MULTI-PLAYER H_∞ DIFFERENTIAL GAME USING ON-POLICY AND OFF-POLICY REINFORCEMENT LEARNING

by

PEILIANG AN

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2021

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Dr. Yan Wan, for motivating and encouraging me. I am grateful for all the invaluable advice and guidance that she has provided me with since the first day when I came to the university.

I also want to thank the members of my supervising committee, Dr. Qilian Liang and Dr. Yuze Alice Sun. I appreciate their time to evaluate my project.

Many thanks to the University of Texas at Arlington, for giving me the opportunity to gain knowledge and train skills in electrical engineering.

To my collaborators in the Dynamical Networks & Control Lab, I am blessed with, and proud of, having been working with you all for the past two years.

I also would like to extend my appreciation to my parents, for their unconditional support and love. They have characterized me with the determination to go through the ups and downs.

Finally, I appreciate the support from the National Science Foundation under grants 1714519 and 1839804, Army Research Office under grant W911NF-20-1-0132, and the Office of Naval Research under grant N00014-18-1-2221.

July 30, 2021

ABSTRACT

MULTI-PLAYER H_∞ DIFFERENTIAL GAME USING ON-POLICY AND OFF-POLICY REINFORCEMENT LEARNING

PEILIANG AN, MSc

The University of Texas at Arlington, 2021

Supervising Professor: Dr. Yan Wan

This work studies a multi-player H_∞ differential game for systems of general linear dynamics. In this game, multiple players design their control inputs to minimize their cost functions in the presence of worst-case disturbances. We first derive the optimal control and disturbance policies using the solutions to Hamilton-Jacobi-Isaacs (HJI) equations. We then prove that the derived optimal policies stabilize the system and constitute a Nash equilibrium solution. Two integral reinforcement learning (IRL) -based algorithms, including the policy iteration IRL and off-policy IRL, are developed to solve the differential game online. We show that the off-policy IRL can solve the multi-player H_∞ differential game online without using any system dynamics information. Simulation studies are conducted to validate the theoretical analysis and demonstrate the effectiveness of the developed learning algorithms.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1 Introduction

Differential games [1, 2, 3, 4] have attracted increasing attentions in the control community due to their wide applications in multi-robot systems [5, 6]. Differential games provide a formal mathematical framework to study the coordination, conflict and control of dynamical systems that involve multiple decision-makers (or players) [7, 1, 2, 3, 4]. Two types of differential games, including the two-player zero-sum games and multi-player nonzero-sum games, have been studied [1, 4]. The two-player zero-sum games can be used to solve the pursuit-evasion type of problems, i.e., there is a single performance index that one player tries to minimize while the other tries to maximize [2, 8]. The two-player zero-sum games have also been used to solve the H_∞ control of systems subject to additive external disturbances [8, 1]. The other type of differential games, i.e., the multi-player nonzero-sum games, have been developed to solve the leader-follower optimal tracking type of problems, where there can generally exist more than two players and each player tries to minimize its individual performance index [3]. In this work, we study a new type of differential game, called the multi-player H_∞ differential game, which takes features of the two differential games aforementioned. In the multi-player H_∞ game, each player seeks to minimize its performance index in the presence of a worst-case disturbance. This game provides a theoretical framework for optimal controller design of multi-player systems subject to external disturbances. Per the knowledge of the authors, there are very limited studies till now that study the multi-player H_∞ differential game

[9, 10]. Properties of such systems, e.g., stability and Nash equilibrium have not been thoroughly analyzed.

Finding Nash equilibrium solutions to differential games is not an easy task [3]. In particular, solving zero-sum differential games relies on solving Hamilton-Jacobi-Isaacs (HJI) equations, and solving nonzero-sum differential games relies on solving Hamilton-Jacobi-Bellman (HJB) equations. It has been shown that solving these equations directly in an analytical way is extremely difficult [11]. In addition, solving these equations also requires the information of system dynamics, which is not always available in real applications.

Reinforcement learning (RL) has emerged as an efficient numerical tool for solving optimal control problems online. The use of RL in control theory is documented in [12] for continuous-time linear systems, [13, 14] for discrete-time linear systems, [15, 16] for continuous-time nonlinear systems, and [17] for discrete-time nonlinear systems. Of our interests, RL-based algorithms have also been developed for differential games. Interested readers please refer to [18, 19, 20, 8] for two-player zero-sum games, and [21, 22] for multi-player nonzero-sum games. In particular, an off-policy integral RL (IRL) was developed in [22] to solve the multi-player nonzero-sum games without requiring any information of the system dynamics. In this work, we study both on-policy and off-policy IRL solutions to the new multi-player H_∞ differential game.

The contributions of this work are three-fold. First, we formulate the multi-player H_∞ differential game subject to the worst-case external disturbance, and show that the solution to the game stabilizes the system and constitutes a Nash equilibrium. Second, we develop a policy iteration-based learning algorithm to solve the game online, using partial system dynamics information. Third, we further develop an off-policy IRL algorithm that requires no information of the system dynamics.

The results are documented in paper [23] published in the 16^{th} IEEE International Conference on Control & Automation.

The remainder of the thesis is structured as follows. Chapter 2 formulates the multi-player H_∞ differential game and provides preliminaries to facilitate the analysis. In Chapter 3, properties of the multi-player H_∞ game are studied, and two IRL-based algorithms are developed to find the optimal solutions online. Chapter 4 presents simulation studies and Chapter 5 concludes the work.

CHAPTER 2

Problem Formulation and Preliminaries

In this chapter, we formulate the multi-player H_∞ differential game for a system of general linear dynamics. We then provide preliminaries to facilitate the analysis in Chapter 3.

2.0.1 Problem Formulation

Consider a general N-player linear time-invariant dynamical system given by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}_j \mathbf{u}_j + \sum_{j=1}^{N} \mathbf{C}_j \mathbf{d}_j, \tag{2.1}$$

where $\mathbf{x} = \mathbf{x}(t) \in \mathbf{R}^n$ is the state vector, $\mathbf{u}_j = \mathbf{u}_j(t) \in \mathbf{R}^m$ is the control input for player j, and the $\mathbf{d}_j = \mathbf{d}_j(t) \in \mathbf{R}^q$ is the adversarial disturbance input for player j. $\mathbf{A}$, $\mathbf{B}_j$, and $\mathbf{C}_j$ are the drift, control input, and disturbance input dynamics, respectively. It is assumed that the system (2.1) is stabilizable. Many engineering systems are governed by dynamics (2.1), for example, the aircraft launching, where $\mathbf{x}$ is the aircraft speed, $\mathbf{u}_j$ and $\mathbf{d}_j$ are the control thrust force and the disturbance force of the controller j, respectively.

Define the cost function to be optimized for player i $(i = 1, 2, \cdots, N)$ as

$$\begin{aligned} & J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) \\ & \quad = \int_0^\infty r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) dt \\ & \quad = \int_0^\infty \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{R}_{ij} \mathbf{u}_j - \gamma^2 \sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \right) dt, \end{aligned} \tag{2.2}$$

where $\mathbf{u}_{-i}$ and $\mathbf{d}_{-i}$ are the sets of control and disturbance policies for all players other than player i. $\mathbf{Q}_i$ and $\mathbf{R}_{ij}$ $(i \neq j)$ are positive semi-definite matrices, and $\mathbf{R}_{ii}$ are positive definite matrices.

The value function of player i is defined as

$$\begin{aligned} & V_i(\mathbf{x}(t)) \\ & = \int_t^\infty r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) d\tau \\ & = \int_t^\infty \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{R}_{ij} \mathbf{u}_j - \gamma^2 \sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \right) d\tau. \end{aligned} \tag{2.3}$$

Define the multi-player H_∞ differential game as

$$V_i^*(\mathbf{x}(0)) = \min_{\mathbf{u}_i} \max_{\mathbf{d}_i} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}), \tag{2.4}$$

where $V_i^*(\mathbf{x}(0))$ is the optimal value for player i. In the multi-player H_∞ game, each player tries to minimize its cost function by choosing a control policy $\mathbf{u}_i$, while the disturbance $\mathbf{d}_i$ seeks to maximize this cost. Each player has access to the full state of the system.

The problem is to find the optimal control and disturbance policies $\mathbf{u}_i^*$ and $\mathbf{d}_i^*$ such that

$$\mathbf{u}_i^* =_{\mathbf{u}_i} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}),$$

$$\mathbf{d}_i^* =_{\mathbf{d}_i} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}).$$

2.0.2 Preliminaries

[11] The system (2.1) is said to have L_2-gain less than or equal to γ if the following disturbance attenuation condition is satisfied for all $\mathbf{d}_j \in L_2[0, \infty)$ with $\mathbf{x}(0) = \mathbf{0}$:

$$\frac{\int_t^\infty \|\mathbf{z}(\tau)\|^2 d\tau}{\int_t^\infty \left(\sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \right) d\tau} \leq \gamma^2,$$

where $\|\mathbf{z}(t)\|^2 = \mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j\mathbf{R}_{ij}\mathbf{u}_j,\ \mathbf{d}_j(t)$ is the disturbance input, and γ is the amount of attenuation.

It is assumed that γ in (2.2) satisfies $\gamma \geq \gamma^*$, where γ^* is the smallest γ, also know as H_∞ gain for system (2.1) [1], which satisfies the disturbance attenuation condition.

[1] Policies $\{\mathbf{u}_1^*, \mathbf{d}_1^*, \mathbf{u}_2^*, \mathbf{d}_2^*, \cdots, \mathbf{u}_N^*, \mathbf{d}_N^*\}$ are said to constitute a Nash equilibrium solution to the N-player H_∞ game if the following inequality holds:

$$\begin{aligned} &J_i(\mathbf{x}(0), \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i, \mathbf{d}_{-i}^*) \\ &\quad \leq J_i^*(\mathbf{x}(0), \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*) \\ &\quad \leq J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*), \quad \forall \mathbf{u}_i, \forall \mathbf{d}_i, \forall i. \end{aligned} \tag{2.5}$$

CHAPTER 3

Multi-player H_∞ Differential Game

This chapter derives the optimal solution to the N-player H_∞ differential game. Chapter 3.0.1 studies the stability and Nash equilibrium of the game. Two IRL-based algorithms are then developed in Chapter 3.0.2 to solve the differential game online.

3.0.1 Stability and Nash Equilibrium

Differentiating the value function $V_i(\mathbf{x}(t))$ defined in (2.3), one can obtain the Bellman equation as follows,

$$\begin{aligned}&\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{R}_{ij} \mathbf{u}_j - \gamma^2 \sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \\ &+ \nabla V_i \left(\mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}_j \mathbf{u}_j + \sum_{j=1}^{N} \mathbf{C}_j \mathbf{d}_j \right) = 0,\end{aligned} \tag{3.1}$$

where $\nabla V_i = \partial V_i / \partial \mathbf{x}$. The boundary condition for this partial differential equation is $V_i(\mathbf{0}) = 0$. A solution to (3.1) is the value function $V_i(\mathbf{x})$ for the feedback control policy $\mathbf{u}_i = \mathbf{u}_i(V_i(\mathbf{x}))$ and disturbance policy $\mathbf{d}_i = \mathbf{d}_i(V_i(\mathbf{x}))$.

Define the Hamiltonian function associated with the value function (2.3) as

$$\begin{aligned} &H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) \\ &\quad = r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) \\ &\qquad + \nabla V_i \left(\mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}_j \mathbf{u}_j + \sum_{j=1}^{N} \mathbf{C}_j \mathbf{d}_j \right) \\ &\quad = \mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{R}_{ij} \mathbf{u}_j - \gamma^2 \sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \\ &\qquad + \nabla V_i \left(\mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}_j \mathbf{u}_j + \sum_{j=1}^{N} \mathbf{C}_j \mathbf{d}_j \right). \end{aligned} \tag{3.2}$$

At the equilibrium point, applying the stationary conditions

$$\frac{\partial H_i}{\partial \mathbf{u}_i} = 0 \quad \text{and} \quad \frac{\partial H_i}{\partial \mathbf{d}_i} = 0$$

yields the optimal control and disturbance policies as functions of $V_i(\mathbf{x})$:

$$u_i^* = \mathbf{u}_i^*(V_i(\mathbf{x})) = -\frac{1}{2}\mathbf{R}_{ii}^{-1}\mathbf{B}_i \nabla V_i, \tag{3.3}$$

$$\mathbf{d}_i^* = \mathbf{d}_i^*(V_i(\mathbf{x})) = \frac{1}{2\gamma^2}\mathbf{C}_i \nabla V_i. \tag{3.4}$$

Therefore, the value function $V_i(\mathbf{x})$ in (2.3) is only a function of the state $\mathbf{x}(t)$. Moreover, the Hamiltonian function H_i attains a saddle point at the stationary point since $\partial^2 H_i / \partial \mathbf{u}_i^2 = 2\mathbf{R}_{ii} > 0$ and $\partial^2 H_i / \partial \mathbf{d}_i^2 = -2\gamma^2 < 0$.

Substituting (3.3) and (3.4) into the Bellman Equation (3.1), the following Hamilton-Jacobi-Isaacs (HJI) equation is obtained:

$$\begin{aligned} &\mathbf{x}\mathbf{Q}_i\mathbf{x} + \frac{1}{4}\sum_{j=1}^{N} \nabla V_j \mathbf{B}_j \mathbf{R}_{jj}^{-1} \mathbf{R}_{ij} \mathbf{R}_{jj}^{-1} \mathbf{B}_j \nabla V_j - \frac{1}{4\gamma^2}\sum_{j=1}^{N} \nabla V_j \mathbf{C}_j \mathbf{C}_j \nabla V_j \\ &\quad + \nabla V_i \left(\mathbf{A}\mathbf{x} - \frac{1}{2}\sum_{j=1}^{N} \mathbf{B}_j \mathbf{R}_{jj}^{-1} \mathbf{B}_j \nabla V_j + \frac{1}{2\gamma^2}\sum_{j=1}^{N} \mathbf{C}_j \mathbf{C}_j \nabla V_j \right) = 0. \end{aligned} \tag{3.5}$$

Since the attenuation condition in Definition 2.0.2 is satisfied, the HJI equation (3.5) has a positive semi-definite solution $V_i^*(\mathbf{x}(t))$ [1].

Note that for the optimal policies $\mathbf{u}_i^*$, $\mathbf{d}_i^*$ and the corresponding V_i^*, the HJI equation satisfies

$$H_i(\mathbf{x}, \nabla V_i^*, \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*) = 0. \tag{3.6}$$

Assume the control and disturbance policies are optimal for all players other than player i. Then for any admissible policies $\mathbf{u}_i(\mathbf{x})$ and $\mathbf{d}_i(\mathbf{x})$, and any positive semi-definite value function $V_i(\mathbf{x})$, one has the following equation:

$$\begin{aligned} & H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i, \mathbf{u}_{-i}^*, \mathbf{d}_i, \mathbf{d}_{-i}^*) \\ & \quad = H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*) \\ & \qquad + (\mathbf{u}_i - \mathbf{u}_i^*)^T \mathbf{R}_{ii} (\mathbf{u}_i - \mathbf{u}_i^*) - \gamma^2 (\mathbf{d}_i - \mathbf{d}_i^*)^T (\mathbf{d}_i - \mathbf{d}_i^*). \end{aligned} \tag{3.7}$$

Taking $\mathbf{u}_{-i} = \mathbf{u}_{-i}^*$ and $\mathbf{d}_{-i} = \mathbf{d}_{-i}^*$, the Hamiltonian function in (3.2) can be written as

$$\begin{aligned} & H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i, \mathbf{u}_{-i}^*, \mathbf{d}_i, \mathbf{d}_{-i}^*) \\ & \quad = \mathbf{x} \mathbf{Q}_i \mathbf{x} + \sum_{j \neq i} \mathbf{u}_j^* \mathbf{R}_{ij} \mathbf{u}_j^* + \mathbf{u}_i \mathbf{R}_{ii} \mathbf{u}_i \\ & \qquad - \gamma^2 \sum_{j \neq i} \|\mathbf{d}_j^*\|^2 - \gamma^2 \mathbf{d}_i \mathbf{d}_i \\ & \qquad + \nabla V_i \left(\mathbf{A}\mathbf{x} + \sum_{j \neq i} \mathbf{B}_j \mathbf{u}_j^* + \mathbf{B}_i \mathbf{u}_i + \sum_{j \neq i} \mathbf{C}_j \mathbf{d}_j^* + \mathbf{C}_i \mathbf{d}_i \right) \end{aligned}$$

$$\begin{aligned}
&= \mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_j \mathbf{u}_j^* \mathbf{R}_{ij} \mathbf{u}_j^* - \gamma^2 \sum_j \|\mathbf{d}_j^*\|^2 \\
&\quad + \nabla V_i \left(\mathbf{A}\mathbf{x} + \sum_j \mathbf{B}_j \mathbf{u}_j^* + \sum_j \mathbf{C}_j \mathbf{d}_j^* \right) \\
&\quad + \mathbf{u}_i \mathbf{R}_{ii} \mathbf{u}_i - \mathbf{u}_i^* \mathbf{R}_{ii} \mathbf{u}_i^* - \gamma^2 \mathbf{d}_i \mathbf{d}_i + \gamma^2 \mathbf{d}_i^* \mathbf{d}_i^* \\
&\quad + \nabla V_i (\mathbf{B}_i \mathbf{u}_i - \mathbf{B}_i \mathbf{u}_i^* + \mathbf{C}_i \mathbf{d}_i - \mathbf{C} \mathbf{d}_i^*) \\
&= H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*) + \mathbf{u}_i \mathbf{R}_{ii} \mathbf{u}_i - \mathbf{u}_i^* \mathbf{R}_{ii} \mathbf{u}_i^* \\
&\quad - \gamma^2 \mathbf{d}_i \mathbf{d}_i + \gamma^2 \mathbf{d}_i^* \mathbf{d}_i^* + (\mathbf{u}_i - \mathbf{u}_i^*) \mathbf{B}_i \nabla V_i \\
&\quad + (\mathbf{d}_i - \mathbf{d}_i^*) \mathbf{C}_i \nabla V_i.
\end{aligned} \tag{3.8}$$

According to (3.3) and (3.4), one has

$$\mathbf{B}_i \nabla V_i = -2\mathbf{R}_{ii} \mathbf{u}_i^* \quad \text{and} \quad \mathbf{C}_i \nabla V_i = 2\gamma^2 \mathbf{d}_i^*.$$

As such, (3.8) can be further rewritten as

$$\begin{aligned}
&H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i, \mathbf{u}_{-i}^*, \mathbf{d}_i, \mathbf{d}_{-i}^*) \\
&\quad = H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*) + \mathbf{u}_i \mathbf{R}_{ii} \mathbf{u}_i \\
&\qquad - \mathbf{u}_i^* \mathbf{R}_{ii} \mathbf{u}_i^* - \gamma^2 \mathbf{d}_i \mathbf{d}_i + \gamma^2 \mathbf{d}_i^* \mathbf{d}_i^* \\
&\qquad - 2(\mathbf{u}_i - \mathbf{u}_i^*) \mathbf{R}_{ii} \mathbf{u}_i^* + 2\gamma^2 (\mathbf{d}_i - \mathbf{d}_i^*) \mathbf{d}_i^* \\
&\quad = H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i^*, \mathbf{u}_{-i}^*, \mathbf{d}_i^*, \mathbf{d}_{-i}^*) \\
&\qquad + (\mathbf{u}_i - \mathbf{u}_i^*) \mathbf{R}_{ii} (\mathbf{u}_i - \mathbf{u}_i^*) - \gamma^2 (\mathbf{d}_i - \mathbf{d}_i^*)(\mathbf{d}_i - \mathbf{d}_i^*).
\end{aligned}$$

This result is next employed to show that the optimal policies given by (3.3) and (3.4) in terms of coupled HJI solution $V_i^*(\mathbf{x})$ constitute a Nash equilibrium solution.

Suppose $V_i^*(\mathbf{x})$ are smooth continuous positive semi-definite functions that solve the HJI equations (3.5). The control and disturbance policies $\mathbf{u}_i^*$ and $\mathbf{d}_i^*$ are given by (3.3) and (3.4). Then the following two statements (a) and (b) hold.

(a). The closed-loop system

$$
\begin{aligned}
\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}_j \mathbf{u}_j^* + \sum_{j=1}^{N} \mathbf{C}_j \mathbf{d}_j^* \\
&= \mathbf{A}\mathbf{x} - \frac{1}{2} \sum_{j=1}^{N} \mathbf{B}_j \mathbf{R}_{jj}^{-1} \mathbf{B}_j^T \nabla V_j^* + \frac{1}{2\gamma^2} \sum_{j=1}^{N} \mathbf{C}_j \mathbf{C}_j^T \nabla V_j^*
\end{aligned}
\tag{3.9}
$$

is asymptotically stable.

(b). Policies $\{u_i^*, d_i^*\}$ constitute a Nash solution.

(a). With γ satisfying the attenuation condition, one has

$$
\begin{aligned}
& V_i(\mathbf{x}) \\
&= \int_t^{\infty} \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{R}_{ij} \mathbf{u}_j - \gamma^2 \sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \right) d\tau \geq 0,
\end{aligned}
$$

where $V_i(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Select $V_i(\mathbf{x})$ as the Lyapunov function candidates. Differentiating $V_i(\mathbf{x})$ yields

$$
\begin{aligned}
\dot{V}_i(\mathbf{x}) &= (\nabla V_i) \left(\mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}_j \mathbf{u}_j + \sum_{j=1}^{N} \mathbf{C}_j \mathbf{d}_j \right) \\
&= - \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{R}_{ij} \mathbf{u}_j - \gamma^2 \sum_{j=1}^{N} \|\mathbf{d}_j\|^2 \right) \leq 0,
\end{aligned}
$$

where $\dot{V}_i(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{0}$. Therefore, $V_i(\mathbf{x})$ are Lynapunov functions and the system (3.9) is asymptotically stable.

(b). Since the system (3.9) is asymptotically stable, one has $\mathbf{x}(t) \to \mathbf{0}$, and thus $V_i(\mathbf{x}(t)) \to 0$, as time $t \to \infty$. The cost function (2.2) can be rewritten as

$$\begin{aligned}
&J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) \\
&= \int_0^\infty \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^N \mathbf{u}_j\mathbf{R}_{ij}\mathbf{u}_j - \gamma^2 \sum_{j=1}^N \|\mathbf{d}_j\|^2 \right) dt \\
&\quad + \int_0^\infty \dot{V}_i\, dt - V_i(\mathbf{x}(\infty)) + V_i(\mathbf{x}(0)) \\
&= \int_0^\infty \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^N \mathbf{u}_j\mathbf{R}_{ij}\mathbf{u}_j - \gamma^2 \sum_{j=1}^N \|\mathbf{d}_j\|^2 \right) dt \\
&\quad + \int_0^\infty \nabla V_i \left(\mathbf{A}\mathbf{x} + \sum_{j=1}^N \mathbf{B}_j\mathbf{u}_j + \sum_{j=1}^N \mathbf{C}_j\mathbf{d}_j \right) dt \\
&\quad + V_i(\mathbf{x}(0)) \\
&= \int_0^\infty H_i(\mathbf{x}, \nabla V_i, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{d}_i, \mathbf{d}_{-i}) dt + V_i(\mathbf{x}(0)).
\end{aligned}$$

Now let $V_i(\mathbf{x}) = V_i^*(\mathbf{x})$ satisfy the HJI equation (3.5), and $\mathbf{u}_{-i}$, $\mathbf{d}_{-i}$ choose the optimal policies. By Theorem 3.0.1 one has

$$\begin{aligned}
&J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}^*_{-i}, \mathbf{d}_i, \mathbf{d}^*_{-i}) \\
&= \int_0^\infty H_i(\mathbf{x}, \nabla V_i^*, \mathbf{u}_i, \mathbf{u}^*_{-i}, \mathbf{d}_i, \mathbf{d}^*_{-i}) dt + V_i^*(\mathbf{x}(0)) \\
&= \int_0^\infty \Big((\mathbf{u}_i - \mathbf{u}_i^*)\mathbf{R}_{ii}(\mathbf{u}_i - \mathbf{u}_i^*) - \gamma^2(\mathbf{d}_i - \mathbf{d}_i^*) \\
&\qquad (\mathbf{d}_i - \mathbf{d}_i^*) \Big) dt + V_i^*(\mathbf{x}(0)),
\end{aligned}$$

which implies that (2.5) is satisfied and hence the system is in Nash equilibrium.

3.0.2 Approximated Solutions Using IRL

In Chapter 3.0.1, we develop the optimal policies for the multi-player H_∞ differential game. As one may notice, the key to finding the policies is solving $V_i^*(x)$ from the HJI Equation (3.5), which is, however, extremely difficult analytically [3]. As such, we propose two IRL-based algorithms to solve the HJI equation numerically.

3.0.2.1 On-Policy IRL

The value function (2.3) can be written as

$$\begin{aligned} &V_i(\mathbf{x}(t)) \\ &= \int_t^{t+T} r_i(\mathbf{x}(\tau), \mathbf{u}_i(\tau), \mathbf{u}_{-i}(\tau), \mathbf{d}_i(\tau), \mathbf{u}_{-i}(\tau))d\tau \\ &\quad + V_i(\mathbf{x}(t+T)), \end{aligned} \tag{3.10}$$

where T is the time interval. Assume that there exits a weight vector $\mathbf{W}_i$ such that the value function can be approximated as

$$V_i(\mathbf{x}) = \mathbf{W}_i \phi_i(\mathbf{x}), \tag{3.11}$$

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

With the approximated value function, the optimal control and disturbance policies can then be determined using RL, in particular, the Policy Iteration (PI) algorithm [1, Page 474]. The PI algorithm constitutes two iterative steps: Policy Evaluation step, to evaluate the value function by (3.10) and (3.11), and Policy Improvement step, to find the optimal policies based on current value function by (3.3) and (3.4). This PI algorithm for the multi-player H_∞ differential game is summarized in Algorithm 1.

Algorithm 1 Policy iteration algorithm for multi-player H_∞ differential game

```
1: Initialize each player with admissible policies $\mathbf{u}_i^{(1)}$ and $\mathbf{d}_i^{(1)}$, $\forall i \in N$.
2: For each iteration $k$, find the value function $V_i^{(k)}(t)$ by
       $V_i^{(k)}(\mathbf{x}(t)) = \int_t^{t+T} r_i\left(\mathbf{x}, \mathbf{u}_i^{(k)}, \mathbf{u}_{-i}^{(k)}, \mathbf{d}_i^{(k)}, \mathbf{d}_{-i}^{(k)}\right) d\tau + \mathbf{W}_i^{(k-1)}\phi_i(\mathbf{x}(t+T)).$   (3.12)
3: Update the weight vector $\mathbf{W}_i^{(k)}$ according to the estimated $V_i^{(k)}(\mathbf{x}(t))$ using the least-squares method,
       $\mathbf{W}_i^{(k)}\phi_i(\mathbf{x}(t)) = V_i^{(k)}(\mathbf{x}(t)).$   (3.13)
```

```
4: Update the policies $\mathbf{u}_i^{(k+1)}$ and $\mathbf{d}_i^{(k+1)}$ for all players as
```

$$\begin{aligned} \mathbf{u}_i^{(k+1)} &= -\frac{1}{2}\mathbf{R}_{ii}^{-1}\mathbf{B}_i\frac{\partial V_i^{(k)}}{\partial \mathbf{x}}, \\ \mathbf{d}_i^{(k+1)} &= \frac{1}{2\gamma^2}\mathbf{C}_i\frac{\partial V_i^{(k)}}{\partial \mathbf{x}}. \end{aligned} \tag{3.14}$$

```
5: Repeat procedures $2 - 4$ until convergence.
```

3.0.2.2 Off-policy IRL

The on-policy algorithm requires the knowledge of the system dynamics, i.e., matrices $\mathbf{B}_i$ and $\mathbf{C}_i$, for learning the optimal policies. In addition, the behavior policies $\mathbf{u}_i$ and $\mathbf{d}_i$ are required to be adjustable at every policy improvement step.

This subchapter develops an off-policy IRL algorithm to learn the optimal policies without any information of the system dynamics. The off-policy IRL learns the optimal policies of the game online while the game is being played based on fixed behavior policies $\mathbf{u}_i$ and $\mathbf{d}_i$, which are used to generate system data [11]. This result is developed for the case when players have identical dynamics, i.e., $\mathbf{B}_j = \mathbf{B}$ and $\mathbf{C}_j = \mathbf{C}$, for all $j = 1, 2, \cdots, N$.

We write the system dynamics in the following form:

$$\begin{aligned} \dot{\mathbf{x}} =& \mathbf{A}\mathbf{x} + \sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j^{(k)} + \sum_{j=1}^{N}\mathbf{C}\mathbf{d}_j^{(k)} \\ &+ \sum_{j=1}^{N}\mathbf{B}\left(\mathbf{u}_j - \mathbf{u}_j^{(k)}\right) + \sum_{j=1}^{N}\mathbf{C}\left(\mathbf{d}_j - \mathbf{d}_j^{(k)}\right), \end{aligned} \tag{3.15}$$

where $\mathbf{u}_j^{(k)}$ and $\mathbf{d}_j^{(k)}$ are the policies to be updated for the optimal solutions.

Differentiation the value $V_i^{(k)}(\mathbf{x}(t))$ along with the system dynamics (3.15) and using (3.1), (3.14) yield

$$
\begin{aligned}
&\dot{V}_i^{(k)}(\mathbf{x}(t)) \\
&= \nabla V_i^{(k)} \left(\mathbf{A}\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}\mathbf{u}_j^{(k)} + \sum_{j=1}^{N} \mathbf{C}\mathbf{d}_j^{(k)} \right) \\
&\quad + \nabla V_i^{(k)} \left(\sum_{j=1}^{N} \mathbf{B} \left(\mathbf{u}_j - \mathbf{u}_j^{(k)} \right) + \sum_{j=1}^{N} \mathbf{C} \left(\mathbf{d}_j - \mathbf{d}_j^{(k)} \right) \right) \\
&= - \left(\mathbf{x}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{(k)} \mathbf{R}_{ij} \mathbf{u}_j^{(k)} - \gamma^2 \sum_{j=1}^{N} \| \mathbf{d}_j^{(k)} \|^2 \right) \\
&\quad - 2\mathbf{u}_i^{(k+1)} \mathbf{R}_{ii} \sum_{j=1}^{N} \left(\mathbf{u}_j - \mathbf{u}_j^{(k)} \right) \\
&\quad + 2\gamma^2 \mathbf{d}_i^{(k+1)} \sum_{j=1}^{N} \left(\mathbf{d}_j - \mathbf{d}_j^{(k)} \right).
\end{aligned}
\tag{3.16}
$$

Integrating (3.16) from both sides gives the following off-policy IRL Bellman equation:

$$
\begin{aligned}
&V_i^{(k)}(\mathbf{x}(t+T)) - V_i^{(k)}(\mathbf{x}(t)) \\
&\qquad = \int_t^{t+T} \Bigg(- \mathbf{x}\mathbf{Q}_i\mathbf{x} - \sum_{j=1}^{N} \mathbf{u}_j^{(k)} \mathbf{R}_{ij} \mathbf{u}_j^{(k)} \\
&\qquad\qquad\qquad + \gamma^2 \sum_{j=1}^{N} \| \mathbf{d}_j^{(k)} \|^2 \Bigg) d\tau \\
&\qquad\quad + \int_t^{t+T} \Bigg(- 2\mathbf{u}_i^{(k+1)} \mathbf{R}_{ii} \sum_{j=1}^{N} \left(\mathbf{u}_j - \mathbf{u}_j^{(k)} \right) \\
&\qquad\qquad\qquad + 2\gamma^2 \mathbf{d}_i^{(k+1)} \sum_{j=1}^{N} \left(\mathbf{d}_j - \mathbf{d}_j^{(k)} \right) \Bigg) d\tau.
\end{aligned}
\tag{3.17}
$$

Note that for any fixed admissible control and disturbance policies $\mathbf{u}_i$ and $\mathbf{d}_i$, (3.17) can be solved for value function $V_i^{(k)}$ and the updated policies $\mathbf{u}_i^{(k+1)}$ and $\mathbf{d}_i^{(k+1)}$ simultaneously. To this end, three neural networks (NNs), i.e., the critic NN, the actor

NN, and the disturber NN, are used here for approximating the value function and the updated control and disturbance policies respectively:

$$\begin{aligned} V_i^{(k)}(\mathbf{x}) &= \mathbf{W}_i^{(k)}\phi_i(\mathbf{x}), \\ \mathbf{u}_i^{(k+1)}(\mathbf{x}) &= \mathbf{W}_{u,i}^{(k+1)}\sigma_i(\mathbf{x}), \\ \mathbf{d}_i^{(k+1)}(\mathbf{x}) &= \mathbf{W}_{d,i}^{(k+1)}\psi_i(\mathbf{x}), \end{aligned} \tag{3.18}$$

where $\phi_i(\mathbf{x})$, $\sigma_i(\mathbf{x})$ and $\psi_i(\mathbf{x})$ provide suitable basis function vectors, and $\mathbf{W}_i^{(k)}$, $\mathbf{W}_{u,i}^{(k+1)}$ and $\mathbf{W}_{d,i}^{(k+1)}$ are weight matrices with proper dimensions.

The implementation of the off-policy IRL algorithm is described in Algorithm 2.

Algorithm 2 Off-policy IRL algorithm for multi-player H_∞ differential game

```
1: Initialize each player with admissible policies $\mathbf{u}_i^{(1)}$ and $\mathbf{d}_i^{(1)}$, $\forall i$.
2: For each iteration $k$, solve (3.17) for $V_i^{(k)}$, $\mathbf{u}_i^{(k+1)}$, and $\mathbf{d}_i^{(k+1)}$ simultaneously.
3: Update $\mathbf{W}_i^{(k)}$, $\mathbf{W}_{u,i}^{(k+1)}$ and $\mathbf{W}_{d,i}^{(k+1)}$ according to the derived $V_i^{(k)}$, $\mathbf{u}_i^{(k+1)}$, $\mathbf{d}_i^{(k+1)}$ by
   (3.18) using the least-squares method.
4: Repeat procedures $2 - 3$ until convergence.
```

CHAPTER 4

Simulation Studies

In this chapter, the two proposed algorithms are applied to a linear system example to validate the theoretical analysis.

Consider a three-player H_∞ game with a linear system described by the following dynamics:

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 0.25 \\ 1 & 0 \end{bmatrix} \mathbf{x} + \sum_{j}^{3} \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_j + \sum_{j}^{3} \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{d}_j, \tag{4.1}$$

where $\mathbf{x} = [x_1, x_2]$.

The parameters in the value function (2.3) are selected as:

$$\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\mathbf{R}_{12} = \mathbf{R}_{13} = \mathbf{R}_{21} = \mathbf{R}_{23} = \mathbf{R}_{31} = \mathbf{R}_{32} = 1,$$

$$\mathbf{R}_{11} = 2, \mathbf{R}_{22} = 3, \mathbf{R}_{33} = 5,$$

and $\gamma = 5$. The reinforcement learning interval T is chosen to be 0.1.

The on-policy PI algorithm (Algorithm 1) is implemented first. We select the basis function $\phi_i = [x_1^2, x_1 x_2, x_2^2]$ with weight vector $\mathbf{W}_i = [W_{i1}, W_{i2}, W_{i3}]$, where $i = 1, 2, 3$. Figure 4.1 and 4.2 show the evolution of the system states and value function weights.

Figure 4.1 shows that the system states converge to $\mathbf{0}$ when the optimal policies are applied to the system (4.1). Moreover, Figure 4.2 verifies the convergence of value function weights, from which the optimal policies can be derived.

Then we simulate the off-policy IRL algorithm (Algorithm 2). Here, three NNs are selected as follows: the critic NN $\phi_i = [x_1^2, x_1x_2, x_2^2]$ with a weight vector $\mathbf{W}_i = [W_{i1}, W_{i2}, W_{i3}]$; the actor NN $\sigma_i = [x_1, x_2]$ with a weight vector $\mathbf{W}_{u,i} = [W_{u,i1}, W_{u,i2}]$; the disturber NN $\psi_i = [x_1, x_2]$ with a weight vector $\mathbf{W}_{d,i} = [W_{d,i1}, W_{d,i2}]$, where $i = 1, 2, 3$. The simulation results are shown in Figure 4.3 and Figure 4.4.

Figure 4.4 shows that the value function weights converge in limited time using the proposed off-policy IRL algorithm, and the converged values are identical to the ones derived from the on-policy algorithm. In addition, the system states converge to $\mathbf{0}$, which validate the asymptotic stability of the closed-loop system.In addition, we find that the HJI Equation (3.5) holds after substituting the derived value function, which verifies the correctness of the derived solutions (3.13), (3.14) and (3.18).

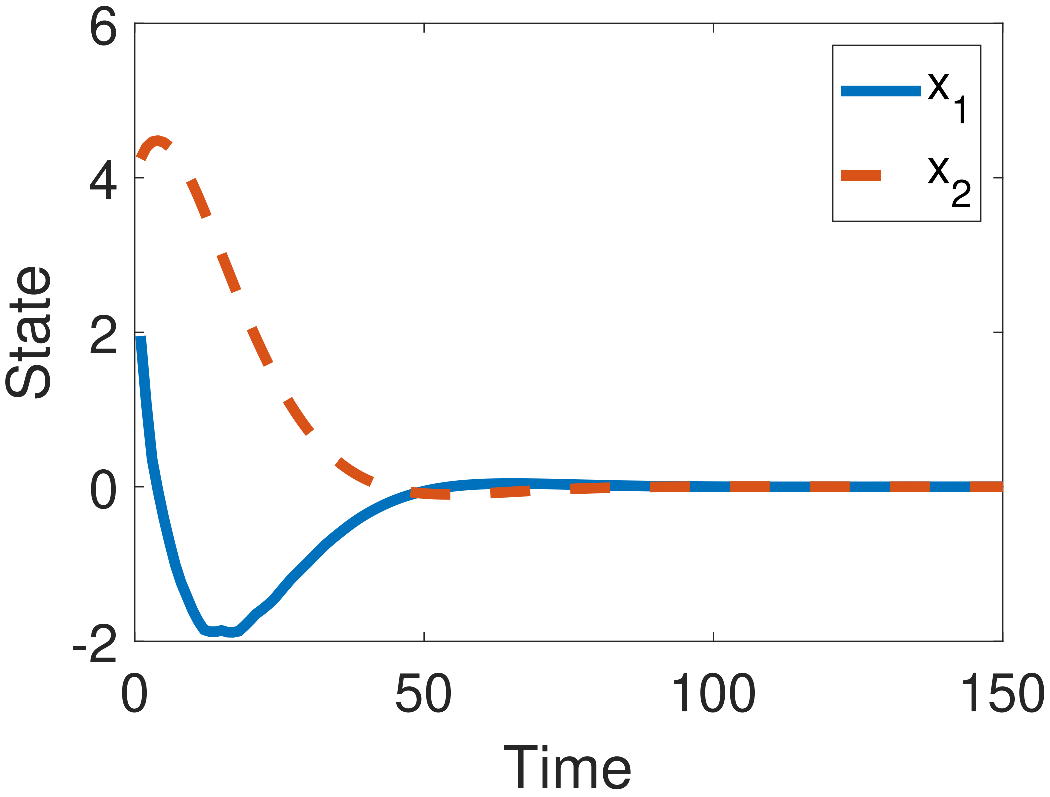

Figure 4.1: The evolution of the system states using on-policy IRL.

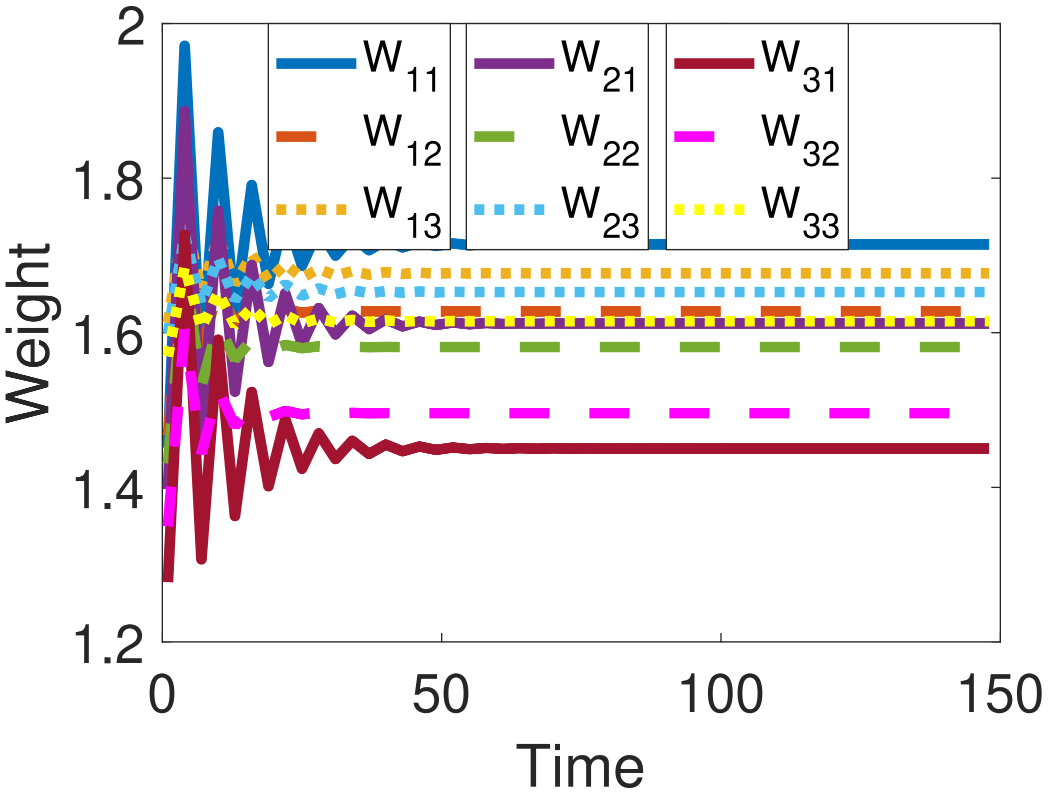

Figure 4.2: The derived value function weights using on-policy IRL.

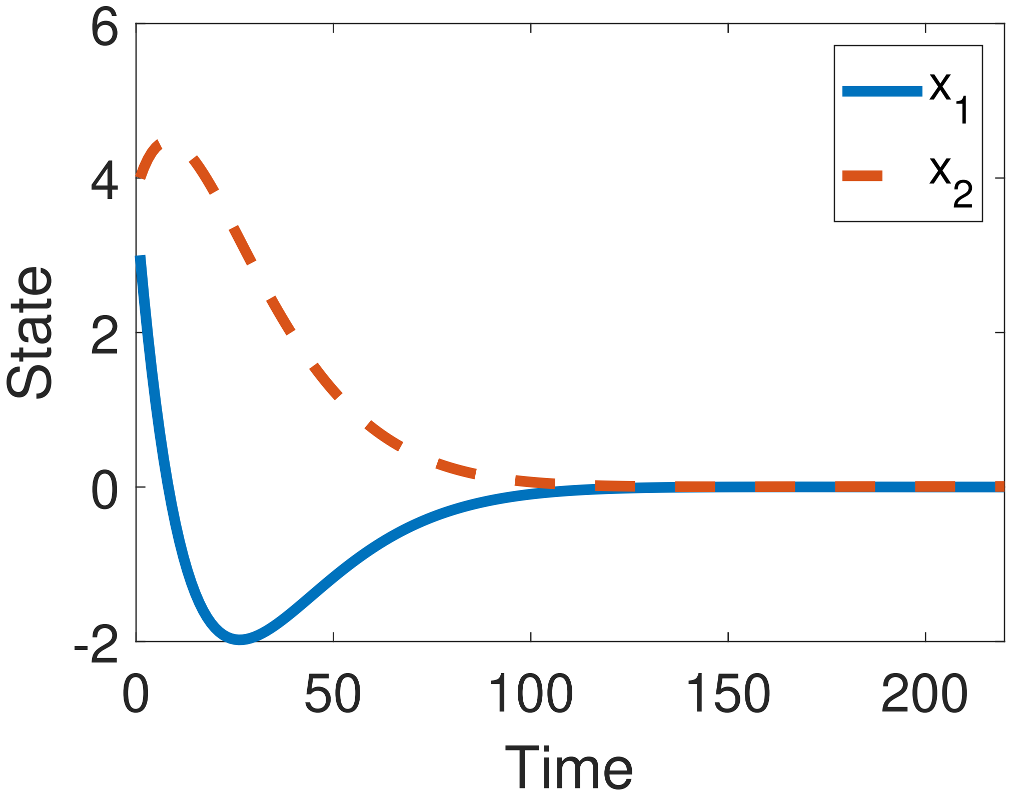

Figure 4.3: The evolution of the system states using off-policy IRL.

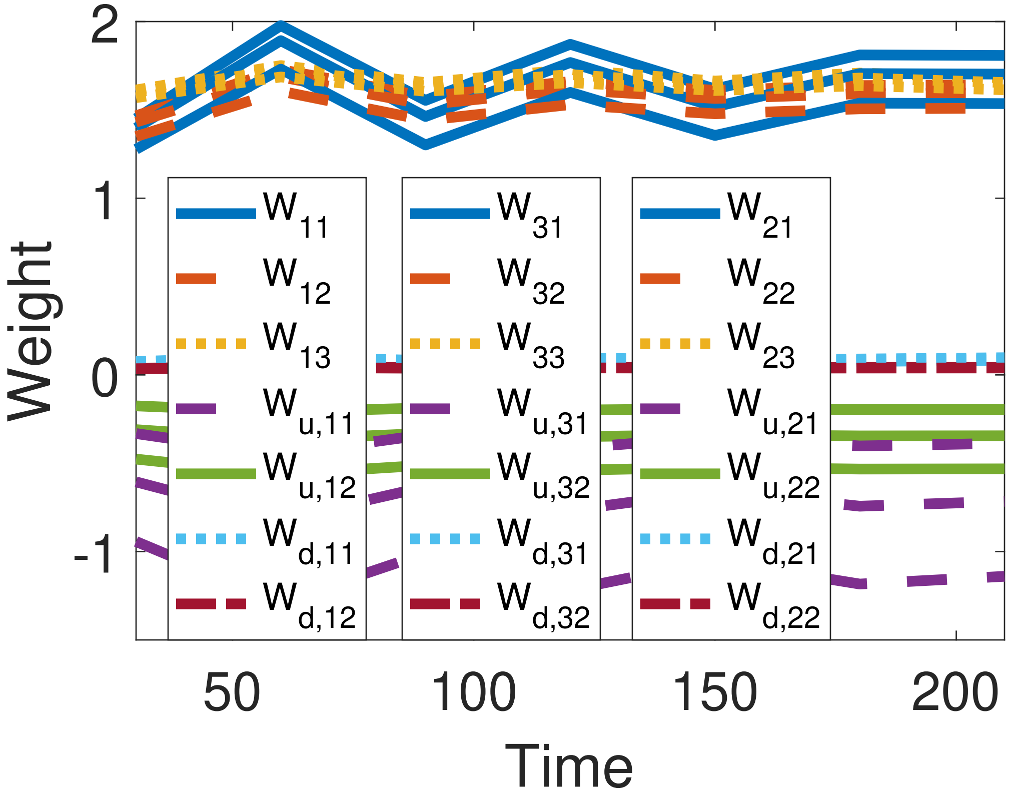

Figure 4.4: The derived value function weights using off-policy IRL.

CHAPTER 5

CONCLUSIONS

This work studies a new differential game that takes features of two existing games, i.e., two-player zero-sum and multi-player nonzero-sum games, to solve the optimal control problems of multi-player systems subject to external disturbances. We showed that the optimal solutions to this differential game can be found by solving the HJI equation, and the derived optimal solutions can make the system asymptotically stable and in Nash equilibrium. Moreover, to solve the differential games online, we designed two IRL-based algorithms, including the policy iteration and off-policy IRLs. In particular, the designed off-policy IRL can find the Nash solutions without using any information of the system dynamics. In the future, this work can be generalized to systems with general nonlinear dynamics, and the designed algorithms can be applied to the real-world applications.

REFERENCES

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

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

[3] 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.

[4] M. Liu, Y. Wan, F. Lewis, and V. G. Lopez, “Adaptive optimal control for stochastic multi-player differential games using on-policy and off-policy reinforcement learning,” *accepted by IEEE Transactions on Neural Network and Learning Systems*, 2020.

[5] A. Perelman, T. Shima, and I. Rusnak, “Cooperative differential games strategies for active aircraft protection from a homing missile,” *Journal of Guidance, Control, and Dynamics*, vol. 34, no. 3, pp. 761–773, 2011.

[6] T. Mylvaganam, M. Sassano, and A. Astolfi, “A differential game approach to multi-agent collision avoidance,” *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 4229–4235, 2017.

[7] A. W. Starr and Y.-C. Ho, “Nonzero-sum differential games,” *Journal of optimization theory and applications*, vol. 3, no. 3, pp. 184–206, 1969.

[8] M. Liu, Y. Wan, F. L. Lewis, and V. G. Lopez, “Stochastic two-player zero-sum learning differential games,” in *Proceedings of IEEE 15th International Conference on Control and Automation (ICCA)*, Edinburgh, Scotland, 2019.

[9] R. Song, Q. Wei, and B. Song, “Neural-network-based synchronous iteration learning method for multi-player zero-sum games,” *Neurocomputing*, vol. 242, pp. 73–82, 2017.

[10] H. Jiang, H. Zhang, J. Han, and K. Zhang, “Iterative adaptive dynamic programming methods with neural network implementation for multi-player zero-sum games,” *Neurocomputing*, vol. 307, pp. 54–60, 2018.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] D. Vrabie and F. Lewis, "Integral reinforcement learning for online computation of feedback nash strategies of nonzero-sum differential games," in *Proceedings of IEEE Conference on Decision and Control (CDC)*, Atlanta, GA, 2010.

[22] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 704–713, 2017.

[23] P. An, M. Liu, Y. Wan, and F. L. Lewis, "Multi-player h differential game using on-policy and off-policy reinforcement learning," in *2020 IEEE 16th International Conference on Control Automation (ICCA)*, 2020, pp. 1137–1142.

BIOGRAPHICAL STATEMENT

Peiliang An was born in Qinhuangdao, Hebei, China in 1993. In June of 2016, he received his Bachalor's degree in Mechanical Engineering from Huazhong University of Science and Technology. In June of 2018, he received his Master's degree in Mechanical and Aerospace Engineering from The University of California, Irvine. Peiliang joined Dynamical Networks and Control Lab under Dr. Wan supervision at the University of Texas at Arlington in 2019 to pursue further research in multi-agent systems and UAV applications. His research interests include autonomous systems, robotics and multi-agent systems. He worked as a graduate teaching and research assistant at The University of Texas at Arlington.