2019 IEEE 15th International Conference on Control and Automation (ICCA)
Edinburgh, Scotland, July 16-19, 2019

Stochastic Two-Player Zero-Sum Learning Differential Games

Mushuang Liu[1], Yan Wan[1], Frank L. Lewis[2], and Victor G. Lopez[1]

Abstract— **The two-player zero-sum differential game has been extensively studied, partially because its solution implies the H_∞ optimality. Existing studies on zero-sum differential games either assume deterministic dynamics or the dynamics corrupted by additive noise. In realistic environments, high-dimensional environmental uncertainties often modulate system dynamics in a more complicated fashion. In this paper, we study the stochastic two-player zero-sum differential game governed by more general uncertain linear dynamics. We show that the optimal control policies for this game can be found by solving the Hamilton-Jacobi-Bellman (HJB) equation. We prove that with the derived optimal control policies, the system is asymptotically stable in the mean, and reaches the Nash equilibrium. To solve the stochastic two-player zero-sum game online, we design a new policy iteration (PI) algorithm that integrates the integral reinforcement learning (IRL) and an efficient uncertainty evaluation method—multivariate probabilistic collocation method (MPCM). This algorithm provides a fast online solution for the stochastic two-player zero-sum differential game subject to multiple uncertainties in the system dynamics.**

I. Introduction

Game theory has been widely used in multi-player systems to obtain decisions that optimize individual payoffs [1]–[6]. In the standard 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 players' payoff functions [4]–[6]. The two-player zero-sum differential game has received much attention since it provides the H_∞ optimal solution [6]. The Nash equilibrium solution of the two-player zero-sum differential game relies on solving the Hamilton-Jacobi-Bellman (HJB) equation for nonlinear systems or the game algebraic Riccati equation (GARE) for linear systems. However, solving these equations is generally extremely difficult or even impossible [5]. Moreover, solving the differential game from HJB or GARE equations requires the full knowledge of the system dynamics, and is an offline process.

Reinforcement learning (RL), a subarea of machine learning, 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 [7]. Bridging between optimal control and adaptive control, RL-based on-line solutions have been designed for a range of system types, including linear continuous-time systems [8], linear discrete-time systems [9], [10], nonlinear continuous-time systems [11], [12], and nonlinear discrete-time systems [13].

Of interest to this paper, the RL method has been used to find the Nash equilibrium solutions online for two-player zero-sum differential games [14]–[19]. For discrete-time systems, paper [14] presented an adaptive dynamic programming (ADP)-based RL algorithm to solve the H_∞ control problem. The solution requires the knowledge of system dynamics. To deal with unknown system dynamics, a model-free Q-learning iteration algorithm was introduced in [15], [16]. For continuous-time systems, paper [17] introduced the idea of integral RL (IRL) to solve the two-player zero-sum differential game, and presented an ADP-based learning algorithm. This algorithm has two iterative loops and hence is time-consuming: one player first learns to optimize its control policy with the other player's policy fixed, and then when the first player's policy converges, the second player also begins to find its optimal control policy. To improve the learning efficiency, paper [18] proposed an single-loop iteration algorithm which updates the control policies of the two players simultaneously. In addition, to deal with unknown system dynamics, paper [19] presented a model-free IRL algorithm using the Q-learning method. All these aforementioned studies assume a time-invariant and deterministic system dynamics.

Modern dynamic systems often operate in uncertain environments. Their dynamics can be modulated by high-dimensional uncertainties, which complicate the decision process. Such stochastic optimal control problems have been studied in e.g., [9], [20]–[22]. For a linear system with additive noise and quadratic cost, the optimal control solution that minimizes the expected cost function can be found analytically [20]. However, for general stochastic systems with multiple uncertainties, simulation-based uncertainty evaluation methods need to be utilized. In addition, the uncertainties, if exploited, can benefit the optimal decision-making [23]–[25]. For instance, unmanned aircraft vehicle (UAV) dynamics are modulated by uncertain weather, and the optimal path planning can benefit from exploring probabilistic weather forecasts, which can be modeled as stochastic processes with known statistical information [26].

The most widely used uncertainty evaluation methods are the Monte Carlo (MC) method and its variants including the Makrov Chain MC and Sequential MC [27]–[29]. However, the MC-based methods require a large number of simulations

*We thank the ONR Grant N00014-18-1-2221, and NSF grants 1714519 and 1839804 for the support of this work.

[1] Mushuang Liu, Yan Wan, and Victor G. Lopez are with the Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX 76019, USA. Email: mushuang.liu@mavs.uta.edu, yan.wan@uta.edu (contact author), and victor.lopezmejia@mavs.uta.edu

[2] Frank L. Lewis is with UTA Research Institute, University of Texas at Arlington, Fort Worth, Texas, USA., and Qian Ren Consulting Professor, State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. Email: lewis@uta.edu.

to estimate the expected cost function accurately, which makes it unrealistic for online algorithms. To deal with this challenge, paper [23] developed an efficient uncertainty evaluation method, named multivariate probabilistic collocation method (MPCM), which accurately estimates 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). Papers [24], [30] further integrated the MPCM method with the discrete-time RL to solve stochastic discrete-time optimal control problems online. Here in this paper, we utilize the MPCM method to solve the continuous-time stochastic two-player zero-sum differential learning game. Per knowledge of the authors, this is the first study in the field of multi-player differential games that considers general uncertainties in the dynamics.

This paper brings together game theory, reinforcement learning, and effective uncertainty evaluation to obtain fast online solutions for the stochastic two-player zero-sum differential game with general uncertain linear dynamics. The main contributions of this paper are three-fold. The first contribution lies in the introduction of the stochastic zero-sum differential game formulation with more general uncertain dynamics. This game formulation for the first time captures broad uncertain impacts, such as stochastic environments and random agent intentions [31], [32]. The second contribution lies in the analysis of game properties, including the stability and Nash equilibrium. The third contribution is a novel stochastic policy iteration (PI) algorithm that integrates MPCM and IRL to provide an fast online solution for the stochastic zero-sum game.

This paper is organized as follows. Section II introduces the preliminaries to facilitate the analysis in the paper and formulates the stochastic two-player zero-sum differential game. Section III studies the properties of the stochastic two-player zero-sum game, and proposes an online solution that integrates MPCM and IRL to solve the game in real time. Section IV presents the simulation studies, and Section V concludes the paper.

II. Problem Formulation and Preliminaries

In this section, we first formulate the stochastic two-player zero-sum game for systems of general linear dynamics. Preliminaries are then introduced to facilitate the analysis in the paper.

A. Problem Formulation

Consider a generic two-player linear system with a time-varing uncertain 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 $\mathbf{x} = \mathbf{x}(t) \in \mathbf{R}^n$ is the system state vector, $\mathbf{u} = \mathbf{u}(t) \in \mathbf{R}^p$ is the control input, and $\mathbf{d} = \mathbf{d}(t) \in \mathbf{R}^q$ 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 $\mathbf{a}(t)$, $a_p(t)$ $(p = 1, ..., m)$, changes independently over time with pdf $f_{A_p}(a_p(t))$.

This game formulation can describe a wide range of system dynamics modulated by uncertainties. One 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 an uncertain 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

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

where $\mathbf{Q}$ and $\mathbf{R}$ are positive semidefinite and positive definite matrices, respectively. $\mathbf{R}$ is a symmetric matrix. 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

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

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.

Consider the problem of finding the optimal control policy $\mathbf{u}^*$ and $\mathbf{d}^*$ such that

$$\begin{aligned} \mathbf{u}^* &= \operatorname*{argmin}_{\mathbf{u}} J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}), \\ \mathbf{d}^* &= \operatorname*{argmax}_{\mathbf{d}} J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}). \end{aligned}$$

B. Preliminaries

Definition 1. [33] The equilibrium solution of a system is said to be stable in the mean (norm) if for any $\epsilon > 0$ there exists a $\delta(t_0, \epsilon) > 0$, such that for any initial condition satisfying $\|\mathbf{x}_0\| < \delta(\epsilon)$,

$$E\{\sup_{t \geq t_0} \|\mathbf{x}(t)\|\} < \epsilon.$$

It is assumed that the system described in Equation (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. [33] 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. [34] 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 $\mathbf{d} \in L_2[0,\infty)$ with $\mathbf{x}(0)=\mathbf{0}$:

$$\frac{\int_t^\infty \|\mathbf{z}(\tau)\|^2 d\tau}{\int_t^\infty \|\mathbf{d}(\tau)\|^2 d\tau} \leq \gamma^2,$$

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

It is assumed that γ in Equation (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 stabilizable [4, Page 450].

Definition 4. [4, Page 445] Policies $\{\mathbf{u}_1^*, \mathbf{u}_2^*, ...\mathbf{u}_N^*\}$ are said to constitute a Nash equilibrium solution for the N-player game if the following equation is satisfied for $\forall \mathbf{u}_i, \forall i \in N$.

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

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

Lemma 1. *[33, Theorem II] Consider a system $\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{a}(t), t)$, 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 $(a-d)$, then the equilibrium solution of the system is asymptotically stable in the mean.*

a. $\tilde{V}(\mathbf{0}) = 0$.

b. $\tilde{V}(\mathbf{x}(t))$ *is continuous with both* $\mathbf{x}$ *and* t*, and the first partial derivatives in these variables exist.*

c. $\tilde{V}(\mathbf{x}(t)) \geq b\|\mathbf{x}\|$ *for some* $b > 0$.

d. $E[\dot{\tilde{V}}(\mathbf{x}(t))]$ *is 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 Nash equilibrium of the stochastic game, and section III-B develops an IRL-based online solution to solve the differential game.

A. Stability and Nash Equilibrium for Stochastic Two-Player Zero-Sum Game

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

$$r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E[\frac{\partial V^T}{\partial \mathbf{x}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})] = 0. \quad (5)$$

with the Hamiltonian function

$$\begin{aligned} &H(\mathbf{x}, \mathbf{u}, \mathbf{d}, \frac{\partial V}{\partial \mathbf{x}}) \\ &= r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E[\frac{\partial V^T}{\partial \mathbf{x}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})]. \end{aligned} \quad (6)$$

The optimal control policies $\mathbf{u}^*$ and $\mathbf{d}^*$ can be derived by employing the stationary conditions in the Hamiltonian function [4, Page 447],

$$\begin{aligned} \frac{\partial H^T}{\partial \mathbf{u}} = 0 \rightarrow \mathbf{u}^* = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^T\frac{\partial V^*}{\partial \mathbf{x}}, \\ \frac{\partial H^T}{\partial \mathbf{d}} = 0 \rightarrow \mathbf{d}^* = \frac{1}{2\gamma^2}\mathbf{C}^T\frac{\partial V^*}{\partial \mathbf{x}}. \end{aligned} \quad (7)$$

Substituting Equation (7) into the Bellman Equation (5), the following Hamilton-Jacobi-Bellman (HJB) equation is obtained.

$$\begin{aligned} &H(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X^*) \\ &= \mathbf{x}^T\mathbf{Q}\mathbf{x} + E[{V_X^*}^T A(\mathbf{a}) - \frac{1}{4}{V_X^*}^T\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T V_X^* \\ &\quad + \frac{1}{4\gamma^2}{V_X^*}^T\mathbf{C}\mathbf{C}^T V_X^*] = 0, \quad V(\mathbf{0}) = 0, \end{aligned} \quad (8)$$

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

Lemma 2. *For any admissible control policies* $\mathbf{u}$ *and* $\mathbf{d}$*, let* $V \geq 0$ *be the corresponding solution to the Bellman equation* (5)*, then the following equation holds [4, Lemma 10.2-1].*

$$\begin{aligned} H(\mathbf{x}, \mathbf{u}, \mathbf{d}, V_X) =& H(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X) + (\mathbf{u} - \mathbf{u}^*)^T\mathbf{R}(\mathbf{u} - \mathbf{u}^*) \\ &- \gamma^2(\mathbf{d} - \mathbf{d}^*)^T(\mathbf{d} - \mathbf{d}^*), \end{aligned} \quad (9)$$

where $\mathbf{u}^*$ *and* $\mathbf{d}^*$ *are described in Equation* (7)*, and* $V_X = \frac{\partial V}{\partial \mathbf{x}}$.

Proof: Combining Equations (6) and (7), the Hamiltonian function can further be written as

$$\begin{aligned} &H(\mathbf{x}, \mathbf{u}, \mathbf{d}, V_X) \\ &= r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E[V_X^T(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})] \\ &= \mathbf{x}^T\mathbf{Q}\mathbf{x} + E[V_X^T(A(\mathbf{a})\mathbf{x})] + V_X^T(\mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d}) \\ &\quad + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2 \\ &= \mathbf{x}^T\mathbf{Q}\mathbf{x} + E[V_X^T(A(\mathbf{a})\mathbf{x})] \\ &\quad + (\frac{1}{2}V_X^T\mathbf{B}\mathbf{R}^{-1} + \mathbf{u}^T)R(\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^T V_X + \mathbf{u}) \\ &\quad - \gamma^2\|(\mathbf{d} - \frac{1}{2\gamma^2}\mathbf{C}^T V_X)\|^2 \\ &\quad - \frac{1}{4}V_X^T\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T V_X + \frac{1}{4\gamma^2}V_X^T\mathbf{C}\mathbf{C}^T V_X \\ &= H(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X) + (\mathbf{u} - \mathbf{u}^*)^T\mathbf{R}(\mathbf{u} - \mathbf{u}^*) \\ &\quad - \gamma^2(\mathbf{d} - \mathbf{d}^*)^T(\mathbf{d} - \mathbf{d}^*). \end{aligned} \quad (10)$$

Theorem 1. *Let* V *be a smooth function satisfying the HJB equation* (8)*, then the following statements hold.*

1). The system (1) *is asymptotically stable in the mean with the control policies* $\mathbf{u}^*$ *and* $\mathbf{d}^*$ *described in Equation* (7).

2). The pair of strategies ($\mathbf{u}^*$ *and* $\mathbf{d}^*$*) derived in Equation* (7) *provides a saddle point solution to the game, and the system is in Nash equilibrium with this strategy pair.*

Proof: 1) Stability. Choose the Lyapunov function candidate as $\tilde{V} = \int_t^\infty(\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2)d\tau$. Since the attenuation condition is satisfied [34], one has

$$\tilde{V} = \int_t^\infty(\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2)d\tau > 0. \quad (11)$$

Denote the derivation of $\tilde{V}$ with respect to time t as $\dot{\tilde{V}}$, then the expectation of $\dot{\tilde{V}}$ is

$$\begin{aligned}E[\dot{\tilde{V}}] &= E[\frac{\partial \tilde{V}}{\partial \mathbf{x}}\dot{\mathbf{x}}] \\ &= E[V_X(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})] \\ &= H(\mathbf{x}, \mathbf{u}, \mathbf{d}, V_X) - (\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2) \\ &= H(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X) + (\mathbf{u}-\mathbf{u}^*)^T\mathbf{R}(\mathbf{u}-\mathbf{u}^*) \\ &\quad - \gamma^2(\mathbf{d}-\mathbf{d}^*)^T(\mathbf{d}-\mathbf{d}^*) \\ &\quad - (\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2).\end{aligned} \tag{12}$$

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

$$E[\dot{\tilde{V}}] = -(\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2) < 0. \tag{13}$$

Therefore $\tilde{V}$ is a Lyapunov function for $\mathbf{x}$. According to Lemma 1, the system described in Equation (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

$$\begin{aligned}&J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}) \\ &= E[\int_0^\infty (\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2)dt] + V(\mathbf{x}(0)) \\ &\quad + \int_0^\infty \dot{V}dt \\ &= E[\int_0^\infty r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + V_X^T(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})] \\ &\quad + V(\mathbf{x}(0)) \\ &= E[\int_0^\infty (\mathbf{u}-\mathbf{u}^*)^T\mathbf{R}(\mathbf{u}-\mathbf{u}^*) - \gamma^2(\mathbf{d}-\mathbf{d}^*)^T(\mathbf{d}-\mathbf{d}^*)] \\ &\quad + V(\mathbf{x}(0)).\end{aligned} \tag{14}$$

The last equality is obtained by combining Equation (6) and Lemma 2.

It can be seen from Equation (14) 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.

B. Online Learning Solution

To find the optimal control policies using Equation (7), a smooth function that satisfies the HJB equation (Equation (8)) needs to be found in closed-form. However, solving Equation (8) analytically is extremely difficult or even impossible. As such, we integrate IRL and an efficient uncertainty sampling method called MPCM to provide an fast online learning algorithm to approximate the solution to the HJB equation.

The IRL Bellman equation can be written as

$$V(\mathbf{x}(t)) = E[\int_t^{t+T} r(\mathbf{x}(\tau), \mathbf{u}(\tau), \mathbf{d}(\tau))d\tau + V(\mathbf{x}(t+T))], \tag{15}$$

where T is the time interval.

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

$$V(\mathbf{x}) = \mathbf{W}^T\phi(\mathbf{x}), \tag{16}$$

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

With the value function approximation (VFA), one can find the optimal control policies from the policy iteration (PI) algorithm by iteratively conducting two steps: policy evaluation and policy improvement [4, Page 474]. The policy evaluation step is designed to evaluate the value function $V(\mathbf{x})$ using Equation (15), given the current control policies. The policy improvement step is to find the optimal control policy based on the current approximated value function using Equation (7). For the stochastic system, the policy evaluation step involves the uncertainty evaluation, which is typically solved using the Monte Carlo method and its variants. However, the Monte Carlo methods are not computationally effective to be used for online solutions.

Here we utilize an efficient uncertainty sampling method, called multivariate probabilistic collocation method (MPCM) [35]. Rooted in quadrature rules, MPCM is designed to smartly select a limited number of samples to evaluate the output mean for a system mapping subject to uncertain input parameters. Simulations are then run at these samples to produce a reduced-order mapping, which has the same expected value of the original system. 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, Theorem 2] Consider a system mapping modulated by m independent uncertain parameters:*

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

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

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

with $E[G(a_1, ..., a_m)] = E[G'(a_1, ..., a_m)]$, where $\Omega_{q_1,...,q_m} \in \mathbb{R}$ are coefficients. The MPCM reduces the number of simulations from $2^m\prod_{p=1}^m n_p$ to $\prod_{p=1}^m n_p$.

Define $G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a}) = \int_t^{t+T} r(\mathbf{x}(\tau), \mathbf{u}(\tau), \mathbf{d}(\tau))d\tau + V(\mathbf{x}(t+T))$. Given an admissible state $\mathbf{x}(t)$ and control policies $\mathbf{u}(t)$ and $\mathbf{d}(t)$, the value function described in Equation (15) can be represented as the output mean of the system mapping $G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})$ subject to uncertain input parameters $\mathbf{a}$ (i.e., $V(\mathbf{x}) = E[G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})]$). The mapping can be approximated using 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 $E[G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})]$. 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,

$$G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a}) = \sum_{q_1=0}^{2n_1-1}\sum_{q_2=0}^{2n_2-1}\cdots\sum_{q_m=0}^{2n_m-1}\psi_{q_1,\ldots,q_m}(\mathbf{x},\mathbf{u},\mathbf{d})\prod_{p=1}^{m}a_p^{q_p}. \quad (17)$$

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

$$V(\mathbf{x}(t)) = E[G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})] = E[G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})], \quad (18)$$

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

$$G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a}) = \sum_{q_1=0}^{n_1-1}\sum_{q_2=0}^{n_2-1}\cdots\sum_{q_m=0}^{n_m-1}\Omega_{q_1,\ldots,q_m}(\mathbf{x},\mathbf{u},\mathbf{d})\prod_{p=1}^{m}a_p^{q_p}. \quad (19)$$

The PI algorithm that integrates IRL and MPCM for the stochastic two-player zero-sum game is described in Algorithm 1.

Algorithm 1 Policy iteration algorithm for stochastic two-player zero-sum game

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

2: Select $\prod_{p=1}^{m} n_p$ MPCM sample points according to the pdfs $f_{A_p}(a_p)$ and the MPCM procedure [35, Section II]. Denote each selected MPCM sample as $\mathcal{A}^l$, where $l = 1,\ldots,\prod_{p=1}^{m} n_p$.

3: For each iteration j, find the value function $\mathcal{V}_j^l(t)$ at each sample $\mathcal{A}^l$, using the following Bellman equation given the current control policy $\mathbf{u}_j$ and $\mathbf{d}_j$.

$$\mathcal{V}_j^l(t) = W_{j-1}^T\phi(\mathbf{x}^l(t+T)) + \int_t^{t+T} r^l(\mathbf{x}(\tau),\mathbf{u}_j(\tau),\mathbf{d}_j(\tau))d\tau. \quad (20)$$

4: Find the reduced polynomial mapping from a_p to the value function according to Lemma 3.

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

where a_p and $G'_{V_j(t)}(\mathbf{x},\mathbf{u}_j,\mathbf{d}_j,\mathbf{a})$ take the value of $\mathcal{A}^l$ and $\mathcal{V}_j^l(t)$ respectively. The coefficients $\Omega_{V\,q_1,\ldots,q_m}(\mathbf{x},\mathbf{u}_j,\mathbf{d}_j)$ can be determined using the least-squares method.

5: Find the value function $V_j(\mathbf{x}(t))$ from the mean output of the reduced-order mapping according to the MPCM procedure [35, Section II].

$$V_j(\mathbf{x}(t)) = E[G'_{V_j(t)}(\mathbf{x},\mathbf{u}_j,\mathbf{d}_j,\mathbf{a})]. \quad (21)$$

6: Update the value function coefficients W_j according to the estimated $V_j(\mathbf{x}(t))$ using the least-squares method.

$$W_j^T\phi(\mathbf{x}(t)) = V_j(\mathbf{x}(t)). \quad (22)$$

7: Update the control policy $\mathbf{u}_{j+1}$ and $\mathbf{d}_{j+1}$ as

$$\mathbf{u}_{j+1} = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^T\frac{\partial V_j^*}{\partial \mathbf{x}}, \qquad \mathbf{d}_{j+1} = \frac{1}{2\gamma^2}\mathbf{C}^T\frac{\partial V_j^*}{\partial \mathbf{x}}. \quad (23)$$

8: Repeat procedures $3-7$.

Theorem 2. *Consider a stochastic zero-sum game shown in Equations* (1)-(4)*, the uncertainty in the system dynamics a_p following a time-invariant pdf $f_{A_p}(a_p)$, then the optimal control policy $\mathbf{u}$ and $\mathbf{d}$ derived from Algorithm 1 is the optimal control policy.*

Proof: To prove this theorem, we need to show that the two optimal control policies, which 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. Since Lemma 3 has proved that $E[G'_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})] = E[G_{V(t)}(\mathbf{x},\mathbf{u},\mathbf{d},\mathbf{a})]$, the equivalence of the two optimal policies can be proved from a contradiction method following a similar argument as described in [24, Theorem 1].

IV. Illustrative Examples

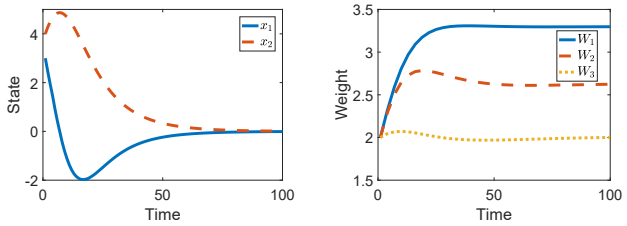
In this section we conduct a simulation study to illustrate and verify the algorithm and results developed in this paper.

Consider the two-player uncertain system with the following dynamics:

$$\dot{\mathbf{x}} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\mathbf{u} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\mathbf{d}, \quad (24)$$

where a_1, a_2, a_3, and a_4 are four uncertain parameters, with their values changing over time. The distributions of the four uncertain parameters are: $f(a_1) = \frac{1}{2}$, $0 < a_1 < 2$; $f(a_2) = 2$, $0 < a_2 < 0.5$; $f(a_3) = 1$, $0.5 < a_3 < 1.5$; and $f(a_4) = \frac{1}{2}$, $-1 < a_4 < 1$ respectively. The parameters in the value function are selected as $\mathbf{Q} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, $R = 1$, and $\gamma = 5$. Figures 1(a) and 1(b) show the evolution of the system state and the learned value function weights respectively, using the designed PI algorithm presented in Algorithm 1.

It can be seen that the system state converges to 0 in the limit of large time, which validates the stability of the system. In addition, the value function weights converge quickly with time, indicating the effectiveness of the proposed algorithm.



(a) (b)

Fig. 1. Stochastic two-player zero-sum game. (a) The evolution of the system state, and (b) the learned value function weights.

V. Conclusion

This paper studies the stochastic two-player zero-sum differential game with a general uncertain linear dynamics. Optimal control policies are obtained from the Hamiltonian function. The system properties, including the stability and Nash equilibrium, are analyzed with the derived optimal policies. In addition, an online PI-based learning algorithm that solves the stochastic two-player zero-sum differential game is designed by integrating the IRL and an efficient uncertainty sampling method named MPCM. This algorithm permits finding the online solution of the stochastic two-player zero-sum differential game in highly uncertain environments.

References

[1] R. B. Myerson, *Game theory*. Harvard university press, 2013.

[2] M. J. Osborne *et al.*, *An introduction to game theory*. Oxford University Press, 2004, vol. 3, no. 3.

[3] M. Shubik, "Game theory in the social sciences: concepts and solutions," 2006.

[4] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.

[5] 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 Systems*, vol. 37, no. 1, pp. 33–52, 2017.

[6] T. Başar and P. Bernhard, *H_∞ optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

[7] F. L. Lewis and D. Liu, *Reinforcement learning and approximate dynamic programming for feedback control*. John Wiley & Sons, 2013, vol. 17.

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

[9] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, 2009.

[10] 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.

[11] 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.

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

[13] B. Kiumarsi and F. L. Lewis, "Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 1, pp. 140–151, 2015.

[14] A. Al-Tamimi, M. Abu-Khalaf, and F. L. Lewis, "Adaptive critic designs for discrete-time zero-sum games with application to h_∞ control," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 1, pp. 240–247, 2007.

[15] 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, 2007.

[16] 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.

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

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

[19] H. Li, D. Liu, D. Wang, and X. Yang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 706–714, 2014.

[20] D. P. Bertsekas and S. Shreve, *Stochastic optimal control: the discrete-time case*, 2004.

[21] H. J. Kappen, "An introduction to stochastic control theory, path integrals and reinforcement learning," in *Cooperative Behavior in Neural Systems*, vol. 887, no. 1, 2007, pp. 149–181.

[22] L. Xie, D. Popa, and F. L. Lewis, *Optimal and robust estimation: with an introduction to stochastic control theory*. CRC press, 2007.

[23] Y. Zhou, Y. Wan, S. Roy, C. Taylor, C. Wanke, D. Ramamurthy, and J. Xie, "Multivariate probabilistic collocation method for effective uncertainty evaluation with application to air traffic flow management," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 10, pp. 1347–1363, 2014.

[24] 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 Systems Letters*, vol. 1, no. 1, pp. 98–103, 2017.

[25] 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," *Mathematics and Computers in Simulation*, vol. 159, pp. 93–118, 2019.

[26] M. A. Pinheiro, M. Liu, Y. Wan, and A. Dogan, "On the analysis of on-board sensing and off-board sensing through wireless communication for uav path planning in wind fields," in *Proceedings of AIAA Scitech*, San Diego, CA, 2019.

[27] N. Kantas, A. Lecchini-Visintini, and J. Maciejowski, "Simulation-based bayesian optimal design of aircraft trajectories for air traffic management," *International Journal of Adaptive Control and Signal Processing*, vol. 24, no. 10, pp. 882–899, 2010.

[28] M. Prandini, J. Hu, J. Lygeros, and S. Sastry, "A probabilistic approach to aircraft conflict detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 4, pp. 199–220, 2000.

[29] A. L. Visintini, W. Glover, J. Lygeros, and J. Maciejowski, "Monte carlo optimization for conflict resolution in air traffic control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp. 470–482, 2006.

[30] 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 *Proceedings of IEEE Asian Control Conference (ASCC)*, Gold Coast, QLD, Australia, 2017.

[31] J. Xie, Y. Wan, Y. Zhou, S.-L. Tien, E. P. Vargo, C. Taylor, and C. Wanke, "Distance measure to cluster spatiotemporal scenarios for strategic air traffic management," *Journal of Aerospace Information Systems*, vol. 12, no. 8, pp. 545–563, 2015.

[32] M. Liu, Y. Wan, S. Li, and F. Lewis, "Learning and uncertainty-exploited directional antenna control for robust aerial networking," *submitted to American Control Conference (ACC)*, 2019.

[33] J. Bertram and P. Sarachik, "Stability of circuits with randomly time-varying parameters," *IEEE Transactions on Circuit Theory*, vol. 6, no. 5, pp. 260–270, 1959.

[34] H. Modares, F. L. Lewis, and Z.-P. Jiang, "h_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2550–2562, 2015.

[35] Y. Zhou, Y. Wan, S. Roy, C. Taylor, C. Wanke, D. Ramamurthy, and J. Xie, "Multivariate probabilistic collocation method for effective uncertainty evaluation with application to air traffic flow management," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 10, pp. 1347–1363, 2014.