathrm{T}}}\mathbf{R}_{ij}\mathbf{u}_j^{(s)} \\
&\quad -\sum_{j=1}^{N}2\mathbf{u}_i^{(s+1)^{\mathrm{T}}}\mathbf{R}_{ii}\left(\mathbf{u}_j-\mathbf{u}_j^{(s)}\right). \qquad (39)
\end{aligned}$$

The second equality is obtained by combining the Hamiltonian functions (29) and (37).

Integrating both sides in (39), one has

$$\begin{aligned}
&V_i^{(s)}(\mathbf{x}(t+T))-V_i^{(s)}(\mathbf{x}(t)) \\
&\quad = E\left[\int_t^{t+T}-\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x}+\sum_{j=1}^{N}\mathbf{u}_j^{(s)^{\mathrm{T}}}\mathbf{R}_{ij}\mathbf{u}_j\right)d\tau\right] \\
&\quad\quad -\int_t^{t+T}\left(\sum_{j=1}^{N}2\mathbf{u}_i^{(s+1)^{\mathrm{T}}}\mathbf{R}_{ii}\left(\mathbf{u}_j-\mathbf{u}_j^{(s)}\right)\right)d\tau. \qquad (40)
\end{aligned}$$

For any fixed admissible behavior control policy $\mathbf{u}_j$ ($j = 1, 2, \ldots, N$), (40) can be solved for the value function $V_i^{(s)}$ and the optimal control policy $\mathbf{u}_i^{(s+1)}$ simultaneously, using the following NNs:

$$\begin{aligned}
V_i^{(s)}(\mathbf{x}) &= \mathbf{W}_i^{(s)^{\mathrm{T}}}\phi_i(\mathbf{x}) \\
\mathbf{u}_i^{(s+1)}(\mathbf{x}) &= \mathbf{W}_{\mathbf{u},i}^{(s+1)^{\mathrm{T}}}\phi_{\mathbf{u},i}(\mathbf{x}). \qquad (41)
\end{aligned}$$

The detailed algorithm that integrates off-policy IRL and the MPCM for the multiplayer nonzero-sum game is described in Algorithm 4.

Theorem 6: Consider the stochastic multiplayer nonzero-sum game shown in (5)–(8). The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume the following: 1) VFA in (41) holds; 2) the relation between the value function $V_i(\mathbf{x}(t))$ and the uncertain parameters $\mathbf{a}(t)$ can be approximated by a polynomial system mapping (42) with the form of (15); and 3) Algorithm 4 converges. Then, the solution found from off-policy IRL described in Algorithm 4 is the optimal solution.

Proof: See Appendix H. □

Remark 4: Algorithms 3 and 4 integrate IRL and the MPCM to solve the multiplayer nonzero-sum game with uncertain parameters in the system dynamics. These two algorithms find the Nash solutions accurately with computational efficiency. The potential applications of the two algorithms include the control of transportation networks and the cooperative control of multiple robots with individual goals, in real environments modulated by uncertain parameters.

V. Illustrative Examples

In this section, we conduct simulation studies to illustrate and verify the above-mentioned analysis.

Algorithm 4 Off-Policy IRL for Multiplayer Nonzero-Sum Game With Uncertain System Dynamics

1: Initialize the players with initial state $\mathbf{x}(0)$ and admissible control policies $\mathbf{u}_i(0)$.

2: Apply the MPCM procedure [35, Section II] to select a set of samples for the uncertain vector $\mathbf{a}(t)$.

3: For each iteration s, find the value of

$$\int_t^{t+T}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x}+\sum_{j=1}^{N}\mathbf{u}_j^{(s)^{\mathrm{T}}}\mathbf{R}_{ij}\mathbf{u}_j\right)d\tau+V_i^{(s)}(\mathbf{x}(t+T))$$

at each MPCM sample.

4: Find the mean output of mapping $G^o_{V_i^{(s)}}(\cdot)$ subject to uncertain parameters $\mathbf{a}(t)$ using the MPCM [35],

$$\begin{aligned}
G^o_{V_i^{(s)}}\left(\mathbf{x},\mathbf{u}_i^{(s)},\mathbf{u}_{-i}^{(s)},\mathbf{a}\right) &= V_i^{(s)}(\mathbf{x}(t+T)) \\
&\quad +\int_t^{t+T}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x}+\sum_{j=1}^{N}\mathbf{u}_j^{(s)^{\mathrm{T}}}\mathbf{R}_{ij}\mathbf{u}_j\right)d\tau.
\end{aligned}$$

5: Solve the following equation for $V_i^{(s)}(\mathbf{x})$ and $\mathbf{u}_i^{(s+1)}$, respectively.

$$\begin{aligned}
&V_i^{(s)}(\mathbf{x}(t))-\int_t^{t+T}\left(\sum_{j=1}^{N}2\mathbf{u}_i^{(s+1)^{\mathrm{T}}}\mathbf{R}_{ii}\left(\mathbf{u}_j-\mathbf{u}_j^{(s)}\right)\right)d\tau \\
&\qquad = E\left[G^o_{V_i^{(s)}}\left(\mathbf{x},\mathbf{u}_i^{(s)},\mathbf{u}_{-i}^{(s)},\mathbf{a}\right)\right]. \qquad (42)
\end{aligned}$$

6: Repeat procedures 3–5.

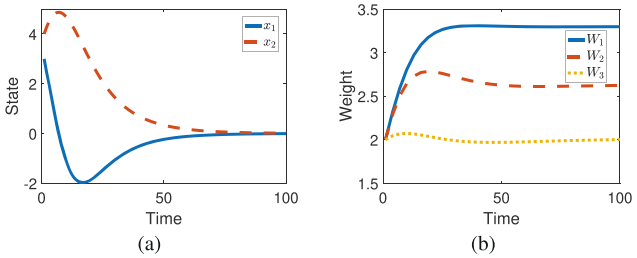

Fig. 1. Solution of two-player zero-sum game derived from Algorithm 1. (a) Evolution of system states. (b) Updates of value function weights.

A. Two-Player Zero-Sum Game

We first simulate the two-player zero-sum game with the uncertain system dynamics described as follows:

$$\dot{\mathbf{x}} = \begin{bmatrix} a_1(t) & a_2(t) \\ a_3(t) & a_4(t) \end{bmatrix}\mathbf{x}+\begin{bmatrix}1\\0\end{bmatrix}\mathbf{u}+\begin{bmatrix}1\\0\end{bmatrix}\mathbf{d}$$

where $a_1(t)$, $a_2(t)$, $a_3(t)$, and $a_4(t)$ are four random variables with time-varying values. The four random variables follow the uniform distributions: $f(a_1(t)) = 1/2$, $0 < a_1(t) < 2$; $f(a_2(t)) = 2$, $0 < a_2(t) < 0.5$; $f(a_3(t)) = 1$, $0.5 < a_3(t) < 1.5$; and $f(a_4(t)) = 1/2$, $-1 < a_4(t) < 1$. The parameters in the value function are selected as $\mathbf{Q} = \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$, $R = 1$, and $\gamma = 5$. The basis function is $\phi = [x_1^2, x_1x_2, x_2^2]^{\mathrm{T}}$, with the weight vector $\mathbf{W} = [W_1, W_2, W_3]^{\mathrm{T}}$. Fig. 1(a) and (b) shows the

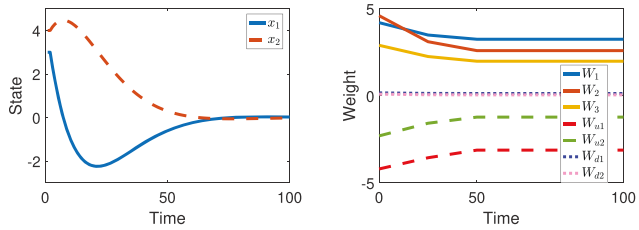

Fig. 2. Solution of two-player zero-sum game derived from Algorithm 2. (a) Evolution of system states. (b) Updates of NN weights.

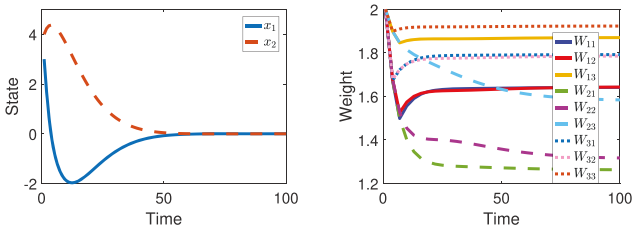

Fig. 3. Solution of multiplayer nonzero-sum game derived from Algorithm 3. (a) Evolution of system states. (b) Updates of value function weights.

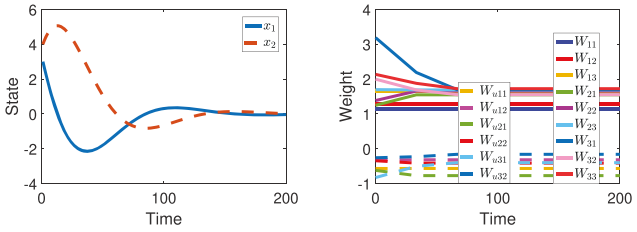

Fig. 4. Solution of multiplayer nonzero-sum game derived from Algorithm 4. (a) Evolution of system states. (b) Updates of NN weights.

evolution of the system states and the derived value function weights, respectively, using the on-policy PI algorithm (see Algorithm 1). It can be seen that the system states converge to $\mathbf{0}$ in the limit of large time with the derived control policies, and the value function weights converge quickly with the proposed algorithm.

We also conduct a comparative study to show the performance improvement of Algorithm 1 over the MC method, typically used to address uncertainty in decision. Here, the MC method is used to evaluate the value function, i.e., the mean value $E[\cdot]$ in (13), at each time step. The numbers of samples used by the MPCM and the MC to estimate each value function are 16 and 10000, respectively, to obtain a converged mean value. The NN weight derived by the MPCM is $\mathbf{W} = [3.29, 2.62, 2.00]^{\mathrm{T}}$, which is close to $\mathbf{W} = [3.16, 2.61, 2.09]^{\mathrm{T}}$ obtained using the MC method. The accurate estimation of the value function and the significant reduction of computational load demonstrate the value of using the proposed integrated RL and the MPCM algorithm to facilitate decisions for this game.

We then simulate the off-policy IRL algorithm described in Algorithm 2. Fig. 2(a) and (b) shows the evolution of system states and NN weights, respectively. Note that in the off-policy IRL, three NNs, including critic NN, actor NN, and disturbance NN, are utilized. The critic NN is $\mathbf{W} = [W_1, W_2, W_3]^{\mathrm{T}}$ with the basis function $\phi = [x_1^2, x_1x_2, x_2^2]^{\mathrm{T}}$, the actor NN is $\mathbf{W}_u = [W_{u1}, W_{u2}]^{\mathrm{T}}$ with the basis function $\phi_{\mathbf{u}} = [x_1, x_2]^{\mathrm{T}}$, and the disturbance NN is $\mathbf{W}_d = [W_{d1}, W_{d2}]$ with the basis function $\phi_{\mathbf{d}} = [x_1, x_2]^{\mathrm{T}}$. It can be seen that the system states converge to $\mathbf{0}$ in the limit of large time with the proposed off-policy IRL algorithm. In addition, the derived value function weight vector $[W_1, W_2, W_3]$ of the two algorithms are identical, which validates Theorem 3.

B. Multiplayer Nonzero-Sum Game

We then simulate the multiplayer nonzero-sum game discussed in Section IV, where the number of players $N = 3$. The system dynamic is described as follows:

$$\dot{\mathbf{x}} = \begin{bmatrix} a_1(t) & a_2(t) \\ a_3(t) & a_4(t) \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_1 + \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_2 + \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_3$$

where $a_1(t)$, $a_2(t)$, $a_3(t)$, and $a_4(t)$ are four randomly time-varying variables with the same pdfs described in Section V-A. The parameters in the value function are selected as $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $R_{12} = R_{13} = R_{21} = R_{23} = R_{31} = R_{32} = 1$, $R_{11} = 2$, $R_{22} = 3$, and $R_{33} = 5$. The value function weight vectors for the three players are $\mathbf{W}_1 = [W_{11}, W_{12}, W_{13}]^{\mathrm{T}}$, $\mathbf{W}_2 = [W_{21}, W_{22}, W_{23}]^{\mathrm{T}}$, and $\mathbf{W}_3 = [W_{31}, W_{32}, W_{33}]^{\mathrm{T}}$, respectively. Fig. 3(a) shows the evolution of system states, and Fig. 3(b) shows the learned value function weights.

We also simulate the off-policy IRL algorithm described in Algorithm 4. Fig. 4(a) and (b) shows the evolution of system states and NN weights, respectively. Note that in the off-policy algorithm, each player has two NNs: one for the critic NN and the other for the actor NN. It can be seen from Fig. 4(a) and (b) that the off-policy IRL algorithm works well for the multiplayer nonzero-sum game. The system states converge to $\mathbf{0}$ in the limit of large time, and the derived value function weights are the same with the on-policy algorithm, validating Theorem 6.

VI. Conclusion

This article studies multiplayer differential games for systems with randomly time-varying parameters. Two games, including two-player zero-sum and multiplayer nonzero-sum games, are formulated, respectively, with general uncertain linear dynamics. The optimal control policies for the two games are obtained from the corresponding Hamiltonian functions. The system properties, including the stability and the Nash equilibrium, are analyzed. In addition, we develop IRL-based online learning algorithms for each game to find optimal control solutions in real time. To evaluate the value functions with multidimensional uncertainties, an efficient uncertainty evaluation method, called the MPCM, is utilized to significantly reduce the computational cost. We integrate the MPCM with both on-policy and off-policy IRLs for each game and prove that the proposed algorithms find the correct Nash equilibrium solutions. Moreover, we show that the solutions derived from the on-policy and off-policy algorithms are identical under general uncertain linear system dynamics. This study provides new effective online learning methods to solve differential

games of general uncertain linear systems. The solution can be widely applied to stochastic systems, where uncertain player intentions or environmental factors modulate the system dynamics in a complicated fashion. In the future work, we will generalize the current work to heterogeneous players and study dynamical graphical games with stochastic dynamics. We will also thoroughly implement the proposed algorithms in real-world applications, such as UAV traffic management and autonomous driving in uncertain environments.

APPENDIX

A. Proof of Lemma 2

Combining (10) and (11), the Hamiltonian function can be further written as

$$\begin{aligned}
&H(\mathbf{x},\mathbf{u},\mathbf{d},V_X)\\
&\quad= r(\mathbf{x},\mathbf{u},\mathbf{d})+E\big[V_X^{\mathrm{T}}(A(\mathbf{a})\mathbf{x}+\mathbf{B}\mathbf{u}+\mathbf{C}\mathbf{d})\big]\\
&\quad= \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+E\big[V_X^{\mathrm{T}}(A(\mathbf{a})\mathbf{x})\big]+V_X^{\mathrm{T}}(\mathbf{B}\mathbf{u}+\mathbf{C}\mathbf{d})\\
&\qquad+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\\
&\quad= \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+E\big[V_X^{\mathrm{T}}(A(\mathbf{a})\mathbf{x})\big]\\
&\qquad+\left(\frac{1}{2}V_X^{\mathrm{T}}\mathbf{B}\mathbf{R}^{-1}+\mathbf{u}^{\mathrm{T}}\right)\mathbf{R}\left(\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}V_X+\mathbf{u}\right)\\
&\qquad-\gamma^2\left\|\left(\mathbf{d}-\frac{1}{2\gamma^2}\mathbf{C}^{\mathrm{T}}V_X\right)\right\|^2\\
&\qquad-\frac{1}{4}V_X^{\mathrm{T}}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}V_X+\frac{1}{4\gamma^2}V_X^{\mathrm{T}}\mathbf{C}\mathbf{C}^{\mathrm{T}}V_X\\
&\quad= H(\mathbf{x},\mathbf{u}^*,\mathbf{d}^*,V_X)+(\mathbf{u}-\mathbf{u}^*)^{\mathrm{T}}\mathbf{R}(\mathbf{u}-\mathbf{u}^*)\\
&\qquad-\gamma^2\|\mathbf{d}-\mathbf{d}^*\|^2
\end{aligned}$$

which derives Lemma 2.

B. Proof of Theorem 1

1) Stability: Choose the Lyapunov function candidate as $\tilde{V}(\mathbf{x}(t))=\int_t^\infty\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right)d\tau$. Since the attenuation condition is satisfied, there always exists a positive definite matrix $\mathbf{P}$, such that $\tilde{V}(\mathbf{x}(t))=\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$ [45, p. 337]. As such, one has

$$\tilde{V}(\mathbf{x}(t))=\int_t^\infty\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right)d\tau\geq 0 \tag{43}$$

and $\tilde{V}(\mathbf{x}(t))=0$ if and only if $\mathbf{x}=\mathbf{0}$. Denote the derivation of $\tilde{V}$ with respect to time t as $\dot{\tilde{V}}$, and then, the expectation of $\dot{\tilde{V}}$ is

$$\begin{aligned}
E\left[\dot{\tilde{V}}(\mathbf{x}(t))\right]&=E\left[\frac{\partial\tilde{V}}{\partial\mathbf{x}}\dot{\mathbf{x}}\right]\\
&=E\big[V_X(A(\mathbf{a})\mathbf{x}+\mathbf{B}\mathbf{u}+\mathbf{C}\mathbf{d})\big]\\
&=H(\mathbf{x},\mathbf{u},\mathbf{d},V_X)-\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right)\\
&=H(\mathbf{x},\mathbf{u}^*,\mathbf{d}^*,V_X)+(\mathbf{u}-\mathbf{u}^*)^{\mathrm{T}}\mathbf{R}(\mathbf{u}-\mathbf{u}^*)\\
&\quad-\gamma^2\|\mathbf{d}-\mathbf{d}^*\|^2\\
&\quad-\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right).
\end{aligned}$$

The last equality is obtained from Lemma 2. Selecting $\mathbf{u}=\mathbf{u}^*$ and $\mathbf{d}=\mathbf{d}^*$, one has

$$E\left[\dot{\tilde{V}}(\mathbf{x}(t))\right]=-\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right)\leq 0$$

and $E\left[\dot{\tilde{V}}(\mathbf{x}(t))\right]=0$ if and only if $\mathbf{x}=\mathbf{0}$. Therefore, $\tilde{V}$ is a Lyapunov function for $\mathbf{x}$. According to Lemma 1, the system described in (1) is asymptotically stable in the mean.

2) Nash Equilibrium: Since the system is asymptotically stable in the mean, we have $E\{\|\mathbf{x}(t)\|\}=0$ holds when $t\to\infty$. Therefore, the cost function can be rewritten as

$$\begin{aligned}
&J(\mathbf{x}(0),\mathbf{u},\mathbf{d})\\
&\quad=E\left[\int_0^\infty\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right)dt\right.\\
&\qquad\left.+V(\mathbf{x}(0))+\int_0^\infty\dot{V}dt\right]\\
&\quad=E\left[\int_0^\infty\left(r(\mathbf{x},\mathbf{u},\mathbf{d})+V_X^{\mathrm{T}}(A(\mathbf{a})\mathbf{x}+\mathbf{B}\mathbf{u}+\mathbf{C}\mathbf{d})\right)dt\right]\\
&\qquad+V(\mathbf{x}(0))\\
&\quad=E\left[\int_0^\infty\left((\mathbf{u}-\mathbf{u}^*)^{\mathrm{T}}\mathbf{R}(\mathbf{u}-\mathbf{u}^*)-\gamma^2\|\mathbf{d}-\mathbf{d}^*\|^2\right)dt\right]\\
&\qquad+V(\mathbf{x}(0)).
\end{aligned}\tag{44}$$

The last equality is obtained by combining (10) and Lemma 2.

It can be seen from (44) that $J(\mathbf{x}(0),\mathbf{u}^*,\mathbf{d})\leq J(\mathbf{x}(0),\mathbf{u}^*,\mathbf{d}^*)\leq J(\mathbf{x}(0),\mathbf{u},\mathbf{d}^*)$, and thus, the Nash equilibrium is obtained.

C. Proof of Theorem 2

The control and disturbance policies derived by evaluating the original value function mapping $G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})$ is optimal according to Theorem 1 and (17). As such, to prove this theorem, we only need to show that the two optimal solutions that are found by evaluating the reduced-order mapping $G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})$ and the original value function mapping $G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})$ are the same. Lemma 3 proved that $E\left[G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})\right]=E\left[G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})\right]$, and hence, the equivalence of the two optimal solutions can be proved from a contradiction method following a similar argument, as described in [37, Th. 1].

D. Proof of Theorem 3

It has been proven that for a deterministic system dynamics, the solutions derived from the off-policy IRL and on-policy IRL are identical for the two-player zero-sum game [44]. As such, for each MPCM sample point $\mathcal{A}^l$, $l=1,2,\ldots,\prod_{p=1}^m n_p$, the value functions and optimal solutions derived from the on-policy and off-policy IRL algorithms are identical. Note that the expected value function is the weighted average of the value functions derived at each sample point (see Lemma 3 and [35]). As such, the expected value function derived from the two algorithms is identical, and hence, the off-policy solution is the optimal control policy.

E. Proof of Lemma 4

Combining (29) and (30), Lemma 4 can be obtained naturally following a similar procedure as described in Lemma 2.

F. Proof of Theorem 4

1) Stability: Choose the Lyapunov function candidate for player i as $\tilde{V}_i = \int_t^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^N \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right) d\tau$, and then, one has

$$E\left[\tilde{V}_i\right] = E\left[\int_t^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^N \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right) d\tau\right] \geq 0. \quad (45)$$

The derivation of $\tilde{V}_i$ with time t is derived as

$$\begin{aligned} E\left[\dot{\tilde{V}}_i\right] &= E\left[\frac{\partial \tilde{V}_i}{\partial \mathbf{x}}\dot{\mathbf{x}}\right] \\ &= E\left[V_X\left(\mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^N \mathbf{B}\mathbf{u}_j\right)\right] \\ &= -\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} - \sum_{j=1}^N \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j \\ &\leq 0. \end{aligned}$$

Therefore, $\tilde{V}_i$ is a Lyapunov function for $\mathbf{x}$, and the system described in (5) is asymptotically stable in the mean [42].

2) Nash Equilibrium: Since the system is asymptotically stable in the mean, we have $E\{\|\mathbf{x}(t)\|\} = \mathbf{0}$ holds when $t \to \infty$. Therefore, the cost function can be rewritten as

$$\begin{aligned} &J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}) \\ &= E\left[\int_0^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^N \mathbf{u}_i^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_i\right) dt\right] \\ &\quad + V_i(\mathbf{x}(0)) + E\left[\int_0^\infty \dot{V}_i dt\right] \\ &= V_i(\mathbf{x}(0)) + E\left[\int_0^\infty \left(\sum_{j=1}^N \left(\mathbf{u}_j - \mathbf{u}_j^*\right)^{\mathrm{T}}\mathbf{R}_{ij}\left(\mathbf{u}_j - \mathbf{u}_j^*\right)\right.\right. \\ &\qquad + \frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}} \sum_{j=1}^N \mathbf{B}\left(\mathbf{u}_j - \mathbf{u}_j^*\right) \\ &\qquad \left.\left. + 2\sum_{j=1}^N \left(\mathbf{u}_j^*\right)^{\mathrm{T}}\mathbf{R}_{ij}\left(\mathbf{u}_j - \mathbf{u}_j^*\right)\right) dt\right]. \end{aligned}$$

The second equality is obtained by combining (29) and Lemma 1.

Assume that all other players' control policies are optimal, i.e., $\mathbf{u}_{-i} = \mathbf{u}_{-i}^*$, and then, we have

$$\begin{aligned} &J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}^*) \\ &\quad = V_i(\mathbf{x}(0)) + E\left[\int_0^\infty (\mathbf{u}_i - \mathbf{u}_i^*)^{\mathrm{T}}\mathbf{R}_{ii}(\mathbf{u}_i - \mathbf{u}_i^*) dt\right]. \quad (46) \end{aligned}$$

It can be seen from (46) that $J_i(\mathbf{x}(0), \mathbf{u}_i^*, -\mathbf{u}_i^*) < J_i(\mathbf{x}(0), \mathbf{u}_i, -\mathbf{u}_i^*)$ holds for every player i, which proves the Nash equilibrium.

G. Proof of Theorem 5

This proof follows a similar procedure as described in Theorem 2.

H. Proof of Theorem 6

For the multiplayer nonzero-sum game with deterministic system dynamics, the solutions derived from the off-policy IRL and on-policy IRL have been proved to be identical [12]. The proof for the game with uncertain system dynamics, then, follows a similar argument, as described in Theorem 3.

REFERENCES

[1] R. B. Myerson, *Game Theory*. Cambridge, MA, USA: Harvard Univ. Press, 2013.

[2] M. J. Osborne *et al.*, *An Introduction to Game Theory*. Oxford, U.K.: Oxford Univ. Press, 2004.

[3] M. Shubik, *Game Theory in the Social Sciences: Concepts and Solutions*. Cambridge, MA, USA: MIT Press, 1984.

[4] 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, Aug. 2012.

[5] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ, USA: Wiley, 2012.

[6] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Syst.*, vol. 37, no. 1, pp. 33–52, Feb. 2017.

[7] T. Başar and P. Bernhard, *H_∞ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. New York, NY, USA: Springer, 2008.

[8] D. Vrabie and F. Lewis, "Adaptive dynamic programming for online solution of a zero-sum differential game," *J. Control Theory Appl.*, vol. 9, no. 3, pp. 353–360, Aug. 2011.

[9] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, Mar. 2007.

[10] J.-H. Kim and F. L. Lewis, "Model-free H_∞ control design for unknown linear discrete-time systems via Q-learning with LMI," *Automatica*, vol. 46, no. 8, pp. 1320–1326, 2010.

[11] D. Vrabie and F. Lewis, "Integral reinforcement learning for online computation of feedback Nash strategies of nonzero-sum differential games," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Atlanta, GA, USA, Dec. 2010, pp. 3066–3071.

[12] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 704–713, Mar. 2017.

[13] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, vol. 17. Hoboken, NJ, USA: Wiley, 2013.

[14] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, Feb. 2009.

[15] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 3rd Quart., 2009.

[16] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167–1175, 2014.

[17] 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.

[18] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, Apr. 2009.

[19] B. Kiumarsi and F. L. Lewis, "Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 140–151, Jan. 2015.

[20] H.-N. Wu and B. Luo, "Simultaneous policy update algorithms for learning the solution of linear continuous-time H_∞ state feedback control," *Inf. Sci.*, vol. 222, pp. 472–485, Feb. 2013.

[21] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 3, pp. 706–714, Jul. 2014.

[22] D. P. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete Time Case*. Belmont, MA, USA: Athena Scientific, 1996.
[23] H. J. Kappen, "An introduction to stochastic control theory, path integrals and reinforcement learning," *Cooperat. Behav. Neural Syst.*, vol. 887, no. 1, pp. 149–181, 2007.
[24] L. Xie, D. Popa, and F. L. Lewis, *Optimal and Robust Estimation: With an Introduction to Stochastic Control Theory*. Boca Raton, FL, USA: CRC Press, 2007.
[25] S. Mohseni-Bonab, A. Rabiee, S. Jalilzadeh, B. Mohammadi-Ivatloo, and S. Nojavan, "Probabilistic multi objective optimal reactive power dispatch considering load uncertainties using Monte Carlo simulations," *J. Oper. Autom. Power Eng.*, vol. 3, no. 1, pp. 83–93, 2015.
[26] Y. Matsuno, T. Tsuchiya, J. Wei, I. Hwang, and N. Matayoshi, "Stochastic optimal control for aircraft conflict resolution under wind uncertainty," *Aerosp. Sci. Technol.*, vol. 43, pp. 77–88, Jun. 2015.
[27] N. Kantas, A. Lecchini-Visintini, and J. M. Maciejowski, "Simulation-based Bayesian optimal design of aircraft trajectories for air traffic management," *Int. J. Adapt. Control Signal Process.*, vol. 24, no. 10, pp. 882–899, Sep. 2010.
[28] M. Prandini, J. Hu, J. Lygeros, and S. Sastry, "A probabilistic approach to aircraft conflict detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 4, pp. 199–220, Dec. 2000.
[29] A. L. Visintini, W. Glover, J. Lygeros, and J. Maciejowski, "Monte Carlo optimization for conflict resolution in air traffic control," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 4, pp. 470–482, Dec. 2006.
[30] M. D. Mckay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, May 1979.
[31] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," *Manage. Sci.*, vol. 35, no. 11, pp. 1367–1392, Nov. 1989.
[32] S. Heinrich, "Multilevel Monte Carlo methods," in *Proc. Int. Conf. Large-Scale Sci. Comput.*, Sozopol, Bulgaria, 2001, pp. 1–39.
[33] J. S. Hesthaven, B. Stamm, and S. Zhang, "Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods," *ESAIM, Math. Model. Numer. Anal.*, vol. 48, no. 1, pp. 259–283, 2014.
[34] T. Hachisuka *et al.*, "Multidimensional adaptive sampling and reconstruction for ray tracing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 1, Aug. 2008.
[35] Y. Zhou *et al.*, "Multivariate probabilistic collocation method for effective uncertainty evaluation with application to air traffic flow management," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 10, pp. 1347–1363, Oct. 2014.
[36] J. Xie, Y. Wan, K. Mills, J. J. Filliben, Y. Lei, and Z. Lin, "M-PCM-OFFD: An effective output statistics estimation method for systems of high dimensional uncertainties subject to low-order parameter interactions," *Math. Comput. Simul.*, vol. 159, pp. 93–118, May 2019.
[37] J. Xie, Y. Wan, K. Mills, J. J. Filliben, and F. L. Lewis, "A scalable sampling method to high-dimensional uncertainties for optimal and reinforcement learning-based controls," *IEEE Control Syst. Lett.*, vol. 1, no. 1, pp. 98–103, Jul. 2017.
[38] J. Xie, Y. Wan, and F. L. Lewis, "Strategic air traffic flow management under uncertainties using scalable sampling-based dynamic programming and Q-learning approaches," in *Proc. 11th Asian Control Conf. (ASCC)*, Gold Coast, QLD, Australia, Dec. 2017, pp. 1116–1121.
[39] J. Xie *et al.*, "Distance measure to cluster spatiotemporal scenarios for strategic air traffic management," *J. Aerosp. Inf. Syst.*, vol. 12, no. 8, pp. 545–563, Aug. 2015.
[40] M. Liu, Y. Wan, and F. L. Lewis, "Adaptive optimal decision in multi-agent random switching systems," *IEEE Control Syst. Lett.*, vol. 4, no. 2, pp. 265–270, Apr. 2020.
[41] M. Liu, Y. Wan, S. Li, F. L. Lewis, and S. Fu, "Learning and uncertainty-exploited directional antenna control for robust long-distance and broadband aerial communication," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 593–606, Jan. 2020.
[42] J. Bertram and P. Sarachik, "Stability of circuits with randomly time-varying parameters," *IRE Trans. Circuit Theory*, vol. 6, no. 5, pp. 260–270, 1959.
[43] F. Kozin, "A survey of stability of stochastic systems," *Automatica*, vol. 5, no. 1, pp. 95–112, Jan. 1969.
[44] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2550–2562, Oct. 2015.
[45] B. M. Chen, Z. Lin, and Y. Shamash, *Linear Systems Theory: A Structural Decomposition Approach*. New York, NY, USA: Springer, 2004.

Mushuang Liu (Student Member, IEEE) received the B.S. degree in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX, USA.

Her research interests include distributed decisions for multiagent systems, uncertain systems, multiplayer games, graphical games, reinforcement learning, and their applications to UAV traffic management and UAV networking.

Yan Wan (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Washington State University, Pullman, WA, USA, in 2008. She then did her post-doctoral training at the University of California at Santa Barbara, Santa Barbara, CA, USA.

She is currently an Associate Professor with the Electrical Engineering Department, The University of Texas at Arlington, Arlington, TX, USA. Her research has led to over 170 publications and successful technology transfer outcomes. Her research interests lie in the modeling and control of large-scale dynamical networks, cyber–physical systems, stochastic networks, learning control, networking, uncertainty analysis, algebraic graph theory, and their applications to UAV networking, UAV traffic management, epidemic spread, complex information networks, and air traffic management.

Dr. Wan received several prestigious awards, including the NSF CAREER Award, the RTCA William E. Jackson Award, the U.S. Ignite and GENI Demonstration Awards, the IEEE WCNC and ICCA Best Paper Awards, and the Tech Titan of the Future-University Level Award.

Frank L. Lewis (Life Fellow, IEEE) received the bachelor's degree in physics/electrical engineering and the M.S.E.E. degree from Rice University, Houston, TX, USA, the M.S. degree in aeronautical engineering from the University of West Florida, Pensacola, FL, USA, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA.

He is currently a UTA Distinguished Scholar Professor, a UTA Distinguished Teaching Professor, and the Moncrief-ODonnell Chair with the Research Institute, The University of Texas at Arlington, Arlington, TX, USA. He works in feedback control, intelligent systems, cooperative control systems, and nonlinear systems. He is author of seven U.S. patents, numerous journal special issues, 420 journal articles, and 20 books.

Dr. Lewis is also a Founding Member of the Board of Governors of the Mediterranean Control Association. He is also a member of the National Academy of Inventors and a fellow of International Federation of Automatic Control (IFAC), American Association for the Advancement of Science (AAAS), the U.K. Institute of Measurement Control, and PE Texas. He is also a Chartered Engineer in U.K. He received the Fulbright Research Award, the NSF Research Initiation Grant, the ASEE Terman Award, the International Neural Network Society Gabor Award, the U.K. Institute of Measurement Control Honeywell Field Engineering Medal, the IEEE Computational Intelligence Society Neural Networks Pioneer Award, the AIAA Intelligent Systems Award, the Received Outstanding Service Award from the Dallas IEEE Section, and the Texas Regents Outstanding Teaching Award in 2013. He was selected as the Engineer of the Year by the Ft. Worth IEEE Section and listed in Ft. Worth Business Press Top 200 Leaders in Manufacturing.

Victor G. Lopez (Student Member, IEEE) received the B.S. degree from the Universidad Autónoma de Campeche, Campeche, Mexico, in 2010, the M.S. degree from the Research and Advanced Studies Center (CINVESTAV), Guadalajara, Mexico, in 2013, and the Ph.D. degree from The University of Texas at Arlington (UTA), Arlington, TX, USA.

He was also a Lecturer with the Western Technologic Institute of Superior Studies (ITESO), Guadalajara, in 2015. He is currently a Post-Doctoral Fellow with Research Institute, UTA, where he is also an Adjunct Professor with the Electrical Engineering Department. His research interests include cyber–physical systems, game theory, distributed control, reinforcement learning, and robust control.