DISTRIBUTED OPTIMAL POLICIES FOR MULTI-AGENT SYSTEMS UNDER UNCERTAINTIES

by

MUSHUANG LIU

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

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2020

ACKNOWLEDGEMENTS

I deeply thank my supervising professor Dr. Yan Wan for constantly motivating and encouraging me, and also for all the invaluable advice she provided me during all these years. With the same emphasis I thank my coadviser, Dr. Frank L. Lewis, for his constant support and assistance to perform my research.

I also want to thank the members of my dissertation committee, Dr. Jonathan Bredow, Dr. Robert Magnusson and Dr. Ramtin Madani, for their helpful comments and unreserved willingness to devote their time to evaluate my research.

I am grateful to all of my collaborators and friends at UTA, Dr. Victor Gabriel Lopez Mejia, Dr. Songwei Li, Peiliang An, Chenyuan He, Lu Zhao, and Bishrut Subedi, for their invaluable support and friendship throughout this experience.

Finally, I would like to express my deep gratitude to my parents for their advice, encouragement, and unconditional support. It is much easier to move forward when I know that I always have their back up.

April 23, 2020

ABSTRACT

DISTRIBUTED OPTIMAL POLICIES FOR MULTI-AGENT SYSTEMS UNDER UNCERTAINTIES

Mushuang Liu, Ph.D.

The University of Texas at Arlington, 2020

Supervising Professor: Yan Wan

Multi-agent systems (MAS) have attracted increasing attention in the past years due to their wide applications in mobile robots, sensor networks, autonomous driving systems, etc. Along with this trend, developing distributed optimal policies for MAS under uncertainties has become indispensable. Multi-dimensional uncertainties often modulate system dynamics in a complicated fashion, which leads to computational challenges for real-time control. In many practical MAS, each agent also has its own interest to optimize beyond a global objective. Developing distributed optimal control for agents with self-interests is needed. To address the above challenges, this dissertation contributes in two major directions for MAS: 1) computationally-effective real-time optimal policies under multi-dimensional uncertainties, and 2) distributed optimal policies in networked MAS, including graphical games.

In the first direction, we develop a framework to solve optimal control problems for MAS that involve multi-dimensional uncertainties. Two types of uncertain systems are investigated: 1) MAS subject to uncertain agent intentions, and 2) MAS operating in uncertain environments. For the first type, we use stochastic switching

models to capture the uncertain intentions and develop an online optimal control solution that integrates an effective uncertainty sampling method called multivariate probabilistic collocation method (MPCM), reinforcement learning, and random state estimation. For the second type, we formulate and investigate two new stochastic differential games, where the system parameters are modulated by multi-dimensional uncertainties. Effective online learning algorithms are designed to solve these games with computational efficiency.

In the second direction, we study the distributed control, Nash optimality, and robustness properties of networked MAS. We point out that in existing graphical game formulations, being global Nash and being distributed are two contradicting properties. We then propose a new graphical game formulation, which promises the existence of solutions that are both distributed and Nash. In addition, we develop a new Lyapunov-based analytical framework for the robustness of networked MAS measured by gain and phase margins. The effect of communication graph topology on the stability margins is analyzed.

Beyond the theoretical contributions, we also apply the developed solutions to diverse practical applications, including air-to-air unmanned aerial vehicle (UAV) communications, UAV traffic management (UTM), and damage pattern estimations in composite materials.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Multi-agent systems (MAS) have attracted increasing attention in the past years due to their wide applications in mobile robots, sensor networks, autonomous driving systems, etc. Along with this trend, developing distributed optimal policies for MAS under uncertainties has become indispensable. Multi-dimensional uncertainties often modulate system dynamics in a complicated fashion, which leads to computational challenges for real-time control. In many practical MAS, each agent also has its own interest to optimize beyond a global objective. Developing distributed optimal control for agents with self-interests is needed. To address the above challenges, this dissertation contributes results in two major directions for MAS: 1) computationally-effective real-time optimal policies under multi-dimensional uncertainties, and 2) distributed optimal policies in networked MAS, including differential graphical games. These results are documented in the following papers [2, 4–22].

1. Computationally-Effective Real-Time Optimal Policies under Multi-Dimensional Uncertainties

   The dynamics of modern dynamical systems are often modulated by multi-dimensional uncertainties in a complicated fashion. The uncertainties can arise either from the agents' uncertain intentions, or from uncertain environment conditions, such as probabilistic weather forecasts. They lead to challenges for real-time control, considering the significant computational load needed to evaluate these uncertainties. To deal with these challenges, we developed in our previous work an effective

uncertainty evaluation method, called multivariate probabilistic collocation method (MPCM) [23]. The MPCM estimates accurately the mean output of a system modulated by multi-dimensional uncertainties with very small computational load. In this dissertation, we further investigate the use of MPCM in developing optimal control solutions for MAS of multi-dimensional uncertainties. Two types of uncertain systems are studied: 1) MAS subject to uncertain agents' intentions, and 2) MAS operating in uncertain environments.

MAS subject to Uncertain Intentions Driven by the emergence of internet of-things (IOT) applications, mobile agents play increasingly important roles in optimal decision processes. Random mobility models (RMMs) have been widely used to capture the uncertain intentions of mobile agents. However, decision-making for agents of RMMs has not been explored. Motivated by this need, we develop a novel online optimal control framework for MAS governed by RMM, based on the modules of RL and MPCM, and the properties of random switching models. To overcome the issues caused by unavailable state information, we also develop a novel estimator that integrates the unscented Kalman filter (UKF) and MPCM. This work lays a foundation to analyze the behaviors of randomly moving mobile agents, and enables decision-making for agents with uncertain intentions.

MAS operating in Uncertain Environments Control-theoretic differential games have been recently explored to study the interactions of multiple agents with individual payoffs. However, most existing studies on differential games assume deterministic dynamics, which is not practical in real environments that involve uncertainties. We propose two new stochastic differential games, including two-player zero-sum and multi-player nonzero-sum stochastic games, the system dynamics of which are modulated by multi-dimensional time-varying parameters to capture the effects of uncertain environments. This is the first time in the literature that uncertain environments are

considered in differential games. Effective online learning algorithms are then designed to solve these games with computational efficiency, based on the modules of on-policy and off-policy integral RL (IRL), and the MPCM. This work provides a method to analyze the effects of uncertain environments on MAS of self-interests.

2. Distributed Optimal Policies in Networked MAS

In many practical systems, agents communicate on a graph, and determine their policies based on limited information constrained by the network topology. We study properties of distributed optimal policies for agents of self-interests in networked MAS, including distributed control, Nash optimality, and robustness properties. In this progress, new analytical frameworks for differential graphical games and stability margins of networked MAS are developed respectively.

Differential Graphical Games To solve the distributed optimal control for networked MAS of self-interests, differential graphical games have been recently explored in the literature. We point out that in existing graphical game formulations, being global Nash and being distributed are two contradicting properties. To address this issue, we propose in this dissertation a novel differential graphical game formulation, and prove that this formulation promises the existence of solutions that are both distributed and Nash. This work provides a new perspective on distributed optimal control of networked MAS, and makes it possible to achieve global Nash equilibrium in networked MAS using only local information.

Stability Margins of Networked MAS Despite the abundant literature on networked MAS, few work is concerned with the robustness properties of networked MAS. We take the perspective that the robustness of a MAS is affected by communication structure and hence the design of MAS should consider communication structure to achieve desired robustness levels. We develop a new Lyapunov-based analytical framework that determines the stability margins of networked MAS of local linear

quadratic regulator (LQR) design. For the first time in the literature, we extend the phase and gain margins analysis from single-agent systems to MAS, and show that the directed tree communication graph promises the best stability margins among all communication graph topology. This work lays foundations in analyzing the effects of communication graph topology on the robustness of networked MAS, and enables communication structure design in improving the robustness of networked MAS.

1.2 Organization of this Dissertation

The structure of this dissertation is shown in Figure 1.1. The content of each chapter is listed as follows.

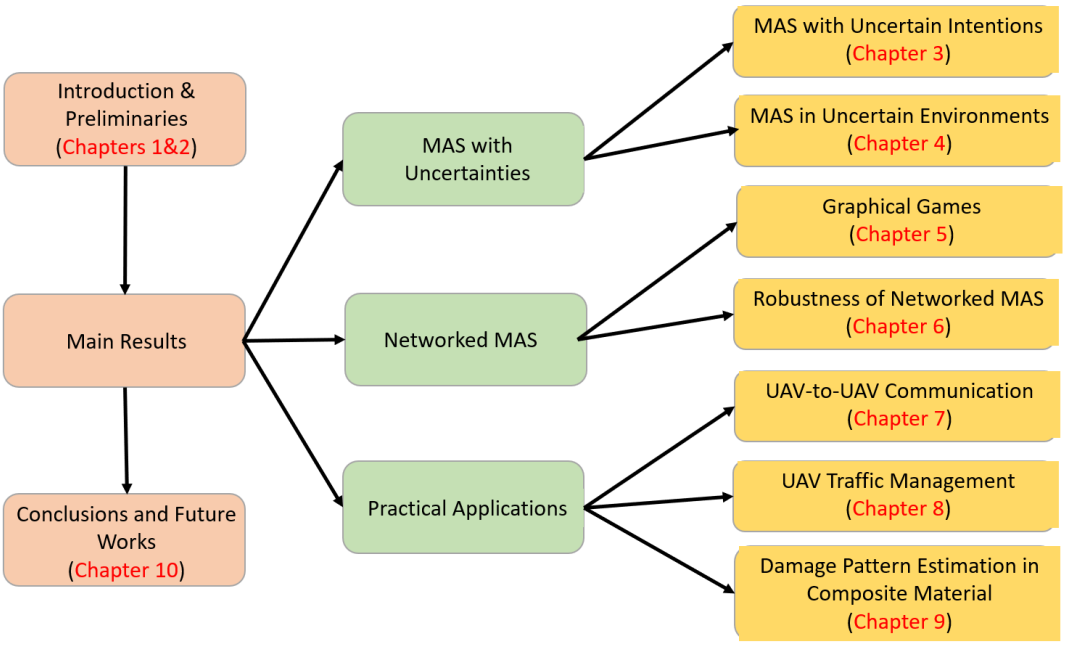


Figure 1.1. Organization of this dissertation.

In Chapter 2, we introduce some preliminaries to facilitate the analysis in this dissertation. In particular, an effective uncertainty evaluation method, the MPCM

(documented in the book chapter [8]), and an effective state estimation method for nonlinear systems, the UKF, are reviewed respectively.

In Chapter 3, we develop a novel online optimal control solution to adaptively find the optimal policies for MAS with random switching mobility models, based on the modules of RL and MPCM. We also develop a novel estimator that integrates the UKF and MPCM to provide online estimation solutions for these agents. Efficiency and accuracy of the proposed solutions are analyzed. A concrete communication and antenna control co-design problem for a multi-unmanned aerial vehicle (UAV) network is analyzed in the end to illustrate and validate the results. The results are documented in paper [4] published in IEEE Control System Letters.

In Chapter 4, we study stochastic multi-player 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 multi-player nonzero-sum games. We then show that optimal control policies, which constitute Nash equilibrium solutions, can be derived from the corresponding Hamiltonian functions. Stability is proved using the Lyapunov type of analysis. In order to solve the stochastic differential games online, we integrate RL and MPCM. Two learning algorithms, including the on-policy and off-policy IRL, are designed for the formulated games respectively. We show that the proposed learning algorithms can effectively find Nash equilibrium solutions for the stochastic multi-player differential games. The results are documented in paper [6] published in IEEE Transactions on Neural Networking and Learning Systems.

In Chapter 5, we study multi-agent differential graphical games in linear dynamical networks. We prove that the best response strategy, most widely-used in solving graphical games, can constitute Nash, but does not provide distributed solutions. On the other hand, the minmax strategy, which is recently developed in solving graph-

ical games, can provide distributed solutions, but prevents the agents from reaching a global Nash equilibrium. In this chapter, we propose a novel differential graphical game formulation, which promises the existence of solutions that 1) can constitute global Nash equilibrium, and 2) are distributed in the sense that each agent only uses the state information of its own and its direct neighbors. Stability and Nash equilibrium properties of the proposed graphical game are proven, respectively. Simulation studies are conducted to illustrate the theoretical results. The results are documented in paper [7] submitted to IEEE Transactions on Control of Network Systems.

In Chapter 6, we develop a theoretical framework for the analysis of stability margins, i.e., phase margin and gain margin, of networked MAS. It is well-known that a single-agent LQR system can guarantee 60° phase margin and infinite gain margin. However, for networked MAS, there exist no theoretical results on guaranteed stability margins, due to the complexity caused by the interplay of communication structure and agent dynamics. In this chapter, we analyze the effect of communication graph topology on the robustness properties of networked cooperative tracking systems with local LQR designs. For such systems, we provide closed-form expressions of phase and gain margins modulated by their graph topology, following a Lyapunov type of analysis. We further derive upper bounds of phase and gain margins for MAS of general graph topology, through a structural analysis based on the algebraic graph theory. Results show that for networked MAS of local LQR design, the robustness performance is upper bounded by that of a single-agent system, in terms of guaranteed phase and gain margins. In addition, we prove that the directed tree communication topology promises the best stability margin performance for networked MAS among all possible communication topology, and has the same guaranteed gain and phase margin as single-agent LQR systems. Simulation studies are conducted to validate

the theoretical results. The results are documented in paper [9] submitted to IEEE Transactions on Automatic Control.

In Chapter 7, we consider applying the optimal control for MAS under uncertain intentions to the aerial communication application. The aerial communication using directional antennas (ACDA) system is a promising solution to enable long-distance and broad-band UAV-to-UAV networking. The automatic alignment of directional antennas allows the transmission energy to focus in certain direction and significantly extends the communication range and rejects interference. Robust automatic alignment of directional antennas is not easy to achieve, considering practical issues such as the limited on-board sensing devices due to the physical constraints of UAV payload and power supplies, uncertain and varying UAV movement patterns, and unstable GPS and unknown communication environments. In this chapter, we develop RL-based online antenna control solutions for the ACDA system to conquer these challenges. The control solution adopts an uncertain UAV mobility modeling and intention estimation framework to capture and predict the uncertain intentions of UAV maneuvers and hence permit robust tracking. To account for an unstable GPS environment, the control solution features a learning of communication channel models to provide additional measurement signals in GPS-denied settings. A novel stochastic optimal control solution for nonlinear random switching dynamics is developed that integrates RL, MPCM, and UKF. Simulation studies are conducted to illustrate and validate the proposed solutions. The results are documented in paper [5] published in IEEE Transactions on Vehicle Technology.

In Chapter 8, we study the UAV traffic management(UTM) application . We provide statistical analysis of RMMs equipped with physical sense-and-avoid protocols. In particular, we propose a new modeling framework of random mobility models equipped with physical sense-and-avoid protocols to capture the flexible, variable, and

uncertain movement patterns of UAVs subject to the separation safety constraints. For the random direction (RD) RMM equipped with a commonly used sense-and-avoid (S&A) protocol, named sense-and-stop (S&S), we provide its statistical properties including stationary location distribution and stationary inter-vehicle distance distribution, using the Markov analysis. This study provides knowledge on the impact of S&A protocols to critical UAV networking statistics. In addition, we define collision probabilities and airspace capacity concepts for UAVs based on the inter-vehicle distance distribution, and derive their closed-form expressions. This analytical framework mathematically bridges local autonomy with global airspace capacity, and allows the impact analysis of local autonomy configurations for effective UAV airspace capacity management. The results are documented in paper [10] submitted to IEEE Transactions on Intelligent Transportation Systems.

In Chapter 9, we study damage pattern estimation in composite materials. Composite materials play important roles in multi-functional applications, and the diagnosis of damage patterns in composite materials is crucial to avoid "critical events" such as structural or functional failures. The impact of an individual damage in composite materials has been extensively studied, however, the interaction of defects/cracks, which leads to critical fracture paths, has not been understood well. In this chapter, we develop a Bayesian estimation based statistical analysis technique that estimates the damage pattern of a composite material, in particular, the relative positions of defects in the material, by measuring its through-thickness dielectric properties. We first explain the fundamental dielectric principle that leads to the detection of defect patterns. A capacitance model is then built to measure the material permittivity, and the relationship between the dielectric permittivity and relative positions are found using COMSOL Multiphysics®. The interaction effects between defects observed in the simulation are interpreted using the fundamental dielectric

principle. A Bayesian estimation based statistical analysis model is then developed to estimate the relative positions of defects in composite materials from the measured global dielectric properties. The results are documented in paper [14] published in 2019 SPIE on Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, and Civil Infrastructure.

CHAPTER 2

PRELIMINARIES

In this chapter, we introduce some preliminaries to facilitate the development in this dissertation. To prepare for the development of optimal decision under uncertainty, we introduce an effective uncertainty evaluation method, the MPCM. To prepare for the state estimation in nonlinear systems, we review an effective nonlinear estimation method, the UKF.

2.1 Effective Uncertainty Evaluation Method

This section formulates the uncertainty evaluation problem and introduces MPCM for effective uncertainty evaluation.

2.1.1 Problem Formulation

We first formulate an general uncertainty evaluation problem as follows. Consider a system $G(\cdot)$ modulated by m-dimensional uncertainties, $a_1, a_2, \cdots, a_m$, which are considered as inputs to a black box (see Figure 2.1). The output of the system, y, is the system performance of interest. The goal is to correctly estimate the mean output of the systems, i.e., $E[y]$, given the statistical information of the uncertain inputs, e.g., pdfs of a_p, $f_{A_p}(a_p)$.

2.1.2 The MPCM

The MPCM is an effective uncertainty evaluation method, which estimates the mean output of a system modulated by multi-dimensional uncertainties accurately

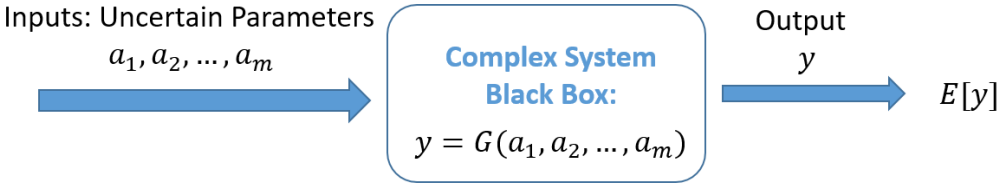


Figure 2.1. Estimation of mean output for a system modulated by m-dimensional uncertainties. .

and effectively. The MPCM smartly selects a limited number of sample points according to the Gaussian Quadrature rules, and runs simulations at these samples to produce a reduced-order mapping, which has the same mean output as that of the original mapping [23]. The main properties of the MPCM are summarized as follows.

Theorem 1. *[23, Theorem 2] Consider a system mapping modulated by m independent uncertain parameters:*

$$G\left(a_1, a_2, \cdots, a_m\right) = \sum_{q_1=0}^{2n_1-1} \sum_{q_2=0}^{2n_2-1} \cdots \sum_{q_m=0}^{2n_m-1} \psi_{q_1,q_2,\cdots,q_m} \prod_{p=1}^{m} a_p^{q_p}, \tag{2.1}$$

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

$$G'\left(a_1, a_2, \cdots, a_m\right) = \sum_{q_1=0}^{n_1-1} \sum_{q_2=0}^{n_2-1} \cdots \sum_{q_m=0}^{n_m-1} \Omega_{q_1,q_2,\cdots,q_m} \prod_{p=1}^{m} a_p^{q_p}, \tag{2.2}$$

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

Theorem 1 shows that the MPCM reduces the number of simulations from $2^m \prod_{p=1}^{m} n_p$ to $\prod_{p=1}^{m} n_p$, where m is the number of uncertain parameters. Despite the dramatic reduction of computation by a factor of 2^m, MPCM evaluates the mean

output of system (2.1) precisely. The MPCM design procedure is summarized as follows [23].

Algorithm 1 MPCM

Step 1: Simulation point selection

1: Compute the orthonormal polynomials, $h_p^{n_p}(a_p)$, of degree n_p for the random variable a_p, for $p = 1, 2, \cdots, m$ according to Steps 2-7:

2: **For** $p = 1$ to m

3: Initialize $H_p^{-1}(a_p) = h_p^{-1}(a_p) = 0$ and $H_p^0(a_p) = h_p^0(a_p) = 1$.

4: **For** $q_p = 1$ to n_p

$$\begin{aligned} H_p^{k_p}(a_p) =& a_p h_p^{q_p-1}(a_p) - \left\langle a_p h_p^{q_p-1}(a_p), h_p^{q_p-1}(a_p) \right\rangle h_p^{q_p-1}(a_p) \\ & - \left\langle H_p^{q_p-1}(a_p), H_p^{a_p-1}(a_p) \right\rangle^{\frac{1}{2}} h_p^{q_p-2}(a_p), \end{aligned}$$

$$h_p^{q_p}(a_p) = H_p^{q_p}(a_p) / \left\langle H_p^{q_p}(a_p), H_p^{q_p}(a_p) \right\rangle^{\frac{1}{2}}.$$

5: **End**

6: **End**

7: Find the roots of $h_p^{n_p}(a) = 0$ as the n_p PCM simulation points for a_p, denoted as $a_{p(1)}, a_{p(2)}, \cdots, a_{p(n_p)}$.

Step 2: Evaluation of system outputs at selected simulation points

8: For each m-tuple simulation point $(a_{1(r_1)}, a_{2(r_2)}, \cdots, a_{m(r_m)})$ found in Step 1, where $r_p \in \{1, 2, \cdots, n_p\}$, run simulation and find the associated output

9: $G\left(a_{1(r_1)}, x_{2(r_2)}, \cdots, x_{m(r_m)}\right)$.

Step 3: Mean output evaluation

10: Find the coefficients $b_{q_1,q_2,\cdots,q_m}$ in the low-order PCM mapping $G'(a_1, a_2, \cdots, a_m) = \sum_{q_1=0}^{n_1-1} \sum_{q_2=0}^{n_2-1} \cdots \sum_{q_m=0}^{n_m-1} b_{q_1,q_2,\cdots,q_m} \prod_{p=1}^{m} h_p^{q_p}(a_p)$ following:

$$\begin{bmatrix} b_{0,\cdots,0} \\ b_{0,\cdots,1} \\ \vdots \\ b_{n_1-1,\cdots,n_m-1} \end{bmatrix} = \Gamma^{-1} \begin{bmatrix} G\left(a_{1(1)},\cdots,a_{m(1)}\right) \\ G\left(a_{1(1)},\cdots,a_{m(2)}\right) \\ \vdots \\ G\left(a_{1(n_1)},\cdots,a_{m(n_m)}\right) \end{bmatrix},$$

where

$$\Gamma = \begin{bmatrix} h_1^0(a_{1(1)})\cdots h_m^0(a_{m(1)}) & h_1^1(a_{1(1)})\cdots h_m^1(a_{m(1)}) & \cdots & h_1^{n_1-1}(a_{1(1)})\cdots h_m^{n_m-1}(a_{m(1)}) \\ h_1^0(a_{1(1)})\cdots h_m^0(a_{m(2)}) & h_1^1(a_{1(1)})\cdots h_m^1(a_{m(2)}) & \cdots & h_1^{n_1-1}(a_{1(1)})\cdots h_m^{n_m-1}(a_{m(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ h_1^0(a_{1(n_1)})\cdots h_m^0(a_{m(n_m)}) & h_1^1(a_{1(n_1)})\cdots h_m^1(a_{m(n_m)}) & \cdots & h_1^{n_1-1}(a_{1(n_1)})\cdots h_m^{n_m-1}(a_{m(n_m)}) \end{bmatrix}$$

11: The predicted mean output is $b_{0,0,\cdots,0}$.

Here $H_p^{q_p}(a)$ represents the orthogonal polynomial of degree q_p for the uncertain parameter a_p, $h_p^{q_p}(a_p)$ is the normalized orthonormal polynomial, and $\langle x_1(a_p), x_2(a_p)\rangle$ denotes the integration operation $\int x_1(a_p)x_2(a_p)f_{A_p}(a_p)\,da_p$.

Note that the MPCM can accurately predict not only the mean output of the system, but also the cross-statistics, i.e., statistics of cross input-output relationship, up to a certain degree. The MPCM can be applied to cases where the uncertain parameters are correlated, and where only moments, instead of pdfs, are available. Please see [23] for all detailed solutions and properties of the MPCM.

2.2 Effective State Estimation Method

In this section, we introduce the UKF, including its principles, procedures and properties.

2.2.1 Unscented Transformation

Unscented transformation (UT) is the basis of UKF. UT is developed to solve the following problem.

Consider an $L-$dimensional vector random variable $\mathbf{x}$ with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{P}_{xx}$. It is desired to find the mean $\bar{\mathbf{y}}$ and covariance $\mathbf{P}_{yy}$ of a $m-$dimensional vector random variable $\mathbf{y}$, where $\mathbf{y}$ is related to $\mathbf{x}$ by the nonlinear transformation

$$\mathbf{y} = g(\mathbf{x}). \tag{2.3}$$

2.2.1.1 Principles and procedures of UT

The UT is based on the idea that "it is easier to approximate a probability distribution than it is to approximate an arbitrary nonlinear function or transformation" [24]. The procedure of UT can be summarized as follows. A set of points (sigma points) are chosen so that their sample mean and covariance are $\bar{\mathbf{x}}$ and $\mathbf{P}_{xx}$. The nonlinear function $g(.)$ is applied to each point to derive the transformed points. The statistics of $\mathbf{y}$, i.e., $\bar{\mathbf{y}}$ and $\mathbf{P}_{yy}$ is approximated from the mean and covariance of the transformed sigma points. The detailed procedures are described as follows [25–27].

1. Compute the set of sigma points. The $L-$dimensional random variable $\mathbf{x}$ with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{P}_{xx}$ is approximated by $2L+1$ weighted points by

$$\mathcal{X}_0 = \bar{\mathbf{x}} \qquad W_0 = \frac{\lambda}{L+\lambda}, \tag{2.4}$$

$$\mathcal{X}_i = \bar{\mathbf{x}} + \left(\sqrt{(L+\lambda)\mathbf{P}_{xx}}\right)_i \quad W_i = \frac{1}{2(L+\lambda)}, \tag{2.5}$$

$$\mathcal{X}_{i+L} = \bar{\mathbf{x}} - \left(\sqrt{(L+\lambda)\mathbf{P}_{xx}}\right)_i \quad W_{i+L} = \frac{1}{2(L+\lambda)}, \tag{2.6}$$

which assures that

$$\mathbf{P}_{xx} = \sum_{l=0}^{2L} W_l(\mathcal{X}_l - \bar{\mathbf{x}})(\mathcal{X}_l - \bar{\mathbf{x}})^T, \tag{2.7}$$

where $i = 1, ..., L$, $l = 0, ..., 2L$, $(\sqrt{(L+\lambda)\mathbf{P}_{xx}})_i$ is the ith column of the matrix square root of $(L+\lambda)\mathbf{P}_{xx}$, W_i is the weight associated with the selected sigma points, and λ is a scaling parameter.

2. Transform each point through the function $g(.)$,

$$\mathcal{Y}_l = g(\mathcal{X}_l). \tag{2.8}$$

3. Compute $\bar{\mathbf{y}}$ and $\mathbf{P}_{yy}$ from the transformed points $\mathcal{Y}_l$,

$$\bar{\mathbf{y}} = \sum_{l=0}^{2n} W_l \mathcal{Y}_l, \tag{2.9}$$

$$\mathbf{P}_{yy} = \sum_{l=0}^{2n} W_l (\mathcal{Y}_l - \bar{\mathbf{y}})(\mathcal{Y}_l - \bar{\mathbf{y}})^T. \tag{2.10}$$

2.2.1.2 Accuracy of the UT

Consider the prior variable $\mathbf{x}$ as the $\bar{\mathbf{x}}$ being perturbed by a zero-mean disturbance $\delta\mathbf{x}$ with covariance $\mathbf{P}_{xx}$. Then the Taylor series expansion of the nonlinear transformation $g(\mathbf{x})$ about $\bar{\mathbf{x}}$ is

$$g(\mathbf{x}) = g(\bar{\mathbf{x}} + \delta\mathbf{x}) = \sum_{n=0}^{\infty} \left(\frac{(\delta\mathbf{x}\nabla_x)^n g(\mathbf{x})}{n!} \right)_{\mathbf{x}=\bar{\mathbf{x}}}. \tag{2.11}$$

Define the operator $\mathbf{D}^n_{\delta\mathbf{x}} g$ as $\mathbf{D}^n_{\delta\mathbf{x}} g = ((\delta\mathbf{x}\nabla_x)^n g(\mathbf{x}))_{\mathbf{x}=\bar{\mathbf{x}}}$, then the Taylor series expansion of the nonlinear transformation $\mathbf{y} = g(\mathbf{x})$ can be written as

$$\mathbf{y} = g(\mathbf{x}) = g(\bar{\mathbf{x}}) + \mathbf{D}_{\delta\mathbf{x}} g + \frac{1}{2}\mathbf{D}^2_{\delta\mathbf{x}} g + \frac{1}{3!}\mathbf{D}^3_{\delta\mathbf{x}} g + \frac{1}{4!}\mathbf{D}^4_{\delta\mathbf{x}} g + \cdots. \tag{2.12}$$

The true mean of $\mathbf{y}$ is given by

$$\bar{\mathbf{y}} = E[\mathbf{y}] = E[g(\mathbf{x})] = E\left[g(\bar{\mathbf{x}}) + \mathbf{D}_{\delta\mathbf{x}} g + \frac{1}{2}\mathbf{D}^2_{\delta\mathbf{x}} g + \frac{1}{3!}\mathbf{D}^3_{\delta\mathbf{x}} g + \frac{1}{4!}\mathbf{D}^4_{\delta\mathbf{x}} g + \cdots \right]. \tag{2.13}$$

If $\mathbf{x}$ is a symmetrically distributed random variable, then all odd moments are zero. Note that $E[\delta\mathbf{x}\delta\mathbf{x}^T] = \mathbf{P}_{xx}$, and as such, the third term in Equation (2.13) is

$$\begin{aligned}\frac{1}{2}\mathbf{D}^2_{\delta\mathbf{x}}g =& \frac{1}{2}\left((\delta\mathbf{x}\nabla_x)^2 g(\mathbf{x})\right)_{\mathbf{x}=\bar{\mathbf{x}}} = \frac{1}{2}\left((\nabla_x^T\delta\mathbf{x}^T\delta\mathbf{x}\nabla_x)g(\mathbf{x})\right)_{\mathbf{x}=\bar{\mathbf{x}}} \\ =& \frac{1}{2}\left((\nabla_x^T\mathbf{P}_{xx}\nabla_x)g(\mathbf{x})\right)_{\mathbf{x}=\bar{\mathbf{x}}}.\end{aligned} \tag{2.14}$$

Given this, the mean $\bar{\mathbf{y}}$ can be reduced to

$$\bar{\mathbf{y}} = g(\bar{\mathbf{x}}) + \frac{1}{2}\left((\nabla_x^T\mathbf{P}_{xx}\nabla_x)g(\mathbf{x})\right)_{\mathbf{x}=\bar{\mathbf{x}}} + E\left[\frac{1}{4!}\mathbf{D}^4_{\delta\mathbf{x}}g + \frac{1}{6!}\mathbf{D}^6_{\delta\mathbf{x}}g + \cdots\right]. \tag{2.15}$$

The UT calculates the posterior mean from the propagated sigma points. The sigma points are given by

$$\mathcal{X}_0 = \bar{\mathbf{x}}, \tag{2.16}$$

$$\mathcal{X}_l = \bar{\mathbf{x}} \pm \left(\sqrt{L+\lambda}\right)\sigma_i = \bar{\mathbf{x}} + \sigma_l, \tag{2.17}$$

where $i = 1, ...2L$, σ_i denotes the ith column of the matrix square root of $\mathbf{P}_{xx}$. This implies that $\sum_{i=1}^{L}(\sigma_i\sigma_i^T) = \mathbf{P}_{xx}$. Using the formulation of the sigma points, we can write the propogation of each point through the nonlinear function as a Taylor series expansion about $\bar{\mathbf{x}}$,

$$\mathcal{Y}_0 = g(\bar{\mathbf{x}}), \tag{2.18}$$

$$\mathcal{Y}_l = g(\mathcal{X}_l) = g(\bar{\mathbf{x}}) + \mathbf{D}_{\tilde{\sigma}_l}g + \frac{1}{2}\mathbf{D}^2_{\tilde{\sigma}_l}g + \frac{1}{3!}\mathbf{D}^3_{\tilde{\sigma}_l}g + \frac{1}{4!}\mathbf{D}^4_{\tilde{\sigma}_l}g + \cdots. \tag{2.19}$$

As such, the UT predicted mean is derived as

$$\begin{aligned}\bar{\mathbf{y}}_{UT} &= \frac{\lambda}{L+\lambda}g(\bar{\mathbf{x}}) + \frac{1}{2(L+\lambda)}\sum_{l=1}^{2L}\left(g(\bar{\mathbf{x}}) + \mathbf{D}_{\tilde{\sigma}_l}g + \frac{1}{2}\mathbf{D}^2_{\tilde{\sigma}_l}g + \frac{1}{3!}\mathbf{D}^3_{\tilde{\sigma}_l}g + \frac{1}{4!}\mathbf{D}^4_{\tilde{\sigma}_l}g + ...\right) \\ &= g(\bar{\mathbf{x}}) + \frac{1}{2(L+\lambda)}\sum_{l=1}^{2L}\left(\mathbf{D}_{\tilde{\sigma}_l}g + \frac{1}{2}\mathbf{D}^2_{\tilde{\sigma}_l}g + \frac{1}{3!}\mathbf{D}^3_{\tilde{\sigma}_l}g + \frac{1}{4!}\mathbf{D}^4_{\tilde{\sigma}_l}g + \cdots\right).\end{aligned} \tag{2.20}$$

Since the sigma points are symmetrically distributed around $\bar{\mathbf{x}}$, all the odd moments are zero. As such,

$$\bar{\mathbf{y}}_{UT} = g(\bar{\mathbf{x}}) + \frac{1}{2(L+\lambda)} \sum_{l=1}^{2L} \left(\frac{1}{2} \mathbf{D}^2_{\tilde{\sigma}_l} g + \frac{1}{4!} \mathbf{D}^4_{\tilde{\sigma}_l} g + \frac{1}{6!} \mathbf{D}^6_{\tilde{\sigma}_l} g \cdots \right). \tag{2.21}$$

Because

$$\begin{aligned} &\frac{1}{2(L+\lambda)} \sum_{l=1}^{2L} \frac{1}{2} \mathbf{D}^2_{\tilde{\sigma}_l} g = \frac{1}{2(L+\lambda)} (\nabla_x g)^T \sum_{i=1}^{L} \left(\sqrt{L+\lambda} \sigma_i \sigma_i^T \sqrt{L+\lambda} \right) (\nabla_x g) \\ &= \frac{L+\lambda}{2(L+\lambda)} (\nabla_x g)^T \left(\sum_{i=1}^{L} \sigma_i \sigma_i^T \right) (\nabla_x g) = \frac{1}{2} \left((\nabla_x^T \mathbf{P}_{xx} \nabla_x) g(\mathbf{x}) \right)_{\mathbf{x}=\bar{\mathbf{x}}}, \end{aligned} \tag{2.22}$$

the UT predicted mean can be further simplified to

$$\bar{\mathbf{y}}_{UT} = g(\bar{\mathbf{x}}) + \frac{1}{2} \left((\nabla^T \mathbf{P}_{xx} \nabla_x) g(\mathbf{x}) \right)_{\mathbf{x}=\bar{\mathbf{x}}} + \frac{1}{2(L+\lambda)} \sum_{l=1}^{2L} \left(\frac{1}{4!} \mathbf{D}^4_{\tilde{\sigma}_l} g + \frac{1}{6!} \mathbf{D}^6_{\tilde{\sigma}_l} g \cdots \right). \tag{2.23}$$

Comparing Equations (2.15) and (2.23), we can clearly see that the first three order terms of the true posterior mean and the mean calculated by the UT are the same, i.e., approximation errors only modulate the fourth and higher order terms. In contrast, a linearization approach that calculates the posterior mean as $\bar{\mathbf{y}}_E = g(\bar{\mathbf{x}})$ only captures the true posterior mean to the first order.

2.2.2 The UKF

The UKF is a straightfoward extension of the UT. The procedure of UKF is summarized as follows.

(a). Select Sigma Points. $2L+1$ symmetric weighted sigma points are selected from $\hat{\mathbf{x}}[k]$, the estimator of $\mathbf{x}[k]$.

$$\mathcal{X}_0 = \hat{\mathbf{x}} \qquad W_0 = \frac{\lambda}{L+\lambda}, \tag{2.24}$$

$$\mathcal{X}_i = \hat{\mathbf{x}} + (\sqrt{(L+\lambda)\mathbf{P}_{xx}})_i \quad W_i = \frac{1}{2(L+\lambda)}, \tag{2.25}$$

$$\mathcal{X}_{i+L} = \hat{\mathbf{x}} - (\sqrt{(L+\lambda)\mathbf{P}_{xx}})_i \quad W_{i+L} = \frac{1}{2(L+\lambda)}.$$

where $i = 1, ..., L$, $\left(\sqrt{(L+\lambda)\mathbf{P}_{xx}}\right)_i$ is the ith column of $(L+\lambda)\mathbf{P}_{xx}$, W_i is the weight associated with the selected sigma points, and λ is a scaling parameter

(b). State Prediction. The system state are predicted by instantiating each of the sigma points through the system dynamics $f(.)$.

$$\mathcal{X}_l[k+1|k] = f(\mathcal{X}_l[k], \mathbf{u}[k+1]),$$

where $l = 0, 1, ..., 2L$. Then the priori state estimation can be approximated as a weighted sample mean

$$\hat{\mathbf{x}}[k+1|k] = \sum_{l=0}^{2n} W_l(\mathcal{X}_l[k+1|k]).$$

The corresponding covariance matrix is calculated as

$$\mathbf{P}[k+1|k] = \sum_{l=0}^{2n} W_l(\mathcal{X}_l[k+1|k] - \hat{\mathbf{x}}[k+1|k])(\mathcal{X}_l[k+1|k] - \hat{\mathbf{x}}[k+1|k])^T + \mathbf{Q}[k+1].$$

(c). Measurement Prediction. $2n+1$ sigma points are selected from $\hat{\mathbf{x}}[k|k-1]$ with the error covariance $\mathbf{P}[k|k-1]$.

$$\begin{aligned}
\mathcal{X}_0[k+1|k] &= \hat{\mathbf{x}}[k+1|k], \\
\mathcal{X}_i[k+1|k] &= \hat{\mathbf{x}}[k+1|k] + \sqrt{(L+\lambda)\mathbf{P}[k+1|k]}_i, \\
\mathcal{X}_{i+L}[k+1|k] &= \hat{\mathbf{x}}[k+1|k] - \sqrt{(L+\lambda)\mathbf{P}[k+1|k]}_i,
\end{aligned}$$

with the weights W_0, W_i and W_{i+L} respectively.

The measurement is predicted by instantiating each of the prediction points through the measurement model $h(.)$,

$$\begin{aligned}
\mathcal{Z}_l[k+1|k] &= h(\mathcal{X}_l[k+1|k]), \\
\hat{\mathbf{z}}[k+1|k] &= \sum_{l=0}^{2n} W_l(\mathcal{Z}_l[k+1|k]).
\end{aligned}$$

Correspondingly, the measurement covariance matrix and cross correlation matrix are determined by

$$\mathbf{P}_{zz}[k+1|k] = \sum_{l=0}^{2n} W_l(\mathcal{Z}_l[k+1|k]-\hat{\mathbf{z}}[k+1|k])(\mathcal{Z}_l[k+1|k]-\hat{\mathbf{z}}[k+1|k])^T+\mathbf{R}[k+1], \quad (2.25)$$

$$\mathbf{P}_{xz}[k|k-1] = \sum_{l=0}^{2n} W_l(\mathcal{X}_l[k+1|k] - \hat{\mathbf{x}}[k+1|k])(\mathcal{Z}_l[k+1|k] - \hat{\mathbf{z}}[k+1|k])^T. \quad (2.26)$$

(d). Kalman Gain Update. The Kalman gain can be updated using the covariance information,

$$\mathcal{K} = \mathbf{P}_{xz}[k+1|k]\mathbf{P}_{zz}^{-1}[k+1|k]. \quad (2.27)$$

The estimated state and covariance are thus derived as

$$\hat{\mathbf{x}}[k+1|k+1] = \hat{\mathbf{x}}[k+1|k] + \mathcal{K}(\mathbf{z}[k+1] - \hat{\mathbf{z}}[k+1|k]), \quad (2.28)$$

$$\hat{\mathbf{P}}[k+1|k+1] = \hat{\mathbf{P}}[k+1|k] - \mathcal{K}\mathbf{P}_{zz}[k+1|k]\mathcal{K}^T. \quad (2.29)$$

The properties of the UKF method are summarized as follows.

1. The calculated mean and covariance are correct to the second order [24].

2. The algorithm is suitable for any nonlinear system dynamics.

3. The parameter λ can be selected properly to "fine tune" the higher order moments of the estimation. When $\mathbf{x}$ is Gaussian, λ is usually set to $3 - L$ to capture the fourth-order moment correctly (see [25, 28]).

CHAPTER 3

ADAPTIVE OPTIMAL DECISION IN MULTI-AGENT RANDOM SWITCHING SYSTEMS

3.1 Introduction

Random mobility models [1, 11, 29], including Random Walk, Random Direction, Gauss Markov and Smooth Turn (ST), have been widely used in diverse areas to capture the random movement patterns of mobile agents. Examples include ad hoc networks in wireless communication, random motion of particles in physics, and random UAV mobility in aerospace. These RMMs fall under the general random switching modeling framework: at each randomly selected time point, an agent randomly selects its maneuver of certain statistical properties, and moves with the selected maneuvers until the next selected time point. Driven by the emergence of internet-of-things applications, mobile agents play increasingly important roles in optimal decision processes. In this chapter, we study optimal control and effective estimation for such multi-agent random switching systems.

Optimal controller design for stochastic systems has been studied in e.g., [30]. For linear systems corrupted with additive noise, optimal controls solution that minimize the expected quadratic cost functions can be found analytically. For general stochastic systems with multi-dimensional uncertainties, simulation-based uncertainty evaluation methods need to be utilized. The Monte Carlo (MC) method and its variants including the Markov chain MC and Sequential MC have been widely used to explore the uncertainty space. However, they require a large amount of sample points, and hence, are too time-consuming to be used for online decisions. To address

this challenge, paper [23] developed an effective uncertainty evaluation method, called multivariate probabilistic collocation method (MPCM), and paper [31] integrated it with reinforcement learning (RL) to effectively solve the stochastic optimal control problem online. However, the uncertainties considered in [31] do not capture complex random switching behaviors. In this chapter, we integrate RL and MPCM to provide an online learning-based adaptive optimal control solution for random switching systems of highly flexible, random, and uncertain agent mobility patterns.

In practice, agents' states are not always available for controller design, and thus, effective state estimators are needed. For linear systems with additive noise, KF is the optimal estimator. For nonlinear systems, the sampling-based UKF [26,28] have been used practically. In this chapter, we also describe a practical estimator for multi-agent random switching systems by integrating UKF and MPCM. A communication and control co-design problem for a multi-UAV network governed by ST mobility is studied in the end to illustrate and validate the results.

3.2 Modeling and Problem Formulation

3.2.1 System Model

Consider a group of N agents, each of which moves independently with a general random switching dynamics as follows. At randomly selected time points T_0^i, T_1^i, $T_2^i, ...$, where $0 = T_0^i < T_1^i < ...$, agent i randomly selects its maneuver $\mathbf{a}_i[T_l^i]$ (e.g., velocity, heading direction, or turning center, etc.), and maintains the selected maneuver until the next selected time point. The time duration for agent i to maintain its current maneuver is $\tau_i[T_l^i]$, i.e., $\tau_i[T_l^i] = T_{l+1}^i - T_l^i$. Note that such a general random switching dynamics is constructed using two types of random variables. Type 1 random variable, $\mathbf{a}_i[T_l^i]$, describes the characteristics for each maneuver, and type 2

random variable $\tau_i[T_l^i]$ describes how often the switching of type 1 random variable occurs. The agent dynamics is described as

$$\mathbf{x}_i[k] = f(\mathbf{x}_i[k-1], \mathbf{a}_i[k], \tau_i[T_l^i]), \tag{3.1}$$

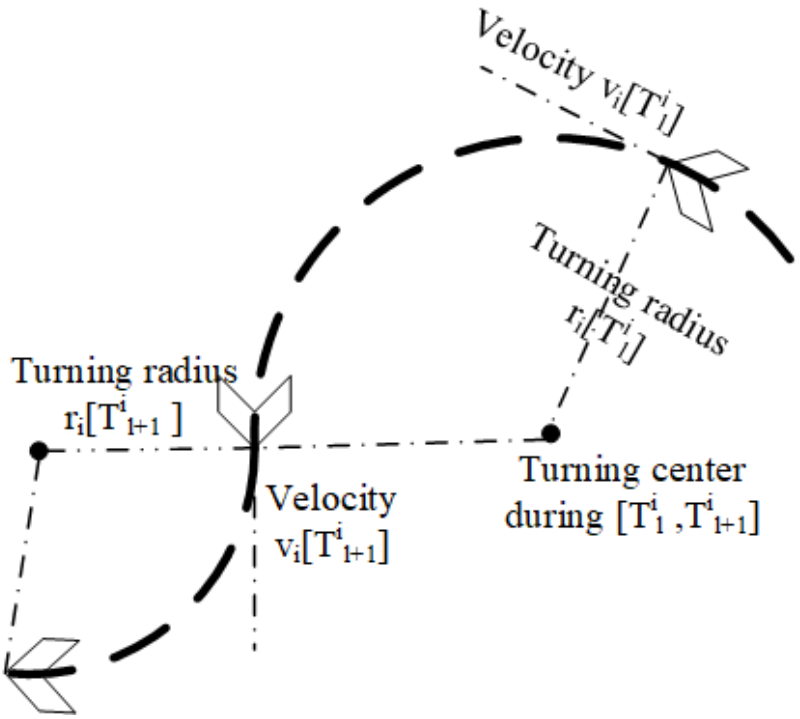
where $\mathbf{x}_i[k] \in \mathbf{R}^n$ is the system state vector of agent i at time instant k, and $f(.)$ captures the general agent dynamics. $\mathbf{a}_i[k] \in \mathbf{R}^m$ is the agent's maneuver at time k, and m is the number of uncertain parameters that describe the statistic properties of the maneuver. Each element of $\mathbf{a}_i[k]$, $a_{i,p}[k]$, where $p \in \{1, ..., m\}$, follows the random switching rule,

$$a_{i,p}[k] = \begin{cases} a_{i,p}[T_l^i], & \text{if } \exists l \in [0, 1, 2, ..), k = T_l^i; \\ a_{i,p}[k-1], & \text{if } \forall l = 0, 1, 2, .., k \neq T_l^i, \end{cases} \tag{3.2}$$

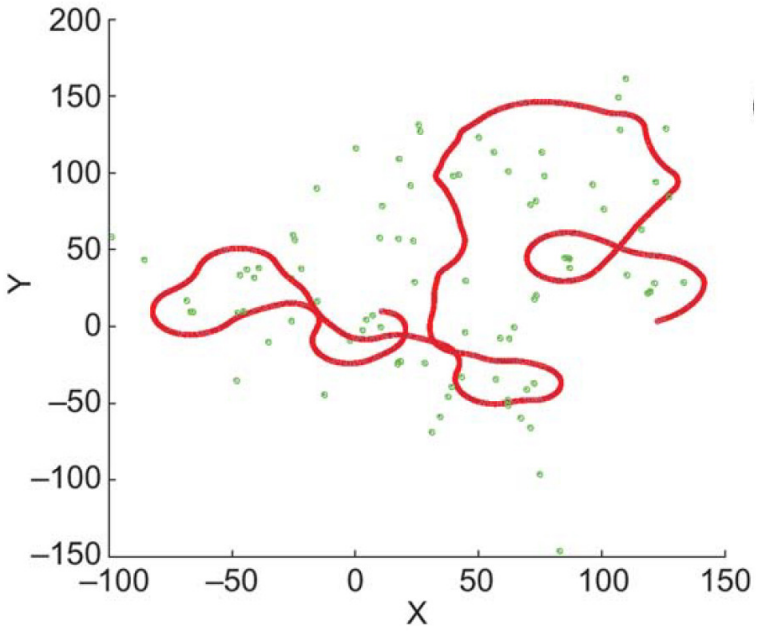
where $a_{i,p}[T_l^i](p = 1, ..., m)$ is the element of the type 1 random variable $\mathbf{a}_i[T_l^i]$, and $a_{i,p}[T_l^i](p = 1, ..., m)$ changes independently over time with pdf $f_{A_p}(a_{i,p}[T_l^i])$. The random variables $(\mathbf{a}_i[T_l^i], \tau_i[T_l^i])$ are independent for each agent i to capture their independent movement patterns.

We use a simple but widely-used UAV RMM, smooth turn RMM [1,29], to illustrate the random switching dynamics. In the ST RMM, each agent selects a velocity $v_i[T_l^i]$ and a turning center with a turning radius $r_i[T_l^i]$ along the line perpendicular to its current heading direction, and then circles around it until the next selected time point. The type 1 random variables $\mathbf{a}_i[T_l^i] = [r_i[T_l^i], v_i[T_l^i]]$ are inversely Gaussian and uniformly distributed respectively, and the type 2 random variable $\tau_i[T_l^i] = T_{l+1}^i - T_l^i$ is exponentially distributed. The switching behavior and sample trajectory are shown in Figs. 3.1(a) and 3.1(b) respectively.

The communication topology among the agents is fixed and represented using an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents $\mathcal{V} = 1, 2, ..., N$, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of communication links. A link $(i, j)(i \neq j)$ means that agents



(a)



(b)

Figure 3.1. Illustration of the ST RMM. (a) Maneuver selection and switching behavior. (b) A sample trajectory (red curve). Green spots are randomly chosen turning centers [1].

i and j can directly communicate with each other, where j is one of the neighbors of agent i.

Each agent has a local measurement model of a general nonlinear form

$$\mathbf{z}_{i|j}[k] = g(\mathbf{x}_i[k]) + \varpi_{z,i}[k], \tag{3.3}$$

where $\mathbf{z}_{i|j}[k]$ is the measured output of agent i by its neighbor j, $g(.)$ is a general nonlinear function, and $\varpi_{z,i}[k]$ is the white Gaussian noise.

3.2.2 Problem Formulation

We consider the following stochastic optimal control problem defined on the network of agents subject to the random switching dynamics. Denote the number of agent i's neighbors as n_i. Each agent i seeks its control policies $\mathbf{u}_{i,j}[k]$, $j \in [1, ..., n_i]$, to optimize a performance cost with its neighbor j according to the measurement $\mathbf{z}_{j|i}[k]$. Each agent i has at least n_i controllers to optimize the cost with the n_i neighbors respectively. This formulation has practical use in a wide range of new mobile networking applications, where the co-design of communication and control components becomes essential. An example is illustrated in Section 8.5.

In general, the expected cost to optimize is

$$J_{i,j} = E[\sum_{k=0}^{\infty} r_{i,j}(\mathbf{x}_i[k], \mathbf{x}_j[k], \mathbf{u}_{i,j}[k], \mathbf{u}_{j,i}[k])]. \tag{3.4}$$

where $r_{i,j}[k]$, $(j = 1, ..., n_i)$ is the cost between agent i and its neighbor j at time k. $\mathbf{u}_{i,j}[k]$ is the control vector of agent i, which seeks to minimize the communication cost with its neighbor j, $J_{i,j}$. The value function $V_{i,j}(\mathbf{x})$ corresponding to the performance index is defined as

$$V_{i,j}[k] = E[\sum_{k'=k}^{\infty} r_{i,j}(\mathbf{x}_i[k'], \mathbf{x}_j[k'], \mathbf{u}_{i,j}[k'], \mathbf{u}_{j,i}[k'])]. \tag{3.5}$$

Consider the problem of finding the optimal control policies $\mathbf{u}^*_{i,j}[k]$ that satisfies

$$\mathbf{u}^*_{i,j}[k] = \underset{\mathbf{u}_{i,j}}{\operatorname{argmin}}\, J_{i,j}[k](\mathbf{x}_i[k], \mathbf{x}_j[k], \mathbf{u}_{i,j}[k], \mathbf{u}_{j,i}[k]). \tag{3.6}$$

3.3 Main Results

3.3.1 Optimal Control in Random Switching Systems

In this subsection, we assume the state information, i.e., $\mathbf{x}_i[k]$ and $\mathbf{x}_j[k]$, is available, and design an adaptive optimal controller to find the optimal policies for agents moving with random switching dynamics.

Consider the value function described in (3.5). Because the uncertain parameters are independent from the system state, the following Bellman equation holds,

$$\begin{aligned} V_{i,j}[k] =& E(\sum_{k'=k}^{\infty} r_{i,j}[k']) = E(r_{i,j}[k] + \sum_{k'=k+1}^{\infty} r_{i,j}[k']) \\ =& E(r_{i,j}[k] + V_{i,j}[k+1]). \end{aligned} \tag{3.7}$$

This Bellman equation can be solved online using RL [32]. In particular, assume that a neural network weight $\mathbf{W}_{i,j}$ exists such that the value function can be approximated as

$$V_{i,j}(\mathbf{x}_i[k], \mathbf{x}_j[k]) = \mathbf{W}^T_{i,j}\phi(\mathbf{x}_i[k], \mathbf{x}_j[k]). \tag{3.8}$$

Using the value function approximation (VFA) method, the optimal control policy can be found from the policy iteration (PI) algorithm [33]. Two main steps are involved in the PI algorithm: 1) policy evaluation, which evaluates the value function at each time step, and 2) policy improvement, which finds the optimal policy based on the current value function.

For random switching systems, the policy evaluation step involves uncertainty evaluation to calculate the expectation of a function as shown in (3.7). This uncertainty evaluation is typically obtained using the Monte Carlo method and its variants,

too slow to be used for on-line decision algorithms. Here we use a multivariate probabilistic collocation method (MPCM) [23] to effectively evaluate the uncertainty. To map to the MPCM framework, we here denote the generic function whose expectation to be evaluated as $G(a_1, ..., a_m)$, which is modulated by uncertain parameters $a_1, a_2, ..., a_m$ with the degree of each parameter up to a certain number. The MPCM accurately evaluates the output mean of G, by smartly selecting a limited number of sample points according to the Gaussian Quadrature rules, evaluating these sample points, and producing the output mean from a reduced-order mapping G'. The main property of MPCM is described in the following lemma. Please refer to [23] for the proof and detailed MPCM design procedures.

Lemma 1. *[23, Theorem 2] Consider a generic 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 ($p \in 1, 2, ..., m$) is an uncertain parameter with 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.

Remark 1. *Lemma 1 shows that the MPCM reduces the number of simulations from $2^m \prod_{p=1}^{m} n_p$ to $\prod_{p=1}^{m} n_p$, where m is the number of uncertain parameters. Despite the significant reduction of computation by 2^m, MPCM accurately predicts the output mean [23]. We note that the degree of an uncertain parameter in G is dependent on specific applications. For a nonlinear system, G is a polynomial approximation*

with properly selected maximal degree for each parameter, $2n_p - 1$. With the increase of maximal degree, the approximation accuracy can be improved, but at the cost of additional computational load and the chance of overfitting.

Here we integrate RL and MPCM to provide an effective online learning-based optimal control algorithm for random switching systems.

To evaluate the value function $V_{i,j}[k]$ at each time instant, one needs to calculate $E(V_{i,j}[k+1])$ according to the Bellman equation (3.7). The value function $V_{i,j}[k+1]$ is determined uniquely by the system states $\mathbf{x}_i[k+1]$ and $\mathbf{x}_j[k+1]$, which can be found from the current states $\mathbf{x}_i[k]$ and $\mathbf{x}_j[k]$, system dynamics $f(.)$, and the random switching behaviors. In particular, given the current states $\mathbf{x}_i[k]$ and $\mathbf{x}_j[k]$, agent i can predict its future state $\mathbf{x}_i[k+1]$ according to its current maneuver $\mathbf{a}_i[k+1]$ using the system dynamics $f(.)$. However, agent i does not know agent j's current maneuver $\mathbf{a}_j[k+1]$, and as such, $\mathbf{x}_j[k+1]$ needs to be estimated by agent i considering its switching behaviors. Denote the switching behavior of agent j at time k as $s_j[k]$. $s_j[k] = 1$ or 0 indicates if the current maneuver switches at time k or not. Denote the value function $V_{i,j}[k]$ when $s_j[k] = 1$ (or $s_j[k] = 0$) as $V^1_{i,j}[k]$ (or $V^0_{i,j}[k]$ correspondingly). When $s_j[k] = 0$, agent j keeps its previous maneuver $\mathbf{a}_j[k]$, and the system state $\mathbf{x}_j[k+1]$ is obtained using $\mathbf{a}_j[k]$, i.e.,

$$V^0_{i,j}[k] = r_{i,j}[k] + V_{i,j}[k+1](\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[k]). \tag{3.9}$$

When $s_j[k] = 1$, agent j chooses a new random maneuver from $\mathbf{a}_j[T^j_l]$ at time k, and in this case, $E(V_{i,j}[k+1])$ needs to be estimated from the characteristics of the random variable $\mathbf{a}_j[T^j_l]$. Define a system mapping subject to uncertain input parameters $\mathbf{a}_j[T^j_l]$: $G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T^j_l]) = r_{i,j}[k] + V_{i,j}[k+1](\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T^j_l])$, then the value function $V^1_{i,j}[k]$ can be estimated from the mean output of the system mapping $G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T^j_l])$ using MPCM, i.e.,

$V^1_{i,j}[k] = E[G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])]$. In particular, a set of samples are selected based on the pdfs of uncertain parameters, and simulations are run at these samples to estimate $E[G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])]$. Under the assumption that each uncertain parameter $a_{j,p}$ has a degree up to $2n_p - 1$, $G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ has the following form,

$$\begin{aligned} &G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j]) \\ &= \sum_{q_1=0}^{2n_1-1} \sum_{q_2=0}^{2n_2-1} \cdots \sum_{q_m=0}^{2n_m-1} \psi^V_{q_1,\ldots,q_m}(\mathbf{x}_i[k+1], \mathbf{x}_j[k]) \prod_{p=1}^{m} a_{j,p}^{q_p}. \end{aligned}$$

According to Lemma 1, the output mean of this system mapping can be estimated from the output of a reduced-order mapping $G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ derived from the MPCM procedure [23, Section II],

$$\begin{aligned} V^1_{i,j}[k] =& E[G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])] \\ =& E[G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])], \end{aligned} \tag{3.10}$$

$$\begin{aligned} &G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j]) \\ &= \sum_{q_1=0}^{n_1-1} \sum_{q_2=0}^{n_2-1} \cdots \sum_{q_m=0}^{n_m-1} \Omega^V_{q_1,\ldots,q_m}(\mathbf{x}_i[k+1], \mathbf{x}_j[k]) \prod_{p=1}^{m} a_{j,p}^{q_p}. \end{aligned} \tag{3.11}$$

The coefficients $\Omega^V_{q_1,\ldots,q_m}(\mathbf{x}_i[k+1], \mathbf{x}_j[k])$ and output mean can be obtained using the evaluated outputs $G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ at each selected simulation point, according to the procedures in [23, Section II-B].

Theorem 2. *The value function described in* (3.7) *can be estimated as*

$$V_{i,j}[k] = PV^0_{i,j}[k] + (1-P)V^1_{i,j}[k], \tag{3.12}$$

where $V^0_{i,j}[k]$ *and* $V^1_{i,j}[k]$ *are described in* (3.9) *and* (3.10) *respectively.* P *is the probability that agent* j *switches its maneuver at time* k*. This probability can be derived from the distribution of* $\tau_j[T_l^j]$*,* $f_\tau(\tau_j[T_l^j])$*.*

Proof. The value function for an agent of random switching dynamics can be derived as

$$\begin{aligned} V_{i,j}[k] =& E(V_{i,j}[k]|s_j[k]=0)P(s_j[k]=0) \\ &+ E(V_{i,j}[k]|s_j[k]=1)P(s_j[k]=1) \\ =& PV_{i,j}^0[k] + (1-P)V_{i,j}^1[k]. \end{aligned} \tag{3.13}$$

(3.9), (3.10) and (3.13) naturally lead to Theorem 2. □

The detailed algorithm that integrates the PI learning algorithm and MPCM is described in Algorithm 2. After initialization in Step 1, Step 2 samples $\mathbf{a}_j[T_l^j]$ to prepare for the uncertainty evaluation procedures in Steps $4-6$. Steps $3-7$ are value function evaluation. In particular, Step 3 evaluates $V_{i,j}^0$, Steps $4-6$ evaluate $V_{i,j}^1$, and Step 7 combines $V_{i,j}^0$ and $V_{i,j}^1$ according to Theorem 2 to find $V_{i,j}[k]$. After value function evaluation, the approximation weights $\mathbf{W}_{i,j}$ and optimal control polices $\mathbf{u}_{i,j}$ are updated respectively in Steps 8 and 9. The detailed procedures for MPCM and PI algorithm can be found in [23, 33] respectively.

Algorithm 2 Policy iteration algorithm for random switching systems

1: Initialize the system with initial states $\mathbf{x}_i[0], \mathbf{x}_j[0]$, and admissible control policies $\mathbf{u}_{i,j}[0]$ and $\mathbf{u}_{j,i}[0]$.

2: Select $\prod_{p=1}^m n_p$ MPCM sample points according to the pdfs $f_{A_p}(a_{j,p}[T_l^j])$ and the MPCM procedure [23, Section II]. Denote each selected MPCM sample as $\mathcal{A}^l$, where $l = 1, ..., \prod_{p=1}^m n_p$.

3: For each iteration s, find the value function when $s_j[k]=0$, $V_{i,j}^{0,(s)}$, using (3.9).

4: Find the value function $\mathcal{V}_{i,j}^{l,(s)}(\mathbf{x}_i[k], \mathbf{x}_j[k])$ at each MPCM sample $\mathcal{A}^l$, using the Bellman equation: $\mathcal{V}_{i,j}^{l,(s)}[k] = r_{i,j}[k] + \mathbf{W}_{i,j}^{(s-1)T}\phi(\mathbf{x}_i[k+1], \mathbf{x}_j[k+1])$.

5: Find the reduced polynomial mapping from $a_{j,p}$ to $G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ described in (3.11) according to Lemma 1. $a_{j,p}$ and $G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ take the value of $\mathcal{A}^l$ and $\mathcal{V}_j^l[k]$ respectively.

6: Find the value function when $s_j[k] = 1$, i.e., $V_{i,j}^{1,(s)}[k]$, by combining (3.10) and the derived system mapping $G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$.

7: Find the value function $V_{i,j}^{(s)}[k]$ by combining Theorem 2, $V_{i,j}^{0,(s)}[k]$ and $V_{i,j}^{1,(s)}[k]$.

8: Update the value function coefficients $W_{i,j}^{(s)}$ according to the estimated $V_{i,j}^{(s)}[k]$: $\mathbf{W}_{i,j}^{(s)T}\phi(\mathbf{x}_i[k], \mathbf{x}_j[k]) = V_{i,j}^{(s)}[k]$.

9: Update the control policy $\mathbf{u}_{i,j}^{(s)}$ as $\mathbf{u}_{i,j}^{(s)} = \text{argmin}_{\mathbf{u}_{i,j}} V_{i,j}^{(s)}[k]$.

10: Repeat procedures $3-9$.

Theorem 3. *Consider the random switching system shown in* (3.1) *with the value function described in* (3.5)*. Assume there exists a unique optimal solution and Algorithm 2 converges. Given the current system states* $\mathbf{x}_i[k]$ *and* $\mathbf{x}_j[k]$*, the solution found by Algorithm 2 is the optimal control policy.*

Proof. The control policy derived by evaluating the value function $V_{i,j}[k] = PV_{i,j}^0[k] + (1-P)V_{i,j}^1[k]$ is optimal according to (3.6), Theorem 2, and the policy iteration properties [33]. As such, to prove this theorem, we are left to show that the two optimal solutions, which are found by evaluating the reduced-order mapping $PV_{i,j}^0[k] + (1-P)G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ and the original value function mapping $PV_{i,j}^0[k] + (1-P)G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])$ are the same. Lemma 1 proves that $E[G'_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])] = E[G_{V_{i,j}}(\mathbf{x}_i[k+1], \mathbf{x}_j[k], \mathbf{a}_j[T_l^j])]$. Therefore, the equivalence of the two optimal solutions can be proved from a contradiction argument described in [31, Theorem 1]. □

Remark 2. *The convergence of Algorithm 2 depends on three numerical solutions: the policy iteration method, the value function approximation, and the MPCM approximation. The policy iteration method has been widely used in dynamic programming and reinforcement learning fields [31, 33], with its convergence conditions provided in [33]. The value function approximation uses neural networks to approximate the*

smooth value function. The assumptions that make this approximation hold are listed in [34]. MPCM works well for both polynomial and non-polynomial system mappings as guaranteed by Lemma 1, with properly selected degrees for the polynomials (see [23] for the details).

3.3.2 State Estimation in Random Switching Systems

In many practical applications, state information $\mathbf{x}_i[k]$ and $\mathbf{x}_j[k]$ may not be available for controller design. In this subsection, we provide a practical online state estimation solution for agents of random switching systems.

Given the previous state $\mathbf{x}_i[k-1]$, the expected current state $E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1])$ can be estimated considering the two possible switching behaviors: $s_i[k-1] = 1$ or 0. When $s_i[k-1] = 1$, agent i chooses a new random maneuver from $\mathbf{a}_i[T_l^i]$ at time $k-1$. As such, the estimation of the expected conditional system state $E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s_i[k-1] = 1)$ involves uncertainty evaluation, which we solve using MPCM, instead of the Monte Carlo methods which are computationally ineffective. In particular, we define $f(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])$ as a system mapping subject to uncertain input parameters $\mathbf{a}_i[T_l^i]$, i.e., $G_i(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])$. The expected system state when $s_i[k-1] = 1$ is then approximated from the mean output of the system mapping $G_i(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])$ using MPCM, i.e.,

$$E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s_i[k-1] = 1) = E[G_i(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])]. \tag{3.14}$$

Theorem 4. *Given the previous system state $\mathbf{x}_i[k-1]$, the expected current state for agent i is estimated as*

$$\begin{aligned} E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1]) &= PE[G_i'(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])] \\ &+ (1-P)f(\mathbf{x}_i[k-1], \mathbf{a}_i[k-1]), \end{aligned} \tag{3.15}$$

where P is the probability that agent i switches its maneuver at time $k-1$. $G_i'(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])$ is a reduced order mapping of $G_i(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])$ derived from MPCM.

Proof. The expected system state at time k for an agent of random switching dynamics can be derived as

$$\begin{aligned} & E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1]) \\ =& E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s_i[k-1]=0)P(s_i[k-1]=0) \\ & + E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s_i[k-1]=1)P(s_i[k-1]=1). \end{aligned} \tag{3.16}$$

In the case of $s_i[k-1]=0$, agent i keeps its previous maneuver. The conditional expected system state $E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s_i[k-1]=0)$ can be found from the previous system state $\mathbf{x}_i[k-1]$ and maneuver $\mathbf{a}_i[k-1]$ as

$$E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s[k-1]=0) = f(\mathbf{x}_i[k-1], \mathbf{a}_i[k-1]). \tag{3.17}$$

In the case of $s_i[k-1]=1$, agent i chooses a new maneuver from $\mathbf{a}_i[T_l^i]$ at time $k-1$. The expected conditional system state $E(\mathbf{x}_i[k]|\mathbf{x}_i[k-1], s_i[k-1]=1)$ can be estimated from the mean output of the reduced-order system mapping $G_i'(\mathbf{x}_i[k-1], \mathbf{a}_i[T_l^i])$ using MPCM according to (3.14) and Lemma 1.

Equations (3.14), (3.16) and (3.17) naturally lead to Theorem 4. □

Theorem 4 provides an accurate and computationally-efficient approach to estimate the expected system state for random switching systems, given the previous state. Next we integrate it with UKF to provide the state estimation solution from the measurement signals $\mathbf{z}_{i|j}[k]$. The system is assumed to be observable. In particular, MPCM and UKF are integrated for a 5-step state estimation procedure. Steps 1 and 2 select initial conditions and MPCM points to initialize Steps 3-5. Steps 3 and 4 find the state estimator for the switching behaviors $s_i[k-1]=0$ and 1 respectively. Step 5 finds the expected state by integrating the two estimators in Steps 3 and 4.

Step 1: Initialize. Initialize $\hat{\mathbf{x}}_i[0]$ and $\mathbf{P}_i[0]$.

Step 2: Select MPCM points. $\prod_{p=1}^{m} n_p$ MPCM sample points are selected according to the pdfs $f_{A_p}(a_{i,p}[T_l^i])$ and the MPCM procedure [23, Section II]. Denote each selected MPCM sample as $\mathcal{A}^l$, where $l = 1, ..., \prod_{p=1}^{m} n_p$.

Step 3: Estimate the system state when $\mathbf{s_i[k-1]=0}$**.** The expected system state $E(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 0)$ can be estimated using UKF through the following four sub-steps [26]: (a) select sigma points from $\hat{\mathbf{x}}_i[k-1]$; (b) predict system state by instantiating the sigma points through the system dynamics $f(.)$; (c) select new sigma points from the predicted state, and predict measurement by instantiating the sigma points through the measurement model $g(.)$; (d) update the Kalman gain and find the expected state $E(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 0)$ and covariance $E(\mathbf{P}_i[k]|\mathbf{P}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 0)$.

Step 4: Estimate the system state when $\mathbf{s_i[k-1]=1}$**.** Uncertainty evaluation is necessary in this step, and the expected system state is derived by integrating MPCM and UKF using the following three sub-steps (a)-(c).

(a). Estimate system state at each selected MPCM point. At each selected MPCM point $\mathcal{A}^l$ $(l = 1, ..., \prod_{p=1}^{m} n_p)$, the system state can be estimated from the UKF procedure shown in ***Step 3***, *(a)-(d)*. Denote the estimated state from UKF at each sample point as $\hat{\mathbf{x}}_i^l[k]$ with the covariance $\mathbf{P}_i^l[k]$ $(l = 1, ..., \prod_{p=1}^{m} n_p)$.

(b). Find the reduced polynomial mappings. Define system mappings $G'_{\mathbf{x}_i}(\hat{\mathbf{x}}_i[k-1], \mathbf{a}_i[T_l^i])$ and $G'_{\mathbf{P}_i}(\hat{\mathbf{x}}_i[k-1], \mathbf{a}_i[T_l^i])$ as the relationships between the expected system state and covariance with the random variable $\mathbf{a}_i[T_l^i]$. According to Lemma 1, the reduced-order mappings can be found respectively as

$$G'_{\mathbf{x}_i}(\hat{\mathbf{x}}_i[k-1], \mathbf{a}_i[T_l^i]) = \sum_{q_1=0}^{n_1-1} \sum_{q_2=0}^{n_2-1} ... \sum_{q_m=0}^{n_m-1} \Omega^{\mathbf{x}}_{q_1,...,q_m}(\hat{\mathbf{x}}_i[k-1]) \prod_{p=1}^{m} a_{i,p}^{q_p},$$

$$G'_{\mathbf{P}_i}(\hat{\mathbf{x}}_i[k-1], \mathbf{a}_i[T_l^i]) = \sum_{q_1=0}^{n_1-1} \sum_{q_2=0}^{n_2-1} ... \sum_{q_m=0}^{n_m-1} \Omega^{\mathbf{P}}_{q_1,...,q_m}(\hat{\mathbf{x}}_i[k-1]) \prod_{p=1}^{m} a_{i,p}^{q_p}.$$

(c). Find the expected system state and covariance. The expected system state and covariance are then found from the mean output of the system mapping according to Lemma 1 and the MPCM design procedure [23], $E(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 1) = E[G'_{\mathbf{x}_i}(\hat{\mathbf{x}}_i[k-1], \mathbf{a}_i[T_l^i])]$, $E(\mathbf{P}_i[k]|\mathbf{P}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 1) = E[G'_{\mathbf{P}_i}(\hat{\mathbf{x}}_i[k-1], \mathbf{a}_i[T_l^i])]$.

Step 5: Estimate the expected system state. The estimated state and covariance are then derived from ***Steps 3*** and ***4*** according to Theorem 4 as

$$\begin{aligned}
&E(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k]) \\
=&PE(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 1]) \\
&+ (1-P)E(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 0]),
\end{aligned}$$

$$\begin{aligned}
&E(\mathbf{P}_i[k]|\mathbf{P}_i[k-1], \mathbf{z}_{i|j}[k]) \\
=&PE(\mathbf{P}_i[k]|\mathbf{P}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 1]) \\
&+ (1-P)E(\mathbf{P}_i[k]|\mathbf{P}_i[k-1], \mathbf{z}_{i|j}[k], s_i[k-1] = 0]).
\end{aligned}$$

As such, the estimator of $\mathbf{x}_i[k]$ is $\hat{\mathbf{x}}_i[k] = E(\mathbf{x}_i[k]|\hat{\mathbf{x}}_i[k-1], \mathbf{z}_{i|j}[k])$, and the expected error covariance is $\mathbf{P}_i[k] = E(\mathbf{P}_i[k]|\mathbf{P}_i[k-1], \mathbf{z}_{i|j}[k])$.

Remark 3. *The performance of the state estimation algorithm is jointly determined by UKF and MPCM. UKF addresses the nonlinear system dynamics and measurement models. MPCM effectively samples the random switching behavior. The accuracy of MPCM is guaranteed by Lemma 1. Note that UKF is not an optimal estimator. It has been practically used to provide approximations to optimal solutions with certain accuracy. Performance analysis on UKF for general systems is limited in the literature. When the measurement model is linear, the estimation error of UKF is bounded when an extra positive definite matrix is added in the calculated covariance*

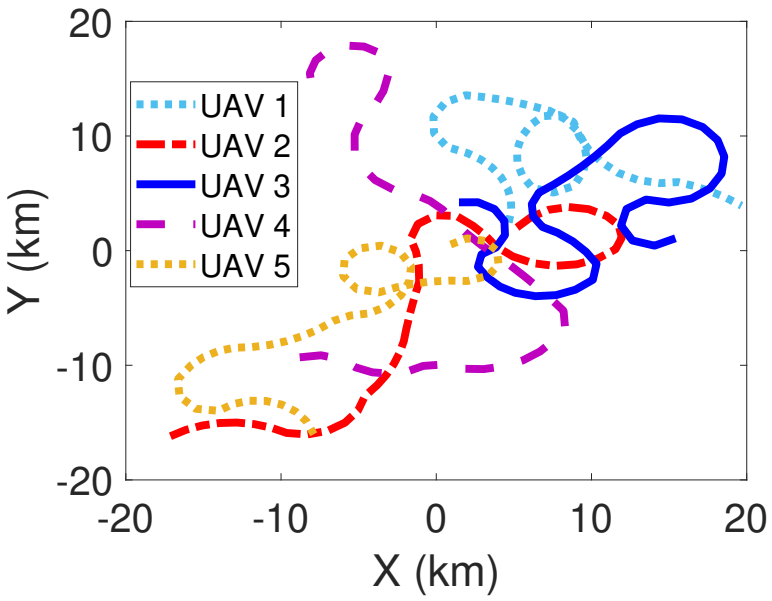
matrix [35]. Here we use the UKF method to address random switching system dynamics. The use of MPCM does not deteriorate the convergence or the optimality of state estimation as guaranteed by Lemma 1.

3.4 Illustrative Examples

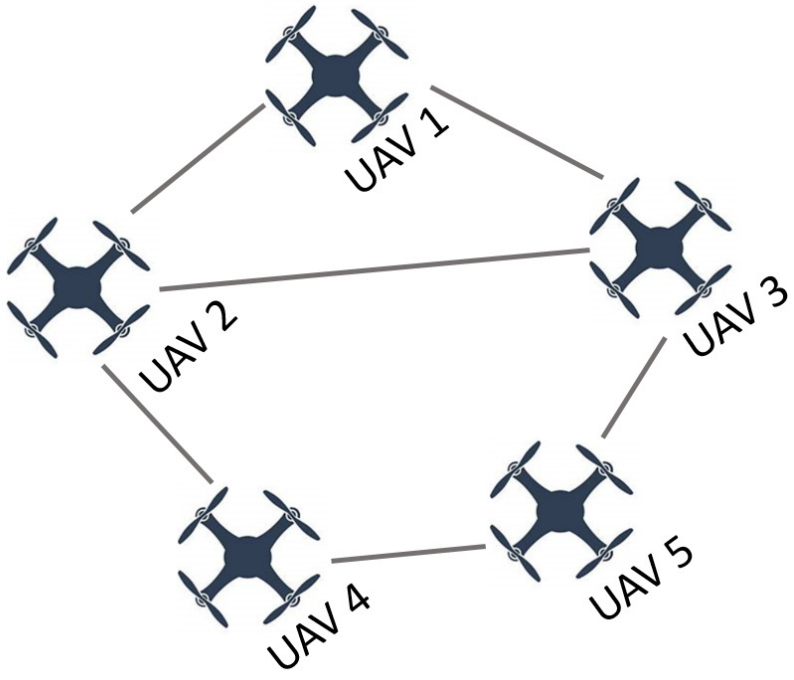
Consider a five-UAV network to support a surveillance-like mission. UAVs move independently according to the ST RMM described in Section 3.2.1. The randomly-generated trajectories of the UAVs are shown in Fig. 3.2(a).

A three-sector directional antenna is mounted on each UAV to communicate with its neighbor UAVs over long distances upon an ID-based fixed communication topology (Fig. 3.2(b)). To maximize the communication performance, each UAV controls the heading directions of its antennas to maximize the the received signal strength indicators (RSSI), which measure the communication channel performance. The cost function in this example is $J_{i,j} = -E[\sum_{k=0}^{N} R_{i,j}[k]]$, where $R_{i,j}$, the RSSI that UAV i receives from its neighbor j is $R_{i,j}[k] = P_{t|dBm} + 20\log_{10}(\frac{\lambda}{4\pi d[k]}) + G_{l|dBi}[k]$ (see [2] for the details), and N is the experimental time. Here $P_{t|dBm}$ is the transmitted signal power, λ is the wavelength, $d[k]$ is the distance between neighboring UAVs, and $G_{l|dBi}[k]$ is the sum of gains of the two antennas, which depends on their heading angles. Measurements are GPS corrupted with Gaussian white noise. The performances of the designed estimators and controllers are simulated for all five UAVs and communication links. We here only show the performance of UAV 3, and its communication link to UAV 2.

We first investigate the computation load reduction of the MPCM through a comparative study. Since two type 1 random variables are involved, the number of uncertain parameters in the system mapping $G(a_1, a_2)$ is $m = 2$. We select $n_1 = 2$ for the degree of $a_1 = v_i[T_l^i]$ and $n_2 = 3$ for $a_2 = r_i[T_l^i]$. With this parameter setting,



(a)



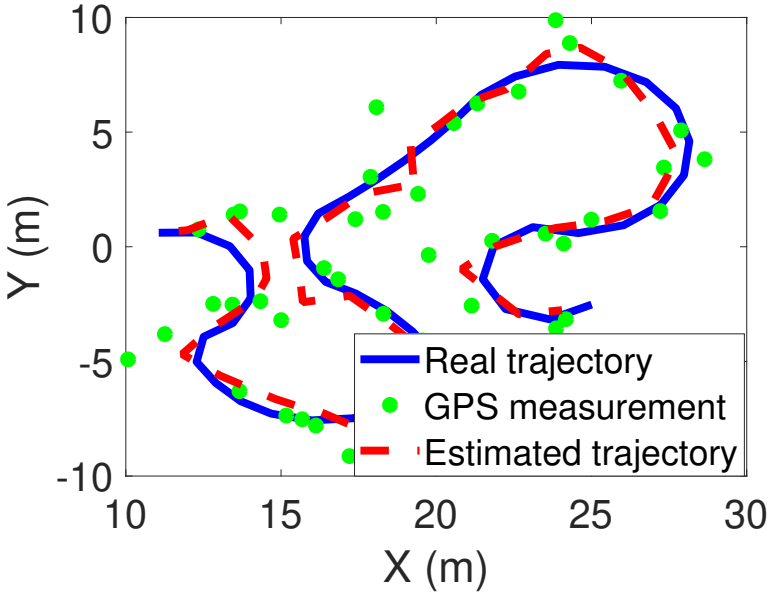
(b)

Figure 3.2. (a) Sample trajectories of the UAVs, (b) Communication topology of the five-UAV network.

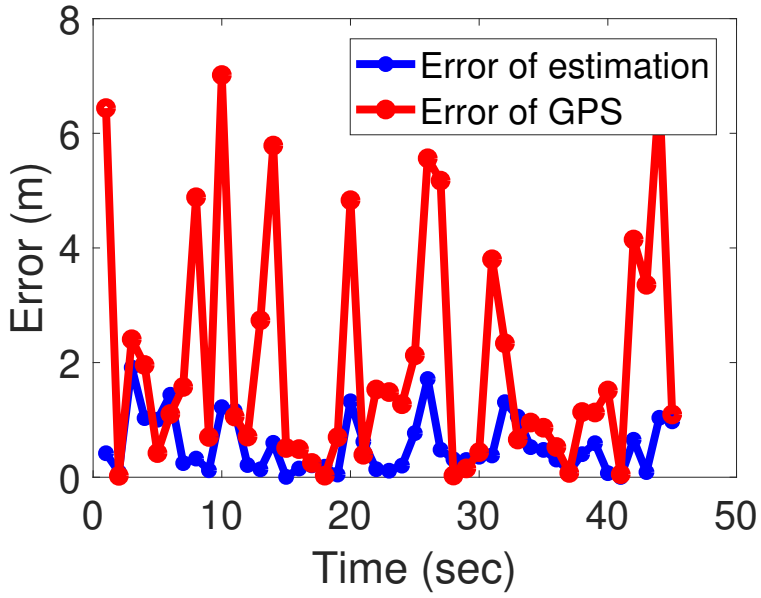
$\prod_{p=1}^{m} n_p = 6$ MPCM points are selected according to the MPCM procedure [23]. For the optimal control solution developed in 3.3.1, the Monte Carlo method requires about 8000 sample points to converge to the output mean, while the MPCM method only needs 6 points to converge to the correct result. The significant reduction of computation load shows the value of MPCM to facilitate online uncertainty evaluation.

We then analyze performance of the state estimator designed in Section 3.3.2 (see Fig. 3.4) . The estimated trajectory matches well with the real UAV trajectory, and the estimation errors are much smaller than the errors of GPS signals, which validates the effectiveness of the estimation solution.

Finally, we simulate the optimal controller designed based on the estimated states. Fig. 3.4(a) shows the controlled heading directions of the directional antenna mounted on UAV 3 to communicate with UAV 2, and Fig. 3.4(b) shows the errors between the controlled and real optimal heading directions. The controlled directional antenna heading direction is very close to the optimal solution, and the errors are within $(-0.2, 0.15)$ rad, which validates effectiveness of the proposed adaptive optimal control solution.

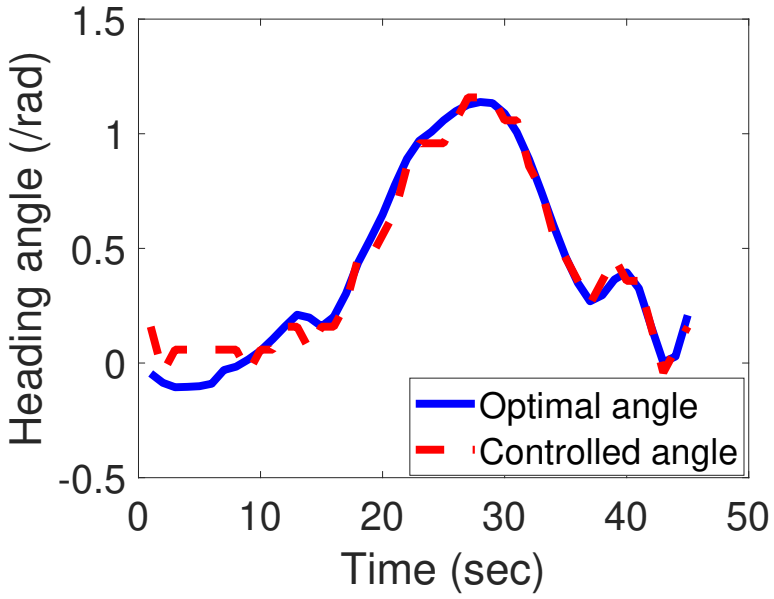


(a)

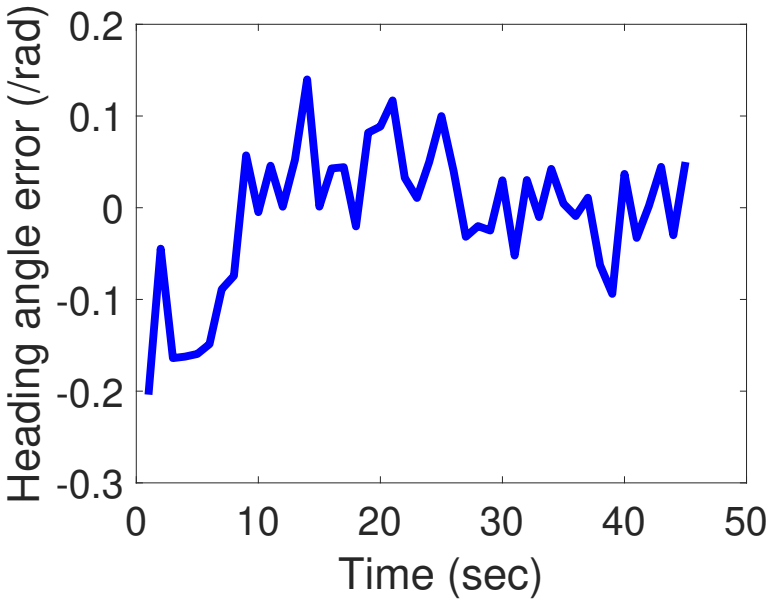


(b)

Figure 3.3. Estimation performance. (a) Trajectory of UAV 3. (b) Estimation errors.
.



(a)



(b)

Figure 3.4. Control performance. (a) Optimal headings of directional antenna on UAV 3 to communicate with UAV 2. (b) Errors between the optimal and the controlled heading angles of the directional antenna.

CHAPTER 4

ADAPTIVE OPTIMAL CONTROL FOR STOCHASTIC MULTI-PLAYER DIFFERENTIAL GAMES USING ON-POLICY AND OFF-POLICY REINFORCEMENT LEARNING

4.1 Introduction

Game theory has been widely used in multi-player systems to obtain decisions that optimize individual payoffs [36–41]. In the traditional game theory, a player finds the best strategy to minimize a static and immediate cost [36–38]. Recently, differential games were combined with control theory to study dynamical systems that involve the evolution of the players' payoff functions [40–42]. Widely used differential games include the two-player zero-sum game, which provides solutions for pursuit-evasion games and H_∞ design for disturbance attenuation [42], and the multi-player nonzero-sum game, which finds applications in e.g., the control of transportation networks and the cooperative control of multiple robots with individual goals [41]. Most existing studies on differential games assume deterministic dynamics. In reality, multi-dimensional uncertainties, such as uncertain player intentions and environmental impacts often modulate system dynamics in a complicated fashion. As such, in this chapter, we formulate and study practical Nash solutions for new stochastic two-player zero-sum and multi-player nonzero-sum games, where the system dynamics are modulated by muti-dimensional time-varying random parameters.

For deterministic differential games, the Nash equilibrium solutions rely on solving Hamiltonian-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 [41]. Moreover, this method requires the full knowledge of system dynamics, and only provides an offline solution. As such, the RL method has been developed to solve the differential games online [43–47]. We also explore RL to develop online solutions in this paper 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 [48–54]. RL has been used to find the Nash equilibrium solutions online for multi-player differential games. In particular, for the two-player zero-sum game, paper [43] 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, paper [55] developed a single-loop iteration algorithm that updates the two players' control policies simultaneously. In addition, to deal with the systems with unknown dynamics, paper [56] developed a model-free IRL for the two-player zero-sum differential game using Q-learning. For the multi-player nonzero-sum game, paper [46] 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, paper [47] established an off-policy IRL to solve the nonlinear continuous-time multi-player 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 [30, 57–60]. The most widely used simulation-based uncertainty evaluation methods are the MC method and its variants including the Makrov Chain MC and Sequential MC [61–63]. However, the MC-based methods require a large number of simulations to estimate the expected cost function accurately, which make them unrealistic for online algorithms. To improve the computational efficiency, other uncertainty sampling methods including Latin Hypercube sampling [64], importance sampling [65], multilevel MC [66], greedy and adaptive sampling [67, 68] have also been developed. However, none of them can estimate expected system outputs accurately with a computational load. To deal with this challenge, papers [23, 69] developed effective uncertainty evaluation methods, named the 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). Papers [70, 71] further integrated the MPCM with the discrete-time RL to solve optimal control problems online for uncertain systems. Here in this chapter, we study the integrated MPCM and IRL to effectively solve stochastic multi-player differential games online.

This chapter, for the first time in the literature per our knowledge, analyzes multi-player 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 chapter are four-fold: 1) The formulation of two-player zero-sum and multi-player nonzero-sum games for systems modulated by time-varying random parameters, which capture stochastic environmental impacts and random player intentions [4, 5, 72]; 2) The analysis of the formulated differential game properties, including stability and Nash equilibrium; 3)

A novel policy iteration algorithm that integrates IRL and an effective uncertainty sampling method: 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 chapter is organized as follows. Section 4.2 formulates the stochastic two-player zero-sum and multi-player nonzero-sum games and presents preliminaries to facilitate the analysis in this chapter. Sections 4.3 and 4.4 study the properties and online solutions of these two stochastic games. Section 4.5 presents the simulation studies that demonstrate performances of the proposed solutions.

4.2 Problem Formulation and Preliminaries

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

4.2.1 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,

$$\dot{\mathbf{x}} = A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d}, \tag{4.1}$$

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, $\mathbf{d} = \mathbf{d}(t) \in \mathbb{R}^q$ 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 $\mathbf{a}(t)$, $a_p(t)$ $(p = 1, 2, \cdots, m)$, 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 (4.1) are ordinary differential equations [73, 74].

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

$$\begin{aligned} J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}) &= E\left[\int_0^\infty r(\mathbf{x}, \mathbf{u}, \mathbf{d}) dt\right] \\ &= E\left[\int_0^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2\right) dt\right], \end{aligned} \tag{4.2}$$

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

$$V(\mathbf{x}(t)) = E\left[\int_t^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2\right) d\tau\right]. \tag{4.3}$$

Define the two-player zero-sum differential game as

$$V^*(\mathbf{x}(0)) = \min_{\mathbf{u}} \max_{\mathbf{d}} J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}), \tag{4.4}$$

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 multi-player 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

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}\mathbf{u}_j, \tag{4.5}$$

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, $\mathbf{A}(\mathbf{a})$ and $\mathbf{B}$ are the drift dynamics and input dynamics, respectively. Each element of $\mathbf{a}(t)$, $a_p(t)$ $(p = 1, 2, \cdots, m)$, 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 (4.5) are ordinary differential equations [73, 74].

The expected cost to optimize for player i is

$$\begin{aligned} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}) &= E\left[\int_0^\infty r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i})dt\right] \\ &= E\left[\int_0^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right) dt\right], \end{aligned} \tag{4.6}$$

where $\mathbf{u}_{-i}$ is the supplementary set of $\mathbf{u}_i$: $\mathbf{u}_{-i} = \{\mathbf{u}_j, j \in (1, 2, \cdots, i-1, i+1, \cdots, N)\}$. $\mathbf{Q}_i$ 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

$$V_i(\mathbf{x}(t)) = E\left[\int_t^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right) d\tau\right]. \tag{4.7}$$

Define the multi-player differential game as

$$V_i^*(\mathbf{x}(0)) = \min_{\mathbf{u}_i} J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}), \tag{4.8}$$

where $V_i^*(\mathbf{x}(0))$ is the optimal value function for player i. In the multi-player 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.

4.2.2 Preliminaries

Definition 1. [73] 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 (4.1) 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. [73] 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. [75] The system (4.1) 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 (4.2) 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. [40] Policies $\{\mathbf{u}_1^*, \mathbf{u}_2^*, \cdots, \mathbf{u}_N^*\}$ are said to constitute a Nash equilibrium solution for the N-player game if the following equation is satisfied for $\forall \mathbf{u}_i, \forall i \in N$.

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

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

Lemma 2. *[73, Theorem 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 (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. $E\left[\dot{\tilde{V}}(\mathbf{x}(t))\right]$ *is negative definite.*

4.3 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 4.3.1 studies the stability and Nash equilibrium of the proposed game, and Section 4.3.2 develops both on-policy and off-policy IRL solutions to solve the game online.

4.3.1 Stability and Nash Equilibrium

With the value function defined in (4.3), the following Bellman equation can be derived by taking derivative of $V(\mathbf{x}(t))$ with respect to time t.

$$r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E\left[\frac{\partial V}{\partial \mathbf{x}}^{\mathrm{T}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})\right] = 0. \tag{4.9}$$

with the Hamiltonian function

$$H(\mathbf{x}, \mathbf{u}, \mathbf{d}, \frac{\partial V}{\partial \mathbf{x}}) = r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + E\left[\frac{\partial V^{\mathrm{T}}}{\partial \mathbf{x}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})\right]. \tag{4.10}$$

The optimal control policies $\mathbf{u}^*$ and $\mathbf{d}^*$ can be derived by employing the stationary conditions in the Hamiltonian function [40, Page 447],

$$\begin{aligned} \frac{\partial H}{\partial \mathbf{u}} = 0 \rightarrow \mathbf{u}^* = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V^*}{\partial \mathbf{x}}, \\ \frac{\partial H}{\partial \mathbf{d}} = 0 \rightarrow \mathbf{d}^* = \frac{1}{2\gamma^2}\mathbf{C}^{\mathrm{T}}\frac{\partial V^*}{\partial \mathbf{x}}. \end{aligned} \tag{4.11}$$

Substituting (4.11) into the Bellman Equation (4.9), the following Hamilton-Jacobi-Bellman (HJB) equation is obtained.

$$\begin{aligned} H(\mathbf{x}, \mathbf{u}^*, \mathbf{d}^*, V_X^*) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + E\Big[{V_X^*}^{\mathrm{T}} A(\mathbf{a})\mathbf{x} - \frac{1}{4}{V_X^*}^{\mathrm{T}}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}V_X^* \\ + \frac{1}{4\gamma^2}{V_X^*}^{\mathrm{T}}\mathbf{C}\mathbf{C}^{\mathrm{T}}V_X^*\Big] = 0, \quad V(\mathbf{0}) = 0, \end{aligned} \tag{4.12}$$

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

Note that the HJB Equation (4.12) contains the randomly time-varying vector, $\mathbf{a}(t)$. Compared to the HJB equation defined in deterministic systems, (4.12) is harder to solve, as it involves the evaluation of uncertainty which can be computationally expensive. In the next subsection, we introduce an effective uncertainty evaluation method, and show its integration with learning methods to solve the HJB Equation (4.12) online.

Lemma 3. *For any admissible control and disturbance policies* $\mathbf{u}$ *and* $\mathbf{d}$*, let* $V \geq 0$ *be the corresponding solution to the Bellman Equation* (4.10)*, then the following equation holds [40, Lemma 10.2-1].*

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

where $\mathbf{u}^*$ *and* $\mathbf{d}^*$ *are described in* (4.11)*, and* $V_X = \frac{\partial V}{\partial \mathbf{x}}$.

Proof. Combining Equations (4.10) and (4.11), the Hamiltonian function can be further written as

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

which derives Lemma 3. □

Theorem 5. *Let $V(\mathbf{x}(t))>0$ be a smooth function satisfying the HJB equation described in* (4.12)*, then the following statements hold.*

1). The system (4.1) *is asymptotically stable in the mean with the policies $\mathbf{u}^*$ and $\mathbf{d}^*$ described in* (4.11)*.*

2). The solution (i.e., policies $\mathbf{u}^*$ and $\mathbf{d}^*$) derived in (4.11) *provides a saddle point solution to the game, and the system is in Nash equilibrium with this solution.*

Proof. 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 $\tilde{V}(\mathbf{x}(t))=\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$ [76, Page 337]. As such, one has

$$\tilde{V}(\mathbf{x}(t))=\int_t^{\infty}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}-\gamma^2\|\mathbf{d}\|^2\right)d\tau\geq 0, \tag{4.13}$$

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 $\dot{\tilde{V}}$, then the expectation of $\dot{\tilde{V}}$ is

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

The last equality is obtained from Lemma 3. 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 2, the system described in (4.1) 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

$$
\begin{aligned}
&J(\mathbf{x}(0), \mathbf{u}, \mathbf{d}) \\
&= E\left[\int_0^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \gamma^2\|\mathbf{d}\|^2\right)dt + V(\mathbf{x}(0)) + \int_0^\infty \dot{V}dt\right] \\
&= E\left[\int_0^\infty \left(r(\mathbf{x}, \mathbf{u}, \mathbf{d}) + V_X^{\mathrm{T}}(A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})\right)dt\right] + V(\mathbf{x}(0)) \\
&= E\left[\int_0^\infty \left((\mathbf{u} - \mathbf{u}^*)^{\mathrm{T}}\mathbf{R}(\mathbf{u} - \mathbf{u}^*) - \gamma^2\|\mathbf{d} - \mathbf{d}^*\|^2\right)dt\right] + V(\mathbf{x}(0)).
\end{aligned}
\tag{4.14}
$$

The last equality is obtained by combining (4.10) and Lemma 3.

It can be seen from (4.14) 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. □

4.3.2 Approximate Solutions using On-Policy and Off-Policy IRL and The MPCM

Solving the HJB Equation (4.12) analytically is extremely difficult or even impossible [75]. 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

$$V(\mathbf{x}(t)) = E\left[\int_t^{t+T} r(\mathbf{x}(\tau), \mathbf{u}(\tau), \mathbf{d}(\tau))d\tau + V(\mathbf{x}(t+T))\right], \tag{4.15}$$

where T is the time interval.

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

$$V(\mathbf{x}) = \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}), \tag{4.16}$$

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

4.3.2.1 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, which evaluates the value function $V(\mathbf{x})$ using (4.15); and policy improvement [40], which finds the optimal solution based on the current approximated value function using (4.11). For systems with uncertain system dynamics, the policy evaluation step involves uncertainty evaluation, which is typically solved by the Monte Carlo method, too slow to be used for online solutions.

Here we utilize an effective uncertainty evaluation method, called the multivariate probabilistic collocation method (MPCM) [23]. To map to the MPCM framework, we denote the generic function whose expectation to be evaluated as $G(a_1, a_2, \cdots, a_m)$, which is modulated by m uncertain parameters, i.e., $a_1, a_2, \cdots, a_m$,

with the degree of each parameter up to $2n_p - 1$, where $p = 1, 2, \cdots, m$. 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 [23].

Lemma 4. *[23, Theorem 2] Consider a system mapping modulated by m independent uncertain parameters:*

$$G(a_1, a_2, \cdots, a_m) = \sum_{q_1=0}^{2n_1-1} \sum_{q_2=0}^{2n_2-1} \cdots \sum_{q_m=0}^{2n_m-1} \psi_{q_1,q_2,\cdots,q_m} \prod_{p=1}^{m} a_p^{q_p}, \tag{4.17}$$

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

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

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

As showed in the above Lemma, 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 [23]. 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 (4.15) can be approximated by the mean output of the system mapping $G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})$ (i.e., $V(\mathbf{x}) = E[G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})]$), 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 $E[G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})]$. 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,

$$G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a}) = \sum_{q_1=0}^{2n_1-1} \sum_{q_2=0}^{2n_2-1} \cdots \sum_{q_m=0}^{2n_m-1} \psi_{q_1,q_2,\cdots,q_m}(\mathbf{x}, \mathbf{u}, \mathbf{d}) \prod_{p=1}^{m} a_p^{q_p}. \tag{4.18}$$

With this mapping, the value function can be estimated from the mean output of a reduced-order mapping according to Lemma 4 as

$$V(\mathbf{x}(t)) = E[G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})] = E\left[G'_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})\right], \tag{4.19}$$

where $G'_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})$ is the reduced-order mapping derived from the MPCM procedure [23],

$$G'_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a}) = \sum_{q_1=0}^{n_1-1} \sum_{q_2=0}^{n_2-1} \cdots \sum_{q_m=0}^{n_m-1} \Omega_{q_1,q_2,\cdots,q_m}(\mathbf{x}, \mathbf{u}, \mathbf{d}) \prod_{p=1}^{m} a_p^{q_p}. \tag{4.20}$$

The PI algorithm that integrates IRL and the MPCM for the two-player zero-sum game with uncertain system dynamics is summarized in Algorithm 3.

Algorithm 3 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 [23, 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 [23], which is the mean output of the mapping $G_{V^{(s)}}(\cdot)$ subject to uncertain parameters $\mathbf{a}(t)$,

$$G_{V^{(s)}}\left(\mathbf{x}, \mathbf{u}^{(s)}, \mathbf{d}^{(s)}, \mathbf{a}\right) = \mathbf{W}^{(s-1)^{\mathrm{T}}}\phi(\mathbf{x}(t+T)) + \int_{t}^{t+T} r\left(\mathbf{x}(\tau), \mathbf{u}^{(s)}(\tau), \mathbf{d}^{(s)}(\tau)\right) d\tau. \tag{4.21}$$

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

$$\begin{aligned} \mathbf{u}^{(s+1)} &= -\frac{1}{2}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V^{(s)}}{\partial \mathbf{x}}, \\ \mathbf{d}^{(s+1)} &= \frac{1}{2\gamma^2}\mathbf{C}^{\mathrm{T}}\frac{\partial V^{(s)}}{\partial \mathbf{x}}. \end{aligned} \tag{4.22}$$

7: Repeat procedures $3-6$.

Theorem 6. *Consider the stochastic two-player zero-sum game shown in* (4.1)-(4.4)*. The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume 1) VFA in* (4.16) *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* (4.21) *with the form of* (4.17)*, and 3) Algorithm 3 converges. Then the policies $\mathbf{u}$ and $\mathbf{d}$ derived from Algorithm 3 are optimal policies.*

Proof. 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 5 and (4.19). As such, to prove this theorem, we only need to show that the two optimal solutions, 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. Lemma 4 proved that $E[G'_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})] = E[G_{V(t)}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{a})]$, and hence the equivalence of the two optimal solutions can be proved from a contradiction method following a similar argument as described in [70, Theorem 1]. □

4.3.2.2 Off-Policy IRL

Algorithm 3 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 subsection, we provide an off-policy IRL algorithm and use three neural networks (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 (4.1) is further written as

$$\dot{\mathbf{x}} = A(\mathbf{a})\mathbf{x} + \mathbf{B}\mathbf{u}^{(s)} + \mathbf{C}\mathbf{d}^{(s)} + \mathbf{B}\left(\mathbf{u} - \mathbf{u}^{(s)}\right) + \mathbf{C}\left(\mathbf{d} - \mathbf{d}^{(s)}\right). \tag{4.23}$$

Here $\mathbf{u}$ 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.

Differentiating the value function $V^{(s)}(\mathbf{x}(t))$ of the system (4.23), one has

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

The second equality is obtained by combining the Hamiltonian function (4.10) and (4.22).

Integrating both sides in (4.24), one has

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

Note that for any fixed admissible control and disturbance behavior policies $\mathbf{u}$ and $\mathbf{d}$, (4.24) 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.

$$\begin{aligned} V^{(s)}(\mathbf{x}) &= \mathbf{W}^{(s)^{\mathrm{T}}}\phi(\mathbf{x}), \\ \mathbf{u}^{(s+1)}(\mathbf{x}) &= \mathbf{W}_{\mathbf{u}}^{(s+1)^{\mathrm{T}}}\phi_{\mathbf{u}}(\mathbf{x}), \\ \mathbf{d}^{(s+1)}(\mathbf{x}) &= \mathbf{W}_{\mathbf{d}}^{(s+1)^{\mathrm{T}}}\phi_{\mathbf{d}}(\mathbf{x}). \end{aligned} \tag{4.24}$$

The detailed algorithm that integrates off-policy IRL and the MPCM is described in Algorithm 4.

Algorithm 4 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 [23, Section II] to select a set of samples for the uncertain vector $\mathbf{a}(t)$.

3: For each iteration s, find the value of

$$V^{(s)}(\mathbf{x}(t+T)) + \int_t^{t+T} \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{(s)^{\mathrm{T}}}\mathbf{R}\mathbf{u}^{(s)} - \gamma^2\|\mathbf{d}^{(s)}\|^2\right) d\tau \tag{4.25}$$

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 [23],

$$\begin{aligned} &G^o_{V^{(s)}}\left(\mathbf{x}, \mathbf{u}^{(s)}, \mathbf{d}^{(s)}, \mathbf{a}\right) \\ &= V^{(s)}(\mathbf{x}(t+T)) + \int_t^{t+T} \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{(s)^{\mathrm{T}}}\mathbf{R}\mathbf{u}^{(s)} - \gamma^2\|\mathbf{d}^{(s)}\|^2\right) d\tau. \end{aligned} \tag{4.26}$$

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

$$\begin{aligned} &E\left[G^o_{V^{(s)}}\left(\mathbf{x}, \mathbf{u}^{(s)}, \mathbf{d}^{(s)}, \mathbf{a}\right)\right] \\ &= V^{(s)}(\mathbf{x}(t)) - \int_t^{t+T} \left(2\mathbf{u}^{(s+1)^{\mathrm{T}}}\mathbf{R}\left(\mathbf{u} - \mathbf{u}^{(s)}\right) - 2\gamma^2\mathbf{d}^{(s+1)^{\mathrm{T}}}\left(\mathbf{d} - \mathbf{d}^{(s)}\right)\right) d\tau. \end{aligned} \tag{4.27}$$

6: Repeat procedures $3-5$.

Theorem 7. *Consider the stochastic two-player zero-sum game shown in* (4.1)-(4.4)*. The uncertainties in the system dynamics* a_p *follow time-invariant pdfs* $f_{A_p}(a_p)$*. Assume 1) VFA in* (4.24) *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* (4.26) *with the form of* (4.17)*, and 3) Algorithm 4 converges. Then the policies derived from off-policy IRL described in Algorithm 4 are optimal policies.*

Proof. It has been proved 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 [75]. As such, for each MPCM sample point $\mathcal{A}^l$, $l = 1, 2, \cdots, \prod_{p=1}^{m} n_p$, 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 4 and [23]). As such, the expected value function derived from the two algorithms is identical, and hence the off-policy solution is the optimal control policy. □

Remark 4. *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 5. *Note that the admissible control and disturbance policies initialize the first step in Algorithm 4. 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 6. *Algorithms 3 and 4 integrate IRL and the 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 are typically solved by Monte Carlo method and its variants, which are time-consuming to use for online solutions. The proposed algorithms find the optimal solutions accurately with computational efficiency, as indicated in Lemma 4 and Theorems 6 and 7. 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.*

4.4 Multi-Player 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 4.4.1, and online learning algorithms are provided in Section 4.4.2.

4.4.1 Stability and Global Nash Equilibrium

Consider the value function described in (4.7), the differential Bellman equation can be found by taking derivative of $V_i(\mathbf{x}(t))$ with respect to time t,

$$r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}) + E\left[\frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}}\left(\mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j\right)\right] = 0. \tag{4.28}$$

The Hamiltonian function is

$$H_i\left(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}}\right) = r_i(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}) + E\left[\frac{\partial V_i^{\mathrm{T}}}{\partial \mathbf{x}}\left(\mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N}\mathbf{B}\mathbf{u}_j\right)\right]. \quad (4.29)$$

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

$$\frac{\partial H_i}{\partial \mathbf{u}_i} = 0 \rightarrow \mathbf{u}_i^* = -\frac{1}{2}\mathbf{R}_{ii}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_i^*}{\partial \mathbf{x}}. \quad (4.30)$$

Substituting (4.30) into the Bellman Equation (4.28), the following Hamilton-Jacobi-Bellman (HJB) equation is obtained.

$$\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + E\left[\frac{1}{4}\sum_{j=1}^{N}\frac{\partial V_j^{*\mathrm{T}}}{\partial \mathbf{x}}\mathbf{B}\mathbf{R}_{jj}^{-\mathrm{T}}\mathbf{R}_{ij}\mathbf{R}_{jj}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_j^*}{\partial \mathbf{x}} + \frac{\partial V_i^{*\mathrm{T}}}{\partial \mathbf{x}}\left(A(\mathbf{a})\mathbf{x} - \frac{1}{2}\sum_{j=1}^{N}\mathbf{B}\mathbf{R}_{jj}^{-1}\mathbf{B}^{\mathrm{T}}\frac{\partial V_j^*}{\partial \mathbf{x}}\right)\right] = 0. \quad (4.31)$$

Lemma 5. *For any admissible control policy $\mathbf{u}_i$, let $V_i \geq 0$ be the corresponding solution to the Bellman Equation* (4.28)*, then the following equation holds.*

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

Proof. Combining (4.29) and (4.30), Lemma 5 can be obtained naturally following a similar procedure as described in Lemma 3. □

Theorem 8. *Let V_i be a smooth function satisfying the HJB Equation* (4.31)*, then the following statements hold.*

1). The system (4.5) *is asymptotically stable in the mean with the control policy $\mathbf{u}_i^*$ described in* (4.30)*.*

2). The control policies $[\mathbf{u}_1^*, \mathbf{u}_2^*, \cdots, \mathbf{u}_N^*]$ *derived in* (4.30) *are global Nash equilibrium policies.*

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

$$E\left[\tilde{V}_i\right] = E\left[\int_t^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^N \mathbf{u}_j^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right) d\tau\right] \geq 0. \tag{4.32}$$

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 (4.5) is asymptotically stable in the mean [73].

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

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

The second equality is obtained by combining (4.29) and Lemma 2.

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

$$J_i(\mathbf{x}(0), \mathbf{u}_i, \mathbf{u}_{-i}^*) = V_i(\mathbf{x}(0)) + E\left[\int_0^\infty (\mathbf{u}_i - \mathbf{u}_i^*)^{\mathrm{T}} \mathbf{R}_{ii}(\mathbf{u}_i - \mathbf{u}_i^*) dt\right]. \tag{4.33}$$

It can be seen from (4.33) 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. □

4.4.2 Approximate Solutions Using On-Policy and Off-Policy IRL and MPCM

The IRL Bellman equation for each player is given as [41]

$$V_i(\mathbf{x}(t)) = E\left[\int_t^{t+T} r_i(\mathbf{x}(\tau), \mathbf{u}_i(\tau), \mathbf{u}_{-i}(\tau))d\tau + V_i(\mathbf{x}(t+T))\right]. \tag{4.34}$$

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

$$V_i(\mathbf{x}) = \mathbf{W}_i^{\mathrm{T}} \phi_i(\mathbf{x}), \tag{4.35}$$

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.

4.4.2.1 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 (4.34) 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.

$$V_i(\mathbf{x}(t)) = E\left[G_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})\right] = E\left[G'_{V_i(t)}(\mathbf{x}, \mathbf{u}_i, \mathbf{u}_{-i}, \mathbf{a})\right], \tag{4.36}$$

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 [23]. The detailed algorithm is described in Algorithm 5.

Algorithm 5 Policy iteration for multi-player nonzero-sum game with uncertain system dynamics

1: Initialize each player with initial state $\mathbf{x}(0)$ and admissible control policy $\mathbf{u}_i(0)$, $i = 1, 2, \cdots, N$.

2: Apply the MPCM procedure [23, Section II] to select a set of samples for the uncertain vector $\mathbf{a}(t)$.

3: For each iteration s, find the value of

$$\int_t^{t+T} r_i\left(\mathbf{x}(\tau), \mathbf{u}_i^{(s)}(\tau), \mathbf{u}_{-i}^{(s)}(\tau)\right) d\tau + \mathbf{W}_i^{(s-1)^{\mathrm{T}}} \phi_i(\mathbf{x}(t+T)) \tag{4.37}$$

at each MPCM sample.

4: Find the value function $V_i^{(s)}(\mathbf{x}(t))$ using the MPCM [23], which is the mean output of the mapping $G_{V_i^{(s)}}(\cdot)$ subject to uncertain parameters $\mathbf{a}(t)$,

$$\begin{aligned} &G_{V_i^{(s)}}\left(\mathbf{x}, \mathbf{u}_i^{(s)}, \mathbf{u}_{-i}^{(s)}, \mathbf{a}\right) \\ &= \int_t^{t+T} r_i\left(\mathbf{x}(\tau), \mathbf{u}_i^{(s)}(\tau), \mathbf{u}_{-i}^{(s)}(\tau)\right) d\tau + \mathbf{W}_i^{(s-1)^{\mathrm{T}}} \phi_i(\mathbf{x}(t+T)). \end{aligned} \tag{4.38}$$

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

$$\mathbf{u}_i^{(s+1)} = -\frac{1}{2}\mathbf{R}_{ii}^{-1}\mathbf{B}^{\mathrm{T}} \frac{\partial V_i^{(s)}}{\partial \mathbf{x}}. \tag{4.39}$$

7: Repeat procedures 3-6.

Theorem 9. *Consider the stochastic multi-player nonzero-sum game shown in Equations* (4.5)-(4.8)*. The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume 1) VFA in* (4.35) *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* (4.38) *with the form of* (4.17)*, and 3) Algorithm 5 converges. Then the solution found from Algorithm 5 is the optimal control solution.*

Proof. This proof follows a similar procedure as described in Theorem 6. □

4.4.2.2 Off-Policy IRL

We introduce auxiliary variable $\mathbf{u}_j^{(s)}$ for the player j, $(j = 1, 2, \cdots, N)$, and rewrite the system dynamics described in (4.5) as

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}\mathbf{u}_j^{(s)} + \sum_{j=1}^{N} \mathbf{B}\left(\mathbf{u}_j - \mathbf{u}_j^{(s)}\right), \tag{4.40}$$

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 (4.40), one has

$$\begin{aligned}
&\dot{V}_i^{(s)}(\mathbf{x}(t)) \\
&= E\left[\frac{\partial V_i^{(s)\mathrm{T}}}{\partial \mathbf{x}}\left(A(\mathbf{a})\mathbf{x} + \sum_{j=1}^{N} \mathbf{B}\mathbf{u}_j^{(s)} + \sum_{j=1}^{N} \mathbf{B}\left(\mathbf{u}_j - \mathbf{u}_j^{(s)}\right)\right)\right] \\
&= -\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} - \sum_{j=1}^{N} \mathbf{u}_j^{(s)\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j^{(s)} - \sum_{j=1}^{N} 2\mathbf{u}_i^{(s+1)\mathrm{T}}\mathbf{R}_{ii}\left(\mathbf{u}_j - \mathbf{u}_j^{(s)}\right).
\end{aligned} \tag{4.41}$$

The second equality is obtained by combining the Hamiltonian function (4.29) and (4.39).

Integrating both sides in (4.41), one has

$$\begin{aligned}
&V_i^{(s)}(\mathbf{x}(t+T)) - V_i^{(s)}(\mathbf{x}(t)) \\
&\quad = E\left[\int_t^{t+T} -\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}_i\mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{(s)\mathrm{T}}\mathbf{R}_{ij}\mathbf{u}_j\right) d\tau\right] \\
&\qquad - \int_t^{t+T}\left(\sum_{j=1}^{N} 2\mathbf{u}_i^{(s+1)\mathrm{T}}\mathbf{R}_{ii}\left(\mathbf{u}_j - \mathbf{u}_j^{(s)}\right)\right) d\tau.
\end{aligned} \tag{4.42}$$

For any fixed admissible behavior control policy $\mathbf{u}_j$ $(j = 1, 2, \cdots, N)$, (4.42) 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.

$$\begin{aligned}
V_i^{(s)}(\mathbf{x}) &= \mathbf{W}_i^{(s)\mathrm{T}}\phi_i(\mathbf{x}), \\
\mathbf{u}_i^{(s+1)}(\mathbf{x}) &= \mathbf{W}_{\mathbf{u},i}^{(s+1)\mathrm{T}}\phi_{\mathbf{u},i}(\mathbf{x}).
\end{aligned} \tag{4.43}$$

The detailed algorithm that integrates off-policy IRL and the MPCM for the multi-player nonzero-sum game is described in Algorithm 6.

Algorithm 6 Off-Policy IRL for multi-player 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 [23, 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 [23],

$$\begin{aligned} & G^o_{V_i^{(s)}} \left(\mathbf{x}, \mathbf{u}_i^{(s)}, \mathbf{u}_{-i}^{(s)}, \mathbf{a} \right) \\ & = V_i^{(s)}(\mathbf{x}(t+T)) + \int_t^{t+T} \left(\mathbf{x}^{\mathrm{T}} \mathbf{Q}_i \mathbf{x} + \sum_{j=1}^{N} \mathbf{u}_j^{(s)\mathrm{T}} \mathbf{R}_{ij} \mathbf{u}_j \right) d\tau. \end{aligned} \tag{4.44}$$

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

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

6: Repeat procedures $3-5$.

Theorem 10. *Consider the stochastic multi-player nonzero-sum game shown in* (4.5)-(4.8)*. The uncertainties in the system dynamics a_p follow time-invariant pdfs $f_{A_p}(a_p)$. Assume 1) VFA in* (4.43) *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* (4.44) *with the form of* (4.17)*, and 3) Algorithm 6 converges. Then the solution found from off-policy IRL described in Algorithm 6 is the optimal solution.*

Proof. For the muti-player nonzero-sum game with deterministic system dynamics, the solution derived from the off-policy IRL and on-policy IRL have been proved to be identical [47]. The proof for the game with uncertain system dynamics then follows a similar argument as described in Theorem 7. □

Remark 7. *Algorithms 5 and 6 integrate IRL and the MPCM to solve the multi-player 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.*

4.5 Illustrative Examples

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

4.5.1 Two-Player Zero-Sum Game

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

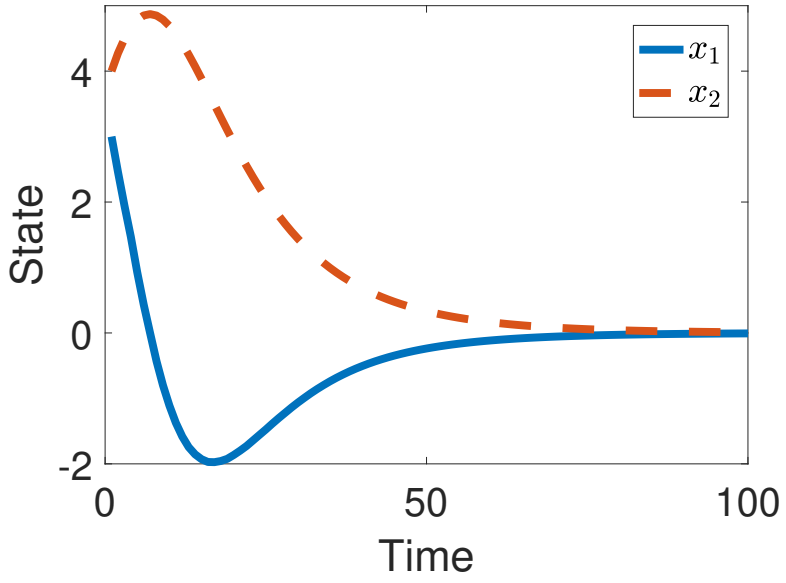
$$\dot{\mathbf{x}} = \begin{bmatrix} a_1(t) & a_2(t) \\ a_3(t) & a_4(t) \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \mathbf{u} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \mathbf{d}.$$

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: $f(a_1(t)) = \frac{1}{2}$, $0 < a_1(t) < 2$; $f(a_2(t)) = 2$, $0 < a_2(t) < 0.5$; $f(a_3(t)) = 1$, $0.5 < a_3(t) < 1.5$; and $f(a_4(t)) = \frac{1}{2}$, $-1 < a_4(t) < 1$. 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

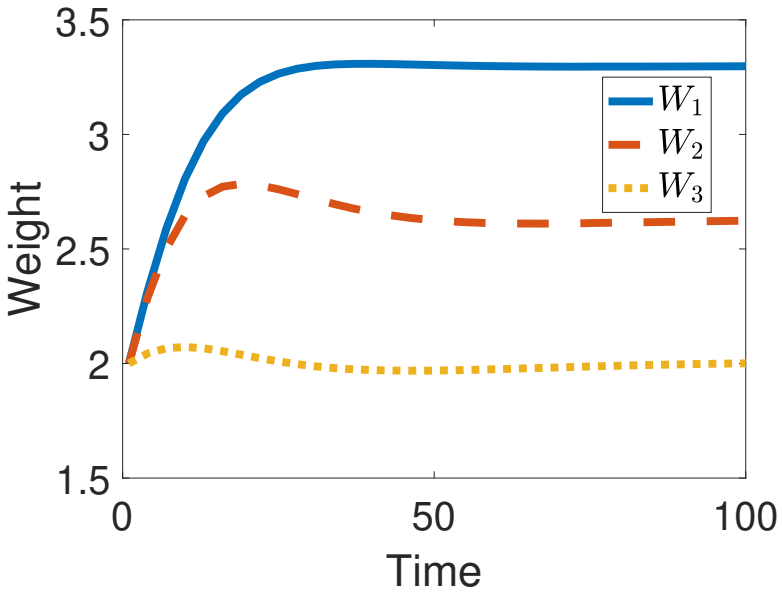
$\mathbf{W} = [W_1, W_2, W_3]^{\mathrm{T}}$. Figures 8.1(a) and 8.1(b) show the evolution of the system states and the derived value function weights respectively, using the on-policy PI algorithm (Algorithm 3). It can be seen that the system states converge to $\mathbf{0}$ in the limit of large time 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 3 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 (4.15), 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 value function and significant reduction of computational load demonstrate the value of using the proposed integrated RL and the MPCM algorithm to facilitate decision for this game.

We then simulate the off-policy IRL algorithm described in Algorithm 4. Figures 8.2(a) and 8.2(b) show the evolution of system states and neural network 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 7.

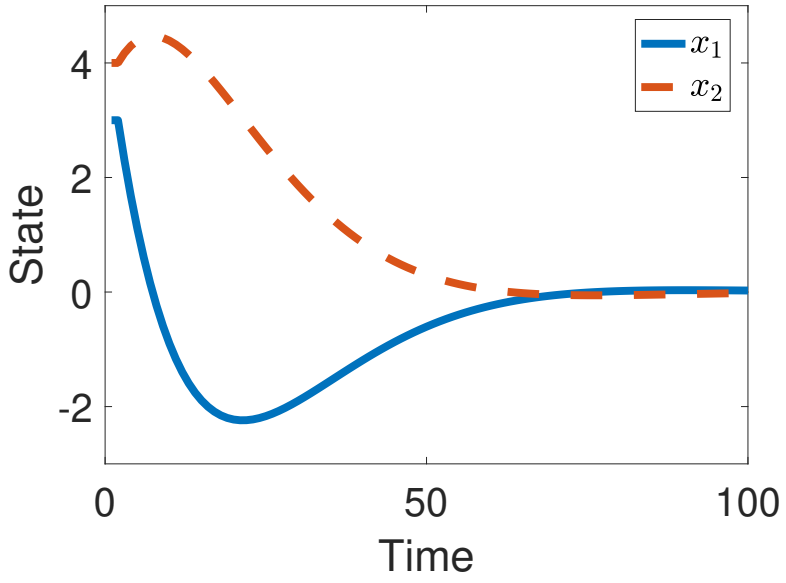


(a)

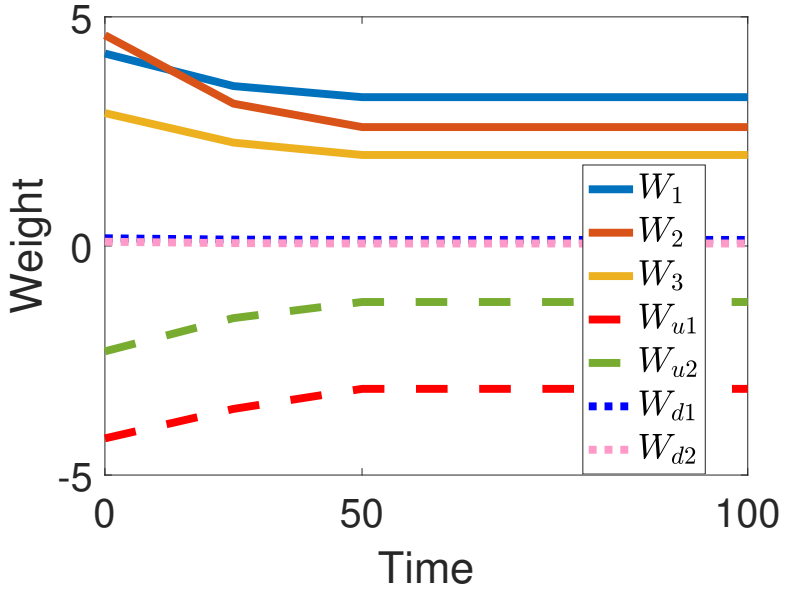


(b)

Figure 4.1. Solution of two-player zero-sum game derived from Algorithm 3. (a) The evolution of system states, and (b) the updates of value function weights.



(a)



(b)

Figure 4.2. Solution of two-player zero-sum game derived from Algorithm 4. (a) The evolution of system states, and (b) the updates of neural network weights.

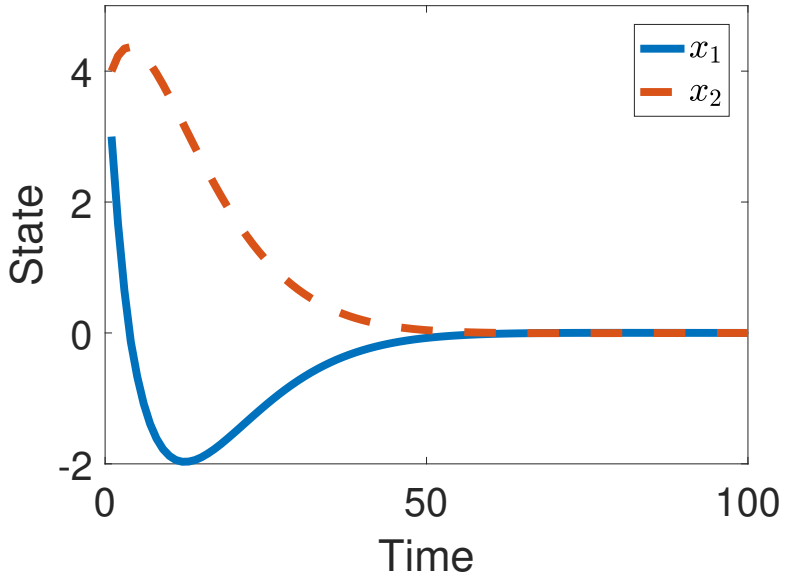
4.5.2 Multi-Player Nonzero-Sum Game

We then simulate the multi-player nonzero-sum game discussed in Section 4.4, where the number of players $N = 3$. The system dynamic is described as follows,

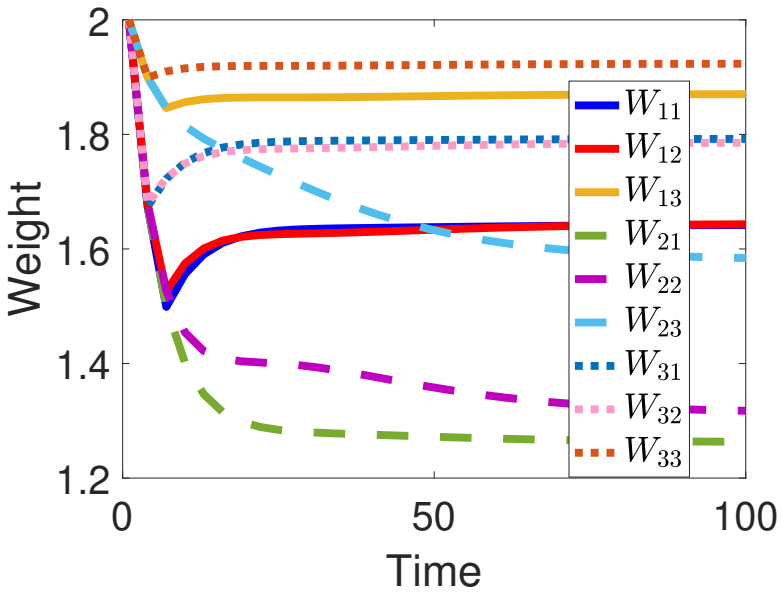
$$\dot{\mathbf{x}} = \begin{bmatrix} a_1(t) & a_2(t) \\ a_3(t) & a_4(t) \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_1 + \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_2 + \begin{bmatrix} 1.3 \\ 0 \end{bmatrix} \mathbf{u}_3,$$

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 4.5.1. The parameters in the value function are selected as $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}_3 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, $R_{12} = R_{13} = R_{21} = R_{23} = R_{31} = R_{32} = 1$, $R_{11} = 2$, $R_{22} = 3$, 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. Figure 4.3(a) shows the evolution of system states, and Figure 4.3(b) shows the learned value function weights.

We also simulate the off-policy IRL algorithm described in Algorithm 6. Figures 4.4(a) and 4.4(b) show 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 the figures that the off-policy IRL algorithm works well for the multi-player 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 10.

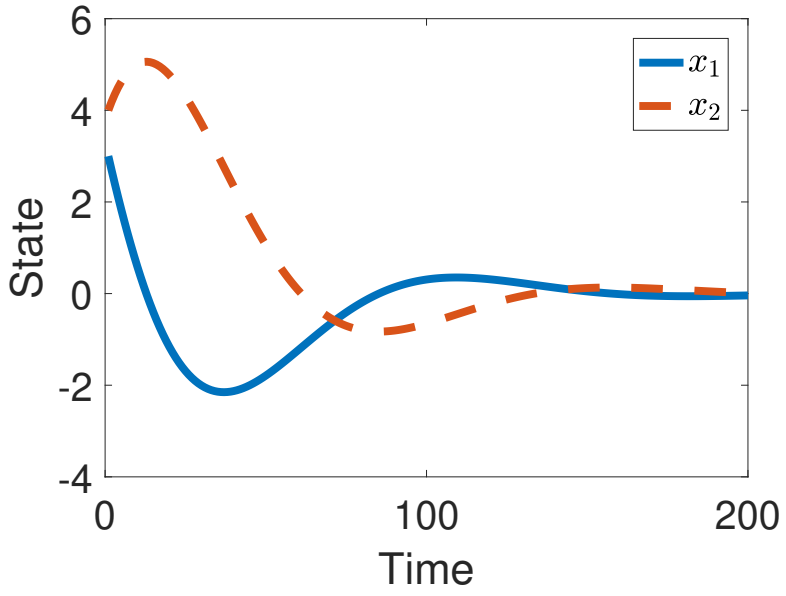


(a)

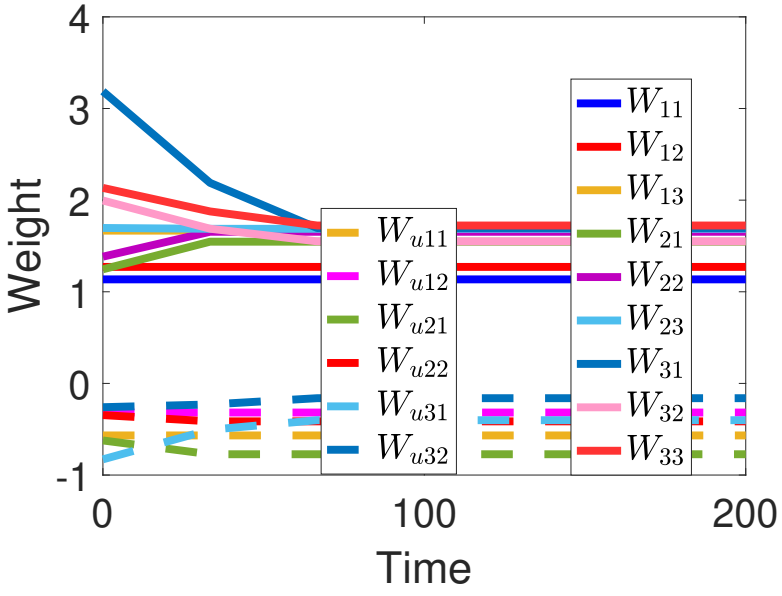


(b)

Figure 4.3. Solution of multi-player nonzero-sum game derived from Algorithm 5. (a) The evolution of system states, and (b) the updates of value function weights.



(a)



(b)

Figure 4.4. Solution of multi-player nonzero-sum game derived from Algorithm 6. (a) The evolution of system states, and (b) the updates of neural network weights.

CHAPTER 5

DIFFERENTIAL GRAPHICAL GAME WITH DISTRIBUTED GLOBAL NASH SOLUTION

5.1 Introduction

Networked multi-agent systems have received extensive attention in the past years because of their wide applications in mobile robots, sensor networks, autonomous driving systems, and so on [77–81]. The consensus control problems in networked MAS aim to design control protocols to make all agents reach a common value or track a reference trajectory (or leader's trajectory) based on the local information of each agent and its neighbors. Consensus control studies do not necessarily impose optimality, and as such, their control policies can be far from being optimal. Consensus optimal control that does not only reach consensus but also guarantees distributed optimal solutions thus recently attracted significant attention [82, 83]. These studies assume a global objective, and agents do not have conflicts of interest among themselves.

On the contrary, game theory provides mathematical formulations to solve optimal decision-making problems for networked MAS, where each agent can have its own interest, or performance index, to optimize [36, 37, 40]. Traditional game theory utilizes static and immediate costs, which cannot capture the evolution of system dynamics [36,37]. Recently, game theory concepts have been integrated with the optimal control theory to develop feedback control solutions as dynamic game strategies, which are called multi-agent differential games [40, 41, 84]. Players in such games are often assumed to have access to the full state of the system. In many applications,

players in a system cannot obtain the complete system state information, and they have to make their decisions based on limited sensing capabilities. In *differential graphical games*, players are connected by a communication graph that captures the information flow, and each player aims to find its optimal control policy based on its own and its neighbors' state information. Differential graphical games have been studied in [41, 85–88], to solve the cooperative optimal tracking problems, and have become one of the most interesting branches in multi-agent differential games.

Pioneering efforts have been made to develop solutions for differential graphical games [19, 85, 86, 88–91]. In one direction, Nash solutions are sought by finding the *best response* for each agent. It was proven in [85, 86] that a global Nash equilibrium can be obtained if all agents use their best response strategies. As the simulations showed in these papers, this approach requires the state information of not only the agents' neighbors, but also their neighbors' neighbors [85,88]. As such, the developed control policies are Nash, but are not truly distributed. In the other direction, *minmax* strategies have been used to achieve distributed control in differential graphical games [19,89–91]. Minmax strategies are based on the idea that each agent prepares itself for the worst-case behavior of its neighbors. To be more specific, each agent assumes that its neighbors act to oppose the agent, by maximizing the agent's cost function. From the perspective of an individual agent, this formulation is the same as H_∞ control, which solves the optimal control problem that minimizes the impact of worse-case disturbances

This chapter studies the distributed and Nash properties for graphical game solutions in linear dynamical networks. The main contributions of this chapter are three-fold. First, we analyze the existing differential graphical game formulations, and prove that the best response strategy can constitute Nash, but does not provide distributed solutions. In addition, the minmax strategy can provide distributed solu-

tions, but prevent the agents to reach a global Nash equilibrium. In other words, no solution that is both Nash and distributed exists in the literature. Second, we develop a novel differential graphical game formulation which promises a solution that is both distributed and can reach global Nash equilibrium. Third, we provide formal proofs for the stability and Nash equilibrium properties of the newly proposed differential graphical game.

This chapter is organized as follows. Section 5.2 analyzes existing graphical games and their solutions, including best response and minmax strategies. Section 5.3 proposes a novel differential graphical game formulation. Section 5.4 analyzes the stability and Nash equilibrium properties of the novel graphical game. Section 5.5 uses illustrative examples to validate the theoretical analysis.

5.2 Differential graphical games

5.2.1 Communication Graph

Consider a set of N agents connected by a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents, $\mathcal{V} = \{1, 2, ..., N\}$, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges. The graph adjacency matrix is $\mathcal{A} = [a_{ij}]$, where $a_{ij} > 0$ if $(j, i) \in \mathcal{E}$, and $a_{ij} = 0$ if $(j, i) \notin \mathcal{E}$. $(j, i) \in \mathcal{E}$ means there exists an edge starting from j to i in the directed graph. It is assumed that the graph is simple, i.e., there are no repeated edges or self-loops (i.e., $(i, i) \notin \mathcal{E}$ for $\forall i$). Denote the set of neighbors of agent i as N_i, i.e., $N_i = \{j : (j, i) \in \mathcal{E}\}$. Denote the in-degree matrix as $\mathcal{D}$, i.e., $\mathcal{D} = diag(d_1, d_2, \cdots, d_N)$, where d_i is the i^{th} row sum of $\mathcal{A}$: $d_i = \sum_j a_{ij}$. Define the graph Laplacian matrix as L, $L = \mathcal{D} - \mathcal{A}$, which has all row sums equal to zero.

5.2.2 Game Settings

Consider N agents with identical dynamics

$$\dot{x}_i = Ax_i + Bu_i, \tag{5.1}$$

where $x_i(t) \in \mathbb{R}^n$ and $u_i \in \mathbb{R}^m$ are the state and control input vectors for agent i respectively, where $i = 1, 2, \cdots, N$. A and B are the drift and input matrices. The dynamics of the leader is given by

$$\dot{x}_0 = Ax_0. \tag{5.2}$$

The communication link between the leader and agent i is represented by the pinning gain $g_i \geqslant 0$. $g_i > 0$ means that the leader node can be observed by agent i. Denote the pinning matrix as G, then $G = diag\{g_i\} \in \mathbb{R}^{N\times N}$. The pair (A, B) is controllable. The graph $\mathcal{G}$ is strongly connected and at least one agent can observe the leader.

The local neighborhood tracking error of agent i is defined as

$$\delta_i = \sum_{j\in N_i} a_{ij}(x_i - x_j) + g_i(x_i - x_0). \tag{5.3}$$

The error dynamics is

$$\begin{aligned} \dot{\delta}_i &= \sum_{j\in N_i} a_{ij}(\dot{x}_i - \dot{x}_j) + g_i(\dot{x}_i - \dot{x}_0) \\ &= A\delta_i + (d_i + g_i)Bu_i - \sum_{j\in N_i} a_{ij}Bu_j. \end{aligned} \tag{5.4}$$

The cost function to minimize for each agent is

$$J_i = \int_0^\infty r_i(\delta_i, \delta_{-i}, u_i, u_{-i})dt, \tag{5.5}$$

where δ_{-i} and u_{-i} are local neighborhood tracking errors and control inputs of the neighbors of agent i, respectively, i.e., $\delta_{-i} = \{\delta_j, j \in N_i\}$, and $u_{-i} = \{u_j, j \in N_i\}$.

$r_i(\delta_i, \delta_{-i}, u_i, u_{-i}) \geqslant 0$ is a scalar function, and $r_i(\delta_i, \delta_{-i}, u_i, u_{-i}) = 0$ if and only if $\delta_i = 0$ and $\delta_{-i} = 0$.

The value function corresponding to the performance index is

$$V_i = \int_t^\infty r_i(\delta_i, \delta_{-i}, u_i, u_{-i}) d\tau. \tag{5.6}$$

The control objective of agent i in the graphical game is to find the optimal policy u_i^* such that

$$u_i^*(t) = \underset{u_i}{\operatorname{argmin}} V_i(\delta(t)).$$

Definition 5 (Best Response)**.** Agent i's best response to fixed policies u_{-i} of his neighbors is the policy u_i^* such that [92, page 191]

$$J_i(u_i^*, u_{-i}) \leq J_i(u_i, u_{-i}) \tag{5.7}$$

for all possible u_i of agent i.

Definition 6 (Global Nash Equilibrium)**.** An N-tuple of policies $\{u_1^*, u_2^*, ..., u_N^*\}$ is said to constitute a global Nash equilibrium solution for an N-player game if for all $i = 1, 2, \cdots, N$ and $\forall u_i$ [92, page 190]

$$J_i^* \triangleq J_i(u_i^*, u_{\mathcal{G}-i}^*) \leq J_i(u_i, u_{\mathcal{G}-i}^*), \tag{5.8}$$

where $u_{\mathcal{G}-i}$ is the set of policies of all the other agents in the graph other than agent i, i.e., $u_{\mathcal{G}-i} = \{u_j : j \in N, j \neq i\}$. The corresponding N-tuple of cost functions $\{J_1^*, J_2^*, \cdots, J_N^*\}$ is known as a global Nash equilibrium outcome of the N-player game.

Remark 8. *Note that differential graphical games, as a branch of differential games, are very different from the widely studied differential games in the literature. In those widely studied differential games, players are assumed to have access to the full state information of all agents in a network, and as a result, the control policies are*

generally not distributed [41, 93]. On the contrary, in differential graphical games, player are linked by a time-invariant communication graph, and each player can only have access to its own and its direct neighbors' system states. As such, distributed control strategies are sought to solve the differential graphical games.

5.2.3 Existing Differential Graphical Games

Existing differential graphical games assume the following cost function [41, 85, 86]

$$J_i = \int_0^\infty r_i(\delta_i, u_i, u_{-i})dt, \tag{5.9}$$

where $r_i(\delta_i, u_i, u_{-i})$ has the quadratic form

$$r_i(\delta_i, u_i, u_{-i}) = \delta_i^{\mathrm{T}} Q_i \delta_i + u_i^{\mathrm{T}} R_i u_i + \sum_{j\in N_i} a_{ij} u_j^{\mathrm{T}} R_{ij} u_j, \tag{5.10}$$

and $Q_i > 0$, $R_i > 0$, and $R_{ij} \geqslant 0$ are constant and symmetric matrices.

The corresponding value function is

$$V_i(t) = \int_t^\infty \left(\delta_i^{\mathrm{T}} Q_i \delta_i + u_i^{\mathrm{T}} R_i u_i + \sum_{j\in N_i} a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) d\tau. \tag{5.11}$$

To find a distributed solution, it is assumed that $V_i(t)$ is only related to δ_i, i.e., $\dot{V}_i = \nabla V_i^{\mathrm{T}} \dot{\delta}_i$, where $\nabla V_i = \frac{\partial V_i}{\partial \delta_i}$. In particular, the value function is assumed to have the quadratic form [41, 85, 86], i.e.,

$$V_i(\delta_i) = \delta_i^{\mathrm{T}} P_i \delta_i, \tag{5.12}$$

where P_i is a symmetric positive definite matrix, and $\nabla V_i = 2P_i\delta_i$.

The Hamiltonian associated with this cost function is

$$\begin{aligned} H_i\left(\delta_i, u_i, u_{-i}, \nabla V_i\right) = \nabla V_i^{\mathrm{T}} \left(A\delta_i + (d_i + g_i)Bu_i - \sum_{j\in N_i} a_{ij} Bu_j \right) \\ + \delta_i^{\mathrm{T}} Q_i \delta_i + u_i^{\mathrm{T}} R_i u_i + \sum_{j\in N_i} a_{ij} u_j^{\mathrm{T}} R_{ij} u_j = 0. \end{aligned} \tag{5.13}$$

Next we study the Nash equilibrium an distributed properties of two existing solutions to graphical games, best responses and minmax strategies.

5.2.3.1 Best responses

As proven in [85], the Nash equilibrium solution can be obtained if all agents use their best response strategies simultaneously. For the graphical game formulated in (5.10), the best response for agent i can be found by letting $\frac{\partial H_i}{\partial u_i} = \mathbf{0}$ [85], where $\mathbf{0}$ is a zero matrix with proper dimensions,

$$u_i^* = -\frac{1}{2}(d_i + g_i)R_i^{-1}B^{\mathrm{T}}\nabla V_i. \tag{5.14}$$

The value function V_i satisfies the following Hamilton-Jacobi (HJ) equation, which is derived by substituting (5.14) into (5.13),

$$\begin{aligned}\delta_i^{\mathrm{T}} Q_i \delta_i + \nabla V_i^{\mathrm{T}} A \delta_i - \frac{(d_i+g_i)^2}{4}\nabla V_i^{\mathrm{T}} B R_i^{-1} B^{\mathrm{T}} \nabla V_i + \frac{1}{2}\sum_{j\in N_j} a_{ij}(d_j+g_j)\nabla V_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} \nabla V_j \\ + \frac{1}{4}\sum_{j\in N_j} a_{ij}(d_j+g_j)^2 \nabla V_j^{\mathrm{T}} B R_j^{-1} R_{ij} R_j^{-1} B^{\mathrm{T}} \nabla V_j = 0.\end{aligned} \tag{5.15}$$

Theorem 11. *Consider the graphical game with the local neighborhood tracking error dynamics* (5.4)*, cost function* (5.10) *and the control policy* (5.14)*. Then there does not generally exist a value function $V_i(\delta_i)$ of the form* (5.12) *that solves the HJ Equation* (5.15) *to provide a global Nash equilibrium solution.*

Proof. Substituting Equation (5.12) into Equation (5.15), the HJ equation becomes

$$\begin{aligned}\delta_i^{\mathrm{T}}\left(Q_i + P_i^{\mathrm{T}} A + A^{\mathrm{T}} P_i - (d_i+g_i)^2 P_i^{\mathrm{T}} B R_i^{-1} B^{\mathrm{T}} P_i\right)\delta_i + \sum_{j\in N_i}\delta_i^{\mathrm{T}}\left(a_{ij}(d_j+g_j)P_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} P_j\right)\delta_j \\ + \sum_{j\in N_i}\delta_j^{\mathrm{T}}\left(a_{ij}(d_j+g_j)P_j^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} P_i\right)\delta_i + \sum_{j\in N_i}\delta_j^{\mathrm{T}}\left(a_{ij}(d_j+g_j)^2 P_j^{\mathrm{T}} B R_j^{-1} R_{ij} R_j^{-1} B^{\mathrm{T}} P_j\right)\delta_j = 0.\end{aligned} \tag{5.16}$$

Note that in the above HJ equation, the second term (i.e., $\sum_{j\in N_i} \delta_i^{\mathrm{T}} \left(a_{ij}(d_j+g_j)P_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} P_j\right) \delta_j$) is the transpose of the third term (i.e., $\sum_{j\in N_i} \delta_j^{\mathrm{T}} \left(a_{ij}(d_j+g_j)P_j^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} P_i\right) \delta_i$). To make the HJ Equation hold for all possible δ_i and δ_j, the following three equations need to hold:

$$Q_i + P_i^{\mathrm{T}} A + A^{\mathrm{T}} P_i - (d_i+g_i)^2 P_i^{\mathrm{T}} B R_i^{-1} B^{\mathrm{T}} P_i = \mathbf{0}, \tag{5.17}$$

$$a_{ij}(d_j+g_j) P_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} P_j = \mathbf{0}, \tag{5.18}$$

$$a_{ij}(d_j+g_j)^2 P_j^{\mathrm{T}} B R_j^{-1} R_{ij} R_j^{-1} B^{\mathrm{T}} P_j = \mathbf{0}. \tag{5.19}$$

It is clear that Equations (5.18) and (5.19) do not generally hold because $a_{ij} > 0$, $(d_j + g_j) > 0$, $P_i > 0$, $R_j > 0$, and $P_j > 0$. As such, there does not generally exist a distributed value function $V_i(\delta_i)$ of the form (5.12) that solves the HJ Equation (5.16) to provide a global Nash equilibrium solution, where each agent only uses the state information of its own and its neighbors. □

5.2.3.2 MinMax strategies

Minmax strategies provide distributed solutions for multi-agent systems [89–91]. In minmax strategies, agent i prepares its policy by assuming that the goals of its neighbors are to maximize $J_i(\delta_i, u_i, u_{-i})$ to oppose it, i.e.,

$$u_i^+ = \underset{u_i}{\operatorname{argmin}} \max_{u_{-i}} J_i(\delta_i, u_i, u_{-i}). \tag{5.20}$$

where u_i^+ is the optimal control policy derived from the minmax strategy for agent i.

With this assumption, the coupled HJ equation becomes decoupled, which hence guarantees a distributed solution in graphical games. However, we show in the next theorem that a minmax strategy does not generally permit Nash in graphical games.

Theorem 12. *Consider the graphical game with the local neighborhood tracking error dynamics* (5.4) *and cost function* (5.9). *Then the solution found by the minmax strategy* (5.20) *can not constitute a global Nash equilibrium in general.*

Proof. From Equation (5.20), one has

$$u_i^+ = \underset{u_i}{\operatorname{argmin}}\, J_i(\delta_i, u_i, v_{-i}^+), \tag{5.21}$$

where

$$v_{-i}^+ = \{v_j^+, j \in N_i\},$$

and

$$v_j^+ = \underset{v_j}{\operatorname{argmax}}\, J_i(\delta_i, u_i, v_{-i}). \tag{5.22}$$

Note that v_j^+ represents the worst-case policy of agent i's neighbor j, from the minmax strategies. v_j^+ is not necessarily the actual control policy employed by agent j, u_j.

The Nash equilibrium solution for agent i is

$$u_i^* = \underset{u_i}{\operatorname{argmin}}\, J_i(\delta_i, u_i, u_{-i}^*), \tag{5.23}$$

where

$$u_{-i}^* = \{u_j^*, j \in N_i\},$$

and

$$u_j^* = \underset{u_j}{\operatorname{argmin}}\, J_j(\delta_j, u_j, u_{-j}). \tag{5.24}$$

Here we show that $v_i^+ \neq u_i^*$ following a contradiction method. Assume $v_i^+ = u_i^*$, then one has $v_j^+ = u_j^*$ for all $j \in N_i$, by comparing (5.21) and (5.23). From the definitions of v_j^+ and u_j^* in (5.22) and (5.24) respectively, v_j^+ is the optimal policy that maximizes the cost J_i, while u_j^* is the optimal policy that minimizes the cost J_j. As such, in general, $v_j^+ = u_j^*$ does not hold, which contradicts the assumption that

$v_i^+ = u_i^*$. Therefore, the solution found from the minmax strategy is generally not Nash in graphical games.

□

Remark 9. *Best responses and minmax strategies are the two existing approaches to solve differential graphical games defined in* (5.9) *and* (5.12)*. However, none of them can find a solution that is both global Nash and distributed in the sense that each agent only uses the state information of its own and its neighbors. In particular, the solution found from best responses can constitute Nash, but it is generally not distributed because the distributed quadratic value function* (5.12) *makes the coupled HJ Equation* (5.16) *unsolvable. On the other hand, the minmax strategy can find distributed solution in differential graphical games by decoupling the HJ equation. However, it does not provide a Nash equilibrium solution because the assumptions on the neighbors' policies, i.e., the goals of agent i's neighbors are to maximize J_i to oppose it, do not generally hold in graphical games.*

5.3 A Novel Differential Graphical Game

Here we propose a novel differential graphical game formulation, which admits a distributed solution that can constitute a global Nash equilibrium. As shown in Theorem 11, the existing graphical game formulation does not permit a distributed solution that can constitute Nash because the coupled HJ Equation (5.16) is not solvable with the distributed quadratic value function (5.12). In the new differential graphical game formulation, we introduce a modified cost function, which includes extra terms with the purpose of decoupling the HJ equation to guarantee a distributed solution.

In this proposed graphical game, the cost function is defined as

$$J_i = \int_0^\infty \sum_{j \in N_i} \left(\delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) dt, \tag{5.25}$$

where $\delta_{ij} = \left[\delta_i^{\mathrm{T}}, \delta_j^{\mathrm{T}} \right]^{\mathrm{T}}$, and $Q_{ij} = \left[Q_i, \tilde{Q}_{ij}; \tilde{Q}_{ij}^{\mathrm{T}}, \hat{Q}_{ij} \right]$. R_i and Q_{ij} are symmetric matrices, and $R_i > 0$, $Q_i \geqslant 0$. Comparing this cost function with the existing cost function in Equation (5.10), one can find that two extra terms for each $j \in N_i$, i.e., $2\delta_i^{\mathrm{T}} \tilde{Q}_{ij} \delta_j$ and $\delta_j^{\mathrm{T}} \hat{Q}_{ij} \delta_j$, are introduced in the proposed cost function for J_i. The interpretation of the formulation (5.25) is that agent i cares not only about its own local error δ_i, but also the local error of its neighbors δ_j. In other words, this is a cooperative formulation of the differential graphical game. We will show the necessity of introducing these extra terms in the next subsection.

The corresponding value function is

$$\begin{aligned} &V_i(\delta_i, \delta_{-i}) \\ &\quad = \int_t^\infty \sum_{j \in N_i} \left(\delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) d\tau. \end{aligned} \tag{5.26}$$

Note that the value function in (5.26) can be regarded as a summation of N_i value functions (denoted as $\tilde{V}_i(\delta_i, \delta_j)$) for each $j \in N_i$. In particular,

$$V_i(\delta_{ij}, \delta_{-i}) = \sum_{j \in N_i} \tilde{V}_i(\delta_i, \delta_j),$$

where $\tilde{V}_i(\delta_i, \delta_j)$ is

$$\tilde{V}_i(\delta_i, \delta_j) = \int_t^\infty \left(\delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) d\tau.$$

The Hamiltonian associated with the value function (5.26) is

$$H_i \left(\delta_i, \delta_{-i}, u_i, u_{-i}, \frac{\partial \tilde{V}_i}{\partial \delta_{ij}} \right) = \sum_{j \in N_i} \left(\frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_{ij}} \dot{\delta}_{ij} + \delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) = 0. \tag{5.27}$$

To find a distributed solution, $V_i(\delta_i, \delta_{-i})$ is expected to be only related to δ_i, which leads to $\frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_j} = \mathbf{0}$. As such, one has

$$\begin{aligned}\frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_{ij}}\dot{\delta}_{ij} &= \begin{bmatrix}\frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_i} & \frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_j}\end{bmatrix}\begin{bmatrix}\dot{\delta}_i \\ \dot{\delta}_j\end{bmatrix} \\ &= \frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_i}\dot{\delta}_i \\ &= \frac{\partial \tilde{V}_i^{\mathrm{T}}}{\partial \delta_i}\left(A\delta_i + (d_i+g_i)Bu_i - \sum_{j\in N_i} a_{ij}Bu_j\right).\end{aligned} \tag{5.28}$$

Substituting Equation (5.28) into (5.27), the Hamiltonian becomes

$$\begin{aligned}&H_i(\delta_i, \delta_{-i}, u_i, u_{-i}, \nabla\tilde{V}_i) \\ &= \sum_{j\in N_i}\left(\nabla\tilde{V}_i^{\mathrm{T}}\left(A\delta_i + (d_i+g_i)Bu_i - \sum_{j\in N_i} a_{ij}Bu_j\right) + \delta_{ij}^{\mathrm{T}}Q_{ij}\delta_{ij} + u_i^{\mathrm{T}}R_iu_i + a_{ij}u_j^{\mathrm{T}}R_{ij}u_j\right) = 0.\end{aligned} \tag{5.29}$$

The best response for agent i can be found by letting $\frac{\partial H_i}{\partial u_i} = \mathbf{0}$ as

$$u_i^* = -\frac{1}{2}(d_i+g_i)R_i^{-1}B^{\mathrm{T}}\nabla\tilde{V}_i, \tag{5.30}$$

where $\tilde{V}_i$ solves the following HJ equation, which is derived by substituting Equation (5.30) into (5.29).

$$\begin{aligned}\sum_{j\in N_i}\Bigg(&\delta_i^{\mathrm{T}}Q_i\delta_i + \nabla\tilde{V}_i^{\mathrm{T}}A\delta_i - \frac{(d_i+g_i)^2}{4}\nabla\tilde{V}_i^{\mathrm{T}}BR_i^{-1}B^{\mathrm{T}}\nabla\tilde{V}_i + 2\delta_i^{\mathrm{T}}\tilde{Q}_{ij}\delta_j \\ &+ \frac{1}{2}\sum_{j\in N_j} a_{ij}(d_j+g_j)\nabla\tilde{V}_i^{\mathrm{T}}BR_j^{-1}B^{\mathrm{T}}\nabla\tilde{V}_j \\ &+ \delta_j^{\mathrm{T}}\hat{Q}_{ij}\delta_j + \frac{1}{4}a_{ij}(d_j+g_j)^2\nabla\tilde{V}_j^{\mathrm{T}}BR_j^{-1}R_{ij}R_j^{-1}B^{\mathrm{T}}\nabla\tilde{V}_j\Bigg) = 0.\end{aligned} \tag{5.31}$$

Lemma 6. *Assume the value function $\tilde{V}_i(\delta_i)$ have the quadratic form, i.e.,*

$$\tilde{V}_i(\delta_i) = \delta_i^{T}\tilde{P}_i\delta_i. \tag{5.32}$$

Then the HJ Equation (5.31) *can be rewritten as the following form*

$$\begin{aligned}
&\sum_{j\in N_i} \delta_i^T \left(\tilde{P}_i^T A + A^T \tilde{P}_i + Q_i - (d_i+g_i)^2 \tilde{P}_i^T B R_i^{-1} B^T \tilde{P}_i\right) \delta_i \\
&\quad + \sum_{j\in N_i} \delta_i^T \left(\tilde{Q}_{ij} + N_i a_{ij}(d_j+g_j) \tilde{P}_i^T B R_j^{-1} B^T \tilde{P}_j\right) \delta_j \\
&\quad + \sum_{j\in N_i} \delta_j^T \left(\tilde{Q}_{ij}^T + N_i a_{ij}(d_j+g_j) \tilde{P}_j^T B R_j^{-1} B^T \tilde{P}_i\right) \delta_i \\
&\quad + \sum_{j\in N_i} \delta_j^T \left(\hat{Q}_{ij} + a_{ij}(d_j+g_j)^2 \tilde{P}_j^T B R_j^{-1} R_{ij} R_j^{-1} B^T \tilde{P}_j\right) \delta_j \\
&= 0.
\end{aligned} \tag{5.33}$$

and the best response for agent i is

$$u_i^* = -(d_i+g_i) R_i^{-1} B^T \tilde{P}_i \delta_i. \tag{5.34}$$

Proof. Since $\tilde{V}_i(\delta_i) = \delta_i^{\mathrm{T}} \tilde{P}_i \delta_i$, one has $\nabla \tilde{V}_i = 2\tilde{P}_i \delta_i$. Substituting $\tilde{V}_i(\delta_i)$ and $\nabla \tilde{V}_i$ into Equation (5.31), one has

$$\begin{aligned}
&\sum_{j\in N_i} \delta_i^{\mathrm{T}} \left(\tilde{P}_i^{\mathrm{T}} A + A^{\mathrm{T}} \tilde{P}_i + Q_i - (d_i+g_i)^2 \tilde{P}_i^{\mathrm{T}} B R_i^{-1} B^{\mathrm{T}} \tilde{P}_i\right) \delta_i \\
&\quad + \sum_{j\in N_i} \delta_i^{\mathrm{T}} \tilde{Q}_{ij} \delta_j + \sum_{j\in N_i} \sum_{j\in N_i} a_{ij}(d_j+g_j) \delta_i^{\mathrm{T}} \tilde{P}_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} \tilde{P}_j \delta_j \\
&\quad + \sum_{j\in N_i} \delta_j^{\mathrm{T}} \tilde{Q}_{ij}^{\mathrm{T}} \delta_i + \sum_{j\in N_i} \sum_{j\in N_i} a_{ij}(d_j+g_j) \delta_j^{\mathrm{T}} \tilde{P}_j^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} \tilde{P}_i \delta_i \\
&\quad + \sum_{j\in N_i} \delta_j^{\mathrm{T}} \left(\hat{Q}_{ij} + a_{ij}(d_j+g_j)^2 \tilde{P}_j^{\mathrm{T}} B R_j^{-1} R_{ij} R_j^{-1} B^{\mathrm{T}} \tilde{P}_j\right) \delta_j \\
&= 0.
\end{aligned} \tag{5.35}$$

Note that

$$\sum_{j\in N_i} \sum_{j\in N_i} a_{ij}(d_j+g_j) \delta_i^{\mathrm{T}} \tilde{P}_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} \tilde{P}_j \delta_j = N_i \sum_{j\in N_i} a_{ij}(d_j+g_j) \delta_i^{\mathrm{T}} \tilde{P}_i^{\mathrm{T}} B R_j^{-1} B^{\mathrm{T}} \tilde{P}_j \delta_j, \tag{5.36}$$

$$\sum_{j\in N_i}\sum_{j\in N_i} a_{ij}(d_j+g_j)\delta_j^{\mathrm{T}}\tilde{P}_j^{\mathrm{T}}BR_j^{-1}B^{\mathrm{T}}\tilde{P}_i\delta_i = N_i\sum_{j\in N_i} a_{ij}(d_j+g_j)\delta_j^{\mathrm{T}}\tilde{P}_j^{\mathrm{T}}BR_j^{-1}B^{\mathrm{T}}\tilde{P}_i\delta_i. \tag{5.37}$$

Substituting (5.36) and (5.37) into (5.35), (5.33) is derived.

Moreover, substituting $\nabla\tilde{V}_i = 2\tilde{P}_i\delta_i$ into (5.30), (5.34) is derived.

Note that the derived solution u_i^* in (5.34) is a distributed solution, in the sense that each agent i only utilizes the state information of its own and its neighbors. No other information needs to be communicated in the game.

□

Theorem 13. *Assume that $(A, \sqrt{Q_i})$ is observable and (A, B) is stabilizable, then there exist positive definite matrices $\tilde{P}_i$ and $\tilde{P}_j$ satisfying the coupled HJ Equation (5.33) for all possible δ_i and δ_j, if $\tilde{Q}_{ij}$ and $\hat{Q}_{ij}$ satisfy the following conditions:*

$$\tilde{Q}_{ij} = -N_i a_{ij}(d_j+g_j)\tilde{P}_i^{T}BR_j^{-1}B^{T}\tilde{P}_j, \tag{5.38}$$

$$\hat{Q}_{ij} = -a_{ij}(d_j+g_j)^2\tilde{P}_j^{T}BR_j^{-1}R_{ij}R_j^{-1}B^{T}\tilde{P}_j. \tag{5.39}$$

where $\tilde{P}_i$ and $\tilde{P}_j$ solve the following algebraic Riccati equations (ARE) respectively.

$$\tilde{P}_i^{T}A + A^{T}\tilde{P}_i + Q_i - (d_i+g_i)^2\tilde{P}_i^{T}BR_i^{-1}B^{T}\tilde{P}_i = \mathbf{0}, \tag{5.40}$$

$$\tilde{P}_j^{T}A + A^{T}\tilde{P}_j + Q_j - (d_j+g_j)^2\tilde{P}_j^{T}BR_j^{-1}B^{T}\tilde{P}_j = \mathbf{0}. \tag{5.41}$$

Proof. It is known that if $(A, \sqrt{Q_i})$ is observable and (A, B) is stabilizable, Equations (5.40) and (5.41) always have positive definite solutions for $\tilde{P}_i$ and $\tilde{P}_j$ respectively. By substituting Equations (5.38)-(5.40) into Equation (5.33) in Lemma 6, the conclusion is derived naturally. □

With the derived positive definite $\tilde{P}_i$, the value function $\tilde{V}_i(\delta_i)$ can be derived as $\tilde{V}_i(\delta_i) = \delta_i^{\mathrm{T}}\tilde{P}_i\delta_i$. As such, we have

$$V_i(\delta_i) = \sum_{j\in N_i}\tilde{V}_i(\delta_i) = N_i\delta_i^{\mathrm{T}}\tilde{P}_i\delta_i \tag{5.42}$$

.

Remark 10. *Note that Equations* (5.38)-(5.41) *always have solutions because Equations* (5.40) *and* (5.41) *are decoupled ARE equations, which always promise positive definite* $\tilde{P}_i$ *and* $\tilde{P}_j$ *[40]. This is in contrast to Equations* (5.17)-(5.19)*, which generally never have solutions. As such, for the proposed graphical game with the performance index defined in Equation* (5.25)*, there always exists a positive definite value function with the quadratic form* (5.42) *that solves the HJ equation* (5.31)*, and thus promises a distributed solution that constitutes Nash.*

Remark 11. *The proposed cost function* (5.25) *introduces extra terms (i.e.,* $\delta_i^T \tilde{Q}_{ij} \delta_j$ *and* $\delta_j^T \hat{Q}_{ij} \delta_j$*) compared to the commonly-used cost function in Equation* (5.10)*. These extra terms cancel out the coupled terms in the HJ equation (i.e.,* $\sum_{j \in N_i} 2a_{ij}(d_j + g_j) P_i^T B R_j^{-1} B^T P_j$ *and* $a_{ij}(d_j+g_j)^2 P_j^T B R_j^{-1} R_{ij} R_j^{-1} B^T P_j$*), and thus permit a distributed solution that is only dependent on the local error* δ_i *as shown in Equation* (5.34)*. There are two more points to clarify regarding the* Q_{ij} *matrix:*

1) Not all elements in the matrix Q_{ij} *can be designed arbitrarily. In particular, the sub-matrix* Q_i *can take arbitrary values as needed, as long as it is a positive semi-definite matrix. However, the other two sub-matrices* $\tilde{Q}_{ij}$ *and* $\hat{Q}_{ij}$ *need be selected according to* (5.38) *and* (5.39)*, respectively, to ensure the existence of the positive definite solution* $\tilde{V}_i$ *to the coupled HJ equation* (5.31)*.*

2) It is not necessary to restrict Q_{ij} *to be a positive semi-definite matrix. Only* Q_i *is required to be positive semi-definite. This is because the effects of the other two sub-matrices in* Q_{ij}*, i.e.,* $\tilde{Q}_{ij}$ *and* $\hat{Q}_{ij}$*, are canceled out by the coupled terms in the HJ equation. The value function* $\tilde{V}_i$ *is guaranteed to be positive definite according to Theorem 13, regardless of whether* Q_{ij} *is positive semi-definite or not.*

To make the procedures of solving the proposed graphical game clearer, an algorithm is now presented and described in Algorithm 7.

Algorithm 7 Procedures for Solving the Novel Graphical Game

Input:

System dynamic matrices A and B;

Communication graph matrices $\mathcal{A}$, $\mathcal{D}$, and G;

Cost function weighting matrices Q_i, R_i, and R_{ij}, $\forall j \in N_i$, and $i = 1, 2, \cdots, N$.

Output:

Optimal control u_i^*, $i = 1, 2, \cdots, N$.

Procedures:

1: Solve $\tilde{P}_i$ from (5.40) for all $i = 1, 2, \cdots, N$.

2: Select $\tilde{Q}_{ij}$ and $\hat{Q}_{ij}$ according to (5.38) and (5.39) respectively, $\forall j \in N_i$, and $i = 1, 2, \cdots, N$.

3: Find u_i^* using (5.34) for all $i = 1, 2, \cdots, N$.

4: **return** u_i^*.

5.4 Stability and Global Nash Equilibrium Analysis

This section studies properties of the proposed solution for the novel graphical game formulation. Asymptotical stability and Nash equilibrium results are proven.

The Hamiltonian in Equation (5.29) can be regarded as a summation of N_i Hamiltonians (denote as $H_{ij}(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i)$) for each $j \in N_i$. In particular,

$$H_i\left(\delta_i, \delta_{-i}, u_i, u_{-i}, \nabla\tilde{V}_i\right) = \sum_{j \in N_i} H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right), \tag{5.43}$$

where

$$\begin{aligned} &H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right) \\ &\quad = \nabla\tilde{V}_i^{\mathrm{T}}\left(A\delta_i + (d_i + g_i)Bu_i - \sum_{j \in N_i} a_{ij}Bu_j\right) + \delta_{ij}^{\mathrm{T}}Q_{ij}\delta_{ij} + u_i^{\mathrm{T}}R_iu_i + a_{ij}u_j^{\mathrm{T}}R_{ij}u_j. \end{aligned} \tag{5.44}$$

Lemma 7. *The Hamiltonian* $H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right)$ *in Equation* (5.44) *can be rewritten as the following form.*

$$\begin{aligned} H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right) &= H^*_{ij}\left(\delta_{ij}, u^*_i, u^*_{-i}, \nabla\tilde{V}_i\right) + (u_i - u^*_i)^T R_i (u_i - u^*_i) \\ &+ a_{ij}(u_j - u^*_j)^T R_{ij}(u_j - u^*_j) \\ &+ \frac{2}{(d_i + g_i)} \sum_{j\in N_i} a_{ij}(u^*_i)^T R_i (u_j - u^*_j) + 2a_{ij}(u^*_j)^T R_{ij}(u_j - u^*_j). \end{aligned} \tag{5.45}$$

Proof. According to Equation (5.44), the optimal Hamiltonian is

$$\begin{aligned} H_{ij}\left(\delta_{ij}, u^*_i, u^*_{-i}, \nabla\tilde{V}_i\right) &= \nabla\tilde{V}_i^{\mathrm{T}} \left(A\delta_i + (d_i + g_i)Bu^*_i - \sum_{j\in N_i} a_{ij} Bu^*_j \right) \\ &+ \delta_{ij}^{\mathrm{T}} Q_{ij}\delta_{ij} + (u^*_i)^{\mathrm{T}} R_i u^*_i + a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u^*_j \\ &= \nabla\tilde{V}_i^{\mathrm{T}} A\delta_i + \delta_{ij}^{\mathrm{T}} Q_{ij}\delta_{ij} + (d_i + g_i)\nabla\tilde{V}_i^{\mathrm{T}} Bu^*_i + (u^*_i)^{\mathrm{T}} R_i u^*_i \\ &+ a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u^*_j - \sum_{j\in N_i} a_{ij}\nabla\tilde{V}_i^{\mathrm{T}} Bu^*_j \\ &= \nabla\tilde{V}_i^{\mathrm{T}} A\delta_i + \delta_{ij}^{\mathrm{T}} Q_{ij}\delta_{ij} - (u^*_i)^{\mathrm{T}} R_i u^*_i + a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u^*_j \\ &- \sum_{j\in N_i} a_{ij}\nabla\tilde{V}_i^{\mathrm{T}} Bu^*_j. \end{aligned} \tag{5.46}$$

The last equality holds because $(d_i+g_i)\nabla\tilde{V}_i^{\mathrm{T}} Bu^*_i = -2(u^*_i)^{\mathrm{T}} R_i u^*_i$ from Equation (5.30). The following equations also hold,

$$(u_i - u^*_i)^{\mathrm{T}} R_i (u_i - u^*_i) = u_i^{\mathrm{T}} R_i u_i + (u^*_i)^{\mathrm{T}} R_i u^*_i - 2(u^*_i)^{\mathrm{T}} R_i u_i, \tag{5.47}$$

$$a_{ij}(u_j - u^*_j)^{\mathrm{T}} R_{ij}(u_j - u^*_j) = a_{ij} u_j^{\mathrm{T}} R_{ij} u_j + a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u^*_j - 2a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u_j, \tag{5.48}$$

$$\frac{2}{(d_i + g_i)} \sum_{j\in N_i} a_{ij}(u^*_i)^{\mathrm{T}} R_i (u_j - u^*_j) = -\sum_{j\in N_i} a_{ij}\nabla\tilde{V}_i^{\mathrm{T}} B(u_j - u^*_j), \tag{5.49}$$

$$a_{ij}(u^*_j)^{\mathrm{T}} R_{ij}(u_j - u^*_j) = a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u_j - a_{ij}(u^*_j)^{\mathrm{T}} R_{ij} u^*_j. \tag{5.50}$$

Combining Equations (5.46)-(5.50), one has

$$\begin{aligned}
&H_{ij}\left(\delta_{ij}, u_i^*, u_{-i}^*, \nabla\tilde{V}_i\right) + (u_i - u_i^*)^{\mathrm{T}} R_i (u_i - u_i^*) + a_{ij}(u_j - u_j^*)^{\mathrm{T}} R_{ij}(u_j - u_j^*) \\
&+ 2a_{ij}(u_j^*)^{\mathrm{T}} R_{ij}(u_j - u_j^*) + \frac{2}{(d_i + g_i)} \sum_{j\in N_i} a_{ij}(u_i^*)^{\mathrm{T}} R_i (u_j - u_j^*) \\
&= \nabla\tilde{V}_i^{\mathrm{T}} A\delta_i + \delta_{ij}^{\mathrm{T}} Q_{ij}\delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j - 2(u_i^*)^{\mathrm{T}} R_i u_i - \sum_{j\in N_i} a_{ij}\nabla\tilde{V}_i^{\mathrm{T}} B u_j \\
&= H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right),
\end{aligned} \tag{5.51}$$

which derives Equation (5.45). □

Theorem 14. *Let Assumptions* 1 *and* 2 *hold. Let* $\tilde{V}_i(\delta_i)$ *have the form* (5.32) *and solve the HJ Equation* (5.31)*. Then the control policy* (5.34) *makes the system* (5.4) *asymptotically stable, i.e., all agents are synchronized to the leader.*

Proof. Because $\tilde{V}_i(\delta_i)$ has the quadratic form $\tilde{V}_i(\delta_i) = \delta_i^{\mathrm{T}} \tilde{P}_i \delta_i$ as shown in Theorem 13 and Remark 10, $\tilde{V}_i(\delta_i) > 0$ is a Lyapunov function candidate. Take derivative of $\tilde{V}_i$ with respect to time t along with the trajectory of the local neighborhood tracking error δ_i, one has

$$\begin{aligned}
\frac{d\tilde{V}_i}{dt} &= \nabla\tilde{V}_i^{\mathrm{T}}\dot{\delta}_i = \nabla\tilde{V}_i^{\mathrm{T}}\left(A\delta_i + (d_i + g_i)Bu_i - \sum_{j\in N_i} a_{ij} B u_j\right) \\
&= H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right) - \left(\delta_{ij}^{\mathrm{T}} Q_{ij}\delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j\right) \\
&= H_{ij}\left(\delta_{ij}, u_i, u_{-i}, \nabla\tilde{V}_i\right) - r_{ij}\left(\delta_i, \delta_j, u_i, u_j\right),
\end{aligned} \tag{5.52}$$

where $r_{ij}\left(\delta_i, \delta_j, u_i, u_j\right) = \delta_{ij}^{\mathrm{T}} Q_{ij}\delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j$.

When $u_i = u_i^*$ and $u_{-i} = u_{-i}^*$, one has $H_{ij}\left(\delta_{ij}, u_i^*, u_{-i}^*, \nabla\tilde{V}_i\right) = 0$, and Equation (5.52) becomes

$$\frac{d\tilde{V}_i}{dt} = -r_{ij}(\delta_i, \delta_j, u_i, u_j) < 0. \tag{5.53}$$

Note that $r_{ij}(\delta_i, \delta_j, u_i, u_j) > 0$ always holds because $\tilde{V}_i(\delta_i) > 0$ always holds according to Theorem 13, and $\tilde{V}_i = \int_t^\infty r_{ij}(\delta_i, \delta_j, u_i, u_j) d\tau$.

As such, $\tilde{V}_i$ is a Lyapunov function, and the system with the optimal control policies u_i^* and u_j^* is asymptotically stable.

□

Theorem 15. *Let Assumptions* 1 *and* 2 *hold. Let* $\tilde{V}_i(\delta_i)$ *take the quadratic form* (5.32) *and solve the HJ Equation* (5.31)*. Then the control policies* $(u_1^*, u_2^*, ..., u_N^*)$ *in* (5.34) *constitute a global Nash equilibrium.*

Proof. Define $J_{ij}(\delta_i, \delta_j, u_i, u_j)$ as the cost between agents i and j, i.e.,

$$J_{ij}(\delta_i, \delta_j, u_i, u_j) = \int_0^\infty \left(\delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) dt,$$

then according to Equation (5.25), the local cost can be written as

$$J_i(\delta_i, \delta_{-i}, u_i, u_j) = \sum_{j \in N_i} J_{ij}(\delta_i, \delta_j, u_i, u_j).$$

Since the system is asymptotically stable, the local neighborhood tracking error $\delta_i(t) \to 0$ when $t \to \infty$. As such, $\tilde{V}_i(\delta_i(\infty)) = 0$, and J_{ij} can be further written as

$$\begin{aligned}
J_{ij}(\delta_i, \delta_j, u_i, u_j) &= \int_0^\infty \left(\delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) dt + \tilde{V}_i(\delta_i(0)) + \int_0^\infty \dot{V}_i dt \\
&= \int_0^\infty \left(\delta_{ij}^{\mathrm{T}} Q_{ij} \delta_{ij} + u_i^{\mathrm{T}} R_i u_i + a_{ij} u_j^{\mathrm{T}} R_{ij} u_j \right) dt + \tilde{V}_i(\delta_i(0)) \\
&+ \int_0^\infty \nabla \tilde{V}_i^{\mathrm{T}} \left(A\delta_i + (d_i + g_i) B u_i - \sum_{j \in N_i} a_{ij} B u_j \right) dt \\
&= \int_0^\infty H_{ij} \left(\delta_{ij}, u_i, u_{-i}, \nabla \tilde{V}_i \right) dt + \tilde{V}_i(\delta_i(0)) \\
&= \int_0^\infty (u_i - u_i^*)^{\mathrm{T}} R_i (u_i - u_i^*) + a_{ij} (u_j - u_j^*)^{\mathrm{T}} R_{ij} (u_j - u_j^*) \\
&\quad + \frac{2}{(d_i + g_i)} \sum_{j \in N_i} a_{ij} (u_i^*)^{\mathrm{T}} R_i (u_j - u_j^*) + 2 a_{ij} (u_j^*)^{\mathrm{T}} R_{ij} (u_j - u_j^*) dt + \tilde{V}_i(\delta_i(0)).
\end{aligned} \tag{5.54}$$

The last equality holds because of Lemma 7.

Now select $u_j = u_j^*$, then

$$J_{ij}(\delta_i, \delta_j, u_i, u_j^*) = \int_0^\infty (u_i - u_i^*)^{\mathrm{T}} R_i (u_i - u_i^*) dt + \tilde{V}_i(\delta_i(0)). \tag{5.55}$$

It is clear that $J_{ij}(\delta_i, \delta_j, u_i^*, u_j^*) \leq J_{ij}(\delta_i, \delta_j, u_i, u_j^*)$ holds for $\forall u_i$, $\forall j \in N_i$, and $i = 1, 2, \cdots, N$, which leads to a global Nash equilibrium. □

5.5 Illustrative Examples

This section develops simulation studies to illustrate the theoretical results developed in this chapter.

5.5.1 Game Settings

Consider a multi-agent system with five agents and one leader connected by a directed graph shown in Figure 5.1. The edge weights a_{ij} are selected as 1 if $(j, i) \in \mathcal{E}$. The leader is pinned to agent 1, i.e., $g_1 = 1$. The Laplacian of this graph is

$$L = \mathcal{D} - \mathcal{A} = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & 0 & 2 & 0 & -1 \\ 0 & -1 & 0 & 2 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{bmatrix}. \tag{5.56}$$

The pinning matrix is

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5.57}$$

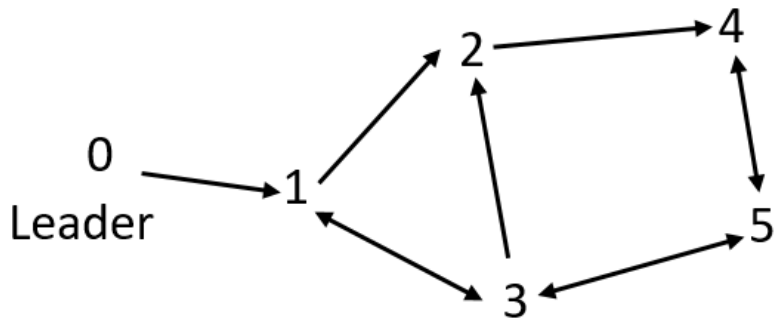


Figure 5.1. The communication graph of five agents and one leader..

The system dynamics for each agent is

$$\dot{x}_i = \begin{bmatrix} \dot{x}_{i,1} \\ \dot{x}_{i,2} \end{bmatrix} = \begin{bmatrix} u_{i,1} \\ u_{i,2} \end{bmatrix}, \tag{5.58}$$

where x_i is a vector of size 2×1, and $i = 1, 2, 3, 4, 5$.

5.5.2 Game Solutions

5.5.2.1 Best responses in existing graphical games

To illustrate the best response strategies for the existing graphical game formulation, we consider the cost function (5.10), where the weighting matrices are selected as

$$R_i = 10I, R_{ij} = 20I, \tag{5.59}$$

and

$$Q_i = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}, \tag{5.60}$$

for all $j \in N_i$ and $i = 1, 2, 3, 4, 5$.

To find P_i using the best response strategies, one needs to solve the coupled HJ Equation (5.16). In particular, to make the HJ equation hold for all δ_i and δ_j,

Equations (5.17)-(5.19) have to hold. Substituting A, B, Q_i, and R_i into Equation (5.17), one can solve P_i as

$$P_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.61}$$

for all $i = 1, 2, 3, 4, 5$.

Substituting P_i, B, and R_i into Equations (5.18) and (5.19), one has

$$\frac{1}{10} \sum_{j \in N_i} a_{ij}(d_j + g_j) = 0, \tag{5.62}$$

$$\frac{1}{5} \sum_{j \in N_i} a_{ij}(d_j + g_j)^2 = 0, \tag{5.63}$$

which clearly do not hold, because $a_{ij} > 0$, $d_j > 0$, and $g_j \geqslant 0$ for all $j \in N_i$.

There exists no distributed value function of the form (5.12) that solves the HJ Equation (5.16). In other words, the best response strategies generally cannot find a distributed solution that can constitute global Nash equilibrium for the existing differential graphical game formulation.

5.5.2.2 Minmax strategies

To use minmax strategies in a graphical game, the cost function is required to be modified to formulate an adversarial form between agent i and its neighbors [19]. To this end, define the cost function as

$$J_i = \int_0^\infty (\delta_i^{\mathrm{T}} Q_i \delta_i + (d_i + g_i) u_i^{\mathrm{T}} R_i u_i - \gamma^2 \sum_{j \in N_i} a_{ij} u_j^{\mathrm{T}} R_{ij} u_j) dt. \tag{5.64}$$

Note that the cost function (5.64) is different from the cost function in existing graphical games (5.10). It is because the minmax strategies assume worst-case neighbors' control policies, which lead to an adversarial cost function naturally. This modification ensures the asymptotically stability of the derived control policies and

good robustness performances. Please refer to [19] for more information about the minmax strategies in differential graphical games.

For the parameters in the cost function (5.64), we select $\gamma = 2$, and the weighting matrices R_i, R_{ij}, and Q_i are selected with the same values of (5.59)-(5.60).

The Hamiltonian associated with this cost function is

$$H_i = \delta_i^{\mathrm{T}} Q_i \delta_i + (d_i + g_i) u_i^{\mathrm{T}} R_i u_i - \gamma^2 \sum_{j \in N_i} a_{ij} u_j^{\mathrm{T}} R_j u_j + 2\delta_i^{\mathrm{T}} P_i \left(A\delta_i + (d_i + g_i) B u_i - \sum_{j \in N_i} a_{ij} B u_j \right). \tag{5.65}$$

Using the minmax strategies (5.20), the optimal control policy for agent i can be obtained by letting $\frac{\partial H_i}{\partial u_i} = \mathbf{0}$ as

$$u_i^+ = -R_i^{-1} B^{\mathrm{T}} P_i \delta_i. \tag{5.66}$$

Similarly, the worst-case policy of agent i's neighbors is

$$v_j^+ = -\frac{1}{\gamma^2} R_j^{-1} B^{\mathrm{T}} P_i \delta_i. \tag{5.67}$$

With above control policies, the HJ equation can be simplified to the following Riccati equation,

$$Q_i + P_i A + A^{\mathrm{T}} P_i - (d_i + g_i) P_i B R_i^{-1} B^{\mathrm{T}} P_i + \frac{1}{\gamma^2} \sum_{j=1}^{N} a_{ij} P_i B R_j^{-1} B^{\mathrm{T}} P_i = \mathbf{0}. \tag{5.68}$$

Substituting A, B, Q_i, R_i, and γ into Equation (5.68), one can solve P_i as

$$P_1 = \begin{bmatrix} 1.633 & 0 \\ 0 & 1.633 \end{bmatrix}, \tag{5.69}$$

and

$$P_i = \begin{bmatrix} 1.512 & 0 \\ 0 & 1.512 \end{bmatrix}, \tag{5.70}$$

for $i = 2, 3, 4, 5$.

It can be seen from Equation (5.66) that the solution derived from minmax strategies is a distributed solution, since the control policy of agent i only depends on the local neighborhood tracking error δ_i. However, this solution is not Nash because it is derived by assuming the worst-case policy of its neighbors, i.e., Equation (5.67), which is not the real policies of its neighbors, i.e., $u_j^* = -R_j^{-1} B^{\mathrm{T}} P_j \delta_j$, where $j \in N_i$.

5.5.2.3 Solution of the proposed novel graphical game

For the proposed cost function shown in Equation (5.25), the weighting matrices Q_i R_i, and R_{ij} are selected with the same values of (5.59) and (5.60).

Substituting A, B, Q_i, and R_i into Equation (5.40), the weighting matrix $\tilde{P}_i$ in the value function $\tilde{V}_i(\delta_i) = \delta_i^{\mathrm{T}} \tilde{P}_i \delta_i$ can be found as

$$\tilde{P}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.71}$$

for all $i = 1, 2, 3, 4, 5$.

The weighting matrices $\tilde{Q}_{ij}$ and $\hat{Q}_{ij}$ are then computed by substituting $\tilde{P}_i$ into Equations (5.38) and (5.39) as

$$\tilde{Q}_{ij} = \begin{bmatrix} -0.2 & 0 \\ 0 & -0.2 \end{bmatrix}, \tag{5.72}$$

for $i = 1$ and $j \in N_i$,

$$\tilde{Q}_{ij} = \begin{bmatrix} -0.4 & 0 \\ 0 & -0.4 \end{bmatrix}, \tag{5.73}$$

for $i = 2, 3, 4, 5$ and $j \in N_i$, and

$$\hat{Q}_{ij} = \begin{bmatrix} -0.8 & 0 \\ 0 & -0.8 \end{bmatrix}, \tag{5.74}$$

for all $i = 1, 2, 3, 4, 5$ and $j \in N_i$.

As such, the matrix Q_{ij} in the cost function (5.25) is $Q_{ij} = \left[Q_i, \tilde{Q}_{ij}; \tilde{Q}_{ij}^{\mathrm{T}}, \hat{Q}_{ij}\right]$, where Q_i, $\tilde{Q}_{ij}$, and $\hat{Q}_{ij}$ take the values of (5.60), (5.72)-(5.74), respectively. Note that the matrix Q_i can be selected arbitrarily, as long as it is a positive semi-definite matrix. The matrices $\tilde{Q}_{ij}$ and $\hat{Q}_{ij}$ are computed via Equations (5.38) and (5.39), by first computing $\tilde{P}_i$ from (5.40).

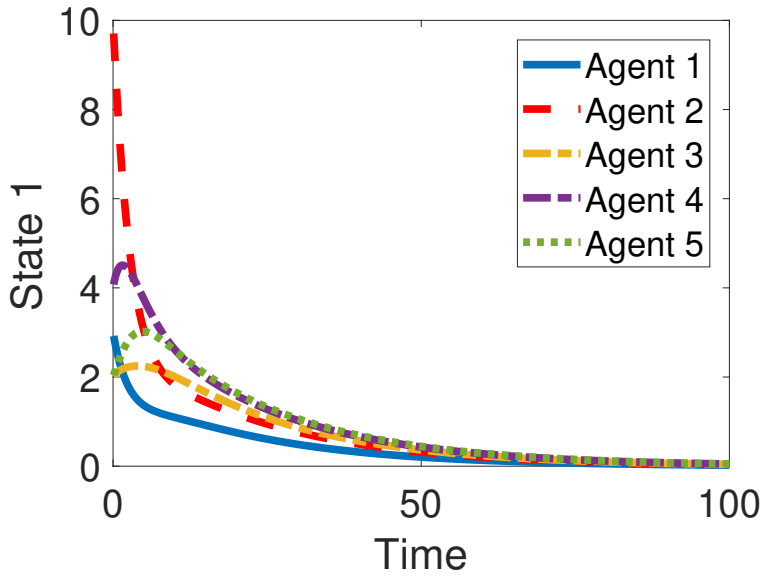
Substituting $\tilde{P}_i$ and B into Equation (5.34), the optimal control policy for agent i can be found as

$$u_i^* = -(d_i + g_i) R_i^{-1} \delta_i. \tag{5.75}$$

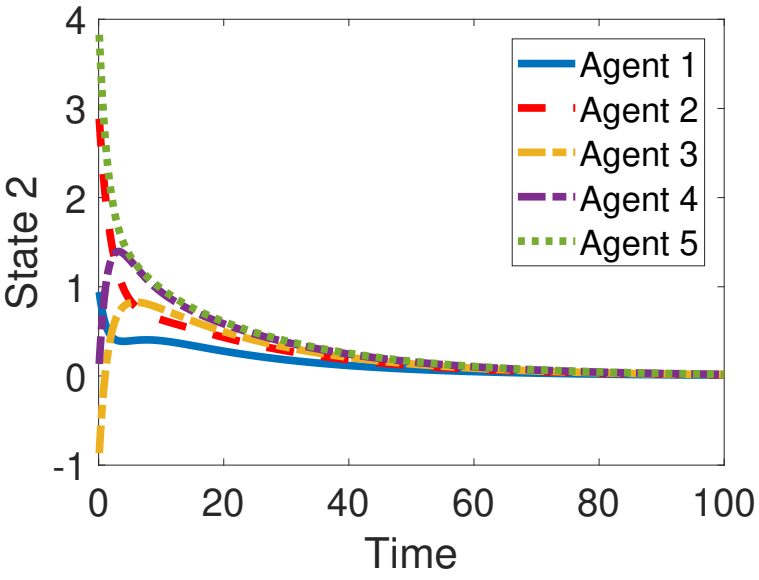
It can be seen that the control policy of agent i is only related to its local neighboring tracking error δ_i, which verifies the distributed properties of the proposed solution.

Substituting $\tilde{P}_i$, $\tilde{Q}_{ij}$, $\hat{Q}_{ij}$, R_j and R_{ij} into Equation (5.33), the HJ equation can be verified to hold for all possible δ_i and δ_j, which indicates the Nash equilibrium.

The state evolution of the five agents with the optimal policies are shown in Figure 5.2. The initial conditions are selected as $x_0 = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$, $x_1 = \left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right]$, $x_2 = \left[\begin{smallmatrix} 10 \\ 3 \end{smallmatrix}\right]$, $x_3 = \left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right]$, $x_4 = \left[\begin{smallmatrix} 4 \\ 0 \end{smallmatrix}\right]$, and $x_5 = \left[\begin{smallmatrix} 2 \\ 4 \end{smallmatrix}\right]$ respectively. It can be seen that the consensus is achieved with sufficiently long simulation time.



(a)



(b)

Figure 5.2. Evolution of (a) state 1, and (b) state 2, for all five agents. Consensus is achieved after long enough time..

CHAPTER 6

On the Robustness of Networked Cooperative Tracking Systems

6.1 Introduction

Stability margins, i.e., gain margin and phase margin, describe the ability of a control system to maintain stability in the presence of perturbation, and have been adopted as the measures for robustness for decades [94]. Studies of stability margins largely focus on single-agent systems, including both single-input single-output (SISO) systems and multi-input multi-output (MIMO) systems. For SISO systems, the scalar Nyquist approach and Bode analysis have been developed to find phase and gain margins [95, 96]. Since the 1970s, a number of attempts have been directed to extend the robustness analysis from SISO systems to MIMO systems [97–104]. In a very first effort of this direction, paper [101] introduced the concept of multiloop robustness subject to simultaneous phase and gain perturbation in multiple loops, and showed that the LQR possesses $\pm 60^\circ$ phase margin, 50% gain reduction, and infinite gain margin, following a Lypunov type of analysis. From the viewpoint of system transfer function matrix, a generalization of the classical scalar Nyquist approach and Bode analysis to MIMO systems was investigated in [102], by exploiting the characteristics of singular values, single vectors, and the spectral norm of the closed-loop system transfer matrix. Based on the singular value analysis, the μ-analysis framework was then established, with the purpose of bounding the stability margins of diagonally perturbed MIMO systems [103–108]. However, all of the aforementioned studies assume a single-agent system, which is limited in scope considering the many networked real-world system applications. The stability margin

analysis for networked MAS is challenging considering the complexity caused by the interplay of communication structure and agent dynamics. In this chapter, we develop a framework to analyze the phase and gain margins of networked MAS, which is a first attempt in the literature per knowledge of the authors.

Networked MAS have attracted extensive attention due to their wide applications in mobile robots, UAVs, sensor networks, and satellite formation [109–112]. In general, networked MAS can be classified into two categories: leaderless consensus systems and leader follower tracking systems, depending on whether a leader exists or not [113, 114]. For the leaderless consensus problem, or commonly referred as the *cooperative regulator problem*, distributed controllers have been designed for agents to achieve consensus by utilizing the information received from their immediate neighbors in the communication network [115–118]. Consensus value is usually a function of agents' initial states dependent on network topology and agent dynamics. For the leader follower consensus problem, or called *cooperative tracking problem*, a leader communicates to at least one agent, and all agents are controlled to synchronize to a desired trajectory generated by the leader [113, 114, 119, 120]. Optimal controller design for cooperative tracking systems has been studied in [114, 115, 119]. In particular, reference [119] developed a local LQR design for agents with identical linear time-invariant dynamics, and showed that the local LQR design guarantees unbounded synchronization regions on arbitrary digraphs containing a spanning tree. Reference [114] developed an optimality criterion that promises the existence of a global optimal controller under certain conditions by the inverse optimality method. Although some properties of the cooperative tracking systems, e.g., optimality and stability, have been studied in the aforementioned works, the analysis of robustness in the presence of perturbation is still missing. In addition, the effects of communication graph topology on robustness properties also remain to be investigated.

This chapter studies the robustness properties of networked cooperative tracking systems using the Lypunov analysis and the algebraic graph theory. The contributions of this chapter are fourfold. First, phase and gain margins of networked cooperative tracking systems are derived in closed form, by analyzing the stability conditions of perturbed systems. Second, graph topology characteristics relating to stability margins are developed, through an eigen-analysis. Third, the upper bounds of phase and gain margins for MAS of general communication graph topology are obtained, by integrating the robustness analysis with the graph topology analysis. Fourth, we prove that the directed tree topology is the most robust among all possible communication graph topology, in the sense that they have the same guaranteed gain and phase margin as those of single-agent LQR systems.

This chapter is organized as follows. Section 6.2 introduces notations and basic definitions. Section 6.3 formulates the cooperative tracking problem, including the system dynamics and local LQR design. Section 6.4 investigates phase and gain margins of the cooperative tracking system. Section 6.5 analyzes the graph topology characteristics relating to stability margins, and finds the upper bounds of stability margins. Section 6.6 conducts simulation studies to validate the results.

6.2 Notations and Definitions

We introduce the following notations and definitions to facilitate the analysis in this chapter [101, 121].

1) The space $\mathcal{L}_2^n$ is defined as the set of all piecewise continuous functions $x : [0, \infty) \to \mathbb{R}^n$ such that

$$\|x\|_{\mathcal{L}_2} = \left(\int_0^\infty x^T(t)x(t)dt \right)^{\frac{1}{2}} < \infty, \tag{6.1}$$

i.e., the space $\mathcal{L}_2^n$ defines the set of all square-integrable function $x(t)$.

2) The extension $\mathcal{L}_{2e}^n$ of $\mathcal{L}_2^n$ is defined by

$$\mathcal{L}_{2e}^n = \{x | x_\tau \in \mathcal{L}_2^n, \forall \tau \geqslant 0\}, \tag{6.2}$$

where $x_\tau(t)$ is a truncation of $x(t)$ defined by

$$x_\tau(t) = \begin{cases} x(t) & 0 \leq t \leq \tau, \\ 0 & t > \tau. \end{cases} \tag{6.3}$$

3) Define the inner-product $\langle x, y \rangle$ for piecewise continuous functions $x(t) \in \mathbb{R}^n$ and $y(t) \in \mathbb{R}^n$ as

$$\langle x, y \rangle = \int_0^\infty x^T(t) y(t) dt. \tag{6.4}$$

3) The term *operator* is reserved for the mapping from functions into functions. For example, a dynamic system may be viewed as an operator that maps input time functions into output time functions.

4) An operator P with $P\mathbf{0} = \mathbf{0}$, where $\mathbf{0}$ is a zero matrix, is said to have finite gain if there exists a constant $k < \infty$ such that

$$\|Px\| < k\|x\| \tag{6.5}$$

for all square-integrable x.

5) For $\lambda \in \mathbb{C}$, we use $Re\{\lambda\}$ to represent the real part of λ.

6) A^* denotes the adjoint of the matrix A, i.e., the complex-conjugate of A^T.

6.3 Cooperative Tracking Systems and Problem Formulation

This section describes the cooperative tracking problem. Agent dynamics are governed by identical continuous linear time-invariant systems defined upon a communication graph. The goal of each agent is to synchronize to the leader dynamics by using the state information of itself and its immediate neighbors.

6.3.1 Communication Graph

Consider a group of N agents connected by a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here $\mathcal{V}$ is the set of agents, $\mathcal{V} = 1, 2, \cdots, N$, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges. The graph adjacency matrix is denoted as $\mathcal{A} = [a_{ij}]$, where $a_{ij} > 0$ if $(j, i) \in \mathcal{E}$, and $a_{ij} = 0$ if $(j, i) \notin \mathcal{E}$. $(j, i) \in \mathcal{E}$ means there exists an edge starting from j to i in the directed graph. It is assumed that the graph is simple, i.e., there is no repeated edge or self-loop (i.e., $(i, i) \notin \mathcal{E}\ \forall i$). Denote the set of neighbors of agent i as N_i, i.e., $N_i = \{j : (j, i) \in \mathcal{E}\}$. Denote the in-degree matrix as $\mathcal{D}$, i.e., $\mathcal{D} = diag(d_1, d_2, \cdots, d_N)$, where d_i is the i^{th} row sum of $\mathcal{A}$: $d_i = \sum_j a_{ij}$. Define the graph Laplacian matrix as L, $L = \mathcal{D} - \mathcal{A}$, which has all row sums equal to zero.

The communication between the leader and agent i is captured by the pining gain $g_i \geqslant 0$. $g_i > 0$ means that the leader node can be observed by agent i. Denote the pining matrix as G, then $G = diag\{g_i\} \in \mathbb{R}^{N \times N}$.

The graph $\mathcal{G}$ contains a spanning tree and the root node can observe information from the leader node.

Assumption 6.3.1 is a necessary condition for solving the cooperative tracking problem [120]. Synchronization to the leader dynamics cannot be achieved without it.

6.3.2 Agents' Dynamics

Consider a group of N agents with identical linear dynamics,

$$\dot{x}_i = Ax_i + Bu_i, \tag{6.6}$$

where $x_i \in \mathbb{R}^n$ is the state vector, $u_i \in \mathbb{R}^m$ is the control input vector, and $i = 1, 2, \cdots, N$. A and B are the drift and input matrices, respectively.

The pair (A, B) is controllable.

The dynamics of the leader, indexed with 0, is given by

$$\dot{x}_0 = Ax_0, \tag{6.7}$$

where $x_0 \in \mathbb{R}^n$ is the state. The leader generates the desired trajectory to which the agents should synchronize. The trajectory of the leader can only be observed by a small set of agents, which is described by the pinning matrix G. Note that the leader dynamics is not required to be stable, i.e., the matrix A can be stable, marginally stable, or even unstable.

6.3.3 Cooperative Tracking Control

The objective of the cooperative tracking problem is to design local distributed controllers u_i, where $i = 1, 2, \cdots, N$, such that $\lim_{t\to\infty}(x_i(t) - x_0(t)) = \mathbf{0}$ for all $i = 1, 2, \cdots, N$, i.e., all agents synchronize to the leader dynamics [119, 120]. Define the synchronization error of agent i as

$$\delta_i = x_i - x_0. \tag{6.8}$$

Then the global synchronization error is

$$\delta = x - \underline{x_0}, \tag{6.9}$$

where x is the global state, $x = [x_1^T, x_2^T, \cdots, x_N^T]^T \in \mathbb{R}^{nN}$, and $\underline{x_0} = \mathbf{1}_N \otimes x_0 \in \mathbb{R}^{nN}$. $\mathbf{1}_N$ is an N-vector of ones, and $\otimes$ is the Kronecker product. The synchronization is achieved if $\lim_{t\to\infty}\delta(t) = \mathbf{0}$.

Define the neighborhood synchronization error for agent i as

$$\varepsilon_i = \sum_{j\in N_i} a_{ij}(x_j - x_i) + g_i(x_0 - x_i). \tag{6.10}$$

The feedback controller for agent i is

$$u_i = K\varepsilon_i, \tag{6.11}$$

where $K \in \mathbb{R}^{m\times n}$ is the feedback control gain. The control protocol (6.11) is distributed in the sense that the control of agent i, i.e., u_i, depends only on its local tracking error ε_i. In other words, each agent only utilizes the state information of its own and its neighbors to synchronize to the leader dynamics.

The overall global closed-loop dynamics [119] is

$$\dot{x} = (I_N \otimes A - (L+G) \otimes BK)\, x + ((L+G) \otimes BK)\, \underline{x_0} \tag{6.12}$$

where I_N is the identity matrix, $I_N \in \mathbb{R}^{N\times N}$. Let $A_c = I_N \otimes A - (L+G) \otimes BK$, and $B_c = (L+G) \otimes BK$, then (6.12) becomes

$$\dot{x} = A_c x + B_c \underline{x_0}, \tag{6.13}$$

from which the global synchronization error dynamics can be derived as

$$\begin{aligned} \dot{\delta} =& \dot{x} - \underline{\dot{x}_0} \\ =& (I_N \otimes A - (L+G) \otimes BK)\, \delta \\ =& A_c \delta. \end{aligned} \tag{6.14}$$

It can be seen from (6.12) and (6.14) that the dynamics of both global state and global synchronization error depend on the graph structure, i.e., $(L+G)$. This implies that even if the local systems (6.6) and (6.7) can be stable for all $i = 1, 2, \cdots, N$, the global system state x and synchronization error δ may still be unstable.

6.3.4 Local LQR Design

Consider the following feedback gain for each agent from the local LQR design

$$K = R^{-1} B^T P. \tag{6.15}$$

Here, P is the positive definite solution of the control algebratic Riccati equation (ARE),

$$A^T P + PA + Q - PBR^{-1}B^T P = \mathbf{0}, \tag{6.16}$$

where $Q = Q^T \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix and $R = R^T \in \mathbb{R}^{m\times m}$ is a positive definite matrix of a diagonal form. This feedback control gain K is the local optimal design for the single-agent LQR problem with the following cost function

$$J_i = \frac{1}{2}\int_0^\infty \left(x_i^T Q x_i + u_i^T R u_i\right) dt \tag{6.17}$$

subject to the local agent dynamics (6.6) and the state feedback controller $u_i = -Kx_i$. For multi-agent systems, it has been shown that the control protocol (6.15) is also an optimal solution with respect to certain global quadratic performance indices $J = \int_0^\infty \left(\delta^T \bar{Q}\delta + u^T \bar{R} u\right) dt$ (see [114, Theorem 1] for the detailed descriptions).

6.3.5 Cooperative Tracking Systems with Perturbation

In order to analyze the robustness of cooperative tracking systems in the presence of perturbation, we consider the following perturbed systems [101].

$$\dot{\hat{x}}_i = A\hat{x}_i + BP\hat{u}_i, \tag{6.18}$$

where

$$\hat{u}_i = K\hat{\varepsilon}_i, \tag{6.19}$$

$$\hat{\varepsilon}_i = \sum_{j\in N_i} a_{ij}(\hat{x}_i - \hat{x}_j) + g_i(\hat{x}_i - x_0), \tag{6.20}$$

and $\hat{x}_i$, $\hat{u}_i$, and $\hat{\varepsilon}$ are the perturbed state, control, and local tracking error for agent i, respectively. The perturbation P is a finite-gain operator that takes a diagonal form, such that the perturbation in the feedback loops are noninteracting, i.e.,

$$Pu_i = \begin{bmatrix} P_1 u_{i,1} \\ P_2 u_{i,2} \\ \vdots \\ P_m u_{i,m} \end{bmatrix}. \tag{6.21}$$

The global synchronization error of the perturbed system is

$$\begin{aligned}\dot{\hat{\delta}} =& \dot{\hat{x}} - \underline{\dot{x}_0} \\ =& I_N \otimes A - (L+G) \otimes BPK\hat{\delta} \\ =& \hat{A}_c\hat{\delta},\end{aligned} \tag{6.22}$$

where $\hat{A}_c = I_N \otimes A - (L+G) \otimes BPK$, and $\hat{\delta}(0) = \delta(0)$.

Figure 6.1 shows an example of the perturbed networked MAS with local LQR design. Four agents are connected by a directed communication graph shown in Figure 6.3(a). The Laplacian matrix L and pining gain matrix G for this example are shown in (6.61) and (6.63), respectively.

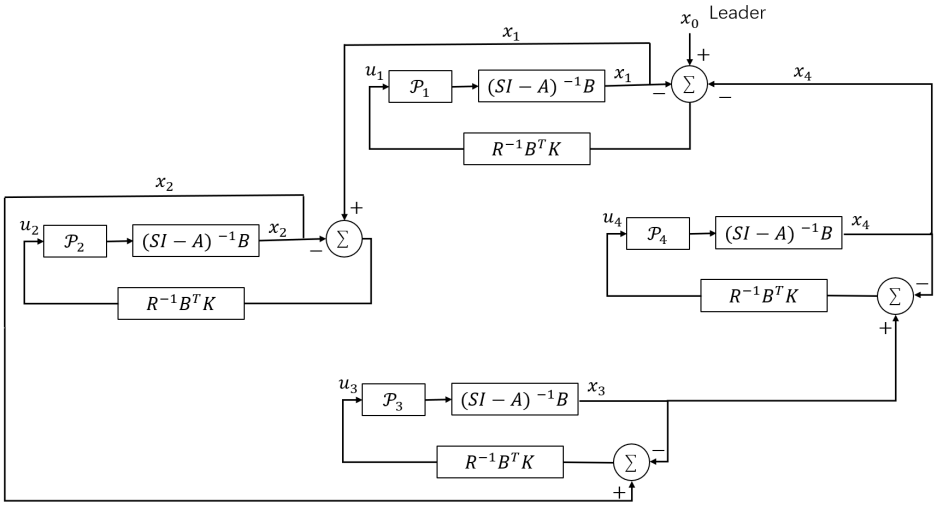


Figure 6.1. An example of perturbed networked MAS with local LQR design.

6.3.6 Problem Formulation

Consider the cooperative tracking problem with the local LQR design described in (6.6)-(6.16). The goal of this chapter is to analyze the robustness performances of the cooperative tracking system, by analyzing the stability conditions of the perturbed

system (6.22) (see Figure 6.1). It is well-known that for a single-agent system, the local LQR design guarantees a $\pm 60^\circ$ phase margin, 50% gain reduction, and infinite gain margin [101, 122]. For networked multi-agent systems, we will show that the phase and gain margins are functions of communication topology.

In particular, we study the following three problems.

1. *Find the phase and gain margins of the networked cooperative tracking system* (6.6)-(6.16)*;*
2. *Analyze the effects of communication graph topology on stability margins;*
3. *Find communication graph topology that promises the best guaranteed stability margins.*

6.4 Robustness of Cooperative Tracking Systems

This section studies the robustness of cooperative tracking systems through investigating stability conditions of the perturbed systems.

Denote the eigenvalues of $L + G$ as λ_i $(i = 1, 2, \cdots, N)$, then the following lemma holds.

Lemma 8. *[120, Lemma 3.3] Under Assumption 6.3.1, the matrix $L + G$ is nonsingular. Moreover, the eigenvalues λ_i satisfy $\lambda_i > 0$, for all $i = 1, 2, \cdots, N$.*

In the next theorem, we provide a necessary and sufficient condition for the stability of the perturbed global synchronization error $\hat{\delta}$.

Theorem 16. *The global synchronization error of the perturbed system* (6.22) *is asymptotically stable if and only if the matrices*

$$A - \lambda_i BPK, \tag{6.23}$$

are asymptotically stable for all $i = 1, 2, \cdots, N$.

Proof. The system (6.22) is asymptotically stable if and only if the matrix $\hat{A}_c$ is Hurwitz, i.e., all eigenvalues of $\hat{A}_c$ have negative real parts. To find the eigenvalues of $\hat{A}_c$, we first put the matrix $L+G$ in its Jordan canonical form through a similarity transformation, i.e.,

$$S^{-1}(L+G)S = J = \begin{bmatrix} J_{n_1}(\bar{\lambda}_1) & & & \\ & J_{n_2}(\bar{\lambda}_2) & & \\ & & \ddots & \\ & & & J_{n_k}(\bar{\lambda}_k) \end{bmatrix}, \tag{6.24}$$

where $S \in \mathbb{R}^{N\times N}$ is a nonsingular matrix, $\sum_{j=1}^{k} n_j = N$ and J_{n_j} is the Jodan block of size n_j, $j = 1, \cdots, k$. Note that the eigenvalues $\bar{\lambda}_j$ are distinct for different j.

Similarly, the matrix $\hat{A}_c$ can be transformed to a block triangular matrix $\bar{A}_c$ through the following transformation,

$$\begin{aligned} \bar{A}_c &= (S \otimes I_n)^{-1} \hat{A}_c (S \otimes I_n) \\ &= (S \otimes I_n)^{-1} \left(I_N \otimes A - (L+G) \otimes BPK\right) (S \otimes I_n) \\ &= I_N \otimes A - J \otimes BPK, \end{aligned} \tag{6.25}$$

where $I_n \in \mathbb{R}^{n\times n}$ is an identity matrix. The last equality holds because of the following two equations (6.26)-(6.27).

$$\begin{aligned} &(S \otimes I_n)^{-1} \left(I_N \otimes A\right) (S \otimes I_n) \\ &= (S^{-1} \otimes I_n^{-1}) \left(I_N \otimes A\right) (S \otimes I_n) \\ &= (S^{-1} I_N \otimes I_n^{-1} A)(S \otimes I_n) \\ &= (S^{-1} I_N S) \otimes (I_n^{-1} A I_n) \\ &= I_N \otimes A, \end{aligned} \tag{6.26}$$

$$
\begin{aligned}
&(S \otimes I_n)^{-1}((L+G) \otimes (BPK))(S \otimes I_n) \\
&= (S^{-1} \otimes I_n^{-1})((L+G)S) \otimes (BPKI_n)) \\
&= (S^{-1}(L+G)S) \otimes I_n^{-1}BPKI_n) \\
&= (S^{-1}(L+G)S) \otimes (BPK) \\
&= J \otimes BPK.
\end{aligned} \tag{6.27}
$$

For the block triangular matrix $\hat{A}_c$, its eigenvalues are the union of eigenvalues in the diagonal blocks. According to (6.25), the diagonal blocks of $\bar{A}_c$ are $A-\lambda_i BPK$, where $i = 1, 2, \cdots, N$. As such, the matrix $\hat{A}_c$ is asymptotically stable if and only if $A - \lambda_i BPK$ are asymptotically stable for all $i = 1, 2, \cdots, N$. □

Theorem 16 shows that the stability of the global synchronization error of the perturbed system (6.22), i.e., $\hat{\delta}$, is determined by the matrices $A - \lambda_i BPK$, $i = 1, 2, \cdots, N$. Different from single-agent systems whose stability is decided uniquely by the matrix $A - BK$, the stability of the cooperative tracking system depends on not only the matrices A, B and K, but also the communication graph topology $L+G$.

In the next theorem, we show that the stability of the matrices $A - \lambda_i BPK$ depends only on the real parts of λ_i, i.e., $A - Re\{\lambda_i\}BPK$, where $i = 1, 2, \cdots, N$.

Theorem 17. *The global synchronization error of the perturbed system* (6.22) *is asymptotically stable if and only if the following systems*

$$
\dot{\zeta}_i = (A - Re\{\lambda_i\}BPK)\,\zeta_i, \tag{6.28}
$$

are asymptotically stable for all $i = 1, 2, \cdots, N$.

Proof. To prove this theorem, we need to show that $A - \lambda_i BPK$ are asymptotically stable if and only if $A - Re\{\lambda_i\}BPK$ are asymptotically stable, where $i = 1, 2, \cdots, N$.

Let $\lambda_i = \alpha + j\beta$ and $\tilde{A} = A - Re\{\lambda_i\}BPK$, then one has

$$\begin{aligned} A - \lambda_i BPK &= A - (\alpha + j\beta)BPK \\ &= \tilde{A} - j\beta BPK. \end{aligned} \tag{6.29}$$

As such, the following equation holds.

$$\begin{aligned} &P(A - \lambda_i BPK) + (A - \lambda_i BPK)^* P \\ &= P\tilde{A} - j\beta PBPK + \tilde{A}^T P + j\beta K^T PB^T P. \end{aligned} \tag{6.30}$$

Substituting (6.15) into (6.30), we have

$$\begin{aligned} &P(A - \lambda_i BPK) + (A - \lambda_i BPK)^* P \\ &= P\tilde{A} + \tilde{A}^T P - j\beta PBPR^{-1}B^T P + j\beta PBR^{-1}PB^T P \\ &= P\tilde{A} + \tilde{A}^T P. \end{aligned} \tag{6.31}$$

The last equality holds because the matrices P and R are both diagonal, which leads to $PR^{-1} = R^{-1}P$.

According to the Lyapunov theory, a matrix $\tilde{A}$ is asymptotically stable if and if there exists a positive definite matrix P such that $P\tilde{A} + \tilde{A}^* P$ is negative definite. As such, according to (6.31), an asymptotically stable matrix $\tilde{A}$ indicates that $A - \lambda_i BPK$ is also asymptotically stable, and vice versa. The proof is complete. □

From Theorem 17, the stability of the global synchronization error of the perturbed system (6.22), i.e., $\hat{\delta}$, can be determined by checking the stability of the local systems (6.28).

6.4.1 Phase and Gain Margins of the Cooperative Tracking System

In this subsection we first find conditions of the perturbation P that guarantees the stability of ζ_i. The phase and gain margin analysis of the cooperative tracking system (6.6)-(6.16) then follows.

Theorem 18. *Consider the cooperative tracking system* (6.6)-(6.16). *If the perturbation P satisfies the following inequality*

$$\langle \bar{u}_i, (2Re\{\lambda_i\}P - I)R^{-1}\bar{u}_i\rangle \geqslant 0 \tag{6.32}$$

for all $\bar{u}_i \in \mathbb{R}^m$ and $i = 1, 2, \cdots, N$, then

1) the following inequality holds,

$$\zeta_i^T(0)P\zeta_i(0) \geqslant \langle \zeta_i, Q\zeta_i\rangle; \tag{6.33}$$

2) if additionally, $[Q^{\frac{1}{2}}, A]$ is detectable, then the systems ζ_i in (6.28) *are asymptotically stable.*

Proof. Denote $\zeta_{i\tau}$ as a truncation of ζ_i, i.e.,

$$\zeta_{i\tau}(t) = \begin{cases} \zeta_i(t) & 0 \leq t \leq \tau, \\ 0 & t > \tau. \end{cases} \tag{6.34}$$

Combining (6.28) and the feedback gain K in (6.15), one has

$$\begin{aligned}
&\zeta_i^T(0)P\zeta_i(0) \\
&= \zeta_i^T(\tau)P\zeta_i(\tau) - \int_0^\tau \frac{d}{dt}\left(\zeta_i^T(t)P\zeta_i(t)\right)dt \\
&= \zeta_i^T(\tau)P\zeta_i(\tau) - \int_0^\tau 2\zeta_i^T(t)P\dot{\zeta}_i(t)dt \\
&\geqslant -\int_0^\tau 2\zeta_i^T(t)P\dot{\zeta}_i(t)dt \\
&= -\int_0^\tau 2\zeta_i^T(t)P\left(A - Re\{\lambda_i\}BPK\right)\zeta_i(t))dt \\
&= -2\langle \zeta_{i\tau}, P(A - Re\{\lambda_i\}BPR^{-1}B^TP)\zeta_{i\tau}\rangle \\
&= \langle \zeta_{i\tau}, (Q - PBR^{-1}B^TP + 2Re\{\lambda_i\}PBPR^{-1}B^TP)\zeta_{i\tau}\rangle \\
&= \langle \zeta_{i\tau}, Q\zeta_{i\tau}\rangle + \langle \zeta_{i\tau}, \left(PB(2Re\{\lambda_i\}P - I)R^{-1}B^TP\right)\zeta_{i\tau}\rangle.
\end{aligned} \tag{6.35}$$

Let $\Pi_i = (2Re\{\lambda_i\}P - I)\, R^{-1}$ and $\bar{u}_i = B^T P \zeta_{i\tau}$, one has

$$\begin{aligned} &\zeta_i^T(0)P\zeta_i(0) - \langle \zeta_{i\tau}, Q\zeta_{i\tau}\rangle \\ &\geqslant \langle \zeta_{i\tau}, PB\Pi_i B^T P\zeta_{i\tau}\rangle \\ &= \langle B^T P\zeta_{i\tau}, \Pi_i B^T P\zeta_{i\tau}\rangle \\ &= \langle \bar{u}_i, \Pi_i \bar{u}_i\rangle. \end{aligned} \tag{6.36}$$

If P satisfies (6.32), then the following inequality holds according to (6.36),

$$\zeta_i^T(0)P\zeta_i(0) \geqslant \langle \zeta_{i\tau}, Q\zeta_{i\tau}\rangle.$$

Taking the limit $\tau \to \infty$, then the first statement in (6.33) follows.

Note that $\zeta_i^T(0)P\zeta_i(0) \geqslant \langle \zeta_i, Q\zeta_i\rangle$ implies that $\langle \zeta_i, Q\zeta_i\rangle$ is bounded. If additionally, $[Q^{\frac{1}{2}}, A]$ is detectable, then ζ_i is square-integrable [101]. Because P has a finite gain and ζ_i is square-integrable, $\dot{\zeta}_i$ is also square-integrable. Since both ζ_i and $\dot{\zeta}_i$ are square-integrable, ζ_i is asymptotically stable [101], which proves the second statement. □

The following theorem dereives the condition on the perturbation P in the frequency domain, for the case when P is a linear operator.

Theorem 19. *Let the perturbation P be a linear time-invariant operator H with a finite-gain and a proper transfer function $H(j\omega)$. If*

$$2Re\{\lambda_i\}H(j\omega)R^{-1} + 2Re\{\lambda_i\}R^{-1}H^*(j\omega) - R^{-1} \geqslant 0 \tag{6.37}$$

holds for all $i = 1, 2, \cdots, N$ and ω, and $[Q^{\frac{1}{2}}, A]$ is detectable, then the systems ζ_i in (6.28) *are asymptotically stable.*

Proof. From (6.37) and the Parseval's theorem [123], we have

$$\begin{aligned}
&\langle \bar{u}_i, (2Re\{\lambda_i\}P - I)\, R^{-1}\bar{u}_i\rangle \\
&= \frac{1}{2}\Big(\langle \bar{u}_i, (2Re\{\lambda_i\}P - I)\, R^{-1}\bar{u}_i\rangle + \langle (2Re\{\lambda_i\}P - I)\, R^{-1}\bar{u}_i, \bar{u}_i\rangle\Big) \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \bar{U}_i^*(j\omega)\Big(Re\{\lambda_i\}\big(H(j\omega)R^{-1} + R^{-1}H^*(j\omega)\big) - R^{-1}\Big)\bar{U}_i(j\omega)d\omega \\
&\geqslant 0,
\end{aligned} \tag{6.38}$$

where $\bar{U}_i(j\omega)$ is the Fourier transform of $\bar{u}_i$, $\bar{U}_i^*(j\omega)$ is the Hermitian of $\bar{U}_i(j\omega)$, and $H^*(j\omega)$ is the Hermitian of $H(j\omega)$.

Since $\langle \bar{u}_i, (2Re\{\lambda_i\}P - I)\, R^{-1}\bar{u}_i\rangle \geqslant 0$, we know that the systems in (6.28) are asymptotically stable from Theorem 18. □

If each element of the perturbation, P_l, $l = 1, 2, \cdots, m$, is linear time-invariant with proper transfer function $H_l(j\omega)$, and

$$Re\{H_l(j\omega)\} \geqslant \frac{1}{2Re\{\lambda_i\}} \tag{6.39}$$

holds for all $i = 1, 2, \cdots, N$, then the systems ζ_i in (6.28) are asymptotically stable.

Proof. Taking $H(j\omega) = \mathrm{diag}(H_1(j\omega), H_2(j\omega), \cdots, H_m(j\omega))$, one has

$$\begin{aligned}
&2Re\{\lambda_i\}\left(r_l^{-1}\left(H_l(j\omega) + H_l^*(j\omega)\right)\right) - r_l^{-1} \\
&= r_l^{-1}\left(2Re\{\lambda_i\}Re\{H_l(j\omega)\} - 1\right) \\
&\geqslant 0
\end{aligned} \tag{6.40}$$

for all $l = 1, 2, \cdots, m$.

As such, the condition (6.37) is satisfied. According to Theorem 19, the systems in (6.28) are asymptotically stable. □

Denote $\underline{\lambda_R}$ as the minimum value of $Re\{\lambda_i\}$ for all $i = 1, 2, \cdots, N$, i.e., $\underline{\lambda_R} = min_{i\in N} Re\{\lambda_i\}$. A guaranteed phase margin of the cooperative tracking system (6.6)-

(6.16) is found in Theorem 20, and a guaranteed gain margin is found in Theorem 21.

Theorem 20. *The cooperative tracking system* (6.6)-(6.16) *has a guaranteed phase margin* $\pm arccos\frac{1}{2\underline{\lambda_R}}$.

Proof. Express $H_l(j\omega)$ in its polar form, i.e., $H_l(j\omega) = e^{j\phi_l(\omega)}$. If a phase shift ϕ_l in the perturbed system satisfies $|\phi_l| \leq \arccos\frac{1}{2\underline{\lambda_R}}$ for all $l = 1, 2, \cdots, m$, i.e., $|\phi_l| \leq \arccos\frac{1}{2Re\{\lambda_i\}}$ for all $l = 1, 2, \cdots, m$, and $i = 1, 2, \cdots, N$, then one has

$$Re\{H_l(j\omega)\} = cos\phi_l \geqslant \frac{1}{2Re\{\lambda_i\}}. \tag{6.41}$$

According to Corollary 6.4.1, the systems in (6.28) are asymptotically stable. □

Theorem 21. *The cooperative tracking system* (6.6)-(6.16) *has a guaranteed gain reduction tolerance* $\frac{1}{2\underline{\lambda_R}}$ *and an infinite gain margin.*

Proof. Consider a linear constant gain a_l in the perturbed systems. If $a_l \geqslant \frac{1}{2\underline{\lambda_R}}$ for all $l = 1, 2, \cdots, m$, i.e., $a_l \geqslant \frac{1}{2Re\{\lambda_i\}}$ for all $l = 1, 2, \cdots, m$, and $i = 1, 2, \cdots, N$, then the systems in (6.28) are asymptotically stable according to Corollary 6.4.1. □

Remark 12. *Compared to the single-agent LQR system, which has a* $\pm 60^\circ$ *phase margin, a* 50% *gain reduction, and an infinite gain margin, the stability margins of the multi-agent cooperative tracking systems depend on characteristics of the communication graph topology,* $\underline{\lambda_R}$. *In particular, larger* $\underline{\lambda_R}$ *leads to larger guaranteed phase and gain margins according to Theorems 20 and 21.*

In the next section, we study properties of $\underline{\lambda_R}$ to further explore guaranteed phase and gain margins of the networked cooperative tracking system.

6.5 Graphical Results on Phase and Gain Margins

In this section, we first study the range of $\underline{\lambda_R}$ following an algebraic graph theory analysis. We show that $0 < \underline{\lambda_R} \leq 1$ holds for general communication graph topology, and then prove that the directed tree graph permits the maximum $\underline{\lambda_R}$, i.e., $\underline{\lambda_R} = 1$ in this case. Finally, we provide graphical results on the guaranteed phase and gain margins.

6.5.1 $\underline{\lambda_R}$ in General Communication Graph Topology

We denote $\mathbf{Z}$ as the set of all real square matrices whose off-diagonal elements are all non-positive.

Lemma 9. *[124, Theorem 4.3] Let $M \in \mathbf{Z}$. Then the following statements are equivalent:*

1) All principles minors of M are positive;

2) The real part of each eigenvalue of M is positive;

3) The inverse M^{-1} exists and $M^{-1} \geqslant 0$.

The links in the communication graph $\mathcal{G}$ are equally weighted, i.e., $a_{ij} = 1$ if $(j, i) \in \mathcal{E}$. Similarly, $g_i = 1$ if agent i can observer the leader.

The next theorem investigates the maximum and minimum values of $\underline{\lambda_R}$ for general communication graph topology.

Theorem 22. *For any communication graph topology satisfying Assumptions 6.3.1 and 6.5.1, the following inequality holds,*

$$0 < \underline{\lambda_R} \leq 1. \tag{6.42}$$

Proof. The lower limit $\underline{\lambda_R} > 0$ is straightforward from Lemma 8. We now show that $\underline{\lambda_R} \leq 1$ holds by using a contradiction method.

Assume $\underline{\lambda_R} > 1$ under contradiction. Then $Re\{\lambda_i\} > 1$ holds for all $i = 1, \cdots, N$. With this assumption, there exists a real number $\beta > 1$, such that $Re\{\lambda_i\} - \beta > 0$ holds for all $i = 1, 2, \cdots, N$. Denote $\alpha_i = \lambda_i - \beta$. α_i is then an eigenvalue of the matrix $L + G - \beta I_N$, i.e.,

$$(L + G - \beta I_N)\omega_i = (\lambda_i - \beta)\omega_i = \alpha_i \omega_i, \tag{6.43}$$

where ω_i is the i^{th} eigenvector of $L + G$, i.e., $(L + G)\omega_i = \lambda_i \omega_i$. Because $\underline{\lambda_R} > 1$, $Re\{\alpha_i\} > 0$ holds for all $i = 1, 2, \cdots, N$. Next we show that there exists at least one α_i such that $Re\{\alpha_i\} \leq 0$, which contradicts the assumption that $\underline{\lambda_R} > 1$.

Since α_i is the eigenvalue of $L+G-\beta I_N$, we study characteristics of the matrix $L + G - \beta I_N$. Note that the matrix G is a diagonal matrix with g_i in the diagonal. According to Assumption 6.5.1, $\beta I_N - G$ has all positive diagonal elements. Denote the minimum diagonal element of $\beta I_N - G$ as γ, then $\gamma > 0$, and $\beta I_N - G$ can be rewritten as $\beta I_N - G = \gamma I_N + E$, where E is a $N \times N$ diagonal matrix with non-negative diagonal elements. As such, the matrix $L + G - \beta I_N$ can be rewritten as

$$\begin{aligned} L + G - \beta I_N &= L - (\beta I_N - G) \\ &= L - \gamma I_N - E, \end{aligned} \tag{6.44}$$

Note that the minimum eigenvalue of the Laplacian matrix L is 0. As such, the minimum eigenvalue of $L - \gamma I$ is negative. As such, there exists at least one principal minor of the matrix $L - \gamma I$ that is negative, according to Lemma 9.

Denote $|M|$ as the negative principal minor of $L - \gamma I$ with the minimum order, i.e., all principal minors of $L - \gamma I$ that have lower orders are positive. Denote the order of $|M|$ as $k (k \leq N)$. Assume M has the following form

$$M = \begin{bmatrix} m & M_{12} \\ M_{21} & M_{22} \end{bmatrix}, \tag{6.45}$$

where m is a scalar, the row vector $M_{12} \in \mathbb{R}^{1\times(k-1)}$, the column vector $M_{21} \in \mathbb{R}^{(k-1)\times 1}$, and the square matrix $M_{22} \in \mathbb{R}^{(k-1)\times(k-1)}$. Since $|M| < 0$, one has

$$\begin{vmatrix} m & M_{12} \\ M_{21} & M_{22} \end{vmatrix} = (m - M_{12}M_{22}^{-1}M_{21})|M_{22}| < 0. \tag{6.46}$$

Since the matrix M_{22} is of $k-1$ order, we have $|M_{22}| > 0$. As such,

$$m - M_{12}M_{22}^{-1}M_{21} < 0. \tag{6.47}$$

Then we consider the principal minor of $L - \gamma I - E$, $|M - \bar{E}|$, where $\bar{E}$ is a submatrix of E. $M - \bar{E}$ has the following form,

$$M - \bar{E} = \begin{bmatrix} m - e & M_{12} \\ M_{21} & M_{22} - E_{22} \end{bmatrix}, \tag{6.48}$$

where $e \geqslant 0$ is a scalar, and E_{22} is a $k-1$ by $k-1$ square diagonal matrix with non-negative elements. The determinant of $M - \bar{E}$ is

$$\begin{aligned} |M - \bar{E}| &= \begin{vmatrix} m - e & M_{12} \\ M_{21} & M_{22} - E_{22} \end{vmatrix} \\ &= \left((m-e) - M_{12}(M_{22} - E_{22})^{-1}M_{21}\right)|M_{22} - E_{22}|. \end{aligned} \tag{6.49}$$

Since $|M - \bar{E}|$ is a principal minor of $L - \gamma I - E$, we can determine the sign of the eigenvalues of $L - \gamma I - E$, i.e., α_i, by checking the sign of $|M - \bar{E}|$. To do so, we consider two cases: 1) $|M_{22} - E_{22}| \leq 0$, and 2) $|M_{22} - E_{22}| > 0$. For the first case, it is straightforward that there exists at least one $\alpha_i \leq 0$ according to Lemma 9, which contradicts the assumption that $\underline{\lambda_R} > 1$. For the second case, let us prove that

$$(m - e) - M_{12}(M_{22} - E_{22})^{-1}M_{21} < 0, \tag{6.50}$$

which leads to the result that $|M - \bar{E}| < 0$ according to (6.49). $|M - \bar{E}| < 0$ indicates that there exists at least one $\alpha_i < 0$ according to Lemma 9, which contradicts the assumption $\underline{\lambda_R} > 1$.

Noticing that e is a non-negative number, it is clear that if the following equation holds,

$$m - M_{12}(M_{22} - E_{22})^{-1}M_{21} < 0, \tag{6.51}$$

then (6.50) holds.

Compare (6.51) and (6.47). Because (6.47) holds, to show (6.51), we only need to show that

$$M_{22}^{-1} \leq (M_{22} - E_{22})^{-1}. \tag{6.52}$$

Here "$\leq$" is element by element comparison. Note that $M_{22} \in \mathbf{Z}$ and $(M_{22} - E_{22}) \in \mathbf{Z}$. As such, $M_{22}^{-1} \geqslant 0$ and $(M_{22} - E_{22})^{-1} \geqslant 0$ hold according to Lemma 9.

Note that the following equality holds.

$$(M_{22} - E_{22})^{-1} = M_{22}^{-1} + M_{22}^{-1}E_{22}(M_{22} - E_{22})^{-1}. \tag{6.53}$$

As such, one has

$$\begin{aligned} M_{22}^{-1} - (M_{22} - E_{22})^{-1} &= -M_{22}^{-1}E_{22}(M_{22} - E_{22})^{-1} \\ &\leq 0 \end{aligned} \tag{6.54}$$

The last inequality holds because $M_{22}^{-1} \geqslant 0$, $E_{22} \geqslant 0$, and $(M_{22} - E_{22})^{-1} \geqslant 0$. As such, (6.52) holds, which leads to (6.51) and (6.50). As such, $|M - \bar{E}| < 0$ is proven, and thus the assumption $\underline{\lambda_R} > 1$ does not hold.

The proof is complete. □

Theorem 22 provides the maximum and minimum values of $\underline{\lambda_R}$ for a cooperative tracking system of general graph topology. In the next subsection, we show that the directed tree graph has the maximum $\underline{\lambda_R}$ among all communication graphs satisfying Assumptions 6.3.1 and 6.5.1.

6.5.2 $\underline{\lambda_R}$ in Directed Tree Topology

Since a larger $\underline{\lambda_R}$ leads to the increase of guaranteed phase and gain margins according to Theorems 20 and 21, here we find classes of special graph topology that promises the maximum $\underline{\lambda_R}$ among all possible communication graphs.

Theorem 23. *For the cooperative tracking system of a directed tree topology $\mathcal{G}$, $\underline{\lambda_R} = 1$ under assumptions 6.3.1 and 6.5.1.*

Proof. For a directed tree, the Laplacian matrix is a lower triangular matrix, i.e.,

$$L = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ -a_{21} & d_2 & 0 & \cdots & 0 \\ -a_{31} & -a_{32} & d_3 & \cdots & 0 \\ & & & \ddots & \\ -a_{N1} & -a_{N2} & -a_{N3} & \cdots & d_N \end{bmatrix}, \tag{6.55}$$

where $a_{ij} = 1$, if and only if $a_{ik} = 0\ \forall k \neq j$. This is because in a directed tree, each node except the root node has one and only one in-degree.

According to Assumption 6.3.1, the root node can observe the leader. The matrix $L + G$ is then a lower triangular matrix with the following form

$$L + G = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_{21} & d_2 & 0 & \cdots & 0 \\ -a_{31} & -a_{32} & d_3 & \cdots & 0 \\ & & & \ddots & \\ -a_{N1} & -a_{N2} & -a_{N3} & \cdots & d_N \end{bmatrix}. \tag{6.56}$$

Therefore, the eigenvalues of $L + G$ are $\lambda_i = 1, d_2, \cdots, d_N$. Because $d_i = 1$ holds for all $i = 2, 3, \cdots, N$, we have $\underline{\lambda_R} = 1$ for a cooperative tracking system of a directed tree topology. □

Comparing with a general graph topology where $\underline{\lambda_R} \leq 1$ according to Theorem 22, it is straightforward to conclude that the directed tree graph promises the maximum $\underline{\lambda_R}$ among all possible communication graphs.

6.5.3 Graphical Results on Phase and Gain Margins

Theorem 24. *For the networked cooperative tracking system defined in* (6.6)-(6.16)*, the guaranteed phase and gain margin performances are bounded by* $\pm 60^\circ$ *phase margin,* 50% *gain reduction, and infinite gain margin.*

Proof. This result is derived naturally from Theorems 20-22. □

Remark 13. *Theorem 24 implies that for the multi-agent cooperative systems with local LQR design, the guaranteed phase and gain margins are bounded by those of the single-agent LQR system. Intuitively, this is understandable because agents do not have direct access to the leader, and have to track the leader based on limited information from their immediate neighbors. A perturbation at agent i may propagate throughout the communication graph, potentially harm the stability of other agents in the network.*

The next theorem shows the robustness result for the cooperative tracking system of the directed tree topology.

Theorem 25. *The directed tree communication graph promises the best phase and gain margin performances among all possible communication graph topology. In other words, the cooperative tracking system of a directed tree topology* $\mathcal{G}$ *has guaranteed* $\pm 60^\circ$ *phase margin,* 50% *gain reduction, and infinite gain margin.*

Proof. This theorem is derived naturally from Theorems 20, 21, and 23. □

Remark 14. *Theorem 25 shows that among all possible communication graphs, directed tree is the special topology that promises the best phase and gain margins, which*

are the same as those of single-agent LQR systems. This result can be understood intuitively as follows. In the directed tree graph, the control of each agent is uniquely decided by its root agent, but not any other agents. Each agent synchronizes to its root node based on the state information received from the root node. This architecture is equivalent to that of the single-agent LQR system. All agents synchronize to the leader as long as each agent synchronizes to its root node. As each agent behaves the same as the single-agent LQR system, the robustness of the whole cooperative system in terms of guaranteed phase and gain margins is also equivalent to the single-agent LQR system.

6.6 Simulation Studies

In this section, we conduct simulation studies to illustrate and validate the theoretical analysis. Consider a group of four agents connected by a directed communication graph with the following agent dynamics

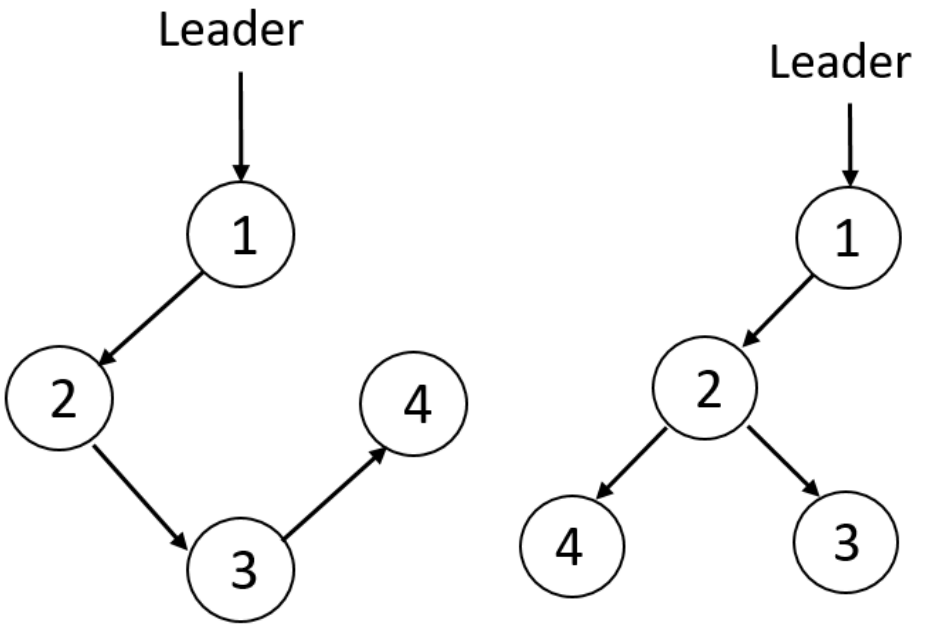
$$\dot{x}_i = \begin{bmatrix} 2 & 4 \\ -2 & 1 \end{bmatrix} x_i + \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} P u_i, \tag{6.57}$$

$$\dot{x}_0 = \begin{bmatrix} 2 & 4 \\ -2 & 1 \end{bmatrix} x_0, \tag{6.58}$$

where $i = 1, 2, 3, 4$, and x_0 denotes the state of the leader. Select the weighting matrices Q and R in the cost function (6.17) as identity matrices, and then the feedback gain K can be calculated accordingly using (6.15).

We consider communication graphs of the following four cases. Cases 1 and 2 in Figure 6.2 are both directed tree communication graphs, and cases 3 and 4 in Figure 6.3 are both non-tree general communication graphs with additional links. Only one agent in case 3 can observe the leader while in case 4, one additional agent can also

observe the leader. The edge weights a_{ij} are selected as 1 if $(j, i) \in \mathcal{E}$, and the pining gain g_i is selected as 1 if agent i can observe the leader.



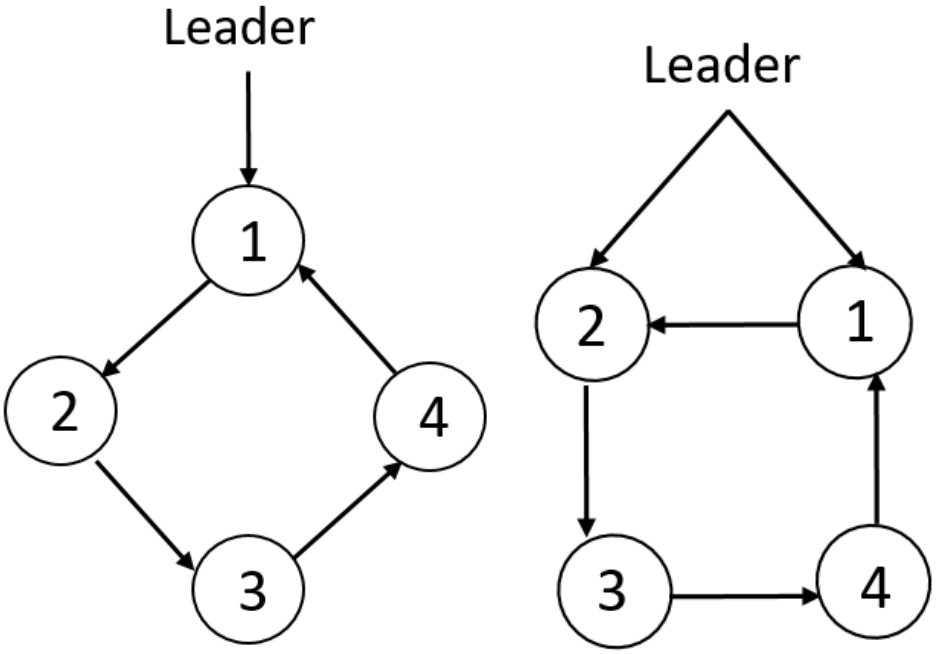
(a) (b)

Figure 6.2. Directed tree communication graph for (a) case 1 and (b) case 2.

The Laplacian matrix of the four graph topology are

$$L_1 = \mathcal{D}_1 - \mathcal{A}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \tag{6.59}$$

$$L_2 = \mathcal{D}_2 - \mathcal{A}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \tag{6.60}$$



(a) (b)

Figure 6.3. General communication graph for (a) case 3, and (b) case 4.

$$L_3 = \mathcal{D}_3 - \mathcal{A}_3 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \tag{6.61}$$

$$L_4 = \mathcal{D}_4 - \mathcal{A}_4 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \tag{6.62}$$

respectively, and the four pining gain matrices are

$$G_1 = G_2 = G_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6.63}$$

$$G_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6.64}$$

respectively. The minimum real part of eigenvalues of $L+G$, i.e., $\underline{\lambda_R}$, for the four cases are $\underline{\lambda_R}_1 = 1$, $\underline{\lambda_R}_2 = 1$, $\underline{\lambda_R}_3 = 0.2$, $\underline{\lambda_R}_4 = 0.3$, respectively. According to Theorems 25, the systems with directed tree communication graph, i.e., cases 1 and 2, should have the best robustness performances among the four cases.

We first simulate the networked MAS with no pertubation. The synchronization errors, i.e., $\delta_i = x_i - x_0$, of the four cases are plotted in Figures 6.4-6.7, respectively. It can be seen from Figures 6.4-6.7 that agents in all of the four cases can all synchronize to the leader without perturbation.

We then add a pertubation P with gain of 0.2 and phase of 0 to the system (6.57). The synchronization errors of the four cases are plotted in Figures 6.8-6.11, respectively. It can be seen from Figures 6.8 and 6.9 that with this pertubation, systems of cases 1 and 2 still achieve synchronization. However, as shown in Figures 6.10 and 6.11, agents in case 3 and 4 fail to synchronize to its leader. These observations validate the result that systems with directed tree communication graph, e.g., Figure 6.2, are more robust than the systems with general communication graphs, e.g. Figure 6.3.

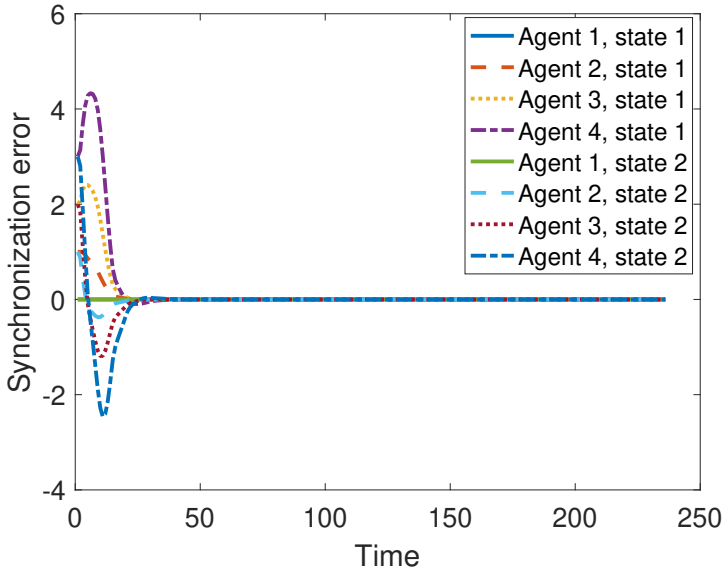


Figure 6.4. Synchronization errors of cooperative tracking system with no perturbation in case 1.

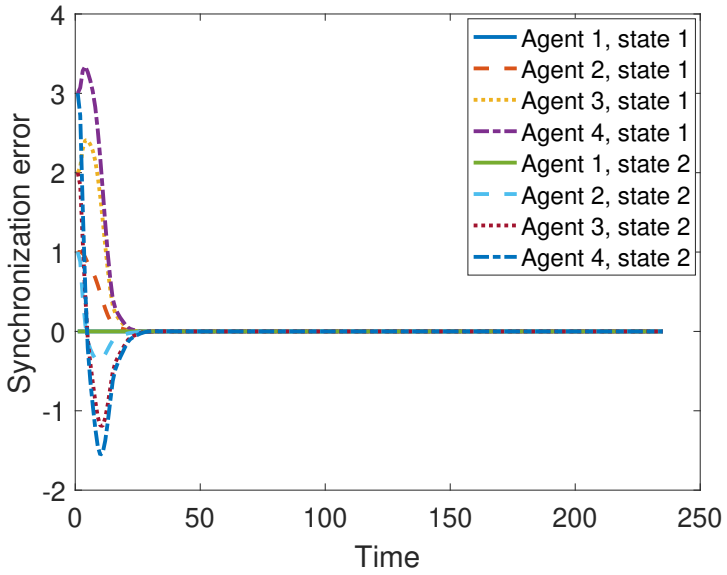


Figure 6.5. Synchronization errors of cooperative tracking system with no perturbation in case 2.

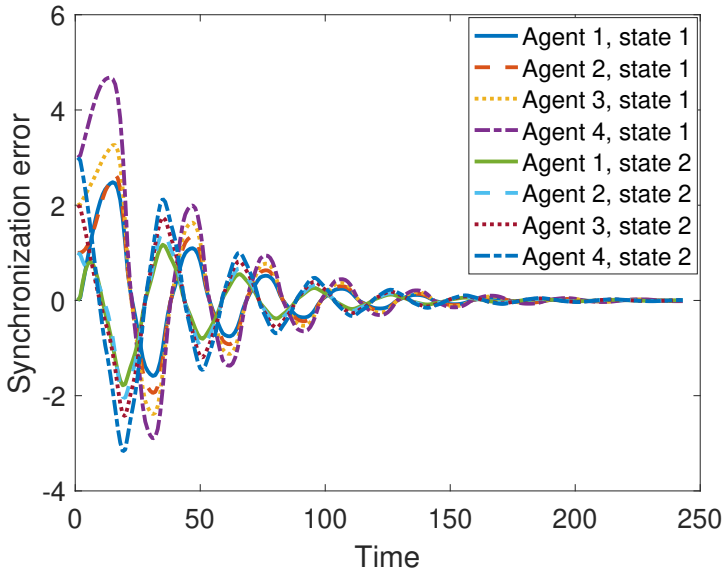


Figure 6.6. Synchronization errors of cooperative tracking system with no perturbation in case 3.

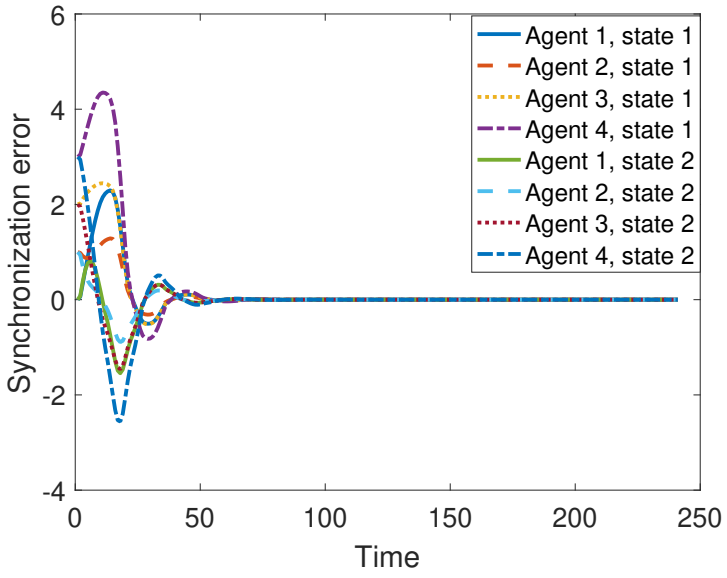


Figure 6.7. Synchronization errors of cooperative tracking system with no perturbation in case 4.

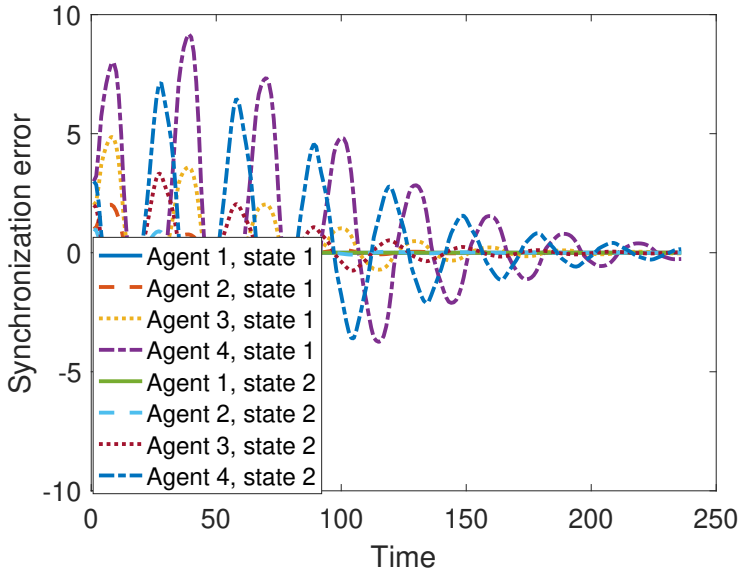


Figure 6.8. Synchronization errors of cooperative tracking system with pertubation of gain 0.2 in case 1.

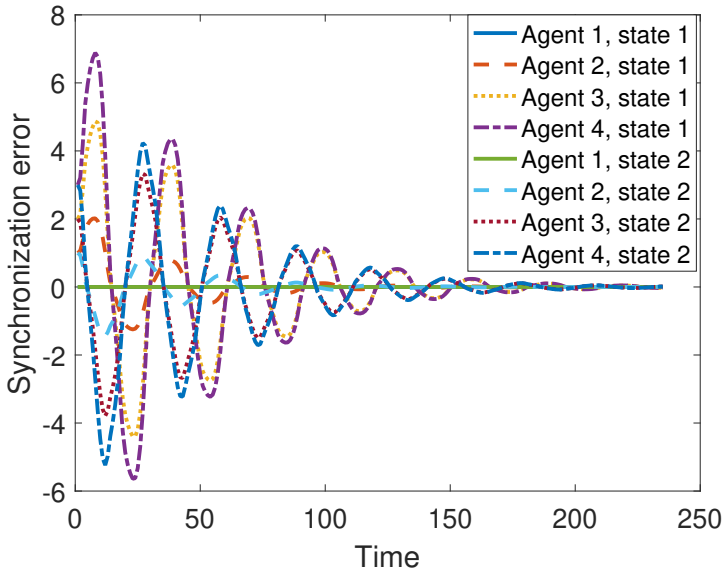


Figure 6.9. Synchronization errors of cooperative tracking systemwith pertubation of gain 0.2 in case 2.

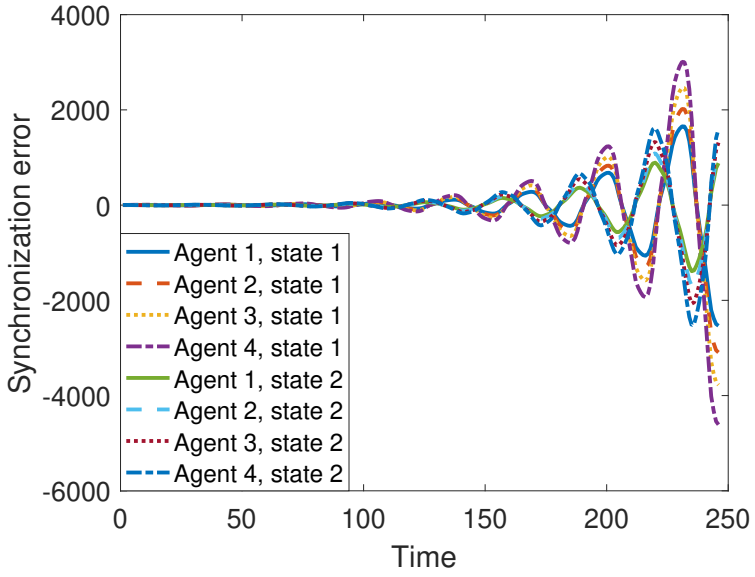


Figure 6.10. Synchronization errors of cooperative tracking system with pertubation of gain 0.2 in case 3.

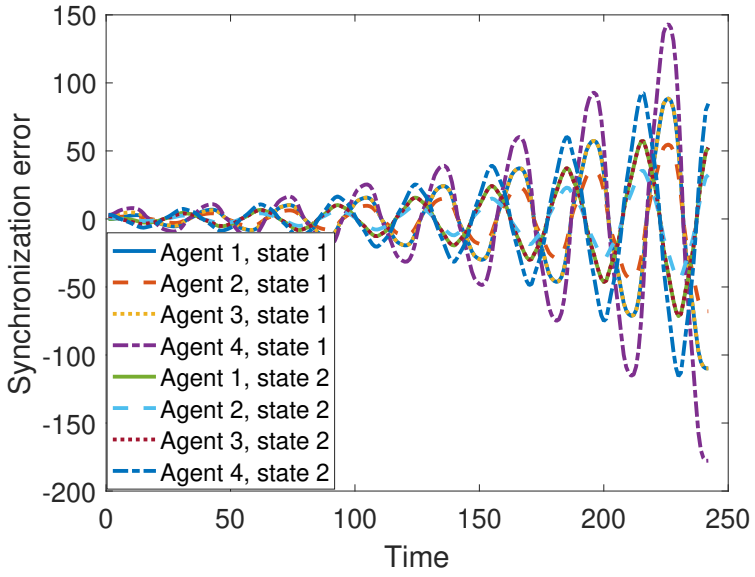


Figure 6.11. Synchronization errors of cooperative tracking system with perturbation of gain 0.2 in case 4.

Finally, to compare the robustness performances of cases 3 and 4, we add a perturbation P with gain of 0.5 and phase of 0. The synchronization errors of all four cases are plotted in Figures 6.12-6.15, respectively. Figures 6.12, 6.13, and 6.15 show that with this perturbation, systems in cases 1, 2, and 4 achieve synchronization successfully. However, as shown in Figure 6.14, agents in case 3 fail to synchronize to the leader. These observation validate the results that the system of case 3 is less robust than the system of case 4. This can be understood intuitively as more agents can directly observe the leader. In all, the system in case 3 has the smallest $\underline{\lambda_R}$ and the worst robustness performances among all four cases. We can conclude that a larger $\underline{\lambda_R}$ results in a better robustness performance, which validates the theoretical analysis in Section 6.4.

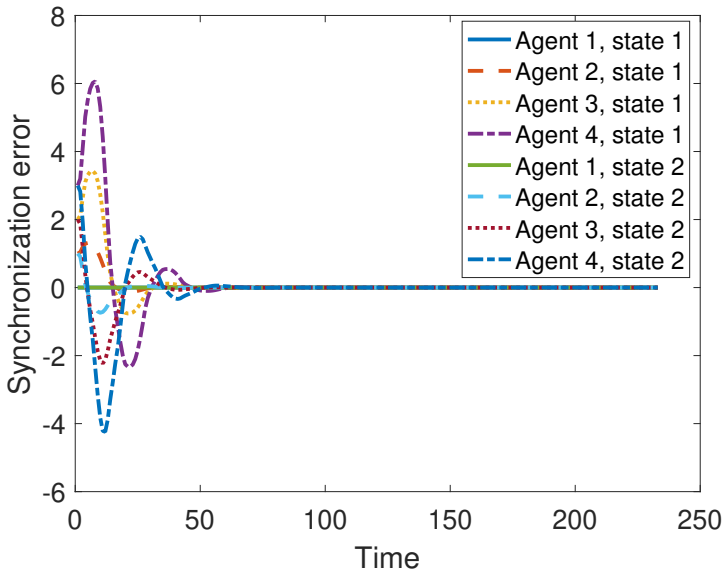


Figure 6.12. Synchronization errors of cooperative tracking system with pertubation of gain 0.5 in f case 1.

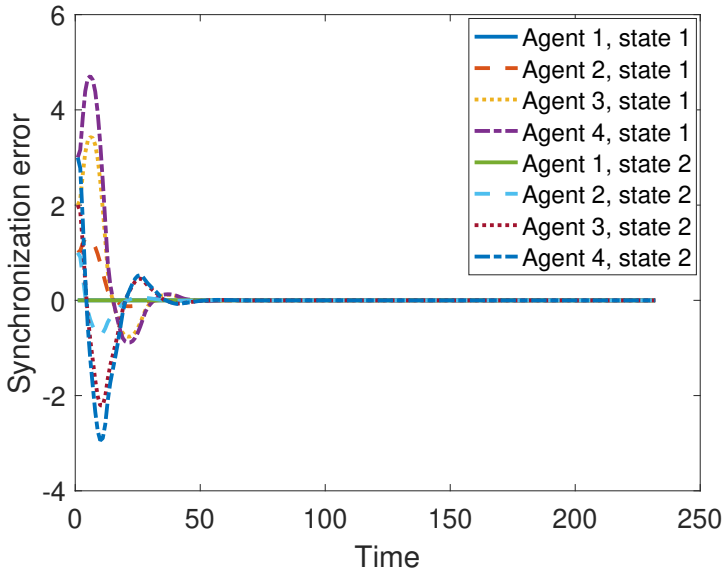


Figure 6.13. Synchronization errors of cooperative tracking system with pertubation of gain 0.5 in case 2.

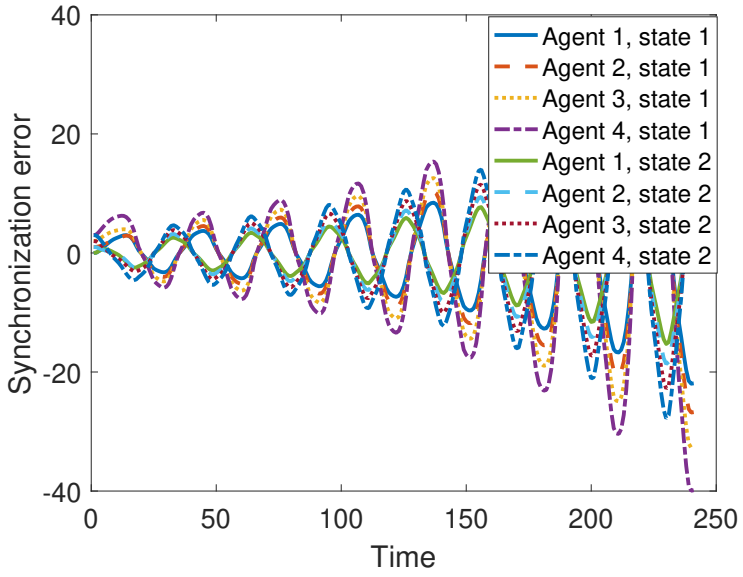


Figure 6.14. Synchronization errors of cooperative tracking system with pertubation of gain 0.5 in case 3.

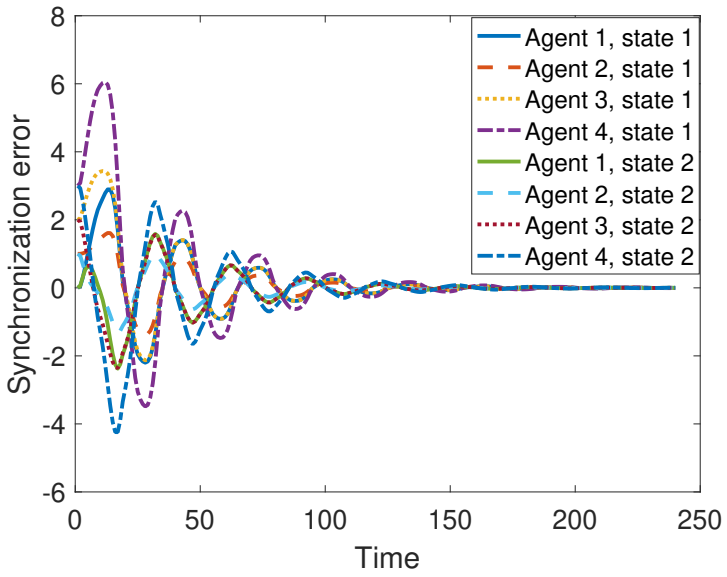


Figure 6.15. Synchronization errors of cooperative tracking system pertubation of gain 0.5 in case 4.

CHAPTER 7

LEARNING AND UNCERTAINTY-EXPLOITED DIRECTIONAL ANTENNA CONTROL FOR ROBUST LONG-DISTANCE AND BROAD-BAND AERIAL COMMUNICATION

7.1 Introduction

UAVs have been widely used in civilian and commercial applications including emergency response, connectivity service, intelligent transportation, precision agriculture, among others [125–127]. Aerial communication among UAVs is expected to play an indispensable role in these applications when multiple UAVs are involved [128–130]. In applications such as emergency response and remote infrastructure health monitoring, the long-distance and broad-band UAV-to-UAV communication capability is desired.

To enable long-distance and broad-band UAV-to-UAV communication, the aerial communication using directional antennas (ACDA) has been developed as a promising solution [2, 131–135]. Through using directional antennas that focus the transmission energy in certain direction, ACDA significantly extends the communication distance and rejects interference, compared to omni-drectional antenna based solutions. With ACDA, UAVs-carried communication infrastructures can be quickly deployed to deliver Wi-Fi services from the air, through which high-rate data such as monitoring streams from remote locations can be transmitted in real-time (see Figure 7.1). The detailed design prototype and hardware components of this ACDA system are described in [2].

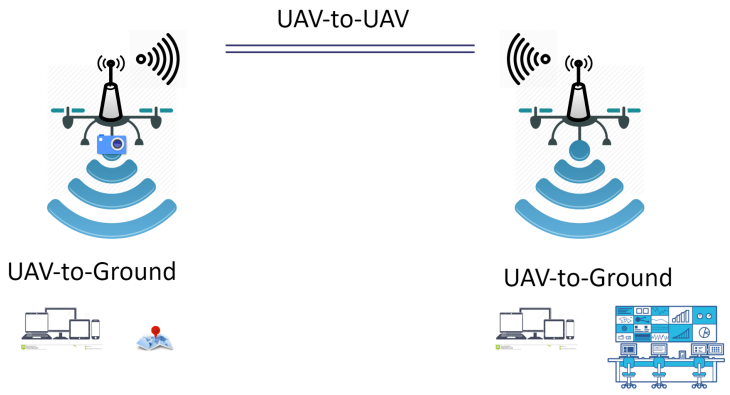


Figure 7.1. Illustration of the broadband long-distance communication infrastructure using controllable UAV-carried directional antennas [2] .

A critical component of the ACDA system is the automatic alignment of directional antennas to maximize the communication performance. Each UAV in the ACDA system carries a rotational plate mounted with a directional antenna [133], which is controlled to align with the directional antenna carried by the other UAV. Robust automatic alignment of directional antennas is not easy to achieve, considering practical issues such as the limited on-board sensing devices due to the physical constraints of UAV payload and power supplies, uncertain and varying UAV mobility, and unstable GPS and unknown communication environments.

There are two general design configurations of the ACDA system, depending on whether the communication channel used for antenna control is omni or not. The first configuration uses a directional antenna-equipped broad-band channel for the transmission of application-oriented data (e.g., real-time video streams), and an additional low-rate omni-directional communication channel for control and command data. In [133], omni-directional antennas are used to transmit the GPS information of the remote UAV for the alignment of antennas. This configuration simplifies the

antennas controller design, as the control channel still functions even if the directional antennas are not in alignment. However, the omni-directional control channel suffers from practical issues such as interference and dissipation over a long communication distance [136,137]. As such, in this chapter, we aim to design the ACDA system using the second configuration where the high-rate application data and low-rate control and command data share the same channel equipped with directional antennas.

Although more practical, this solution that removes the additional control and command channel introduces more challenges to the robustness of antennas control. As control and command data cannot be transmitted if the directional communication channel fails, the antenna control system needs to robustly lock and track the other directional antennas, once the communication channel is established initially. To do that, we develop an uncertain UAV mobility modeling and intention estimation framework to capture and predict the uncertain intentions of the remote UAV's maneuvers. Predictive intentions for robot-robot and human-robot collaborations have been studied in e.g., [138–140]. Most of these studies assume that an agent's intention can be described and modeled in a deterministic and predictable form [138–140]. This is not suitable for UAVs considering their highly flexible and random movement patterns. Probabilistic intentions and their estimation have also been studied in e.g., [141,142], using stochastic models such as Markov chain and Baysien networks. In this chapter, we use random mobility models (RMMs), and in particular, the smooth turn (ST) UAV RMM [1,143] to more realistically capture the uncertain mobility intentions of UAVs. RMMs are a class of random switching models that capture the statistics of random moving objects. The intelligence on RMMs is exploited in this chapter to facilitate robust tracking.

In indoor and many emergency scenarios, GPS signals may be unstable considering environmental disturbances and blockages. In GPS unstable or denied environ-

ment, we need additional measurement signals for antenna control. Received Signal Strength Indicator (RSSI), a communication performance indicator, is a promising measurement signal for ACDA, as it can be measured from ACDA self-equipped directional antennas, and does not require additional localization sensors to be carried by UAVs. In [144], we adopted the RSSI of directional antennas, to compensate unstable GPS signals, under the assumption that the communication environment is perfect. In particular, GPS and directional Wi-Fi RSSI based fusion algorithms were developed to estimate the other UAV's location, which is used to align the headings of directional antennas. However, in an imperfect communication environment, the effects of reflection, refraction and absorption by buildings, obstacles, and interference sources can distort the strongest signal directions. In this case, simply aligning directional antennas using their GPS locations may not lead to the best communication performance (see experimental studies in [2, 145]). In this chapter, we develop a distributed antenna control solution for the goal of maximizing the communication performance, instead of using location-based antenna heading alignment. The solution learns directional Wi-Fi channel models online and provides RSSI as not only alternative measurement signals, but also the goal function for antenna control, in GPS-denied settings.

Our antenna control adopts a novel stochastic optimal control approach that integrates RL for online optimal control, MPCM for effective uncertainty evaluation, and UKF for nonlinear state estimation. On the aspect of optimal control, RL has been developed in [32, 146] for deterministic system dynamics. Paper [31] developed the stochastic optimal control solution that integrates MPCM and RL for systems modulated by uncertain parameters, and paper [147] applied this solution for an air traffic management problem subject to uncertain weather conditions. In this chapter, we study the stochastic optimal control problem for broad random switching

systems. On the aspect of estimation, nonlinear system estimation methods such as Extended Kalman Filter (EKF) and UKF have been widely used typically for known and deterministic systems corrupted with additive noises, but not random switching RMMs. In this chapter we develop a new stochastic optimal control solution for systems that involve nonlinear random switching RMMs and limited measurements, by integrating UKF, RL and MPCM.

The contributions of this chapter are summarized as follows.

1. *The design configuration of the ACDA system using pure directional antennas.* In this ACDA system, the high-rate application data and low-rate control and command data share the same communication channel equipped with directional antennas. This design is more practical compared to the previously developed ACDA systems, which use both directional and omni-directional antennas.
2. *RL-based antenna control.* This solution learns directional channel models online and provides RSSI as not only alternative measurement signals, but also the goal function for antenna control, in GPS-denied settings. In addition, this solution does not require a known and perfect communication channel, which was assumed in the previously developed ACDA systems.
3. *Stochastic UAV intention modeling.* We use RMMs to capture the highly flexible and random movement patterns of UAVs, and develop an online model estimation framework to capture and predict the uncertain intentions of the remote UAV's maneuvers.
4. *Real-time state estimation for random switching systems.* Agents' states in general random switching systems are usually regarded as unpredictable. This solution makes the best prediction out of the agents' intentions coded in the

statistics of the agents' random maneuvers to analyze and further predict the agents' future behaviors.

5. *Real-time distributed optimal control for random switching systems.* This solution integrates RL and MPCM to provide an effective online optimal control solution for agents moving with random switching system models.

The remainder of this chapter is organized as follows. In Section 7.2, we describe the ACDA system shown in Figure 7.1, including both system models and measurement models. The antenna control problem is also formulated. In Section 7.3, we develop the RL based stochastic optimal control solutions. In Section 7.4, an uncertain intention estimation method is provided to estimate the random variables of the remote UAV's uncertain maneuvers. In Section 7.5, simulation studies are conducted.

7.2 Modeling and Problem Formulation

In this section, we first describe the ACDA system model, including the UAV RMM and directional antenna dynamics. We then describe the GPS and RSSI measurement models. The antenna control problems are then formulated.

7.2.1 System Models

We consider two UAVs independently fly in a low-altitude airspace at approximately the same height to fulfill their missions such as search and rescue (see Figure 7.1). The same altitude assumption is reasonable because that 1) the range of flight altitudes for small UAVs is very limited [148]; and 2) the optimal flight altitudes to maximize coverage are proved to be the same for UAVs of the same type [149–151]. On each UAV, a tunable plate attached with a directional antenna is installed and driven by a gear motor [133]. To establish a long-range air-to-air communication

channel to transmit both application data (e.g., surveillance videos) and control and command data, the channel performance needs to be maximized.

7.2.1.1 UAV Random Mobility Model

We use the smooth turn (ST) mobility model ([1,143]) to capture the uncertain intentions of UAVs executing surveillance-like missions (see Figre 7.2). The random maneuvers described by a ST mobility model work as follows. At randomly selected time points T_0^i, T_1^i, $T_2^i, \cdots$, where $0 = T_0^i < T_1^i < \cdots$, UAV i selects a point in the airspace along the line perpendicular to its current heading direction, and then circles around it until the UAV chooses another turning center. The perpendicularity guarantees smooth trajectories [1]. The time duration for UAV i to maintain its current maneuver $\tau_i[T_j^i] = T_{j+1}^i - T_j^i$ follows a memoryless exponential distribution [152].

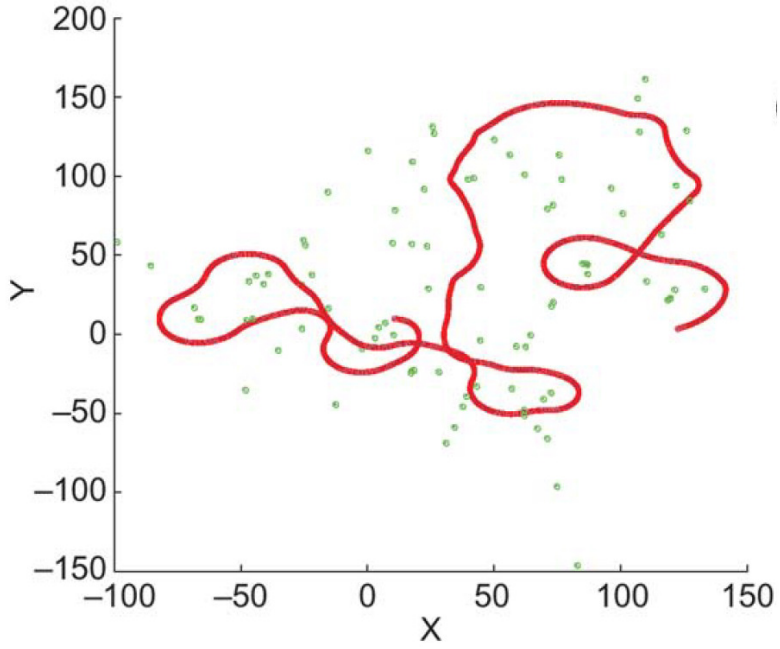
$$f_\tau(\tau_i[T_j^i]) = \lambda_i e^{-\lambda_i \tau_i[T_j^i]}, \tag{7.1}$$

where $1/\lambda_i$ is the mean of $\tau_i[T_j^i]$. The velocity $v_i[T_j^i]$ follows a uniform distribution with the minimum and maximum velocity constraints $v_{i,min} < v_i[T_j^i] < v_{i,max}$,

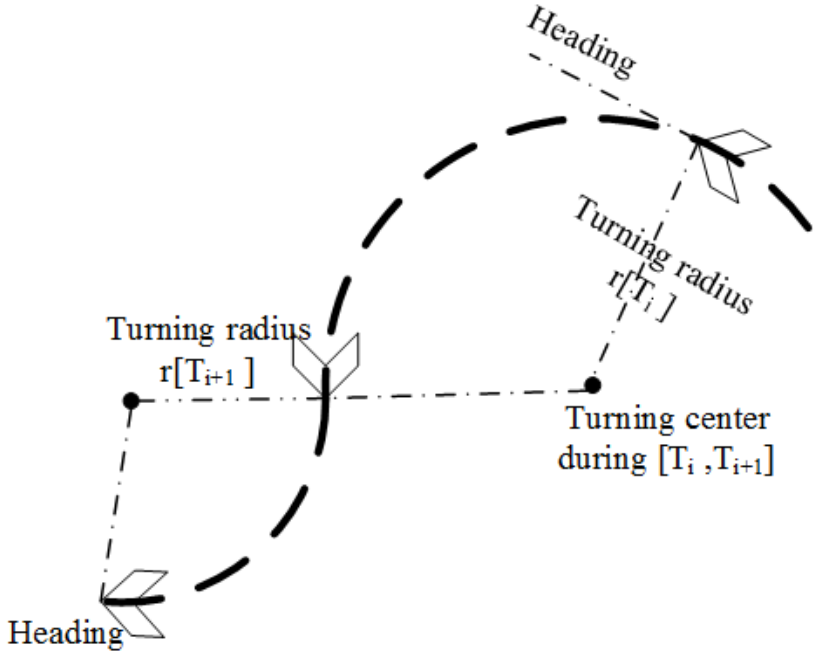
$$f_v(v_i[T_j^i]) = \frac{1}{v_{i,max} - v_{i,min}}. \tag{7.2}$$

The inverse of the turning radius $\frac{1}{r_i[T_j^i]}$ follows the zero-mean Gaussian distribution with variance σ_i^2,

$$f_r(\frac{1}{r_i[T_j^i]}) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2r\sigma_i^2}}. \tag{7.3}$$



(a)



(b)

Figure 7.2. Illustration of the ST RMM: (a) UAV trajectory ensemble (red curve). Green spots are the randomly chosen turning centers [1]; (b) maneuver selection and switching.

Denote the position of UAV i along x and y axes at time instant k as $x_i[k]$ and $y_i[k]$ respectively. The dynamics of UAV i (denote as $f_i(.)$) following the ST uncertain maneuvering intentions are described as

$$\begin{aligned} x_i[k+1] &= x_i[k] + v_i[k]\cos(\phi_i[k])\delta, \\ y_i[k+1] &= y_i[k] + v_i[k]\sin(\phi_i[k])\delta, \\ \phi_i[k+1] &= \phi_i[k] + \omega_i[k]\delta, \end{aligned} \tag{7.4}$$

where δ is the sampling period, $\phi_i[k]$ and $\omega_i[k]$ are the heading angle and angular velocity at time instant k, and

$$\omega_i[k] = \frac{v_i[k]}{r_i[k]}. \tag{7.5}$$

Note that the ST RMM is a random switching model composed of two types of random variables [21]. Type 1 random variables, $v_i[k]$ and $r_i[k]$, describe the characteristics for each maneuver.

$$v_i[k] = \begin{cases} v_i[T_j^i], & \text{if } \exists j \in [0,1,2,..), k = T_j^i \\ v_i[k-1], & \text{if } \forall j = 0,1,2,.., k \neq T_j^i \end{cases} \tag{7.6}$$

$$r_i[k] = \begin{cases} r_i[T_j^i], & \text{if } \exists j \in [0,1,2,..), k = T_j^i \\ r_i[k-1], & \text{if } \forall j = 0,1,2,.., k \neq T_j^i \end{cases} \tag{7.7}$$

The maneuvers' random switching behavior is governed by the type 2 random variable, $\tau_i[T_j^i]$, which describes how often the switching of type 1 random variables occurs.

The two groups of uncertain maneuvers for the UAVs $(v_1[T_j^1], r_1[T_j^1], \tau_1[T_j^1])$ and $(v_2[T_j^2], r_2[T_j^2], \tau_2[T_j^2])$ are independent, as UAV mobility is application-specific, and is not constrained from the communication mission.

7.2.1.2 Directional Antennas Dynamics

The directional antenna installed on each UAV autonomously adjusts its heading angle to establish a robust communication channel between the two UAVs. For UAV i, the heading angle dynamics of its directional antennas is described as

$$\theta_i[k+1] = \theta_i[k] + (\omega_i'[k] + \omega_i[k])\delta, \tag{7.8}$$

where θ_i is the heading angle of antennas i, and ω_i' is the angular velocity of antennas i due to its heading control. Note that the change of θ_i is caused by both the control of antennas i, ω_i', and the movement of UAV i, ω_i.

7.2.2 Measurement Models

We consider two measurement signals for the ACDA system, GPS and RSSI.

7.2.2.1 GPS measurement

If GPS is available, the measured GPS signal of UAV i is denoted as $\mathbf{z}_{G,i}(k)$,

$$\mathbf{z}_{G,i}[k] = \mathbf{H}_G(k)\mathbf{x}_i[k] + \varpi_{G,i}[k], \tag{7.9}$$

where $\mathbf{H}_G = [1,0,0,0;0,1,0,0]$ is the measurement matrix, $\mathbf{x}_i[k] = [x_i[k], y_i[k], \phi_i[k], \theta_i[k]]^T$ is the system state of UAV i, and $\varpi_{G,i}$ is the white Gaussian noise with zero mean and covariance $\mathbf{R}_{G,i}$. GPS signals can be transmitted through the air-to-air communication channel to assist with the control of directional antennas. Denote the relation between the GPS signal and system state as $h_{G,i}$, i.e., $\mathbf{z}_{G,i}[k] = h_{G,i}(\mathbf{x}_i[k])$.

7.2.2.2 RSSI measurement

RSSI measures the signal power received from the transmitting antenna [153], and hence is an important indicator of communication channel performance. In the

ACDA system, RSSI is affected by the relative positions of two UAVs that carry these directional antennas, headings, field radiation patterns of these antennas, and also communication environment. Denote the measured RSSI signal as $z_R[k]$, the relation between the RSSI signal, $z_R[k]$, and the system states, $\mathbf{x}[k]$ ($\mathbf{x}[k] = [\mathbf{x}_1^T[k], \mathbf{x}_2^T[k]]^T$), as $h_R(.)$, i.e., $z_R[k] = h_R(\mathbf{x}[k])$, then $z_R[k]$ is given by the Friis free space equation [153]:

$$\begin{aligned} z_R[k] =& P_{t|dBm}[k] + 20\log_{10}(\lambda) - 20\log_{10}(4\pi) \\ & - 20log_{10}(d[k]) + G_{l|dBi}[k] + \varpi_R[k], \end{aligned} \tag{7.10}$$

where $P_{t|dBm}[k]$ is the transmitted signal power, λ is the wavelength, and $d[k]$ is the distance between the two UAVs at time k, i.e., $d[k] = \sqrt{(x_1[k] - x_2[k])^2 + (y_1[k] - y_2[k])^2}$. $G_{l|dBi}[k]$ is the sum of gains at both the transmitting and receiving sides [154]. The Ubiquiti NanoStation loco M5 directional antennas [155] that we use in the ACDA system is modeled based on the filed pattern of the end-fire array antennas [156],

$$\begin{aligned} G_{l|dBi}[k] =& (G_{t|dBi}^{max} - G_{t|dBi}^{min}) \sin\frac{\pi}{2n} \frac{\sin(\frac{n}{2}(k_a d_a(\cos(\gamma_t[k] - \theta_t[k])) - 1) - \frac{\pi}{n})}{\sin(\frac{1}{2}(k_a d_a(\cos(\gamma_t[k] - \theta_t[k])) - 1) - \frac{\pi}{n})} \\ & + (G_{r|dBi}^{max} - G_{r|dBi}^{min}) \sin\frac{\pi}{2n} \frac{\sin(\frac{n}{2}(k_a d_a(\cos(\gamma_r[k] - \theta_r[k])) - 1) - \frac{\pi}{n})}{\sin(\frac{1}{2}(k_a d_a(\cos(\gamma_r[k] - \theta_r[k])) - 1) - \frac{\pi}{n})} \\ & + G_{t|dBi}^{min} + G_{r|dBi}^{min}, \end{aligned} \tag{7.11}$$

where $G_{t|dBi}^{max}$, $G_{t|dBi}^{min}$, and $G_{r|dBi}^{max}$, $G_{r|dBi}^{min}$ are the maximum and minimum gains of transmitting and receiving antennas. k_a is the wave number, and $k_a = \frac{2\pi}{\lambda}$. n and d_a are design parameters of the directional antenna. $\theta_t[k]$ and $\theta_r[k]$ are the heading angles of the transmitting and receiving antennas at time k, respectively. $\gamma_t[k]$ and $\gamma_r[k]$ are the heading angles of the transmitting and receiving antennas corresponding to the maximal G_l at time k, respectively.

The parameters $G_{t|dBi}^{max}$, $G_{t|dBi}^{min}$,$G_{r|dBi}^{max}$, and $G_{r|dBi}^{min}$, can be obtained from the antenna's datasheet. In ACDA, the two directional antennas are of the same type, and hence $G_{t|dBi}^{max} = G_{r|dBi}^{max}$, and $G_{t|dBi}^{min} = G_{r|dBi}^{min}$. In an imperfect environment (e.g., where

disturbances and interference exist), these parameters in $G_{l|dBi}[k]$ can be environment-specific.

Similarly, in a perfect communication environment, $\gamma_t[k]$ and $\gamma_r[k]$ are achieved when the two antennas are aligned [144]. Affected by the impact of imperfect environment, such as blockages, the desired heading angles can be captured as

$$\gamma_r[k] = \arctan\frac{y_t[k] - y_r[k]}{x_t[k] - x_r[k]} + \theta_{r_{env}}, \tag{7.12}$$

$$\gamma_t[k] = \arctan\frac{y_r[k] - y_t[k]}{x_r[k] - x_t[k]} + \theta_{t_{env}}, \tag{7.13}$$

where $(x_t[k], y_t[k])$ and $(x_r[k], y_r[k])$ are the positions of UAVs that carry the transmitting and receiving antennas respectively, and $\theta_{r_{env}}$ and $\theta_{t_{env}}$ are environment-specific shift angles at the receiver and transmitter sides. $\theta_{r_{env}}$ and $\theta_{t_{env}}$ are zeros in a perfect environment.

7.2.3 Problem Formulation

We aim to design the angular velocities of each directional antenna to maximize the expected RSSI performance of ACDA over a look-ahead window. The RSSI model (as described in Equations (7.10)-(7.13)) contains unknown environment-specific parameters ($G^{max}_{t|dBi}$, $G^{min}_{t|dBi}$, $\theta_{t_{env}}$ and $\theta_{r_{env}}$), and the UAV dynamics contain uncertain parameters ($v_1[k], r_1[k], \tau_1[T^1_j], v_2[k], r_2[k], \tau_2[T^2_j]$).

Here we formulate the problem as a stochastic optimal control problem. Mathematically, considering the random switching system dynamics described in Equations (7.4)-(7.8), the optimal control policy $\mathbf{u}[k]$ is sought to maximize the expected value function, which is the summation of the predicted RSSI signals over a look-ahead window, i.e.,

$$V(\mathbf{x}[k]) = E\{\sum_{l=k}^{k+N} \alpha^{l-k} z_R[l](\mathbf{x}[l], \mathbf{u}[k])\}, \tag{7.14}$$

where $\mathbf{x}[k]$ is the global state, $\mathbf{x}[k] = [\mathbf{x}_1^T[k], \mathbf{x}_2^T[k]]^T$. $\mathbf{u}[k]$ is the control input, $\mathbf{u}[k] = [u_1[k], u_2[k]]^T$, $u_i[k] = [\omega_i'[k]]$. $z_R[l]$ is the RSSI signal at time l, and $\alpha \in (0, 1]$ is a discount factor. Note that the control is decentralized, in the sense that each antenna finds its own optimal control policy, with the assumption that the other antenna adopts its optimal control policy. Each UAV only needs to learn its own environment-specific parameters ($G_{t|dBi}^{max}$, $G_{t|dBi}^{min}$, and $\theta_{t_{env}}/\theta_{r_{env}}$) to find its optimal control policy.

In the rest of this article, we develop the control solution for one UAV, denoted as the local UAV, or UAV 1, and the other UAV as the remote UAV, or UAV 2. The control solution for the other UAV is designed in the same manner.

7.3 Reinforcement Learning based Stochastic Optimal Control for ACDA

In this section, we develop new online solutions to solve the stochastic optimal control problem for the ACDA system described in Section 7.2.3. The solution integrates the uncertainty sampling method MPCM, the adaptive optimal control method RL, and the nonlinear estimation method UKF, to address the challenges including nonlinear and random switching dynamics, unknown RSSI model, limited measurements of system outputs, and online time requirement to derive optimal solutions for random switching systems.

In Section 7.3.1, we describe the solution when GPS is available but the RSSI model is unknown. Online stochastic optimal control solutions are derived and the environment-specific RSSI model is learned. Section 7.3.2 further develops online solutions in both GPS-available and GPS-denied environments, with the learned environment-specific RSSI model.

7.3.1 Stochastic optimal control with unknown RSSI

To develop a decentralized optimal control solution that maximizes the value function (Equation (7.14)) for the nonlinear random switching ACDA dynamics with unknown RSSI model and limited measurements, two main steps are involved: 1) state estimation, and 2) adaptive optimal controller design.

7.3.1.1 State Estimation

The states of both local and remote UAVs need to be estimated. For the local UAV, the trajectory-specific maneuvers ($v_1[k]$, $r_1[k]$, and $\tau_1[T_j^i]$) are known locally, and hence, the local-system states $(x_1[k], y_1[k], \phi_1[k])$ can be estimated utilizing UKF as described in [144, Section 3.1]. We do not repeat the process here to save the space.

For the remote UAV that has random switching dynamics, the RMM-related maneuvers ($v_2[k]$, $r_2[k]$, and $\tau_2[T_j^2]$) are unknown to the local UAV, and hence the remote UAV's states $(x_2[k], y_2[k], \phi_2[k])$ can not be directly estimated using the existing filtering type of methods. We design a new estimation algorithm for the nonlinear and random switching dynamics. Here, a subset of $\mathbf{x}_2[k]$, i.e., $[x_2[k], y_2[k], \phi_2[k]]$ is needed for this estimation, and we use $\mathbf{x}_2[k]$ to represent this subset to simplify presentation, when it does not cause confusion.

Denote the switching behavior of the remote UAV at time k as $s[k]$. $s[k] = 1$ or 0 represent the current maneuver switches at time k or not. Considering the two

possible switching behaviors, the expected conditional current state of UAV 2 given the previous state, $\mathbf{x}_2[k-1]$, can be derived as

$$\begin{aligned} & E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1]) \\ =& E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1], s[k-1]=0)P(s[k-1]=0) \\ & + E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1], s[k-1]=1)P(s[k-1]=1). \end{aligned} \tag{7.15}$$

When $s[k-1]=0$, the remote UAV remains its previous maneuvers $v_2[k-1]$ and $r_2[k-1]$, and thus the expected system state $E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1], s[k-1]=0)$ can be estimated from the system dynamics $f(\mathbf{x}_2[k-1], v_2[k-1], r_2[k-1])$. When $s[k-1]=1$, UAV 2 selects new maneuvers from the two random variables $v_2[T_j^2]$ and $r_2[T_j^2]$. In this case, the estimation of the system state $E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1], s[k-1]=1)$ involves uncertainty evaluation, which is typically solved by Monte Carlo method, too slow to be used for real-time control. Here we use a multivariate probabilistic collocation method (MPCM) [23] to effectively evaluate the uncertainty. MPCM accurately evaluates the output mean of a system mapping subject to uncertain input parameters, by smartly selecting a limited number of sample points according to the Gaussian Quadrature rules. The main property of MPCM is described in the following lemma. Please refer to [23] for the detailed MPCM design procedure.

[23, Theorem 2] Consider a system G modulated by m independent uncertain parameters, a_i, where $i \in \{1, ..m\}$,

$$G(a_1, ..., a_m) = \sum_{j_1=0}^{2n_1-1} \sum_{j_2=0}^{2n_2-1} ... \sum_{j_m=0}^{2n_m-1} \psi_{j_1,...,j_m} \prod_{i=1}^{m} a_i^{j_i}, \tag{7.16}$$

where a_i is an uncertain parameter with the degree up to $2n_i - 1$. n_i is a positive integer for any i. $\psi_{j_1,...,j_m} \in \mathbb{R}$ are the coefficients. Each uncertain parameter a_i

follows an independent pdf $f_{A_i}(a_i)$. The MPCM approximates $G(a_1, ...a_m)$ with the following low-order mapping

$$G'(a_1, ..., a_m) = \sum_{j_1=0}^{n_1-1} \sum_{j_2=0}^{n_2-1} ... \sum_{j_m=0}^{n_m-1} \Omega_{j_1,...,j_m} \prod_{i=1}^{m} a_i^{j_i}, \tag{7.17}$$

with $E[G(a_1, ..., a_m)] = E[G'(a_1, ..., a_m)]$, where $\Omega_{j_1,...,j_m} \in \mathbb{R}$ are coefficients. MPCM reduces the number of simulations from $2^m \prod_{i=1}^{m} n_i$ to $\prod_{i=1}^{m} n_i$.

Define a system mapping subject to uncertain input parameters $v_2[T_j^2]$ and $r_2[T_j^2]$: $G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1]) = f(\mathbf{x}_2[k-1], v_2[T_j^2], r_2[T_j^2])$. When $s[k-1] = 1$, the expected current state $E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1], s[k-1] = 1)$ can be estimated from the mean output of the system mapping $G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])$, i.e., $E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1], s[k-1] = 1) = E[G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])]$, using MPCM according to Lemma 7.3.1.1 and paper [23]. Under the assumption that the two uncertain parameters $v_2[T_j^2]$ and $r_2[T_j^2]$ have a degree up to $2n_1-1$ and $2n_2-1$ respectively, $G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])$ has the following form.

$$G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1]) = \sum_{j_1=0}^{2n_1-1} \sum_{j_2=0}^{2n_2-1} \psi_{j_1,j_2}(\mathbf{x}_2[k-1]) v_2^{j_1}[T_j^2] r_2^{j_2}[T_j^2], \tag{7.18}$$

According to Lemma 7.3.1.1, the output mean of this system mapping can be estimated from the output of a reduced-order mapping $G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])$, i.e., $E[G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])] = E[G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])]$, where the reduced mapping $G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])$ has the following form

$$G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1]) = \sum_{j_1=0}^{n_1-1} \sum_{j_2=0}^{n_2-1} \Omega_{j_1,j_2}(\mathbf{x}_2[k-1]) v_2^{j_1}[T_j^2] r_2^{j_2}[T_j^2]. \tag{7.19}$$

The coefficients $\Omega_{j_1,j_2}(\mathbf{x}_2[k-1])$ and output mean $E[G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])]$ are obtained using the evaluated outputs $G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])$ at each selected simulation point according to the procedures described in [23, Section II-B].

Theorem 26. *Given the previous state* $\mathbf{x}_2[k-1]$ *of the remote UAV* 2*, the expected current state* $E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1])$ *is estimated by the local UAV* 1 *as*

$$\begin{aligned} &E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1]) \\ &= P_2 E[G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])] \\ &\quad + (1-P_2) f(\mathbf{x}_2[k-1], v_2[k-1], r_2[k-1]), \end{aligned} \tag{7.20}$$

where P_2 *is the switching probability of the remote UAV's maneuver at each time instant,* $P_2 = \lambda_2 \delta$*.*

Proof. Let us first find the switching probability P_i. Since the time duration for UAV i to maintain its current maneuver $\tau_i[T_j^i]$ follows exponential distribution as described in Equation (7.1), P_i can be approximated from its exponential distribution as

$$P_i = \lambda_i \delta. \tag{7.21}$$

With the switching probability and the defined system mapping $G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])$, Equation (7.15) can be further written as

$$\begin{aligned} &E(\mathbf{x}_2[k]|\mathbf{x}_2[k-1]) \\ &= P_2 E[G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])] \\ &\quad + (1-P_2) f(\mathbf{x}_2[k-1], v_2[k-1], r_2[k-1]). \end{aligned} \tag{7.22}$$

Since $E[G_2(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])] = E[G_2'(v_2[T_j^2], r_2[T_j^2], \mathbf{x}_2[k-1])]$ according to Lemma 7.3.1.1 and Equations (7.18) and (7.19), Theorem 26 is derived naturally by combining Equations (7.18), (7.19), (7.22) and Lemma 7.3.1.1. □

Theorem 26 provides a general approach to estimate the expected system state of a random switching system with computational efficiency, when the previous system state is given. Here we use this state estimation approach with UKF to estimate the state of the remote UAV from the measurement $\mathbf{z}_{G,2}[k]$. In particular, we integrate

MPCM and UKF for a 5-step state estimation procedure. ***Steps 1*** and ***2*** select initial conditions and MPCM points to initialize ***Steps 3-5***; ***Step 3*** and ***4*** find the state estimators when the switching behavior $s[k-1]=0$ and 1 respectively; ***Step 5*** finds the expected state by integrating the two estimators found in ***Steps 3*** and ***4***.

Step 1: Initialize. Select initial conditions $\hat{\mathbf{x}}_2[0]$ and $\mathbf{P}[0]$ to initialize the system.

Step 2: Select MPCM points. $n_1 n_2$ MPCM simulation point pairs are selected for the random variables $v_2[T_j^2]$ and $r_2[T_j^2]$ according to the MPCM procedure [23, Section II]. Denote the selected MPCM point pairs as $(\mathcal{V}_{j_1}[T_j^2], \mathcal{R}_{j_2}[T_j^2])$, where $j_1 \in \{0, ..., n_1 - 1\}$ and $j_2 \in \{0, ..., n_2 - 1\}$).

Step 3: Estimate system state when $\mathbf{s[k-1]=0}$**.** When $s[k-1]=0$, the remote UAV does not change its maneuver, and hence the conditional expected current state $E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], s[k-1]=0, \mathbf{z}_{G,2}[k])$ can be estimated using UKF as described in sub-steps (a)-(d).

(a). Select Sigma Points. $2n+1$ symmetric weighted sigma points are selected from $\hat{\mathbf{x}}_2[k-1]$, the estimator of $\mathbf{x}_2[k-1]$.

$$\mathcal{X}_0[k-1] = \hat{\mathbf{x}}_2[k-1],$$

and for $i = 1, 2, ...n$,

$$\mathcal{X}_i[k-1] = \hat{\mathbf{x}}_2[k-1] + \sqrt{(n+\kappa)\mathbf{P}[k-1]}_i,$$
$$\mathcal{X}_{i+n}[k-1] = \hat{\mathbf{x}}_2[k-1] - \sqrt{(n+\kappa)\mathbf{P}[k-1]}_i,$$

where $\mathbf{P}[k-1]_i$ is the ith column of the error covariance matrix of $\hat{\mathbf{x}}_2[k-1]$, n is the states' dimension, and $n=3$ here for the remote UAV system. The weights associated with the selected sigma points are $W_0 = \frac{\kappa}{n+\kappa}, W_i = \frac{1}{2(n+\kappa)}$, and $W_{i+n} = \frac{1}{2(n+\kappa)}$ respectively. κ is a scaling parameter usually set to 0 in the general case or set to $3-n$ in the Gaussian case to capture the fourth-order moment correctly [25, 28].

(b). State Prediction. The system state can be predicted by instantiating each of the sigma points through the system dynamics $f_2(.)$ described in Equation (7.4).

$$\mathcal{X}_l[k|k-1] = f_2(\mathcal{X}_l[k-1], r_2[k-1], v_2[k-1]).$$

Then the priori state estimation can be approximated as a weighted sample mean

$$\hat{\mathbf{x}}_2[k|k-1] = \sum_{l=0}^{2n} W_l(\mathcal{X}_l[k|k-1]).$$

The corresponding covariance matrix is calculated as

$$\begin{aligned}\mathbf{P}[k|k-1] = &\sum_{l=0}^{2n} W_l(\mathcal{X}_l[k|k-1] - \hat{\mathbf{x}}_2[k|k-1]) \\ &\times (\mathcal{X}_l[k|k-1] - \hat{\mathbf{x}}_2[k|k-1])^T.\end{aligned}$$

(c). Measurement Prediction. $2n+1$ sigma points are selected from $\hat{\mathbf{x}}_2[k|k-1]$ with the error covariance $\mathbf{P}[k|k-1]$.

$$\begin{aligned}\mathcal{X}_0[k|k-1] &= \hat{\mathbf{x}}_2[k|k-1], \\ \mathcal{X}_i[k|k-1] &= \hat{\mathbf{x}}_2[k|k-1] + \sqrt{(n+\kappa)\mathbf{P}[k|k-1]}_i, \\ \mathcal{X}_{i+n}[k|k-1] &= \hat{\mathbf{x}}_2[k|k-1] - \sqrt{(n+\kappa)\mathbf{P}[k|k-1]}_i,\end{aligned}$$

with the weights W_0, W_i and W_{i+n} respectively.

The GPS measurement is then predicted by instantiating each of the prediction points through the measurement model $h_{G,2}$ (described in Equation (7.9)),

$$\begin{aligned}\mathcal{Z}_l[k|k-1] &= h_{G,2}(\mathcal{X}_l[k|k-1]), \\ \hat{\mathbf{z}}_{G,2}[k|k-1] &= \sum_{l=0}^{2n} W_l(\mathcal{Z}_l[k|k-1]).\end{aligned}$$

Correspondingly, the measurement covariance matrix and cross correlation matrix are determined by

$$\begin{aligned}\mathbf{P}_{ZZ}[k|k-1] =& \sum_{l=0}^{2n} W_l(\mathcal{Z}_l[k|k-1] - \hat{\mathbf{z}}_{G,2}[k|k-1]) \\ & \times (\mathcal{Z}_l[k|k-1]] - \hat{\mathbf{z}}_{G,2}[k|k-1]])^T + \mathbf{R}_{G,2}, \\ \mathbf{P}_{XZ}[k|k-1] =& \sum_{l=0}^{2n} W_l(\mathcal{X}_l[k|k-1] - \hat{\mathbf{x}}_2[k|k-1]) \\ & \times (\mathcal{Z}_l[k|k-1] - \hat{\mathbf{z}}_{G,2}[k|k-1])^T.\end{aligned}$$

(d). Kalman Gain Update. The Kalman gain is then updated using the covariance information,

$$\mathcal{K} = \mathbf{P}_{ZZ}[k|k-1]\mathbf{P}_{XZ}^{-1}[k|k-1].$$

The estimated state and covariance are thus derived as

$$E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k], s[k-1]=0]) = \hat{\mathbf{x}}_2[k|k-1] + \mathcal{K}(\mathbf{z}_{G,2}[k] - \hat{\mathbf{z}}_{G,2}[k|k-1]),$$

$$E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k], s[k-1]=0) = \mathbf{P}[k|k-1] - \mathcal{K}\mathbf{P}_{ZZ}[k|k-1]\mathcal{K}^T.$$

Step 4: Estimate system state when $\mathbf{s[k-1]=1}$**.** When $s[k-1]=1$, the remote UAV changes its maneuvers according to the random variables $v_2[T_j^2]$ and $r_2[T_j^2]$. With the MPCM points selected in ***Step 2***, $(\mathcal{V}_{j_1}[T_j^2], \mathcal{R}_{j_2}[T_j^2])$, the expected state $E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k], s[k-1]=1)$ and covariance $E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k], s[k-1]=1)$ can be estimated using the following three sub-steps (a)-(c) that integrate MPCM and UKF.

(a). Estimate system state at each selected MPCM point. The system state is estimated at each selected MPCM point $(\mathcal{V}_{j_1}[T_j^2], \mathcal{R}_{j_2}[T_j^2])$ by conducting the UKF procedures shown in ***Step 3***, *(a)-(d).* Denote the estimated state from UKF at each MPCM point as $\hat{\mathbf{x}}_{j_1,j_2}[k]$ with the covariance $\mathbf{P}_{j_1,j_2}[k]$.

(b). Find the reduced polynomial mappings. Define the system mappings $G_x(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])$ and $G_P(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])$ as the relationships be-

tween the expected system state and covariance with the random variables $v_2[T_j^2]$ and $r_2[T_j^2]$ respectively. According to Lemma 7.3.1.1, the mean outputs of the two system mappings can be estimated from the outputs of the reduced-order mappings $G'_x(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])$ and $G'_P(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])$ respectively, from MPCM procedures [23, Section II].

$$G'_x(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2]) = \sum_{j_1=0}^{n_1-1}\sum_{j_2=0}^{n_2-1} \Omega_{X_{j_1,j_2}}(\hat{\mathbf{x}}_2[k-1]) v_2^{j_1}[T_j^2] r_2^{j_2}[T_j^2],$$

$$G'_P(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2]) = \sum_{j_1=0}^{n_1-1}\sum_{j_2=0}^{n_2-1} \Omega_{\mathbf{P}_{j_1,j_2}}(\hat{\mathbf{x}}_2[k-1]) v_2^{j_1}[T_j^2] r_2^{j_2}[T_j^2],$$

The coefficients $\Omega_{X_{j_1,j_2}}$ and $\Omega_{\mathbf{P}_{j_1,j_2}}$ and mean outputs can be obtained using the evaluated outputs $G'_x(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])$ and $G'_P(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])$ at each selected MPCM point, according to the procedures in [23, Section II-B]

(c). Find the expected system state and covariance. The expected state and covariance are then found from the system mapping according to Lemma 7.3.1.1 and the MPCM design procedures [23] as

$$E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k], s[k-1]=1) = E[G'_x(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])],$$

$$E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k], s[k-1]=1) = E[G'_P(\hat{\mathbf{x}}_2[k-1], v_2[T_j^2], r_2[T_j^2])].$$

Step 5: Estimate the expected system state. The estimated state and covariance are derived according to Theorem 26.

$$\begin{aligned} &E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k]) \\ =&P_2 E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k], s[k-1]=1]) \\ &+(1-P_2)E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k], s[k-1]=0]), \end{aligned} \tag{7.23}$$

$$\begin{aligned} &E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k]) \\ =&P_2 E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k], s[k-1]=1]) \\ &+(1-P_2)E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k], s[k-1]=0]). \end{aligned} \tag{7.24}$$

As such, the estimate of $\mathbf{x}_2[k]$ is $\hat{\mathbf{x}}_2[k] = E(\mathbf{x}_2[k]|\hat{\mathbf{x}}_2[k-1], \mathbf{z}_{G,2}[k])$, and the expected error covariance is $\mathbf{P}[k] = E(\mathbf{P}[k]|\mathbf{P}[k-1], \mathbf{z}_{G,2}[k])$.

Remark 15. *The above estimation procedure integrates UKF and MPCM to provide a novel and efficient estimation method for nonlinear random switching systems. Note that the ST RMM involves three random variables:* $\tau_i[T_j^i]$, $v[T_j^i]$, *and* $r_i[T_j^i]$. *In the UKF estimation procedure,* $\tau_i[T_j^i]$ *plays a role in determining the switching probability* P_i *as described in Equations* (7.21), (7.23), *and* (7.24). $v[T_j^i]$, *and* $r_i[T_j^i]$ *are random maneuvers, and play roles in the random maneuver sampling procedure (i.e., Step 2) and future state prediction procedure when* $s[k-1] = 1$ *(i.e., Step 4). Note that if the remote UAV's previous maneuver information (*$v_2[k-1]$ *and* $\omega_2[k-1]$*) is unavailable, an additional estimation step is needed before processing* ***Step 3****. In particular,* $v_2[k-1]$ *and* $\omega_2[k-1]$ *need to be estimated from two consecutive previous states* $\hat{\mathbf{x}}_2[k-1]$ *and* $\hat{\mathbf{x}}_2[k-2]$ *as*

$$\hat{v}_2[k-1] = \sqrt{\hat{v}_{2x}^2[k-1] + \hat{v}_{2y}^2[k-1]},$$
$$\hat{\omega}_2[k-1] = (\hat{\theta}_2[k-1] - \hat{\theta}_2[k-2])/\delta,$$

where $\hat{v}_{2x}[k-1]$ *and* $\hat{v}_{2y}[k-1]$ *are the estimated velocities along the* x *and* y *axes respectively,* $\hat{v}_{2x}[k-1] = (\hat{x}_2[k-1] - \hat{x}_2[k-2])/\delta$, *and* $\hat{v}_{2y}[k-1] = (\hat{y}_2[k-1] - \hat{y}_2[k-2])/\delta$.

7.3.1.2 Adaptive optimal control

An online adaptive optimal controller is designed to maximize the expected value function (7.14) with the estimated system state. The existence and uniqueness of the optimal control policy is guaranteed here because of properties of the RSSI model (7.10)-(7.13) (shown in [2, Fig. 17]). In particular, to maximize $z_R[k]$, one needs to find $\theta_t[k]$ and $\theta_r[k]$ to maximize $G_{l|dBi}[k]$ as described in Equation (7.10). $G_{l|dBi}[k]$ is maximized when the heading angles of the two directional antennas are

selected as $\theta_t[k] = \gamma_t[k]$ and $\theta_r[k] = \gamma_r[k]$ respectively, where $\gamma_t[k]$ and $\gamma_r[k]$ are uniquely determined by the positions of two UAVs and environment-related shift angles as shown in Equations (7.12) and (7.13).

Because the uncertain parameters are independent from the system state at time k, the value function for UAV 1 can be further rewritten as

$$\begin{aligned} V_1(\mathbf{x}[k]) =& E[\sum_{l=k}^{k+N} \alpha^{l-k} z_R[l](\mathbf{x}[l], u_1[k], u_2^*[k])] \\ =& E[z_R[k](\mathbf{x}[k], u_1[k], u_2^*[k]) + \sum_{l=k+1}^{k+N} \alpha^{l-k} z_R[l](\mathbf{x}[l], u_1[k], u_2^*[k])]. \end{aligned} \tag{7.25}$$

where $u_2^*[k]$ is the optimal control policy of UAV 2.

The above equation can be solved backward-in-time using dynamic programming, or forward-in-time using RL [32, 146]. Here we use RL, in particular, the policy iteration method, to find the optimal control policy by iteratively conducting two steps: policy evaluation and policy improvement. The policy evaluation step is designed to solve the value function $V_1(\mathbf{x}[k])$ using Equation (7.25), given the current control policy. The policy improvement step is designed to find the best control policy to maximize the value function. The two steps are conducted iteratively until convergence.

Policy Evaluation

$$\begin{aligned} V_{1,j+1}(\mathbf{x}[k]) =& E[z_R[k](\mathbf{x}[k], u_{1,j}[k], u_2^*[k]) \\ &+ \sum_{l=k+1}^{k+N} \alpha^{l-k} z_{j,R}[l](\mathbf{x}[l], u_{1,j}[k], u_2^*[k])] \end{aligned} \tag{7.26}$$

Policy Improvement

$$\begin{aligned} u_{1,j+1}(\mathbf{x}[k]) =& \arg\max_{u_{1,j}[k]} E[z_R[k](\mathbf{x}[k], u_{1,j}[k], u_2^*[k]) \\ &+ \sum_{l=k+1}^{k+N} \alpha^{l-k} z_{j+1,R}[l](\mathbf{x}[l], u_{1,j}[k], u_2^*[k])] \end{aligned} \tag{7.27}$$

where j is the iteration step index, and $z_{j,R}[l](\mathbf{x}[l], u_{1,j}[k], u_2^*[k])$ is the RSSI model with parameters learned in the jth iteration step.

Note that Equation (7.26) involves three unknown parameters for the environment-specific RSSI model ($G_{t|dBi}^{max}$, $G_{t|dBi}^{min}$, and $\theta_{t_{env}}$), which need to be learned. In particular, for each iteration j, three time steps (k, $k+1$ and $k+2$) are needed to come up with three equations to iteratively solve for the three parameters. To solve the nonlinear equations, the Newton's method [157] is utilized here. Newton's method is a root-finding algorithm that iteratively finds better approximations to the roots of a real-valued function. To calculate the value function $V_{1,j+1}(\mathbf{x}[k])$ at each time step (Equation (7.26)), the uncertainty evaluation method needs to be utilized. To reduce the computational cost, we use the MPCM method here. In particular, define a system mapping $G_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])$ as the relationship between the value function and the random variables $v_2[T_j^2]$ and $r_2[T_j^2]$, i.e., $G_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2]) = z_R[k](\mathbf{x}[k], u_1[k], u_2^*[k]) + \sum_{l=k+1}^{k+N} \alpha^{l-k} z_R[l](\mathbf{x}[l], u_1[k], u_2^*[k])$. Then the value function $V_1(\mathbf{x}[k])$ can be estimated by evaluating the mean output of the system mapping using MPCM: $V_1(\mathbf{x}[k]) = E[G_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])]$. According to Lemma 7.3.1.1, the output mean of this system mapping can be obtained using the evaluated outputs of a reduced-order mapping $G'_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])$ at each selected MPCM point, according to the procedures described in [23, Section II-B].

Theorem 27. *Consider the random switching system shown in Equation* (7.4)*, with the value function given by Equation* (7.14)*. Given the current system state* $\mathbf{x}[k]$*, the solution found by applying the policy iteration of RL and approximating the value function using MPCM as shown in Equations* (7.26) *and* (7.27) *is the optimal control policy.*

Proof. Denote the optimal policies derived by evaluating the mean outputs of the reduced order mapping $G'_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])$ and the original value function $G_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])$ as $u_1'^*$ and u_1^* respectively, i.e.,

$$u_1'^* = \underset{u_1}{\operatorname{argmax}}\, E[G'_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])],$$

$$u_1^* = \underset{u_1}{\operatorname{argmax}}\, E[G_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])].$$

To prove this theorem, we need to prove that $u_1'^* = u_1^*$. This is equivalent to proving the following two statements: a) $\nexists u_1'^* \neq u_1^*$ such that

$$E[G'_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] > E[G_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])],$$

and b) $\nexists u_1'^* \neq u_1^*$ such that

$$E[G'_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] < E[G_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])].$$

Here we use a contradiction approach to prove the above two statements. To prove the first statement, we assume there exists $u_1'^* \neq u_1^*$ such that

$$E[G'_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] > E[G_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])].$$

According to Lemma 7.3.1.1, we have

$$\begin{aligned} E[G'_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] &= E[G_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] \\ &> E[G_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])], \end{aligned}$$

which violates the assumption $u_1^* = \operatorname{argmax}_{u_1} E[G_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])]$. Similarly, to prove that the second statement, we assume there exists $u_1'^* \neq u_1^*$ such that $E[G'_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] < E[G_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])]$. According to Lemma 7.3.1.1, we have

$$\begin{aligned} E[G_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] &= E[G'_{V_1}(\mathbf{x}[k], u_1^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])] \\ &> E[G'_{V_1}(\mathbf{x}[k], u_1'^*[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])], \end{aligned}$$

which violates the assumption $u_1'^* = \text{argmax}_{u_1} E[G'_{V_1}(\mathbf{x}[k], u_1[k], u_2^*[k], v_2[T_j^2], r_2[T_j^2])]$. As such, both statements a) and b) are true, and the results $u_1'^* = u_1^*$ are derived naturally. □

Theorem 28. *Consider the random switching system described in Equation* (7.4)*. Given the current system state* $\mathbf{x}[k]$*, the optimal policy found by the decentralized control algorithm (shown in Section 7.3.1.2) maximizes the global value function described in Equation* (7.14)*.*

Proof. Denote the global optimal control policy that maximizes the value function described in Equation (7.14) as $(u_{1,g}^*[k], u_{2,g}^*[k])$. We need to show that $u_1^*[k] = u_{1,g}^*[k]$ and $u_2^*[k] = u_{2,g}^*[k]$. According to Theorem 27, $u_1^*[k]$ is the optimal solution to Equation (7.14) under the assumption that $u_2[k] = u_2^*[k]$. The global optimal control policy $u_{1,g}^*[k]$ can be regarded as the decentralized optimal solution with the assumption that $u_2[k] = u_{2,g}^*[k]$. We show that for each time k, the optimal solution of UAV 1 is unique for any given $u_2[k]$.

Note that given any heading angle of the transmitting antenna $\theta_t[k]$, the optimal heading angle of the receiving antenna is $\gamma_r[k]$ to maximize $G_{l|dBi}[k]$ in Equation (7.11). The desired heading angle $\gamma_r[k]$, which is described in Equation (7.12), is decided uniquely by the positions of the two UAVs and the environment, instead of the transmitting antennas' heading angle. In such case, we have

$$\underset{u_1[k]}{\text{argmax}}\, z_R[k](\mathbf{x}[k], u_1[k], u_2^*[k]) = \underset{u_1[k]}{\text{argmax}}\, z_R[k](\mathbf{x}[k], u_1[k], u_{2,g}^*[k]),$$

$$\underset{u_2[k]}{\text{argmax}}\, z_R[k](\mathbf{x}[k], u_1^*[k], u_2[k]) = \underset{u_2[k]}{\text{argmax}}\, z_R[k](\mathbf{x}[k], u_{1,g}^*[k], u_2[k]),$$

which lead to the result that $u_1^*[k] = u_{1,g}^*[k]$ and $u_2^*[k] = u_{2,g}^*[k]$. The proof is completed. □

7.3.2 Using the learned RSSI model in both GPS-available and GPS-denied environments

With the learned RSSI model, the optimal solution can then be obtained in both GPS-available and GPS-denied environments. In a GPS-denied environment, the RSSI is the only measurement signal. In this case, the optimal control solution can be found following a similar procedure as shown in Section 7.3.1, by replacing $\mathbf{z}_{G,2}[k]$ and $h_{G,2}$ with $z_R[k]$ and h_R. In the GPS-available environment, GPS and RSSI measurements can be fused to estimate the system states, using a fuzzy-logic based fusion algorithm [144] to improve the reliability. The details are not repeated here.

Remark 16. *Note that RSSI is often calibrated for localization, in order to correct the environmental effects [158,159]. This calibration can be captured by a calibrated propagation constant [158], which is environment-related and is usually found by conducting experiments in the testing area prior to implementing the localization algorithm. In our study, this calibrated propagation constant is captured by the environment-related parameters in the RSSI model, i.e.,* $G^{max}_{t|dBi}, G^{min}_{t|dBi}$, *and* $\theta_{t_{env}}/\theta_{r_{env}}$. *In other words, the learning process we proposed in this chapter, which learns the environment-related parameters, can be regarded as an RSSI calibration process in the literature. With the learned parameters, the RSSI model is calibrated and then used in the antenna alignment algorithm.*

Remark 17. *Above distributed antenna control solution assumes a pair of UAVs in the ACDA system. When multiple UAVs are involved, the communications among UAVs can be realized using controllable multi-sector directional antennas or phased array antennas [4,160]. In such case, the communication network can be regarded as a collection of UAV pairs. As such, the study on the communication link between a pair of UAVs developed in this chapter is an important building block for a network of more than two UAVs.*

7.4 Remote UAV Uncertain Intention Estimation

In this section we provide an online uncertain intention estimation method to estimate the characteristics of the remote UAV's uncertain maneuver intentions. The estimation procedure includes two major steps, which is adopted from [161]: 1) estimation of the trajectory-specific maneuvers at each time instant, and 2) estimation of the pdfs of uncertain variables: $f_v(v_2[T_j^2])$, $f_r(r_2[T_j^2])$, and $f_\tau(\tau_2[T_j^2])$. The uncertain intention estimation solution provided in [161] is offline. We here enhance it to an online process to reduce computational costs.

7.4.1 Estimation of Trajectory-specific Maneuvers

We develop two estimation procedures to estimate the two types of random variables in the ST RMM respectively.

7.4.1.1 Estimation of type 1 random variables

Type 1 random variables (i.e., velocity $v_2[T_j^2]$ and turn radius $r_2[T_j^2]$) describe the movement characteristics of each maneuver, and can be estimated from the system states. Given two consecutive system states ($\mathbf{x}_2[k-1] = (x_2[k-1], y_2[k-1], \theta_2[k-1])$ and $\mathbf{x}_2[k] = (x_2[k], y_2[k], \theta_2[k])$), the type 1 random variables in the remote UAV system are estimated as

$$\hat{v}_2[k] = \sqrt{\hat{v}_{2x}^2[k] + \hat{v}_{2y}^2[k]},$$

$$\hat{r}_2[k] = \frac{\hat{v}_2[k]}{\hat{\omega}_2[k]},$$

where $\hat{\omega}_2[k]$ is the estimated angular velocity, and $\hat{\omega}_2[k] = (\theta_2[k] - \theta_2[k-1])/\delta$. $\hat{v}_{2x}[k]$ and $\hat{v}_{2y}[k]$ are the estimated velocities in x and y axes respectively, $\hat{v}_{2x}[k] = (\mathbf{x}[k] - x[k-1])/\delta$ and $\hat{v}_{2y}[k] = (y[k] - y[k-1])/\delta$.

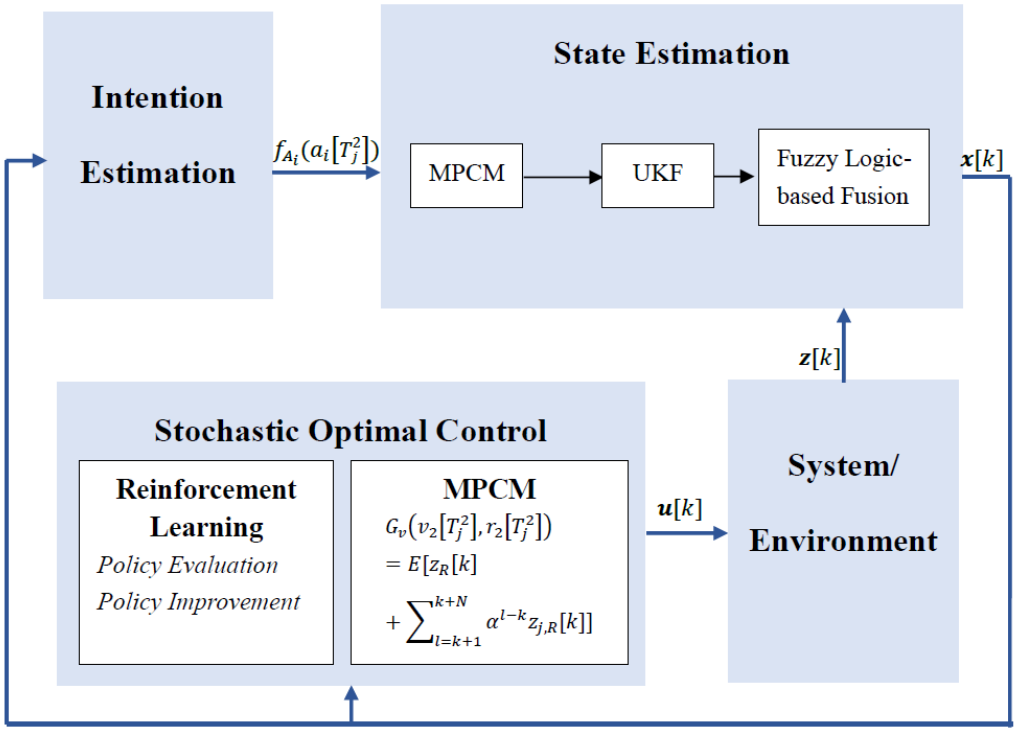


Figure 7.3. Illustration of the proposed algorithm .

7.4.1.2 Estimation of type 2 random variable

The type 2 random variable (i.e., travel time $\tau_2[T_j^2]$) describes how often the maneuvers are switched, and thus is estimated from the change of type 1 random variables. Different from [161] which uses the change of turn radius to find the length of each travel time, we here use the change of the angular velocity $\omega_2[T_j^2]$, which is affected by velocity $v_2[T_j^2]$ and turn radius $r_2[T_j^2]$. Therefore, to estimate $\tau_2[T_j^2]$, we scan the angular velocity $\omega_2[k]$ from $k = T_j^2$ at each time instant, until the change of $\omega_2[k]$ exceeds a threshold ω_2^{thrd}. The travel time interval at T_j^2 is estimated as $\tau_2[\hat{T_j^2}] = k - T_j^2$. The determination of ω_2^{thrd} has a significant impact on the estimation performance. In general, a smaller threshold improves the estimation accuracy but decreases the predictability of the underlining model. Please refer to [161] for the detailed discussion about the threshold selection.

7.4.2 Estimation of pdfs of Uncertain Intention Variables

The pdfs of uncertain intention variables in the remote UAV system can be estimated from the trajectory-specific maneuvers. In particular, assuming that the random variables $v_2[T_j^2]$, $\frac{1}{r_2[T_j^2]}$, and $\tau_2[T_j^2]$ follow the uniform, Gaussian, and Poisson distributions respectively, then the parameters in the distributions: v_{2min} and v_{2max} (minimum and maximum velocity constraints), μ_2 and σ_2 (mean and variance of $\frac{1}{r_2[T_j^2]}$), and λ_2 (expected value of $\tau_2[T_j^2]$), can be estimated from the following three steps.

Step 1: Estimate the velocity pdf. Denote the expectation and variance of velocity as μ_v and σ_v^2 respectively. μ_v and σ_v^2 can be estimated recursively as

$$\begin{aligned}\hat{\mu}_v[k] &= \frac{1}{k}\sum_{j=1}^{k}\hat{v}_2[j] = \frac{1}{k}(\sum_{j=1}^{k-1}\hat{v}_2[j] + \hat{v}_2[k]) \\ &= \frac{1}{k}\left((k-1)\hat{\mu}_v[k-1] + \hat{v}_2[k]\right) \\ &= \frac{k-1}{k}\hat{\mu}_v[k-1] + \frac{1}{k}\hat{v}_2[k],\end{aligned} \tag{7.28}$$

$$\begin{aligned}\hat{\sigma}_v^2[k] &= \frac{1}{k-1}\sum_{j=1}^{k}(\hat{v}_2[j] - \hat{\mu}_v[k])^2 \\ &= \frac{1}{k-1}(\sum_{j=1}^{k-1}(\hat{v}_2[j] - \hat{\mu}_v[k])^2 + (\hat{v}_2[k] - \hat{\mu}_v[k])^2) \\ &= \frac{1}{k-1}\left((k-2)\hat{\sigma}_v^2[k-1] + (\hat{v}_2[k] - \hat{\mu}_v[k])^2\right) \\ &= \frac{k-2}{k-1}\hat{\sigma}_v^2[k-1] + \frac{1}{k-1}(\hat{v}_2[k] - \hat{\mu}_v[k])^2.\end{aligned} \tag{7.29}$$

Remark 18. *Note that the sample mean of a random variable (i.e., $\frac{1}{k}\sum_{j=1}^{k}\hat{v}_2[j]$) is the minimum variance unbiased estimator (MVUE), and also, is the maximum likelihood estimator to μ_v [162]. To estimate σ_v^2, here we use the unbiased estimator ($\frac{1}{k-1}\sum_{j=1}^{k}(\hat{v}_2[j] - \hat{\mu}_v[k])^2$). The performance of the online estimation algorithm is as good as the offline proposed in [161] in terms of estimation accuracy. The equivalence*

of the two algorithms is shown in Equations (7.28) *and* (7.29). *Here we enhance the offline method to an online process to reduce the computational costs. The offline method needs to reuse all previous data in the UAV uncertain intention estimation whenever new data arrives, while the online method only utilizes the newest data.*

From the relation between v_{2min}, v_{2max} and μ_v, σ_v^2, the parameters in the velocity's pdf (v_{2min}, v_{2max}) can be estimated as

$$\begin{aligned}\hat{v}_{2min}[k] =& \hat{\mu}_v[k] - \sqrt{3}\hat{\sigma}_v[k] \\ =& \frac{k-1}{k}\hat{\mu}_v[k-1] + \frac{1}{k}\hat{v}_2[k] \\ & - \sqrt{\frac{3(k-2)}{k-1}\hat{\sigma}_v^2[k-1] + \frac{3}{k-1}(\hat{v}_2[k] - \hat{\mu}_v)^2},\end{aligned} \tag{7.30}$$

$$\begin{aligned}\hat{v}_{2max}[k] =& \hat{\mu}_v[k] + \sqrt{3}\hat{\sigma}_v[k] \\ =& \frac{k-1}{k}\hat{\mu}_v[k-1] + \frac{1}{k}\hat{v}_2[k] \\ & + \sqrt{\frac{3(k-2)}{k-1}\hat{\sigma}_v^2[k-1] + \frac{3}{k-1}(\hat{v}_2[k] - \hat{\mu}_v)^2}.\end{aligned} \tag{7.31}$$

Step 2: Estimate the radius pdf. The parameters in the radius pdf (μ_2 and σ_2^2) are estimated recursively using $\hat{r}_2[k]$ following a similar procedure as described in Equations (7.28) and (7.29) as

$$\hat{\mu}_2[k] = \frac{k-1}{k}\hat{\mu}_2[k-1] + \frac{1}{k}\frac{1}{\hat{r}_2[k]}, \tag{7.32}$$

$$\hat{\sigma}_2^2[k] = \frac{k-2}{k-1}\hat{\sigma}_2^2[k-1] + \frac{1}{k-1}(\frac{1}{\hat{r}_2}[k] - \hat{\mu}_2[k])^2. \tag{7.33}$$

Step 3: Estimate the travel time pdf. λ_2 is the only parameter in the Poisson distribution, and can be estimated recursively from the mean of $\hat{\tau}_2[T_j^2]$ as

$$\hat{\lambda}_2[j] = \frac{j\hat{\lambda}_2[j-1]}{j - 1 + \hat{\lambda}_2[j-1]\hat{\tau}_2[T_j^2]}. \tag{7.34}$$

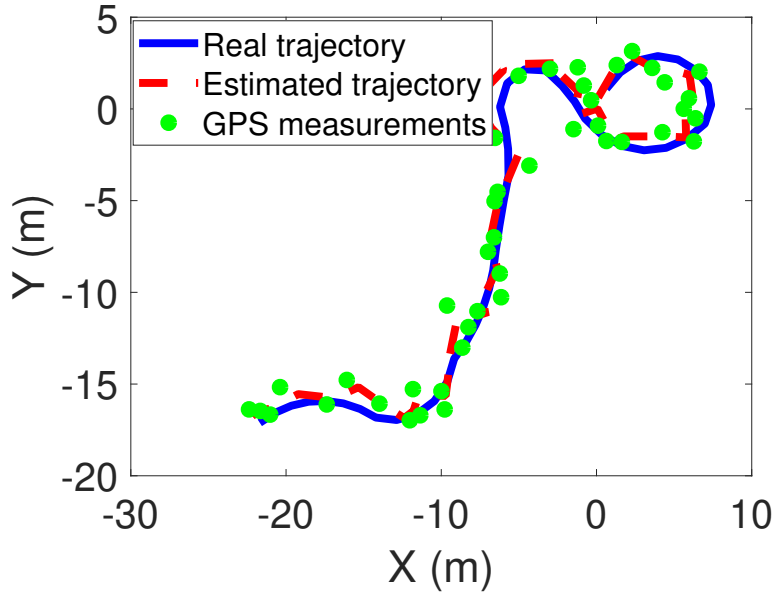
Remark 19. *The uncertain intention estimation procedure can be implemented together with the stochastic optimal control procedure described in Section 7.3. The*

overall algorithm structure is described in Figure 7.3. We also note that because the uncertain intention is estimated from the system states, which are estimated from the measurements, it is suggested to conduct the uncertain intention estimation procedure in a GPS-available environment, which helps to improve the reliability of the estimated system states.

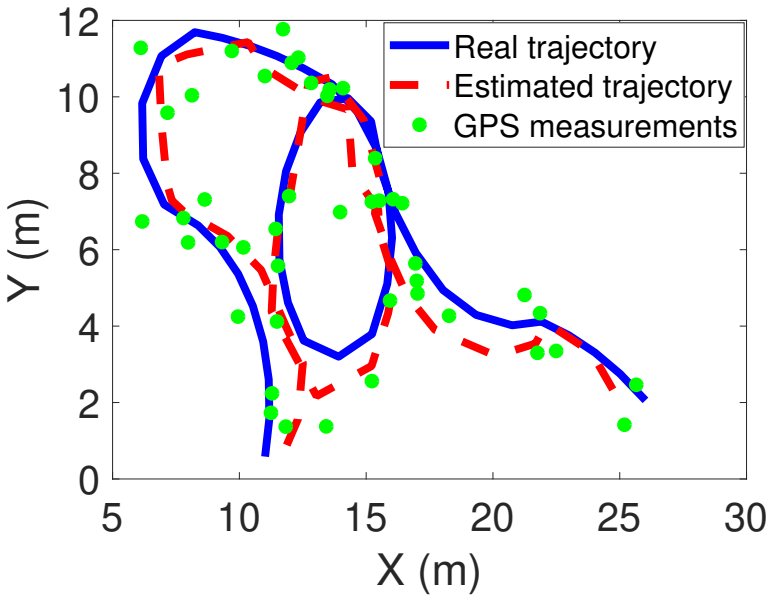
7.5 Simulation Studies

In this section, we conduct simulation studies to illustrate and validate the results and algorithms developed in this chapter. Two UAVs move in a 2-D airspace following the ST RMM independently. Two directional antennas of the same type are mounted on the two UAVs respectively. The design parameters of the directional antennas are selected as $n = 8$, and $d_a = \frac{\lambda}{10}$.

We first simulate the case when the GPS is available but the RSSI model is unknown. Gaussian noises are added to the GPS measurements. Estimation for UAV 1 is based on UKF with known maneuver ($v_1[k]$ and $r_1[k]$), while the estimation for UAV 2 is based on the integration of UKF and MPCM as described in Section 7.3.1 with unknown $v_2[k]$ and $r_2[k]$. Figures 7.4(a) and 7.4(b) show the trajectories of UAV 1 and UAV 2 respectively in one realization with the simulation time $T = 45s$ and sampling period $\delta = 1s$. To find the statistics of the estimation performance, 10 realizations with randomly generated trajectories are conducted. The mean estimation distance errors for the two UAVs are calculated over all realizations as $e_1 = 0.84m$ and $e_2 = 0.89m$ respectively. It can be seen from the simulations that the estimated trajectories for UAVs 1 and 2 are both close to their real trajectories, indicating that the proposed state estimation algorithm performs well in both known and unknown maneuver cases.



(a)

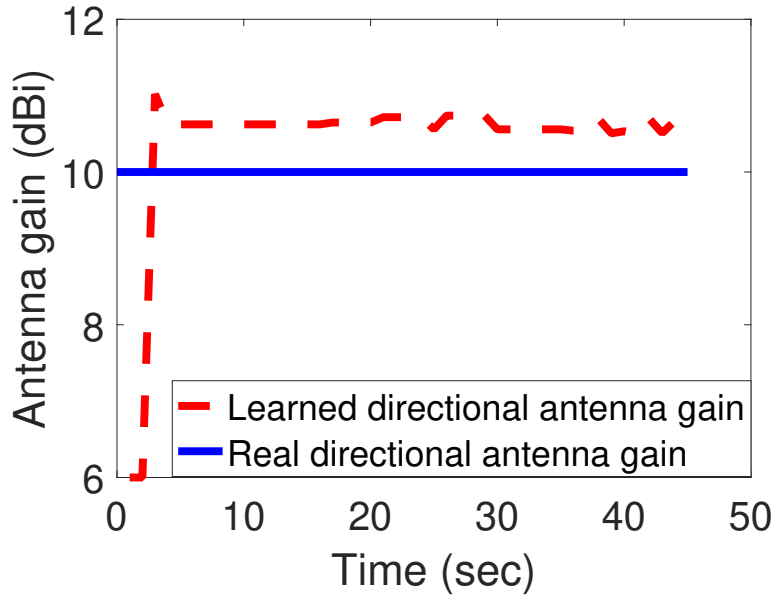


(b)

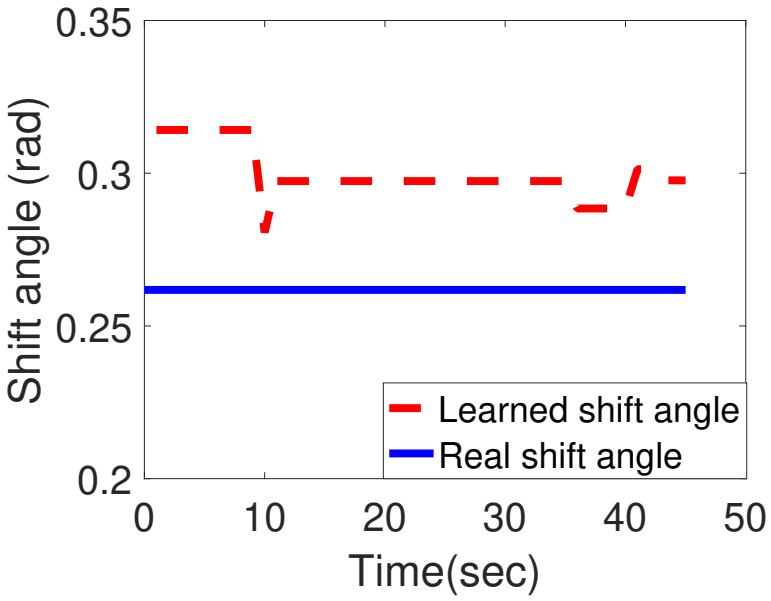
Figure 7.4. (a) Trajectories of UAV 1, and (b) Trajectories of UAV 2. The blue solid curves are real trajectories, red dotted curves are estimated trajectories, and green dots are GPS measurements.

With the estimated states, we simulate the RL-based stochastic optimal control algorithm. To simulate the long-distance communication scenario, the minimum received signal strength is assumed to be 0, and in this case, the directional antennas' minimum gain ($G^{min}_{t|dBi}$) can be calculated accordingly. Figures 7.5(a) and 7.5(b) show the learned environment-specific antennas' maximum gain ($G^{max}_{t|dBi}$) and the shift angle caused by the environment ($\theta_{t_{env}}$) respectively. Gaussian noises are added to the RSSI measurements. To avoid unnecessary divergence, we limit the maximum values of the two parameters. In particular, we assume the directional antenna's maximum gain is no more than the maximum gain given in the data sheet, and the environment-specific shift angle is no more than 20 degrees. As shown in the figures, the learned parameters are very close to their true values, which indicates the effectiveness of the learning algorithm. Figures 7.6(a) and 7.6(b) show the derived optimal heading angles of the local directional antenna and the heading angle errors between the derived and real optimal heading angles in one realization. The small angle errors indicate the good performance of the proposed RL-based stochastic optimal control algorithm.

With the learned RSSI model, we simulate the proposed stochastic optimal control algorithms in both GPS-denied and GPS-available environments. Note that in the GPS-denied environment, RSSI is the only measurement signal, while in the GPS-available case, both GPS and RSSI signals are fused. Figures 7.7 and 7.8 show the performance of state estimation and optimal control algorithms respectively, in both GPS-denied and GPS-available environments.We have simulated 10 realizations with randomly generated UAV trajectories, and calculated the mean estimation errors, RSSI signals, and heading angle errors over all 10 realizations. The system state estimation performance is shown in Table 7.1 and Figure 7.9(a), where "GPS", "RSSI", and "Fusion" represent the UKF-based state estimation using only GPS, only RSSI, and both GPS and RSSI signals respectively. The column "GPS signals" shows the

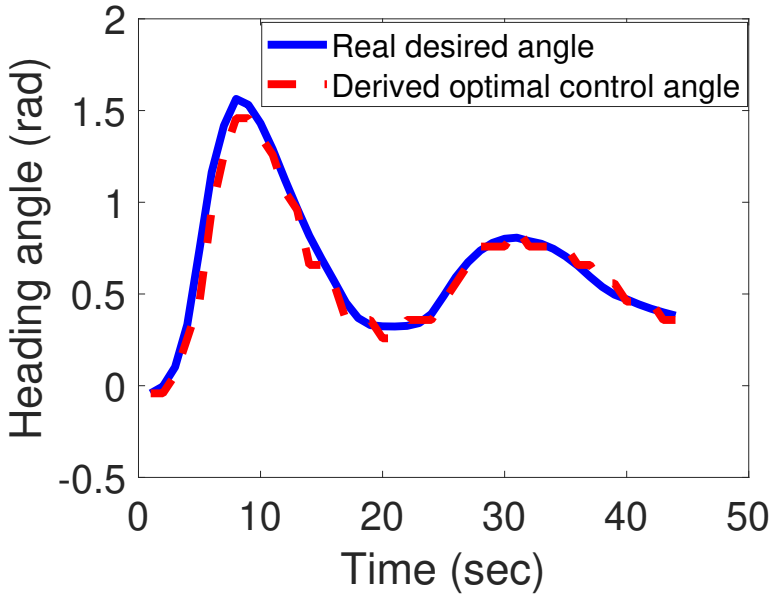


(a)

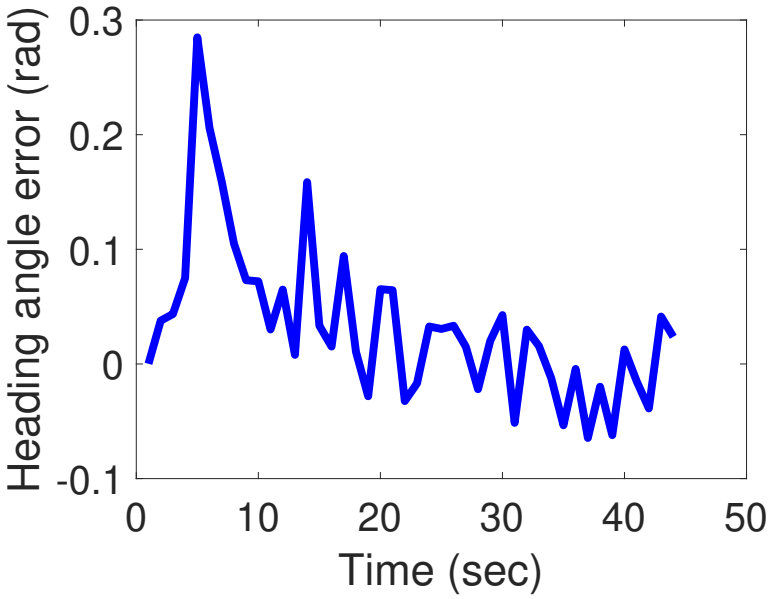


(b)

Figure 7.5. Learned environment-specific (a) maximum directional antenna gain ($G^{max}_{t|dBm}$), and (b) shift angle (θ_{env}) in the RSSI model. The blue solid lines and red dotted curves represent the real and learned parameters respectively.

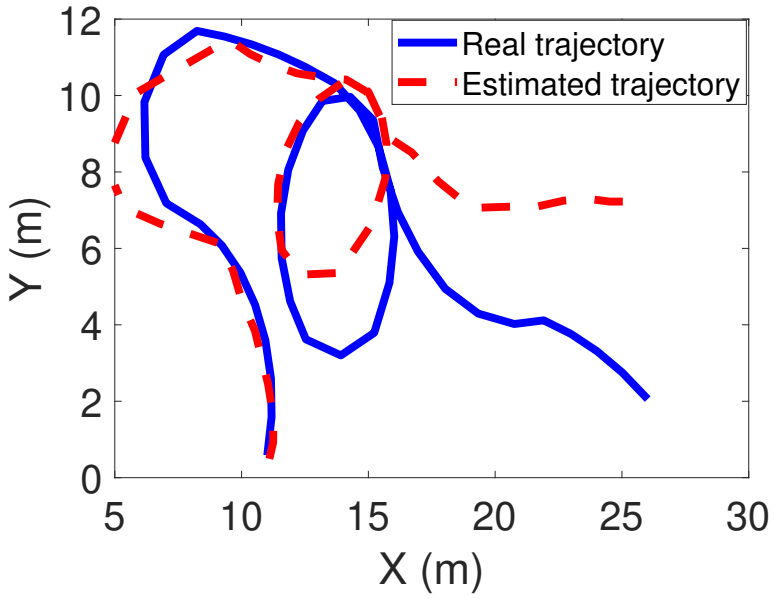


(a)

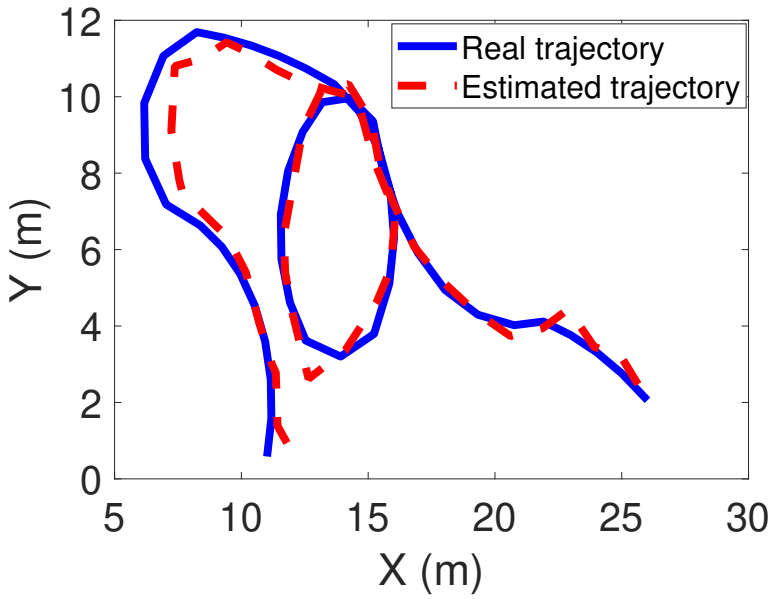


(b)

Figure 7.6. (a) Obtained optimal heading angles with GPS signals and unknown RSSI model. The blue solid curve is the real optimal angles, and the red dotted curve is the obtained optimal angles. (b) Heading angle errors between the derived heading angles and the real optimal heading angles.

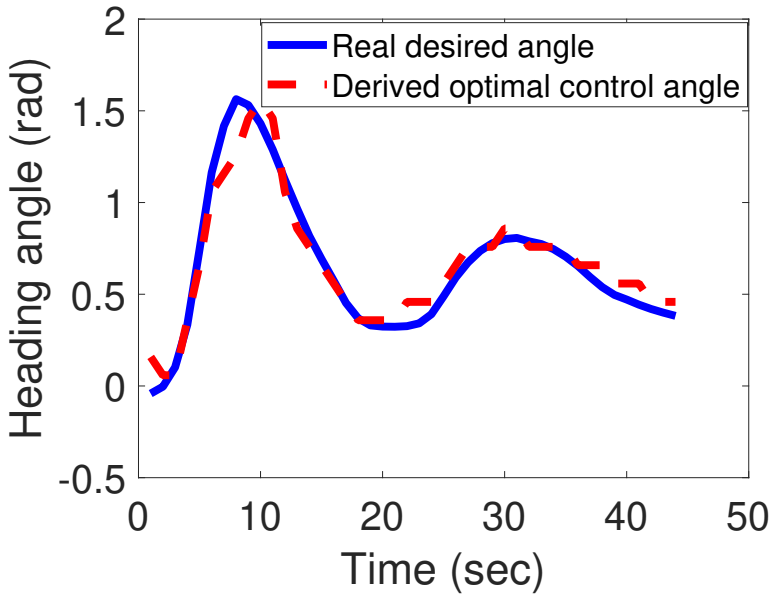


(a)

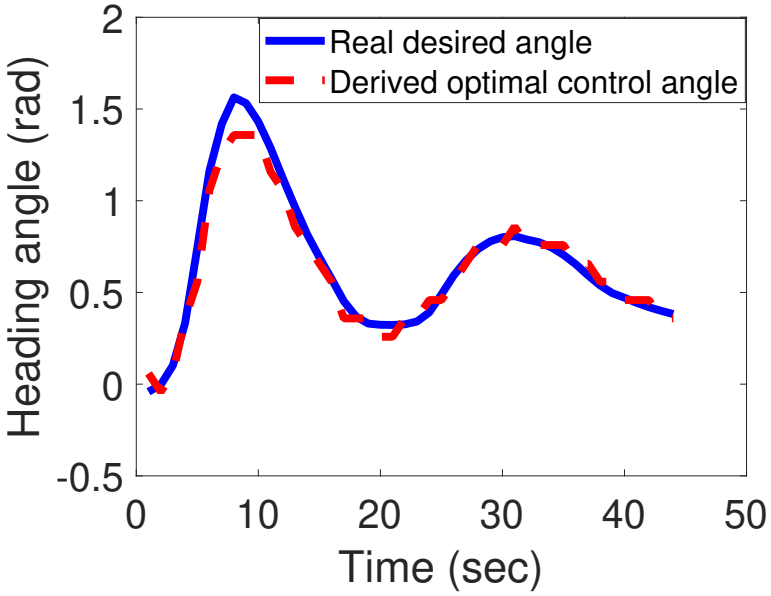


(b)

Figure 7.7. (a) Trajectories of UAV 2 in (a) GPS-denied, and (b) GPS-available environments. The blue solid curves are the real trajectories, and the red dotted curves are the estimated trajectories.



(a)



(b)

Figure 7.8. Obtained optimal heading angles in (a) GPS-denied, and (b) GPS-available environments. The blue solid and red dotted curves are the real optimal heading angles and derived heading angels respectively.

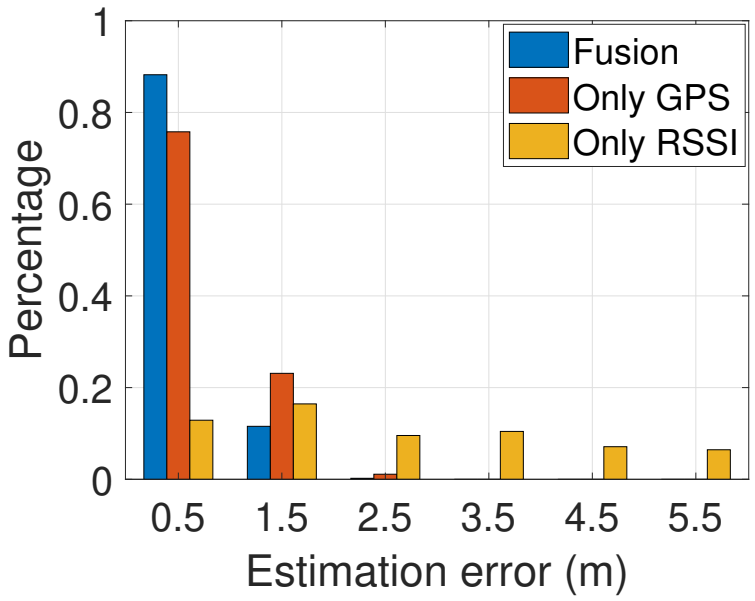
raw GPS measurements. The optimal control performance is shown in Table 7.2 and Figure 7.9(b), where "RSSI" and "Fusion" represent the control algorithms based on only RSSI, and both RSSI and GPS respectively. It can be seen from the tables and plots that: 1) the estimated system states and derived heading angles are very close to their real states and optimal heading angles in both GPS-denied and GPS-available cases, indicating that the proposed algorithms work well in both GPS-available and GPS-denied environments; 2) the estimation errors and heading angle errors in the GPS-available case are much smaller than that in the GPS-denied case, indicating that the fusion of the GPS and RSSI promises a better performance.

To provide comparative studies, we also simulate the GPS alignment-based directional antenna control algorithm developed in [133]. In this algorithm, each directional antenna points towards the GPS location of the other UAV to align the directional antennas, and RSSI is not used as a measurement signal nor value function. The control performance of this GPS alignment-based algorithm is also shown in Table 7.2. The barplots of the controlled heading angle errors are shown in Figure 7.9(b). It can be seen from the tables and plots that the optimal control algorithm developed in our chapter performs much better than the GPS alignment-based algorithm, with larger RSSI signals and less heading angle errors.

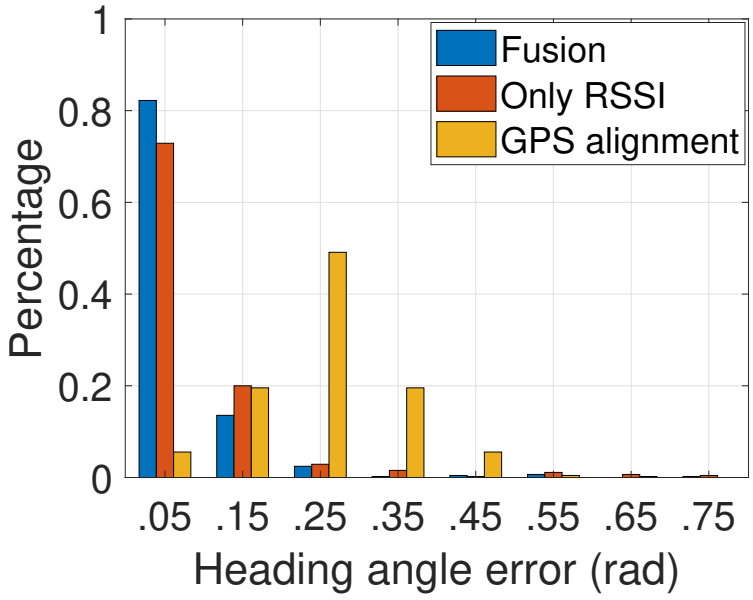
Table 7.1. Estimation Performance

	UKF-based			GPS signals
	GPS	RSSI	Fusion	
Mean distance error (m)	0.89	5.13	0.56	1.25

Finally we simulate the remote UAV uncertain intention estimation algorithm. The total simulation time in this part is set as $T = 10$ minutes, with the sampling



(a)



(b)

Figure 7.9. Barplots of (a) Estimation errors, and (b) Heading angle errors.

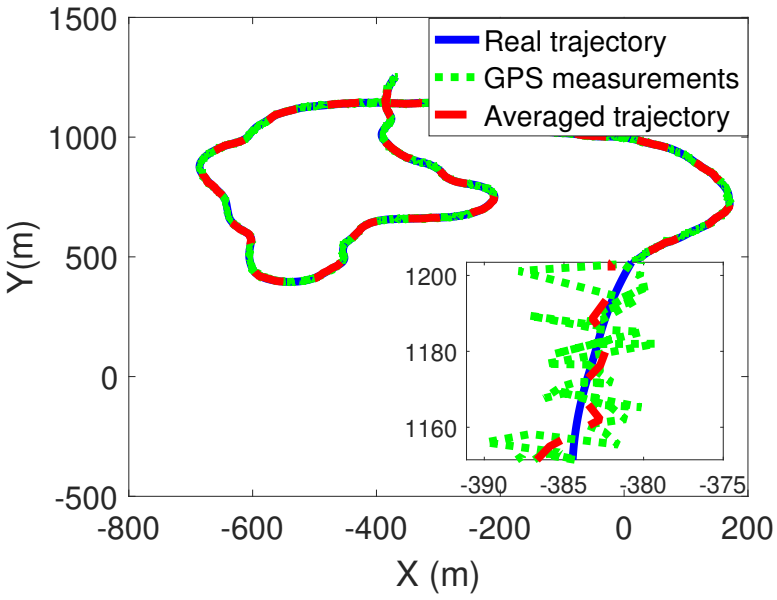
Table 7.2. Control Performance

	RL-based		GPS alignment-based
	RSSI	Fusion	
Mean RSSI (dBm)	-31	-29	-47
Mean angle error (rad)	0.09	0.07	0.3

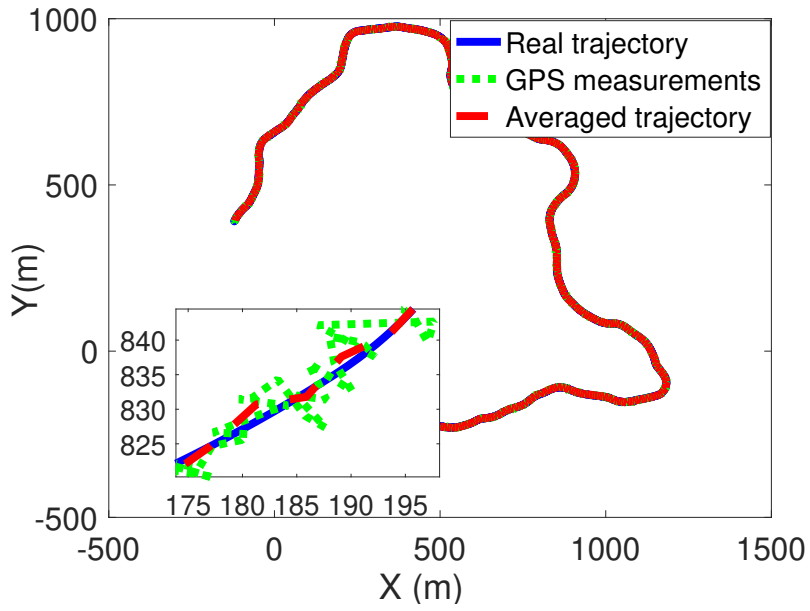
period $\delta = 1s$. The system states are estimated from the GPS measurements by adopting the moving average method. Figures 7.10(a) and 7.10(b) show trajectories of the two UAVs respectively. The performance of the uncertain intention estimation algorithm is shown in Table 7.3. Note that $v_2[T_i^2]$, $\frac{1}{r_2[T_i^2]}$, and $\tau_2[T_i^2]$ follow uniform, Gaussian, and Possion distributions respectively. As such, the parameters to be estimated in their pdfs are: μ_v and σ_v for $v_2[T_i^2]$, μ_2 and σ_2 for $\frac{1}{r_2[T_i^2]}$, and λ_2 for $\tau_2[T_i^2]$ respectively. It can be seen from the table that the estimated means of $v_2[T_i^2]$, $\frac{1}{r_2[T_i^2]}$, and $\tau_2[T_i^2]$ match with their real mean values perfectly, indicating the effectiveness of the proposed estimation algorithm. The estimated variance of $v_2[T_i^2]$ and $1r_2[T_i^2]$ show small biases to their real values, cauessed by Gaussian GPS noises.

Table 7.3. Performance of Online Intention Estimation

Random variables	$v_2[T_i^2]$		$\frac{1}{r_2[T_i^2]}$		$\tau_2[T_i^2]$
Parameters to be estimated	μ_v	σ_v^2	μ_2	σ_2^2	λ_2
Estimated value	12.7	6.3	10^{-4}	10^{-3}	2.07
Real value	12.5	2.1	0	10^{-4}	2



(a)



(b)

Figure 7.10. Trajectories of (a) UAV 1, and (b) Trajectories of UAV 2. The blue solid curves are real trajectories, red dotted curves are estimated trajectories, and green dots are GPS measurements.

CHAPTER 8

STATISTICAL PROPERTIES OF UNMANNED AERIAL VEHICLE NETWORKS SUBJECT TO SENSE-AND-AVOID SAFETY PROTOCOLS

8.1 Introduction

UAV technology has demonstrated its value in broad commercial applications, such as sports coverage, cargo transport, precision agriculture, public safety, on-demand communication provision, and structure health monitoring [133, 163–167]. The global commercial UAV Market is projected to reach $52.30 Billion by 2025 [155]. With the new, small UAV rules released by the Federal Aviation Administration (FAA) in August 2016 [168], we foresee a dense UAV use in the National Airspace System (NAS). Along with this trend, new research directions that deal with multiple UAVs in a dense airspace become urgent, such as UAV networking and UAV traffic management (UTM).

RMMs have been widely used for networking studies. Examples include Random Direction (RD), Random Walk (RW), GUAVs-Markov (GM), and Smooth Turn (ST) developed specifically for fixed-wing UAVs [169–174]. Please refer to our survey paper [169] on the RMMs developed for different UAV applications ranging from search, rescue, and reconnaissance, to patrolling, cargo, and AN backbone. These RMMs capture the random mobility patterns of moving agents, and have commonly been used as the evaluation and design foundation of mobile ad hoc networks (MANET), vehicular ad hoc networks (VANET), and UAV networks (or called flying ad Hoc networks, FANET), from which important statistics can be derived, such as node distribution, inter-vehicle distance distribution, and link/path lifetime [175,176].

We note that all of these existing RMM studies assume the independent movement of mobile agents. This assumption does not hold for UAVs. In particular, in order to maintain airspace safety, UAVs must be equipped with sense and avoid (S&A) capabilities. This S&A feature is a critical difference between FANET and MANET. S&A fundamentally changes the statistics for networking, and its impact should be explored. In this chapter, we develop an analytical framework of RD RMMs equipped with S&A protocols, and analyze its statistical performance, such as node distribution and inter-vehicle distance distribution. We note that the analysis becomes more complicated when the independent movement assumption is removed.

This modeling framework is further used in this chapter for UTM studies. UTM is very different from traditional air traffic management (ATM) [177, 178]. Unlike commercial flights which have pre-defined flight plans and rather deterministic flight trajectories, UAVs in the low-altitude airspace are featured by their highly flexible, variable and uncertain movement patterns. Such features, on the other hand, significantly complicate UTM. Concepts such as "highway in the sky" are borrowed from traditional ATM to simplify the UTM architecture, however, such "infrastructure" constraints limit UAV flexibility and contradict their missions. Little is known about the limitations of airspace capacity subject to highly flexible UAV operations. A capacity concept for UTM was proposed in [179], which assumes unified flow directions for all UAVs. In [180], a phase-transition-based capacity concept was proposed based on simulation studies of randomly generated source-to-destination UAV trajectories.

The modeling framework of RMMs equipped with physical S&A protocols is proposed in this chapter for UAV networking and UTM studies. The modeling framework, first of its kind per knowledge of the authors, succinctly captures the flexible, variable, and uncertain movement patterns of UAVs subject to the separation safety constraints. Further contributions of this chapter are summarized as follows. For

the RD RMM equipped with a commonly used S&A protocol, named sense-and-stop (S&S), we develop statistics that are critical to networking studies, such as stationary node distribution and stationary inter-vehicle distance distribution, using the Markov analysis. This study provides knowledge on the impact of S&A protocols to critical UAV networking statistics. In addition, we define collision probabilities and airspace capacity concepts for UAVs based on the inter-vehicle distance distribution, and derive their expressions. This new UAV airspace capacity concept captures the flexibility of UAV operations as it only relies on the S&A protocols to maintain airspace safety, with no other constraints being enforced. This capacity concept provides us insight on the limitation of airspace density for highly flexible and autonomous UAVs, which are very different from traditional air traffic. This capacity analytical framework mathematically bridges local autonomy with global airspace capacity, and permits insightful impact analysis of local autonomy configurations to achieve effective UAV airspace capacity management.

The remainder of this chapter is organized as follows. Section 8.2 describes both the independent RD RMM and the RD RMM equipped with the S&S protocol. Section 8.3 analyzes the statistical properties of both RMMs in terms of node distribution and inter-vehicle distance distribution. Section 8.4 analyzes the collision probabilities and airspace capacity for both RMMs. Section 8.5 includes the model configuration impact analysis. Simulation studies are included throughout the paper to help illustrate the analytical results.

8.2 The Modeling Framework

In this section, we first describe the independent RD RMM, and then introduce the RD RMM equipped with the S&S protocol to capture the safety constraint of flexible UAV operations.

8.2.1 Independent Random Direction Mobility Model

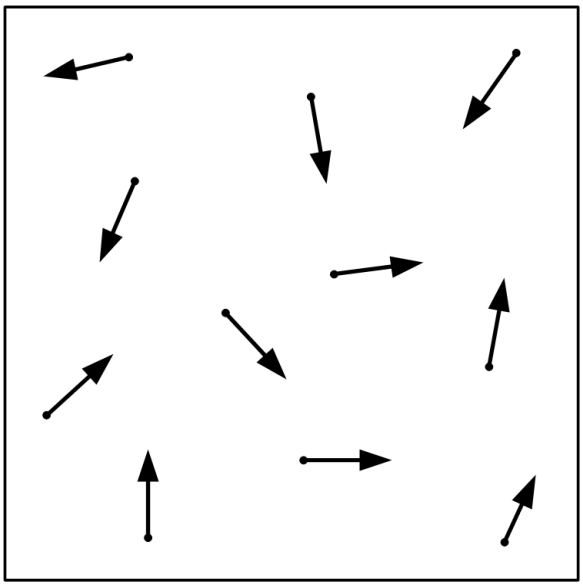
In the independent RD RMM widely used in the literature, UAVs travel independently in an airspace $[0, B)^2$ (Figure 8.1(a)). A comprehensive description of RMMs used for UAVs can be found in the survey paper [169]. At each time instant $1,\ 2, ...,\ k$, UAV i selects a heading direction $\Theta_i[k]$ from $[0, 2\pi)$ randomly, and moves along that direction with a constant heading speed V. $\Theta_i[k]$ is uniformly distributed in $[0, 2\pi)$, $\forall i, k$. $X_i[k]$ and $Y_i[k]$ denote the stochastic processes for UAV i's location along the x and y axes.

We use the widely adopted wrap-around boundary model to avoid the border effect [170]. When a UAV hits the boundary, it wraps around and appears at the opposite side with the same velocity and heading direction (Figure 8.1(b)). This wrap-around model is suitable for large simulation regions and is analysis-friendly. With this boundary model, the inter-vehicle cyclic relative position $S_{i,j}[k] = (\Delta X_{i,j}[k], \Delta Y_{i,j}[k])$ and distance $D_{i,j}[k]$ between two UAVs i and j at time k can be calculated as [181]:

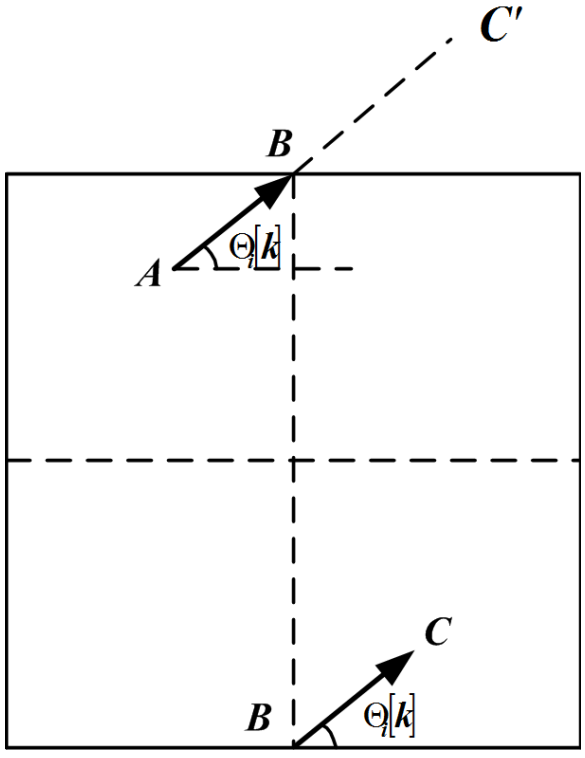
$$\begin{aligned} \Delta X_{i,j}[k] &= \min(|X_i[k] - X_j[k]|, B - |X_i[k] - X_j[k]|), \\ \Delta Y_{i,j}[k] &= \min(|Y_i[k] - Y_j[k]|, B - |Y_i[k] - Y_j[k]|), \\ D_{i,j}[k] &= (\Delta X_{i,j}[k]^2 + \Delta Y_{i,j}[k]^2)^{\frac{1}{2}}. \end{aligned} \tag{8.1}$$

8.2.2 Random Direction Mobility Model Equipped with S&S Protocol

Safety constraints are critical for UAV operations. The FAA "right-of-way" rules state that for vehicles of the same category and operating at the same altitude, the aircraft to the right has the right-of-way [168]. In the literature, many papers focus on the development of S&A safety protocols when two UAVs encounter (see e.g., [182, 183]). However, the successful collision avoidance of two UAVs may lead to a collision that involves other UAVs. As such, it is the purpose of this chapter



(a)



(b)

Figure 8.1. Illustration of (a) a 2-D airspace with UAV mobility captured by the independent RD RMMs, and (b) the wrap-around boundary model.

to exploit the relationship between local S&A protocol's impact and global airspace capacity.

Here, we implement a "right-of-way" rule, named 'sense-and-stop" (S&S), similar to the hovering strategy [184]. Denote the sensing distance (or observing distance) as d_o, which is much smaller than $\frac{B}{2}$. The RD RMM equipped with S&S protocol works as follows: a) when the inter-vehicle distance between two UAVs is greater than sensing distance, i.e., $D_{i,j}[k] > d_o$, each UAV moves independently according to the independent RD RMM; b) when $D_{i,j}[k] \leq d_o$, the vehicle to the left (i.e., with a smaller x location) stops, and the other vehicle follows the independent RD RMM until $D_{i,j}[k] > d_o$. When multiple UAVs are involved, they are considered as a collection of UAV pairs. For any UAV i, if there exists at least one UAV (denoted as j, $j \neq i$) satisfying $D_{i,j}[k] \leq d_o$, and $X_i[k] < X_j[k]$, then UAV i stops. In addition, for any UAV i, if there exists at least one UAV j satisfying $D_{i,j}[k] \leq d_o$ and $X_i[k] = X_j[k]$, then a small noise is added to $X_i[k]$ to differ their locations in x coordinate. As there always exists a UAV with the maximal x location, we can ensure that at least one UAV moves at each time instant k in the airspace and the dead-lock phenomenon does not occur.

8.3 Analysis of Network Statistics

In this section, we study the statistical properties of both RD RMMs critical to the networking studies, in terms of stationary node distribution and inter-vehicle distance distribution.

8.3.1 Stationary Node Distribution

Theorem 29 states that the stationary joint node distribution for the independent RD RMM is uniform, and Theorem 30 states that the stationary node distribution for each UAV following the RD RMM equipped with S&S protocol is uniform.

Theorem 29. *Each of the N UAVs in an airspace $[0, B)^2$ moves independently according to the RD RMM. The stationary joint node distribution is uniform, regardless of the initial node distribution.*

Proof. Define a Morkov process based on the location and heading direction of UAV i when it moves with the independent RD RMM, $\hat{S}_i^b[k] = (X_i[k], Y_i[k], \Theta_i[k])$. The Markov chain $\hat{S}_i^b[k]$ is aperiodic, Φ-irreducible, and Harris recurrent, and therefore, there exists a unique stationary distribution. Following a similar argument as in Paper [170], Proposition 4.2, it can be proved that the stationary node distribution is uniform, based on the transition properties of the Markov process. The stationary node probability distribution function (PDF) is

$$\lim_{k\to\infty} P(X_i[k] < x, Y_i[k] < y, \Theta_i[k] < \theta) = \frac{xy\theta}{2\pi B^2}. \tag{8.2}$$

Since the N UAVs move independently, the joint node distribution is a multiplication of N individual node distributions. A simple argument leads to the conclusion that the N UAVs' stationary node distribution is uniform. □

Theorem 30. *Each of the N UAVs in an airspace $[0, B)^2$ follows the RD RMM equipped with the S&S protocol. The stationary location distribution for each UAV i is uniform, regardless of the initial node distribution.*

Proof. Define a Morkov process based on the location of UAV i when it moves according to the RD RMM equipped with the S&S protocol, $\hat{S}_i^S[k] = (X_i[k], Y_i[k])$. The Markov chain $\hat{S}_i^S[k]$ is aperiodic, Φ-irreducible, and Harris recurrent when N is finite,

and therefore, there exists a unique stationary distribution. To find the stationary location distribution of UAV i, we introduce a set $\mathcal{S}_s[k]$ to hold all the UAV location pairs related to UAV i that satisfy the following condition: there exists at least one UAV (denoted as j, $j \neq i$) satisfying $D_{i,j}[k] \leq d_o$ and $X_i[k] < X_j[k]$. $\bar{\mathcal{S}}_s[k]$ is the complement of $\mathcal{S}_s[k]$. The stationary location distribution along the x axis can be described as follows.

$$\begin{aligned} &\lim_{k\to\infty} P(X_i[k] < x_i) \\ &= \lim_{k\to\infty} P(X_i[k] < x_i|\bar{\mathcal{S}}_s[k])P(\bar{\mathcal{S}}_s[k]) + \lim_{k\to\infty} P(X_i[k] < x_i|\mathcal{S}_s[k])P(\mathcal{S}_s[k]). \end{aligned} \tag{8.3}$$

To prove $\lim_{k\to\infty} X_i[k]$ is uniformly distributed, we only need to show that $\lim_{k\to\infty} P(X_i[k] < x_i|\bar{\mathcal{S}}_s[k]) = x_i$ and $\lim_{k\to\infty} P(X_i[k] < x_i|\mathcal{S}_s[k]) = x_i$. The proof of the first statement $\lim_{k\to\infty} P(X_i[k] < x_i|\bar{\mathcal{S}}_s[k]) = x_i$ is straightforward using Theorem 29, as UAV i moves independently in this case. To prove the second statement, we note that as UAV i stops at time k, $\lim_{k\to\infty} P(X_i[k] < x_i|\mathcal{S}_s[k])$ is the same as the conditional probability right before the UAV enters the stop state. As UAV i moves randomly at that time step, the proof of the first statement leads to the conclusion that $\lim_{k\to\infty} P(X_i[k] < x_i|\mathcal{S}_s[k]) = x_i$. The proofs for location along y axis follow a similar argument. □

Remark 20. Although the location of each individual UAV is uniformly distributed for the RD RMM equipped with S&S protocol, the joint distribution of N UAV locations is not uniform anymore. The introduction of the S&S protocol removes the independence of UAV trajectories, and alters the joint distribution. This property is studied in greater details in Section 8.3.2.

8.3.2 Stationary Inter-vehicle Distance Distribution

In this section, we study the impact of the S&S protocol to stationary inter-vehicle distance distribution. Theorem 31 suggests that the distribution is uniform

for the independent RD RMM, while the uniformity property does not hold for the RD RMM equipped with S&S protocol as shown in Theorem 32.

Theorem 31. *Each of the N UAVs in an airspace $[0, B)^2$ moves independently according to the independent RD RMM. The stationary probabilistic density function (pdf) of the cyclic inter-vehicle distance for two UAVs i and j, $D_{i,j}[k]$, denoted as $f_D^b(d)$, is*

$$f_D^b(d) = \lim_{k\to\infty} f(D_{i,j}[k] = d) = \begin{cases} \frac{2\pi d}{B^2} & 0 \leqslant d < \frac{B}{2} \\ \frac{4(\frac{\pi}{2} - 2arccos(\frac{B}{2d}))d}{B^2} & \frac{B}{2} \leqslant d < \frac{\sqrt{2}B}{2} \end{cases}. \tag{8.4}$$

Proof. Define a Morkov process based on the cyclic relative position between the UAV i and j when they move with the independent RD RMM, $S^b[k] = (\Delta X_{i,j}[k], \Delta Y_{i,j}[k])$. As the deadlock phenomon does not occur, the Markov chain $S^b[k]$ is aperiodic, Φ-irreducible, and Harris recurrent, and therefore, there exists a unique stationary distribution. We first calculate the pdfs for the stationary cyclic relative positions along the x and y axes, $\lim_{k\to\infty} f(\Delta X_{i,j}[k] = \Delta x)$ and $\lim_{k\to\infty} f(\Delta Y_{i,j}[k] = \Delta y)$. We remove the subscript i, j for the simplicity of presentation when it does not cause confusion. Equation (8.1) leads to

$$\begin{aligned} \Delta X[k] &= \begin{cases} |X_i[k] - X_j[k]| & |X_i[k] - X_j[k]| \leqslant \frac{B}{2} \\ B - |X_i[k] - X_j[k]| & |X_i[k] - X_j[k]| > \frac{B}{2} \end{cases} \\ &= \begin{cases} |\Delta X_E[k]| & |\Delta X_E[k]| \leqslant \frac{B}{2} \\ B - |\Delta X_E[k]| & |\Delta X_E[k]| > \frac{B}{2} \end{cases}, \end{aligned} \tag{8.5}$$

where $\Delta X_E[k]$ is the Euclidean relative position between UAVs i and j along the x axis, i.e., $\Delta X_E[k] = X_i[k] - X_j[k]$. Now we find the pdf of $|\Delta X_E[k]|$. Theorem 29 leads to

$$\lim_{k\to\infty} f(X_i[k] = x) = \frac{1}{B} \qquad (0 \leqslant x < B). \tag{8.6}$$

As UAVs i and j move independently, the stationary pdf of $|\Delta X_E[k]|$ can thus be derived from Equation (8.6) as follows:

$$\begin{aligned}
&\lim_{k\to\infty} f(|\Delta X_E[k]| = \Delta x) \\
&= 2\int_0^{B-\Delta x} \lim_{k\to\infty} f(X_i[k] = x + \Delta x, X_j[k] = x)dx \\
&= 2\int_0^{B-\Delta x} \lim_{k\to\infty} f(X_i[k] = x + \Delta x) f(X_j[k] = x)dx \\
&= 2\int_0^{B-\Delta x} \frac{1}{B}\frac{1}{B}dx \\
&= \frac{2(B-\Delta x)}{B^2} \qquad (0 \leqslant \Delta x < B).
\end{aligned} \tag{8.7}$$

Equation (8.7) leads to the stationary pdf of $\Delta X[k]$, according to the relationship between cyclic distance and Euclidean distance (see Equation (8.5) and Figure 8.2(a)):

$$\begin{aligned}
&\lim_{k\to\infty} f(\Delta X[k] = \Delta x) \\
&= \lim_{k\to\infty} f(|\Delta X_E[k]| = \Delta x) + f(|\Delta X_E[k]| = B - \Delta x) \\
&= \frac{2}{B} \qquad (0 \leqslant \Delta x < \frac{B}{2}).
\end{aligned} \tag{8.8}$$

The same argument leads to the uniform stationary distribution of $\Delta Y[k]$.

$$\lim_{k\to\infty} f(\Delta Y[k] = \Delta y) = \frac{2}{B} \qquad (0 \leqslant \Delta y < \frac{B}{2}). \tag{8.9}$$

Since $\Delta X[k]$ and $\Delta Y[k]$ are independent, $\lim_{k\to\infty} f(D[k] = d)$ in Equation (8.4) can be easily derived through integration according to Figure 8.2(b). □

Theorem 32. *Two UAVs in an airspace $[0, B)^2$ follow the RD RMM equipped with the S&S protocol. The stationary pdf of the cyclic inter-vehicle distance, $D[k]$, denoted as $f_D^{S\&S}(d)$, is bounded as follows.*

$$\begin{cases} \frac{\pi}{2} dp_{1min} < f_D^{S\&S}(d) < \frac{\pi}{2} dp_{1max}, \quad (0 \leqslant d \leqslant d_o - V) \\ \frac{\pi}{4} dp_{1max} < f_D^{S\&S}(d)) < \frac{\pi}{2} dp_{1min}, \quad (d_o - V < d \leqslant d_o + V) \\ \frac{\pi}{4} dp_{1min} < f_D^{S\&S}(d) < \frac{\pi}{4} dp_{1max}, \quad (d_o + V < d \leqslant \frac{B}{2}) \\ \left(\frac{\pi}{4} - arccos\left(\frac{B}{2d}\right)\right) dp_{1\,\min} < f_D^{S\&S}(d) \\ < \left(\frac{\pi}{4} - arccos\left(\frac{B}{2d}\right)\right) dp_{1\,\max}, \quad (\frac{B}{2} < d \leqslant \frac{\sqrt{2}B}{2}) \end{cases} \tag{8.10}$$

where the constants $p_{1\,\min}$ and $p_{1\,\max}$ are

$$p_{1\,\min} = \frac{8}{\pi(d_o + V)^2 + B^2}$$
$$p_{1\,\max} = \frac{8}{\pi(d_o - V)^2 + B^2}$$

Proof. We construct a Markov process with states $S^s[k] = (\Delta X[k], \Delta Y[k])$ when the two UAVs move with RD RMM equipped with S&S protocol. The Markov chain $S^s[k]$ is aperiodic, Φ-irreducible, and Harris recurrent, and therefore, there exists a unique stationary distribution. To facilitate the analysis, we further partition the state space into five regions according to their different states transition characteristics (Figure 8.3(a)): Region 1 ($d \leqslant d_o - 2V$), Region 2 ($d_o - 2V < d \leqslant d_o - V$), Region 3 ($d_o - V < d \leqslant d_o$), Region 4 ($d_o < d \leqslant d_o + V$) and Region 5 ($d_o + V < d < \frac{B}{2}$). The five regions form two clusters. In Cluster A ($d \leqslant d_o$, including Regions 1, 2 and 3), one UAV moves and the other stops. In Cluster B ($d > d_o$, including Regions 4 and 5), two UAVs move independently according to the independent RD RMM. We further note that states in Region 1 can only transition from Cluster A, and states in Region 5 can only transition from Cluster B. States in Regions 2, 3 and 4 can transition from both clusters.

Let us first summarize the proof idea. Denote $f_S^i(s)$ as the stationary pdf of the cyclic relative position along the x and y axes in Region i, i.e., $f_S^i(s) = \lim_{k\to\infty} f(\Delta X[k] = \Delta x, \Delta Y[k] = \Delta y)$. We analyze $f_S^i(s)$ in each region respectively. To do that, we first prove that the pdfs for the states in Regions 1 and 5 ($f_S^1(s)$ and $f_S^5(s)$) are uniform, and also identify their relation. The bounds for the pdfs $f_S^2(s)$, $f_S^3(s)$ and $f_S^4(s)$ are then all derived using $f_S^1(s)$. Finally, utilizing the axiom that the sum of pdfs for all parts is 1, the lower and upper bounds of $f_S^1(s)$ are derived. The stationary inter-vehicle distance pdf and probability in Region i, denoted as $f_D^i(d)$ and P^i, can then be found through integration.

In Region 1, one UAV stops and the other moves according to the independent RD RMM. For any position (x_0, y_0) that UAV i stops at, the relative positions between two UAVs along the x and y axes are then $\Delta X[k] = \min(|X_j[k]-x_0|, B-|X_j[k]-x_0|)$ and $\Delta Y[k] = \min(|Y_j[k] - y_0|, B - |Y_j[k] - y_0|)$, respectively. Since $X_j[k]$ and $Y_j[k]$ are both uniformly distributed in the limit of large time (according to Theorem 29), $\Delta X[k]$ and $\Delta Y[k]$ can be easily proved to be also uniformly distributed in the limit, following a similar argument as in the proof of Theorem 31. Hence, $f_S^1(s)$ is uniform with value denoted as p_1. The stationary inter-vehicle distance pdf $f_D^1(d)$ and its probability in Region 1, P^1, can be represented as

$$\begin{aligned} f_D^1(d) &= \frac{1}{2}\pi d p_1 \\ P^1 &= \frac{1}{4}\pi(d_o - 2V)^2 p_1 \end{aligned} \tag{8.11}$$

In Region 5, since the two UAVs follow the RD RMM independently, the stationary inter-vehicle relative positions along the x and y axes are also uniformly distributed according to Theorem 31. Denote the value of $f_S^5(s)$ as p_2. The station-

ary inter-vehicle distance pdf $f_D^5(d)$ and its probability, P^5, can then be represented as

$$\begin{aligned} f_D^5(d) &= \frac{1}{2}\pi d p_2 \\ P^5 &= (\frac{B^2}{4} - \frac{1}{4}\pi(d_o + V)^2)p_2 \end{aligned} \tag{8.12}$$

Next we find the relationship between p_1 and p_2, or $f_S^1(s)$ and $f_S^5(s)$. The most direct method is to solve the pdf $f_S(s)$ from the following equation.

$$f_S(s) = \int f(s', s) f_S(s') ds' \tag{8.13}$$

where $f(s', s)$ is the state transition density kernel, representing the transition probability from states s' to s. $f(s', s)$ in the RD RMM equipped with S&S protocol is a piecewise function. $f(s', s)$ is a constant c when $(\Delta x^2 + \Delta y^2)^{\frac{1}{2}} \leqslant d_o$, and $s' = (\Delta x', \Delta y')$ and $s = (\Delta x, \Delta y)$ satisfy $((\Delta x - \Delta x')^2 + (\Delta y - \Delta y')^2)^{\frac{1}{2}} = V$. When $(\Delta x^2 + \Delta y^2)^{\frac{1}{2}} > d_o$, $f(s', s)$ is contributed by the movements of both UAVs, and can be derived using the intersection area of two circles of radius V centered at the two UAVs, which is complicated to find.

We here use a numerical approach named stationary density look ahead estimator (SDLAE) [185, 186] to find the relationship between p_1 and p_2. For the discrete-time Markov process representation of relative position $S[k]$, the marginal pdf of $S[k]$, represented by Φ_k, satisfies

$$\Phi_{k+1}(s) = \int f(s', s) \Phi_k(s') ds' \tag{8.14}$$

Using the marginal density look ahead estimator (MDLAE) [185, 186], $\Phi_k(s)$ can be approximated as

$$\Phi_k^n(s) := \frac{1}{n} \sum_{j=1}^{n} f(S^j[k-1], s) \tag{8.15}$$

where $S^j[k-1]$, $j \in \{1, 2, ..., n\}$ are n independent samples drawn from the lagged state $S[k-1]$ with the pdf $\Phi^n_{k-1}(s)$. Similarly, the stationary pdf Φ_∞ is approximated as

$$\Phi^n_\infty(s) = \lim_{k\to\infty} \frac{1}{n} \sum_{j=1}^{n} f(S^j[k], s') \tag{8.16}$$

where $(S[k])_{k=1}^{\infty}$ is a time series simulated from $f(s', s)$ and an arbitrary $S[1]$. Notice that $\lim_{n\to\infty} \Phi^n_\infty(s) = f_S(s)$ holds with probability 1. Using this SDLAE approach, the relation between p_1 and p_2 converges to (as shown in Figure 8.3(b))

$$p_1 = 2p_2 \tag{8.17}$$

Note that $f^i_S(s)$ is determined uniquely by the transition density kernel $f(s', s)$ as stated in Equation (8.13), and the relation between p_1 and p_2 is determined uniquely by the relation between the two transition kernels in Region 1 and 5. The transition kernels in Region 1 and 5 can be regarded as a "one-step" transition and a "two-steps" transition respectively, and is not a function of any parameters (e.g., the airspace size B, and the sensing distance d_o). As such, the relation between p_1 and p_2 is also fixed (i.e., not a function of parameters B and d_o), and would not be changed with the parameters.

Then we derive the upper and lower bounds for the stationary inter-vehicle distance distribution in Regions 2, 3, and 4 using the following steps. Step 1: we prove that $f^2_S(s)$ is uniform with density p_1 based on the Markov transition properties. Step 2: through analyzing the source states in Region 3 that transition to Region 2, we express the upper bound of $f^3_S(s)$ using p_1. Steps 3 and 4: following a similar approach, we prove that p_2 and p_1 are the lower and upper bounds of $f^4_S(s)$. In Step 5, utilizing the bounds in Region 4, we prove that p_2 is the lower bound for $f^3_S(s)$.

Step 1*: Find* $f_S^2(s)$. Consider all the points $s' = (\Delta x', \Delta y')$ of distance V to a point $s = (\Delta x, \Delta y)$ that satisfies $d \in [d_o - 3V, d_o - 2V)$ in Region 1, i.e., $((\Delta x - \Delta x')^2 + (\Delta y - \Delta y')^2)^{\frac{1}{2}} = V$ (marked as the circle in Figure 8.4(a)). As s is uniformly distributed in Region 1 with pdf p_1, at the steady-state

$$\begin{aligned} p_1 =& \int_{\widehat{aob}} f(s', s) p_1 ds' + \int_{\widehat{ab}} f(s', s) f_{S'}^2(s') ds' \\ =& f(s', s) \left(\int_{\widehat{aob}} p_1 ds' + \int_{\widehat{ab}} f_{S'}^2(s') ds' \right), \end{aligned} \tag{8.18}$$

where $\widehat{aob}$ represents the superior arc in Region 1 (solid curve in Figure 8.4(a)), and $\widehat{ab}$ represents the inferior arc in Region 2 (dotted curve in Figure 8.4(a)). Equation 8.18 holds because for any states s', s in Cluster A, $f(s', s)$ is a constant.

Similarly, for states s located in $[0, d_o - 3V)$ (e.g., $\sqrt{\Delta x^2 + \Delta y^2} \in [0, d_o - 3V)$) in Region 1, we have

$$p_1 = \oint f(s', s) p_1 ds' = f(s', s) \oint p_1 ds' \tag{8.19}$$

where $\oint$ is the integration sign over the whole circle satisfying $((\Delta x - \Delta x')^2 + (\Delta y - \Delta y')^2)^{\frac{1}{2}} = V$. As Equation (8.18) holds for any angle $\phi \in [0, \phi_{max})$ (shown in Figure 8.4(a)), where $\phi_{max} = 2arccos\frac{-V}{2(d_o - 2V)}$, and any state $s = (\Delta x, \Delta y)$ satisfying $\sqrt{\Delta x^2 + \Delta y^2} \in [d_o - 3V, d_o - 2V)$. The term-to-term comparison between Equations (8.18) and (8.19) leads to

$$f_S^2(s) = p_1 \tag{8.20}$$

Step 2*: Find the upper bound of* $f_S^3(s)$. Consider all the points that can transition to $s = (\Delta x, \Delta y)$ in Region 2. Since $f_S^2(s) = p_1$, $f_S^2(s)$ satisfies the following equation, based on the Markov transition properties (Figure 8.4(b)),

$$\begin{aligned} p_1 =& \frac{2\pi - \phi_2}{2\pi} p_1 + \int_{\widehat{cd}} f(s', s) f_{S'}^3(s') \\ &+ \int_{S_4} f(s', s) f_{S'}^4(s') ds' \end{aligned} \tag{8.21}$$

where ϕ_2 is the central angle decided by the state and V as shown in Figure 8.4(b). The last part in Equation (8.21) represents the transition probability from Region 4 (shaded region in Figure 8.4(b)), which is non-negative. This is because states in Region 4 can change a maximum distance of $2V$ at each time instance to reach s. As this equation holds for any ϕ_2 and any s in Region 2, a comparison between Equations (8.19) and (8.21) leads to

$$f_S^3(s) < p_1 \tag{8.22}$$

Steps $3-5$ prove that $f_S^3(s)$ is lowered bounded by p_2 and $f_S^4(s)$ is lowered bounded by p_2 and upper bounded by p_2. The proofs are documented in the Appendix.

To summarize, $f_S^i(s)$ in all regions satisfy

$$\begin{cases} f_S^i(s) = p_1 & 0 \leqslant d < d_o - V \\ p_2 < f_S^i(s) < p_1 & d_o - V \leqslant d < d_o + V \\ f_S^i(s) = p_2 & d_o + V \leqslant d < \frac{B}{2} \end{cases} \tag{8.23}$$

The stationary probability P^i in each region can thus be derived through integration. Utilizing $p_1 = 2p_2$ and the axiom that $\sum_i P^i = 1$, the bounds for p_1 can be found as

$$\frac{8}{\pi(d_o+V)^2 + B^2} < p_1 < \frac{8}{\pi(d_o-V)^2 + B^2} \tag{8.24}$$

The stationary inter-vehicle distance distribution $f_D^{S\&S}(d)$ can be derived by integrating $f_S^i(s)$ in Equations (8.23) and (8.24). Simple algebra leads to Equation (8.10). □

Remark 21. Theorems 31 and 32 suggest that the S&S protocol impacts the UAV inter-vehicle distance distribution. When the S&S protocol is in place, the inter-vehicle distance distribution is no more uniform.

8.3.3 Numerical Illustration

We conduct simulation studies to illustrate and validate the above theoretical results. Two UAVs follow the independent and the RD RMM equipped S&S protocol, respectively. The airspace size is $100 \times 100m^2$, and UAV velocity is $1m/s$. UAVs are initially randomly distributed, and choose their heading directions uniformly from $[0, 2\pi)$ at every time point $1s, 2s, 3s, ...$ The sensing distance is set as $10m$. The airspace is divided into 100×100 grids, and we count the number of aircarft in each grid for the entire 100000 seconds to approximate the node distribution. Both cases are uniform as shown in Figures 8.5(a) and 8.5(b).

We then simulate the inter-vehicle relative position distribution along the x and y axes for both cases. The results are shown in Figure 8.6. The distribution of inter-vehicle relative position for the independent RD RMM is uniform, while the RD RMM equipped with the S&S protocol is not uniform any more.

Finally, we simulate the inter-vehicle distance distribution. For the independent RD RMM, $f_D^b(d)$ increases in proportion to the distance below $\frac{B}{2}$ (Figure 8.7(a)), as captured in Theorem 31. For the RD RMM equipped with the S&S protocol (Figure 8.7(b)), $f_D^{S\&S}(d)$ increases in proportion to the distance in Regions 1, 2, and 5 below $\frac{B}{2}$, but fluctuates in Regions 3 and 4 between the upper bound (red line) and lower bound (purple line), in accordance to Theorem 32. In addition, the slope in Region 5 is half of that in Region 1, which verifies $p_1 = 2p_2$.

8.4 Collision probabilities and Airspace Capacity

In this section, we first define collisions and stationary collision probabilities between a pair of UAVs and then among an arbitrary number of UAVs. The concpet of airspace capacity follows. Based on the stationary inter-vehicle distance distribu-

tions derived in Section 8.3.2, we analyze the stationary collision probabilities for the independent RD RMM and the RD RMM equipped with the S&S protocol. We also find the closed-form airspace capacities for both RMMs.

8.4.1 Definitions

Denote the collision distance as d_c, where $0 \leq d_c < d_o - V$. For a pair of UAVs, the collision and stationary collision probability are defined as follows.

Definition 7. Collision occurs between a pair of UAVs i and j at time k, if $D_{i,j}[k] \leq d_c$. The stationary collision probability between the two UAVs is defined as

$$\lim_{k\to\infty} \hat{P}_{i,j}[k] = \lim_{k\to\infty} P(D_{i,j}[k] \leqslant d_c) \tag{8.25}$$

To facilitate the comparative analysis in the rest of this sequel, we use $\lim_{k\to\infty} \hat{P}^{b}_{2,i,j}[k]$ and $\lim_{k\to\infty} \hat{P}^{S\&S}_{2,i,j}[k]$ to represent the stationary collision probabilities between a pair of UAVs following the independent RD RMM and the RD RMM equipped with the S&S protocol, respectively.

Similarly, we define collision and stationary collision collision probability for multiple UAVs.

Definition 8. Collision occurs among N UAVs at time k, if and only if there exists at least one pair of UAVs (denoted as i and j) satisfying $D_{i,j}[k] \leqslant d_c$. The stationary collision probability for the N UAVs is thus defined as

$$\begin{aligned} &\lim_{k\to\infty} \hat{P}_N[k] \\ =&\lim_{k\to\infty} P(\exists D_{i,j}[k] \leqslant d_c, i, j \in [1, N], i \neq j) \end{aligned} \tag{8.26}$$

We use $\lim_{k\to\infty} \hat{P}^{b}_{N}[k]$ and $\lim_{k\to\infty} \hat{P}^{S\&S}_{N}[k]$ to represent the stationary collision probabilities for N UAVs following the independent RD RMM and the RD RMM equipped with the S&S protocol, respectively.

We define airspace capacity based on collision probability as follows.

Definition 9. The airspace capacity N_C is defined as the maximum number of UAVs with collision probability not exceeding a pre-defined collision probability threshold $\hat{P}_t$:

$$N_C = \arg\max_N \{\lim_{k\to\infty} \hat{P}_N[k] \leqslant \hat{P}_t\} \tag{8.27}$$

8.4.2 Analysis

We first study the stationary collision probabilities when the UAVs follow independent RD RMM, and then the RM RMM equipped with S&S protocol. We derive the results for a pair of UAVs, and then multiple UAVs respectively. The analysis of airspace capacity follows.

Theorem 33. *Two UAVs in an airspace $[0, B)^2$ follow the independent RD RMM. The stationary collision probability between the two UAVs is*

$$\lim_{k\to\infty} \hat{P}^b_{2,i,j}[k] = \frac{\pi d_c^2}{B^2} \tag{8.28}$$

Proof. According to Definition 7, the stationary collision probability between two UAVs can be derived by integrating the stationary inter-vehicle distance distribution with $D_{i,j}[k] < d_c$.

$$\begin{aligned} \lim_{k\to\infty} \hat{P}^b_{2,i,j}[k] &= \int_0^{d_c} f^b_D(r)dr \\ &= \int_0^{d_c} \frac{2\pi r}{B^2} dr = \frac{\pi d_c^2}{B^2} \end{aligned} \tag{8.29}$$

□

Theorem 34. *Each of the N ($N > 2$) UAVs in an airspace $[0, B)^2$ follows the RD RMM independently. The stationary collision probability among the N UAVs is*

$$\lim_{k\to\infty} \hat{P}^b_N[k] = 1 - (1 - \frac{\pi d_c^2}{B^2})^{\frac{N(N-1)}{2}} \tag{8.30}$$

where d_c satisfies $d_c \ll B$.

Proof. First we consider the case of three UAVs i, j, and l moving independently according to the independent RD RMM. The collision probability among the three UAVs is described according to Definition 8 as

$$\begin{aligned}
&\lim_{k\to\infty} \hat{P}_3^b[k] \\
&= \lim_{k\to\infty} P(D_{i,j}[k] \leqslant d_c \cup D_{i,l}[k] \leqslant d_c \cup D_{j,l}[k] \leqslant d_c) \\
&= 1 - \lim_{k\to\infty} P(D_{i,j}[k] > d_c, D_{i,l}[k] > d_c, D_{j,l}[k] > d_c) \\
&= 1 - \lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c) \\
&\times \lim_{k\to\infty} P(D_{i,j}[k] > d_c, D_{i,l}[k] > d_c)
\end{aligned} \tag{8.31}$$

Note that any two of the three distances are independent, i.e., $f(D_{i,j}[k] = d_1, D_{i,l}[k] = d_2) = f(D_{i,j}[k] = d_1) f(D_{i,l}[k] = d_2)$, and the third UAV distance $D_{j,l}[k]$ can be determined by the other two distances. Therefore, Equation (8.31) is further written as

$$\begin{aligned}
&\lim_{k\to\infty} \hat{P}_3^b[k] \\
&= 1 - \lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c) \\
&\times \lim_{k\to\infty} P(D_{i,j}[k] > d_c) \lim_{k\to\infty} P(D_{i,l}[k] > d_c)
\end{aligned} \tag{8.32}$$

where $\lim_{k\to\infty} P(D_{i,j}[k] > d_c)$ and $\lim_{k\to\infty} P(D_{i,l}[k] > d_c)$ can be derived from the integration of the stationary pdf ($f_D^b(d)$) shown in Equation (8.4), according to Theorem 31.

Then let us find the conditional probability $\lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c)$.

Since the three UAVs move independently with RD RMM, the relative positions of the two independent UAV pairs ($\Delta X_{i,j}[k]$, $\Delta Y_{i,j}[k]$), and ($\Delta X_{i,l}[k]$, $\Delta Y_{i,l}[k]$) are both uniformly distributed in $[0, \frac{B}{2})^2$ according to Theorem 8.3.

$$\begin{aligned} \lim_{k\to\infty} f(\Delta X_{i,j}[k] = \Delta x_1, \Delta Y_{i,j} = \Delta y_1) &= \frac{4}{B^2} \\ \lim_{k\to\infty} f(\Delta X_{i,l}[k] = \Delta x_2, \Delta Y_{i,l} = \Delta y_2) &= \frac{4}{B^2} \end{aligned} \tag{8.33}$$

where Δx_1, Δy_1, Δx_2, and $\Delta y_2 \in [0, \frac{B}{2})$. With the relative positions, the conditional probability $\lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c)$ can be rewritten as

$$\begin{aligned} &\lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c) \\ &= \lim_{k\to\infty} P(\sqrt{\Delta X_{j,l}[k]^2 + \Delta Y_{j,l}[k]^2} > d_c | \\ &\sqrt{\Delta X_{i,j}[k]^2 + \Delta Y_{i,j}[k]^2} > d_c, \sqrt{\Delta X_{i,l}[k]^2 + \Delta Y_{i,l}[k]^2} > d_c) \end{aligned} \tag{8.34}$$

where $\Delta X_{j,l}[k]$ is determined by $\Delta X_{i,j}[k]$ and $\Delta X_{i,l}[k]$, and $\Delta Y_{j,l}[k]$ is determined by $\Delta Y_{i,j}[k]$ and $\Delta Y_{i,l}[k]$ considering the following four cases.

Case 1, UAVs j and l are on the same side of i along both x and y axes. In this case, the relative positions of j, l are $(\Delta X_{j,l}[k], \Delta Y_{j,l}[k])_1 = (|\Delta X_{i,j}[k] - \Delta X_{i,l}[k]|, |\Delta Y_{i,j}[k] - \Delta Y_{i,l}[k]|)$. Therefore, Equation (8.34) becomes

$$
\begin{aligned}
&\lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c) \\
&= \lim_{k\to\infty} P\Big(\big((\Delta X_{i,j}[k] - \Delta X_{i,l}[k])^2 \\
&+ (\Delta Y_{i,j}[k] - \Delta Y_{i,l}[k])^2 \big)^{\frac{1}{2}} > d_c | \sqrt{\Delta X_{i,j}[k]^2 + \Delta Y_{i,j}[k]^2} > d_c, \\
&\sqrt{\Delta X_{i,l}[k]^2 + \Delta Y_{i,l}[k]^2} > d_c \Big) \\
&= \iint_{S_{1C}} \iint_{S_{2C}} \int_{d_c}^{\frac{B}{2}} f(\sqrt{(\Delta x_1 - \Delta x_2)^2 + (\Delta y_1 - \Delta y_2)^2} = r) \\
&\times f(\Delta X_{i,j}[k] = \Delta x_1, \Delta Y_{i,j} = \Delta y_1) \\
&\times f(\Delta X_{i,l}[k] = \Delta x_2, \Delta Y_{i,l} = \Delta y_2) dr ds_2 ds_1 \\
&< \iint_{S_1} \iint_{S_2} \int_{d_c}^{\frac{B}{2}} f(\sqrt{(\Delta x_1 - \Delta x_2)^2 + (\Delta y_1 - \Delta y_2)^2} = r) \\
&\times f(\Delta X_{i,j}[k] = \Delta x_1, \Delta Y_{i,j} = \Delta y_1) \\
&\times f(\Delta X_{i,l}[k] = \Delta x_2, \Delta Y_{i,l} = \Delta y_2) dr ds_2 ds_1 \\
&= \lim_{k\to\infty} P(D_{j,l}[k] > d_c)
\end{aligned}
\tag{8.35}
$$

where $s_1 = (\Delta x_1, \Delta y_1)$, $s_2 = (\Delta x_2, \Delta y_2)$. S_{1C} is the integral region constructed by four lines and one curve: $\Delta x_1 = 0$, $\Delta x_1 = \frac{B}{2}$, $\Delta y_1 = 0$, $\Delta y_1 = \frac{B}{2}$, and $\sqrt{\Delta x_1^2 + \Delta y_1^2} > d_c$ (marked as the shaded area in Figure 8.8(a)). The condition $D_{i,j}[k] > d_c$ is satisfied in S_{1C}. Similarly, S_{2C} is the region constructed by $\Delta x_2 = 0$, $\Delta x_2 = \frac{B}{2}$, $\Delta y_2 = 0$, $\Delta y_2 = \frac{B}{2}$, and $\sqrt{\Delta x_2^2 + \Delta y_2^2} > d_c$, and the condition $D_{i,l}[k] > d_c$ is satisfied in S_{2C}. S_1 is the integral region constructed by four lines: $\Delta x_1 = 0$, $\Delta x_1 = \frac{B}{2}$, $\Delta y_1 = 0$, and $\Delta y_1 = \frac{B}{2}$, and is shown as the shaded region in Figure 8.8(b). S_2 is the region constructed by $\Delta x_2 = 0$, $\Delta x_2 = \frac{B}{2}$, $\Delta y_2 = 0$, and $\Delta y_2 = \frac{B}{2}$.

Note that with the condition $d_c \ll B$, the conditional probability $\lim_{k\to\infty} P(D_{j,l}[k] > d_c | D_{i,j}[k] > d_c, D_{i,l}[k] > d_c)$ can be approximated by $\lim_{k\to\infty} P(D_{j,l}[k] > d_c)$ from Equation (8.35).

Similar analysis can be found in Case 2, where UAVs j and l are on the same side of i along x axis, but on different sides of i along y axis $\big((\Delta X_{j,l}[k], \Delta Y_{j,l}[k])_2 = (|\Delta X_{i,j}[k] - \Delta X_{i,l}[k]|, |\Delta Y_{i,j}[k] + \Delta Y_{i,l}[k]|)\big)$, Case 3, where UAVs j and l are on different sides side of i along x axis, but on the same side of i along y axis$\big((\Delta X_{j,l}[k], \Delta Y_{j,l}[k])_3 = (|\Delta X_{i,j}[k] + \Delta X_{i,l}[k]|, |\Delta Y_{i,j}[k] - \Delta Y_{i,l}[k]|)\big)$, and Case 4, where UAVs j and l are on different sides of i along both x and y axes $\big((\Delta X_{j,l}[k], \Delta Y_{j,l}[k])_4 = (|\Delta X_{i,j}[k] + \Delta X_{i,l}[k]|, |\Delta Y_{i,j}[k] + \Delta Y_{i,l}[k]|)\big)$.

Combining Equations (8.32) and (8.35), under the condition $d_c \ll B$, the collision probability among three UAVs moving with basic RD RMM is

$$\begin{aligned}\lim_{k\to\infty} \hat{P}_3^b[k] =& 1 - \lim_{k\to\infty} P(\forall D_{i,j}[k] > d_c, i, j \in [1,3], i \neq j) \\ =& 1 - (1 - \lim_{k\to\infty} \hat{P}_{2,i,j}^b[k])^3\end{aligned} \tag{8.36}$$

For the N UAVs case, there are a total of $\frac{N(N-1)}{2}$ inter-vehicle distance pairs, and $N-1$ of them are independent. Following a similar argument, it can be proven that under the condition $d_c \ll B$, the stationary collision probability among N UAVs with basic RD RMM is

$$\begin{aligned}\lim_{k\to\infty} \hat{P}_N^b[k] &= \lim_{k\to\infty} P(\exists D_{i,j}[k] \leqslant d_c, i, j \in [1,N], i \neq j) \\ &= 1 - \lim_{k\to\infty} P(\forall D_{i,j}[k] > d_c, i, j \in [1,N], i \neq j) \\ &= 1 - (1 - \lim_{k\to\infty} \hat{P}_{2,i,j}^b[k])^{\frac{N(N-1)}{2}}\end{aligned} \tag{8.37}$$

Substituting Equation (8.28) into Equation (8.37), the stationary collision probability among N UAVs is obtained. □

The next two theorems state the stationary collision probabilities for UAVs that follow the RD RMM equipped with the S&S protocol.

Theorem 35. *Two UAVs in an airspace* $[0, B)^2$ *follow the RD RMM equipped with the S&S protocol. The stationary collision probability between the two UAVs is upper bonded by*

$$\lim_{k\to\infty} \hat{P}_{2,i,j}^{S\&S}[k] < \frac{2\pi d_c^2}{\pi(d_o - V)^2 + B^2} \tag{8.38}$$

Proof. Integrating the upper bound of stationary inter-vehicle distance distribution in Theorem 32, we obtain

$$\begin{aligned} \lim_{k\to\infty} \hat{P}_{2,i,j}^{S\&S}[k] =& \int_0^{d_c} \frac{\pi r}{2} p_1 dr \\ <& \frac{2\pi d_c^2}{\pi(d_o - V)^2 + B^2} \end{aligned} \tag{8.39}$$

□

In the case of multiple UAVs, the independence assumption among UAV pairs is removed when the S&A protocol is in place. Therefore, the collision probability among multiple UAVs is not equal to the simple multiplication of the collision probabilities between each UAV pair. In the next theorem, we derive its upper bound.

Theorem 36. *N UAVs in a airspace* $[0, B)^2$ *follow the RD RMM equipped with the S&S protocol. The stationary collision probability among the N UAVs is upper bounded by*

$$\lim_{k\to\infty} \hat{P}_{N}^{S\&S}[k] < \frac{N(N-1)}{2} \frac{2\pi d_c^2}{\pi(d_o - V)^2 + B^2} \tag{8.40}$$

Proof. According to Definition 8, collision occurs among N UAVs when there exists at lease one pair of UAVs satisfying $D_{i,j}[k] \leqslant d_c$.

$$\begin{aligned}
&\lim_{k\to\infty} \hat{P}_N^{S\&S}[k] = \lim_{k\to\infty} P(\exists D_{i,j}[k] \leqslant d_c, i, j \in [1, N], i \neq j) \\
&= \lim_{k\to\infty} P(D_{1,2}[k] \leqslant d_c \cup D_{1,3}[k] \leqslant d_c \cup ... \cup D_{N-1,N}[k] \leqslant d_c) \\
&< \lim_{k\to\infty} P(D_{1,2}[k] \leqslant d_c) + P(D_{1,3}[k] \leqslant d_c) + ... \\
&\quad + P(D_{N-1,N}[k] \leqslant d_c) \\
&= \binom{N}{2} \lim_{k\to\infty} \hat{P}_{N,i,j}^{S\&S}
\end{aligned} \tag{8.41}$$

where $\hat{P}_{N,i,j}^{S\&S}$ is the collision probability between UAVs i and j when N ($N \geqslant 3$) UAVs move in the airspace. $\binom{N}{2}$ is a 2-combination of UAVs from a N UAV airspace, $\binom{N}{2} = \frac{n(n-1)}{2}$.

We note that $\hat{P}_{N,i,j}^{S\&S}$ may not equal to $\hat{P}_{2,i,j}^{S\&S}$, i.e., the collision probability between UAVs i and j when they move alone in the airspace, due to the existence of other UAVs.

Here we state that $\hat{P}_{N,i,j}^{S\&S}$ can be approximated by $\hat{P}_{2,i,j}^{S\&S}$. $\hat{P}_{N,i,j}^{S\&S}$ is affected by the other UASs in the airspace. However, using the argument similar to the five region analysis in Theorem 32, we note that the existence of other UASs makes the UAS i or j be more likely to "stop" in all five regions, which extends the time duration for UASs i and j to be in these distance regions. When the increased time duration is approximately the same in all possible distances, the collision probability $\hat{P}_{N,i,j}^{S\&S}$ can well be approximated by $\hat{P}_{2,i,j}^{S\&S}$. This assumption holds as the triggering of the extra "stopping" only depends on the inter-vehicle distance between UAS i (or j) with other UASs, regardless of the inter-vehicle distance between i and j. With this approximation, combining Equations (8.41) and (8.38), the upper bound of the stationary collision probability among N UASs is derived as shown in Equation (8.40).

□

The airspace capacities for the independent RD RMM and the RD RMM equipped with the S&S protocol are derived according to Definition 9, based on the collision probabilities.

Theorem 37. *Given a threshold collision probability $\hat{P}_t$, the airspace capacity for the independent RD RMM (N_C^b) and the RD RMM equipped with the S&S protocol ($N_C^{S\&S}$) are expressed as follows.*

$$N_C^b = \left\lfloor \sqrt{\log_{1-\frac{\pi d_c^2}{B^2}}^{(1-\hat{P}_t)^2} + \frac{1}{4}} + \frac{1}{2} \right\rfloor \tag{8.42}$$

$$N_C^{S\&S} > \left\lfloor \sqrt{\frac{\hat{P}_t(\pi(d_o - V)^2 + B^2)}{\pi d_c^2} + \frac{1}{4}} + \frac{1}{2} \right\rfloor \tag{8.43}$$

Where $\lfloor\rfloor$ is the floor operation.

Proof. According to Definition 9, the airspace capacity can be derived from the collision probability analysis (described in Theorems 34 and 36) naturally. □

Remark 22. The analytical framework presented in this chapter provides us valuable insights on the the effectiveness of local S&A protocols to global airspace capacity. A comparison between Theorems 8.5 and 8.7 suggests that surprisingly the S&S protocol is not effective for a highly variable on-demand UAV traffic. In particular, S&S can lead to increased collision probability and hence reduced airspace capacity. Intuitively, this is because the "stopping" protocol enlarges the collision duration if the other vehicle moves toward it.

8.4.3 Numerical Illustration

We simulate N UAVs ($N = 2, 4, 6, 8$) moving in a confined square airspace ($20 \times 20m^2$) following independent RD RMM with the speed of $1m/s$, and then the

RD RMM equipped with the S&S. The sensing distance is set as $d_o = 2m$ and collision distance is $d_c = 1m$. Figure 8.9(a) suggests that the collision probability for the independent RD RMM converges in the limit of large time to the theoretical values. For the RD RMM equipped with the S&S protocol, the theoretical upper bounds characterize the collision properties (8.9(b)). The stationary collision probability increases with the increase of the number of UAVs for both RMMs.

8.5 Impact Analysis of S&S Configurations

The above analytical framework allows us to systematically study the impact of local S&A protocols to global airspace capability. In this section, we study the effect of local S&S configurations, including travel time, sensing distance, collision distance. We also compare the impact of some other S&A protocols with the S&S protocol.

8.5.1 Impact Analysis of Travel Time

Travel time, defined as the time duration for a vehicle to hold its current heading direction [171], is one of the indicators of model randomness. UAVs of different missions may have different travel time statistics. We have proved that in a 1-D airspace, travel time affects the collision probability significantly [11]. Characterizing the relationship between travel time and collision probability helps us understand the effect of S&S protocols for UAVs of different randomness levels.

First, we note that for the independent RD RMM, the change of travel time does not impact the joint uniform node distribution in Theorem 29, the uniform inter-vehicle distribution in Theorem 31, or the collision probability in Theorems 33 and 34. Hence, the airspace capacity remains the same as illustrated in Equation (8.42).

For the RD RMM equipped with S&S protocol, the impact of travel time is shown in Figures 8.10(a), obtained using the SDLAE method. For numerical example

in this section, the parameters are set as $B = 100m$, $d_o = 10/m$, and $V = 1m/s$. Each UAV's node distribution is uniform as shown in Theorem 30. However, extending the travel time reduces the slope of inter-vehicle distance distribution in Region 1, and hence leads to reduced collision probability and larger airspace capacity.

8.5.2 Impact Analysis of Sensing Distance and Collision Distance

Sensing distance is also an important parameter. Convective weather can shorten the sensing distance of UAVs and affect airspace capacity [187]. This effect can be captured by reducing the parameter d_o in the S&A protocol. We here only analyze the impact of sensing distance to collision probability for the RD RMM equipped with S&S protocol, as sensing distance does not affect the independent RD RMM. It can be seen from Figure 8.10(b) that longer sensing distance leads to reduced stationary collision probability in Region 1 and hence larger airspace capacity.

Collision distance, d_c, on the other hand, does not alter the inter-vehicle distance distributions for both RD RMMs. With reduced collision distance, airspace capacity increases as shown in Equations (8.42) and (8.43) for both RD RMMs.

8.5.3 Comparison with other S&A Protocols

In this section, we compare the performance of the S&S protocol with the other two S&A protocols for the highly variable UAV traffic, including sense-and-turn-left (S&T) and sense-and-reverse (S&R).

The two protocols work as follows: when the inter-vehicle distance between two UAVs is greater than the sensing distance, the two UAVs follow the RD RMM independently. When the distance between them is smaller than the sensing distance, the vehicle to the relative right continue its original movement, and the vehicle to the relative left turns left for the S&T protocol, and reverses its direction for the S&R

protocol, until the inter-vehicle distance is greater than the sensing distance again. The inter-vehicle distance distributions between a UAV pair for the independent RD and the RD with the S&S, S&T, and S&R are plotted in Figure 8.11(a). Clearly, both the S&T and S&R protocols reduce the collision probability and lead to larger airspace capacity compared to the independent RD model. Furthermore, the S&R has the best collision avoidance capabilities among them all.

Now we further study the properties of S&R. In order to evaluate the local S&R's impact to collision probability and airspace capacity, we further plot the key metric, $\hat{P}_{N,i,j}$ with the increase of number of UAVs in the airspace, N (see Figure 8.11(b)). Clearly, with the increase of N, $\hat{P}_{N,i,j}$ increases, suggesting other vehicle's impact to the effectiveness of the local protocol.

Appendix

Proofs of Steps $3-5$ for Theorem 32.

Step 3*: Find the lower bound for* $f_S^4(s)$. Consider all the points that can transition a maximum of $2V$ to the point $s=(\Delta x, \Delta y)$ that satisfies $d \in [d_o+V, d_o+2V)$ in Region 5 (Figure 8.12(a)). As s is uniformly distributed in Region 5 with pdf p_2, the following equation holds in the limit of large time.

$$p_2 = \frac{2\pi-\phi_3}{2\pi}p_2 + \int_{s_5} p_2 f(s',s) + \int_{s_4} f_S^4(s) f(s',s) ds' \tag{8.44}$$

where ϕ_3 is the angle determined by the states' position and the boundary of Region 4, s_4 and s_5 are the regions marked in different shades in Figure 8.12(a).

Similarly, for the states that are located in Region 5 and can solely transition from Region 5, we have

$$p_2 = \oint f(s',s) p_2 ds' \tag{8.45}$$

where the integration is for the whole area inside the circle, satisfying

$$\left((\Delta x - \Delta x')^2 + (\Delta y - \Delta y')^2\right)^{\frac{1}{2}} \leqslant 2V \tag{8.46}$$

The comparison between Equations (8.44) and (8.45) leads to the conclusion that

$$f_S^4(s) > p_2 \tag{8.47}$$

Step 4: Find the upper bound for $f_S^4(s)$. As $p_1 = 2p_2$, Equation (8.45) can be further written as

$$\begin{aligned} p_2 =& \frac{2\pi - \phi_3}{2\pi} p_2 + \int_{s_3+s_4+s_5} p_2 f(s', s) ds' \\ =& \frac{2\pi - \phi_3}{2\pi} p_2 + \int_{s_5} p_2 f(s', s) ds' + \int_{s_4} \frac{1}{2} p_1 f(s', s) ds' \\ &+ \int_{s_3} \frac{1}{2} p_1 f(s', s) ds' \end{aligned} \tag{8.48}$$

With the inequality that $\int_{s_3} p_1 f(s', s) ds' < \int_{s_4} p_1 f(s', s) ds'$, Equation (8.48) becomes

$$\begin{aligned} p_2 <& \frac{2\pi - \phi_3}{2\pi} p_2 + \int_{s_5} p_2 f(s', s) ds' + 2 \int_{s_4} \frac{1}{2} p_1 f(s', s) ds' \\ =& \frac{2\pi - \phi_3}{2\pi} p_2 + \int_{s_5} p_2 f(s', s) ds' + \int_{s_4} p_1 f(s', s) ds' \end{aligned} \tag{8.49}$$

Comparing Equations (8.44) and (8.49), we can conclude that

$$f_S^4(s) < p_1 \tag{8.50}$$

Step 5: Find the lower bound for $f_S^3(s)$. Revisit Equation (8.21). Denote the region in the loose shades in Figure 8.12(b) as s_3, and the region in the dense shades as s_4. Utilizing $f_S^4(s) < p_1$ (Equation (8.50)) and the relation that $p_1 = 2p_2$, Equation (8.21) can be further written as

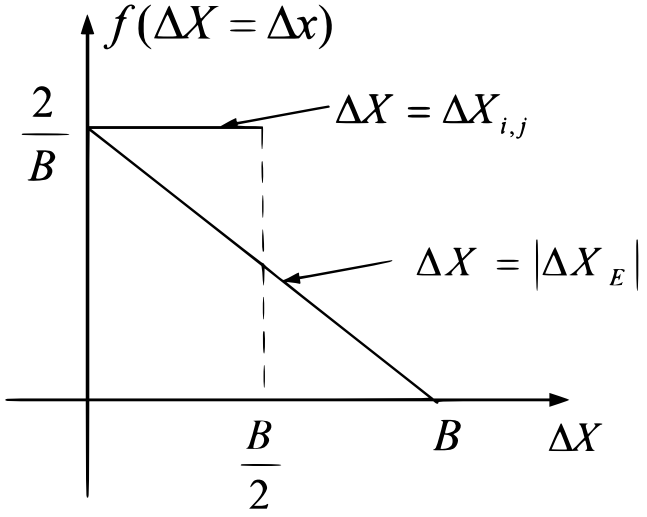
$$p_2 < \frac{2\pi - \phi_2}{2\pi} p_2 + \frac{1}{2} \int_{\widehat{cd}} f(s', s) f_{S'}^3 ds' + \int_{s_4} f(s', s) p_2 ds' \tag{8.51}$$

With the inequality that $\int_{s_4} f(s',s)p_2 ds' < \int_{s_3} f(s',s)p_2 ds'$, Equation (8.51) becomes

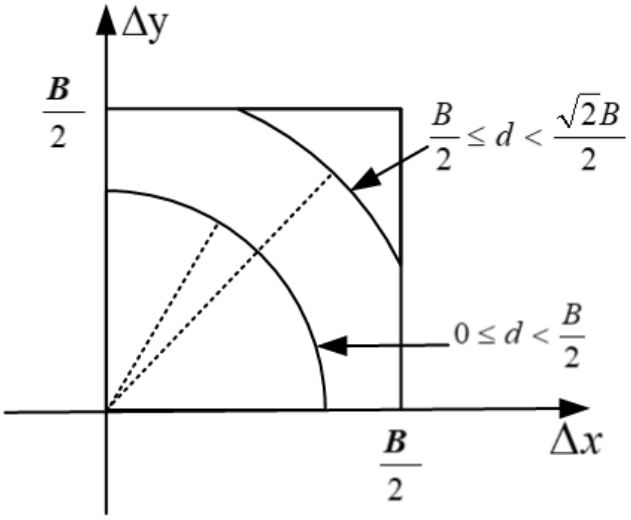
$$\begin{aligned} p_2 <& \frac{2\pi-\phi_2}{2\pi}p_2 + \frac{1}{2}\int_{\widehat{cd}} f(s',s)f^3_{S'}ds' + \frac{1}{2}\int_{s_3+s_4} f(s',s)p_2 ds' \\ =& \frac{2\pi-\phi_2}{2\pi}p_2 + \frac{1}{2}\int_{\widehat{cd}} f(s',s)f^3_{S'}(s') + \frac{1}{2}\frac{\phi_2}{2\pi}p_2 \end{aligned} \tag{8.52}$$

Comparing Equations (8.52) and (8.19), we can easily concluded that

$$f^3_S(s) > p_2 \tag{8.53}$$

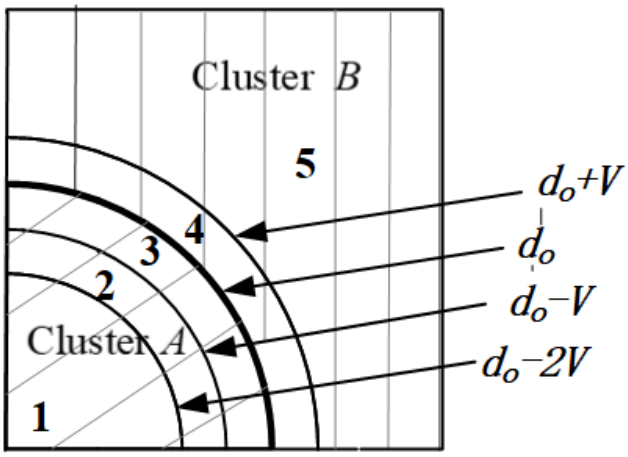


(a)

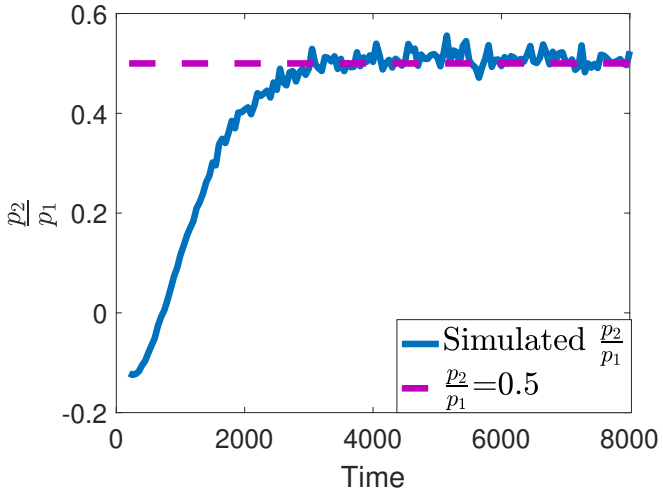


(b)

Figure 8.2. (a). The relationship between pdfs for $|\Delta X_E[k]|$ and $\Delta X[k]$. (b). The range of cyclic distance D.

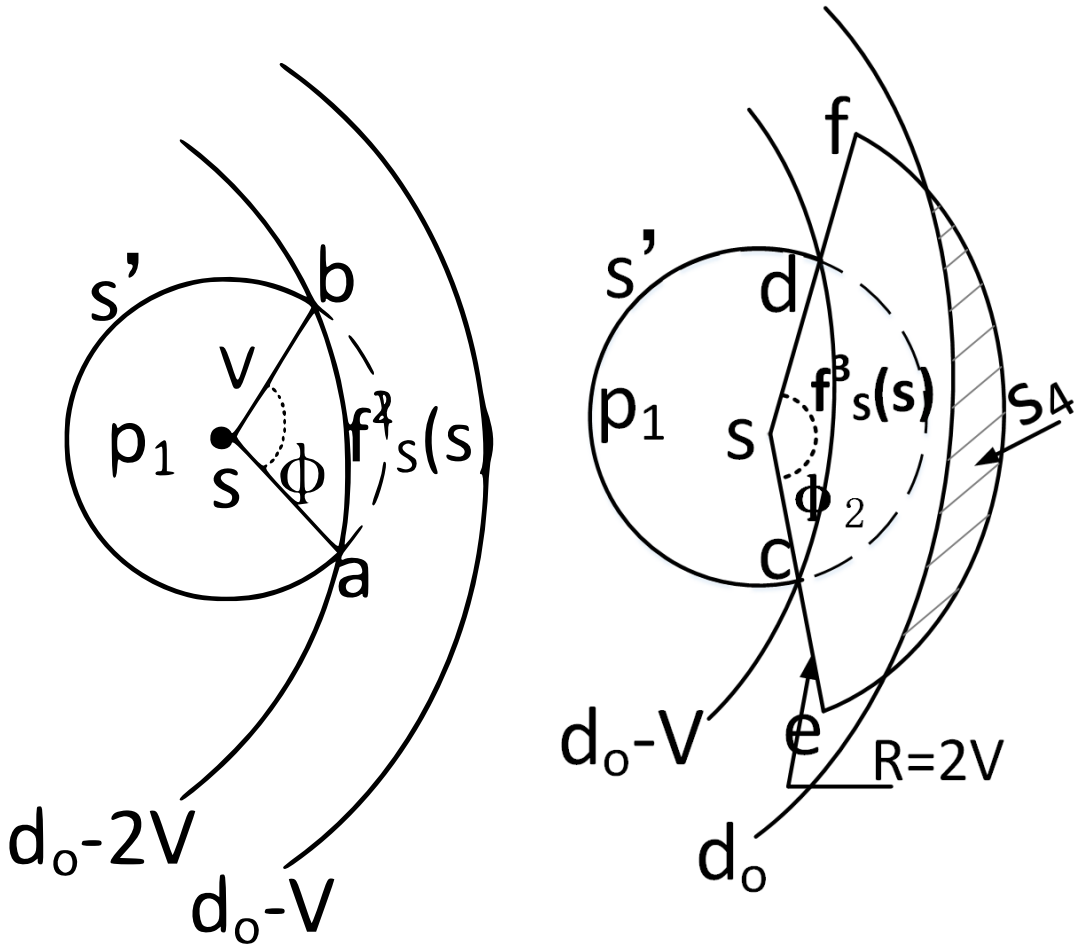


(a)



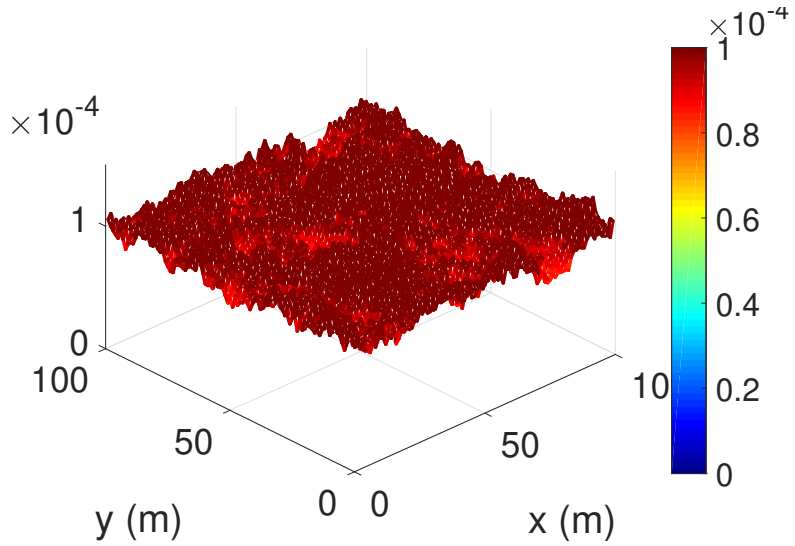
(b)

Figure 8.3. (a) Partition of the state space into 5 regions based on the inter-vehicle distance. Clusters A and B are marked in different shades. (b) Illustration of the SDLAE method to find the relation between p_1 and p_2.

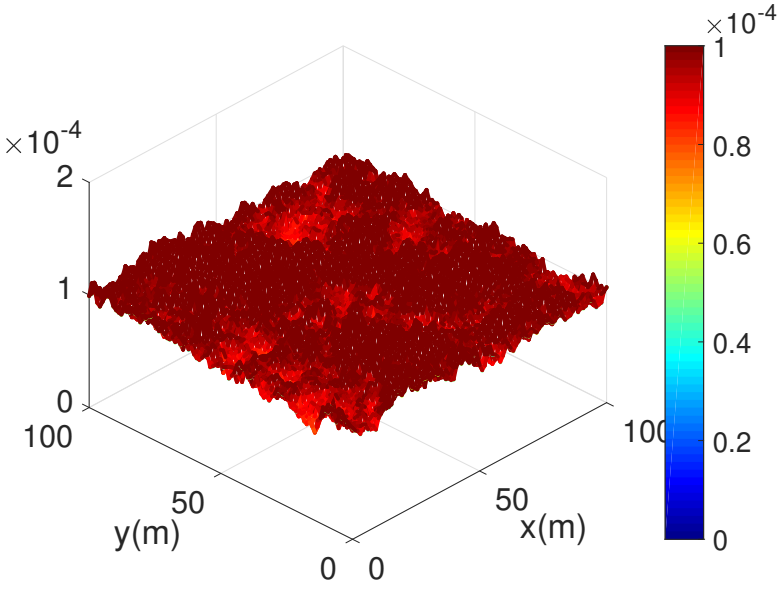


(a) (b)

Figure 8.4. (a) Illustration of Step 1. (b) Illustration of Step 2.

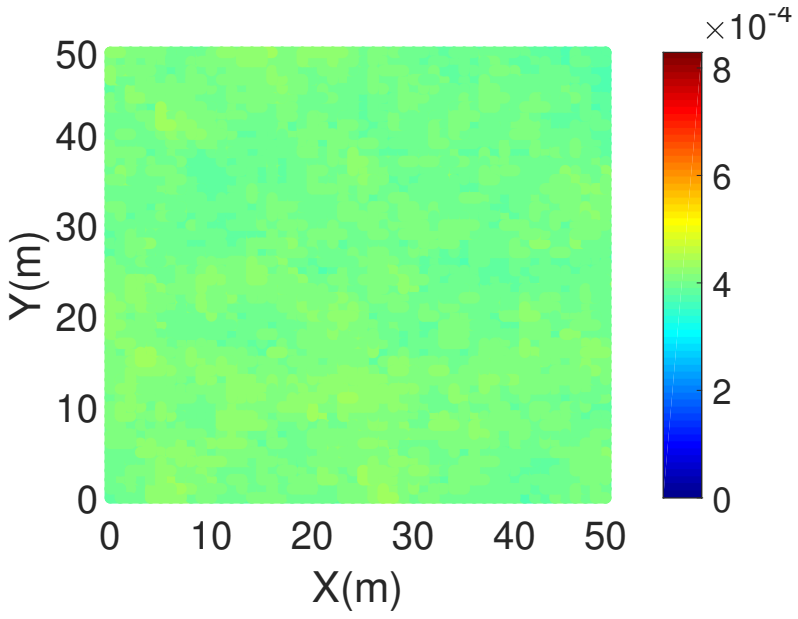


(a)

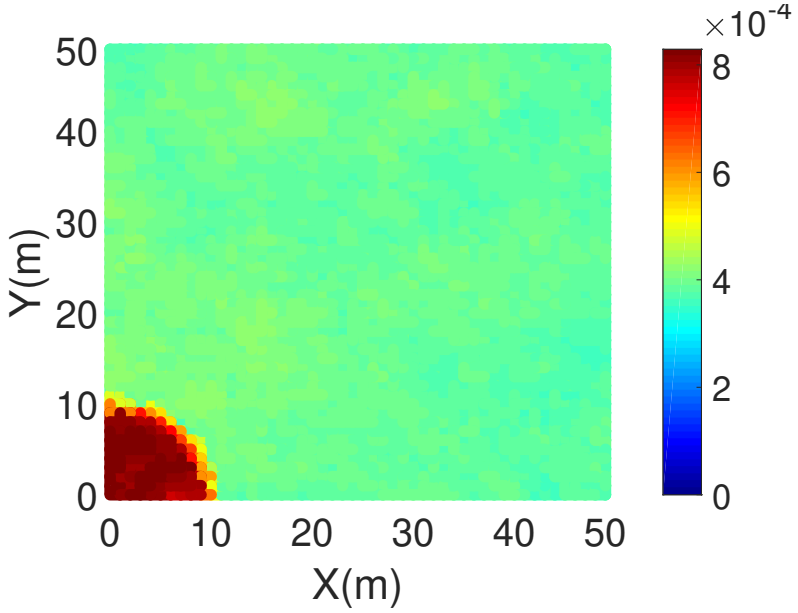


(b)

Figure 8.5. (a) Node distribution of the independent RD RMM. (b) Node distribution of the RD RMM equipped with the S&S protocol.

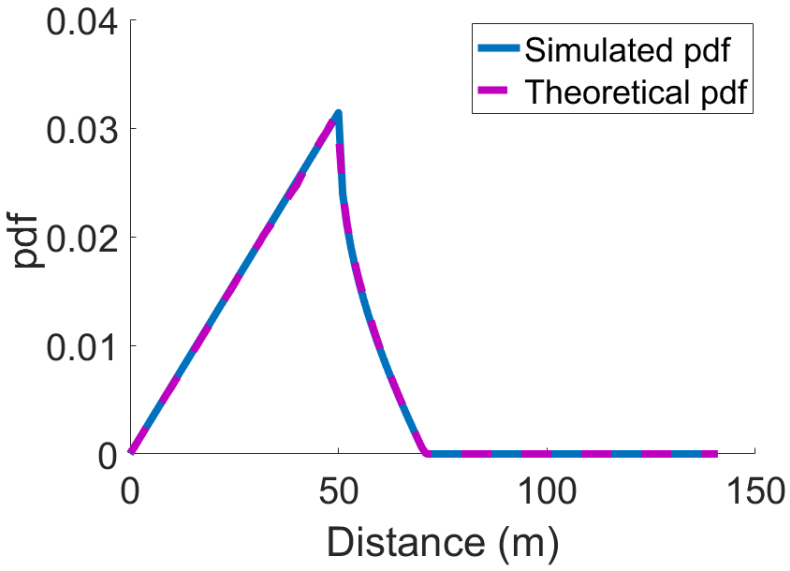


(a)

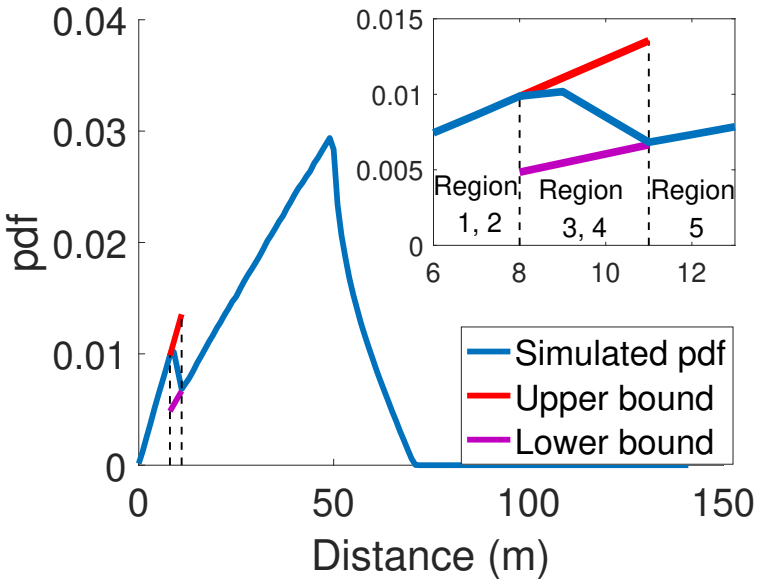


(b)

Figure 8.6. (a) Inter-vehicle relative position distribution of independent RD RMM. (b) Inter-vehicle distance distribution of the RD RMM equipped with the S&S protocol.

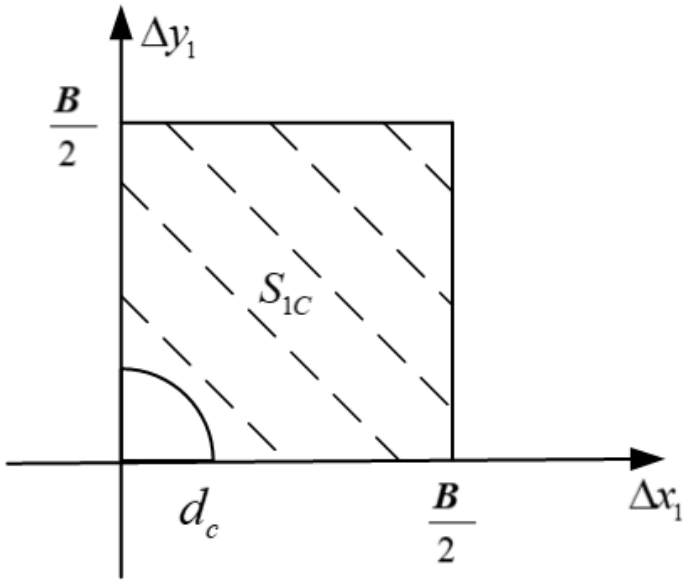


(a)

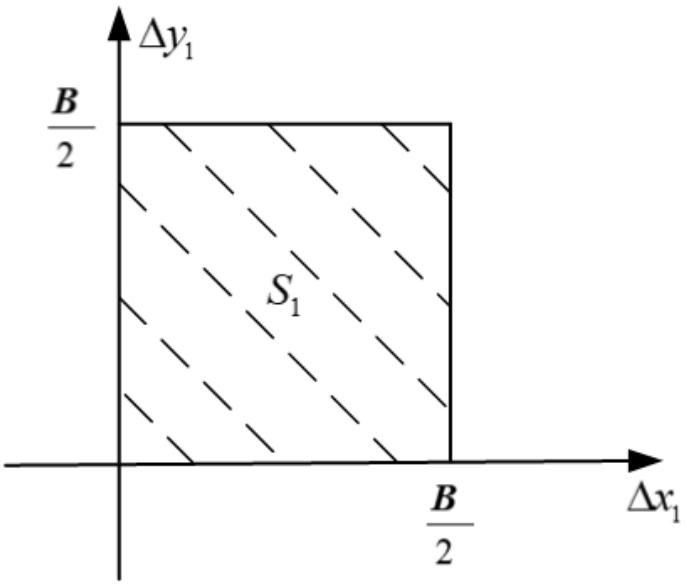


(b)

Figure 8.7. Pdfs of inter-vehicle distance for (a) the independent RD RMM, (b) the RD RMM equipped with the S&S protocol.

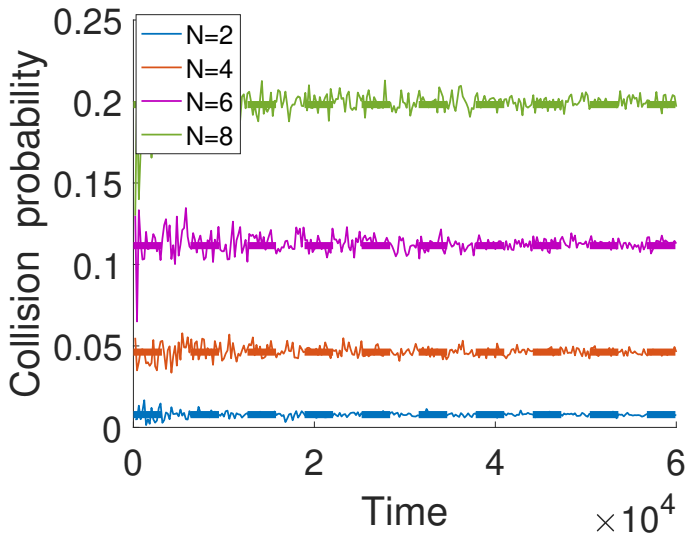


(a)

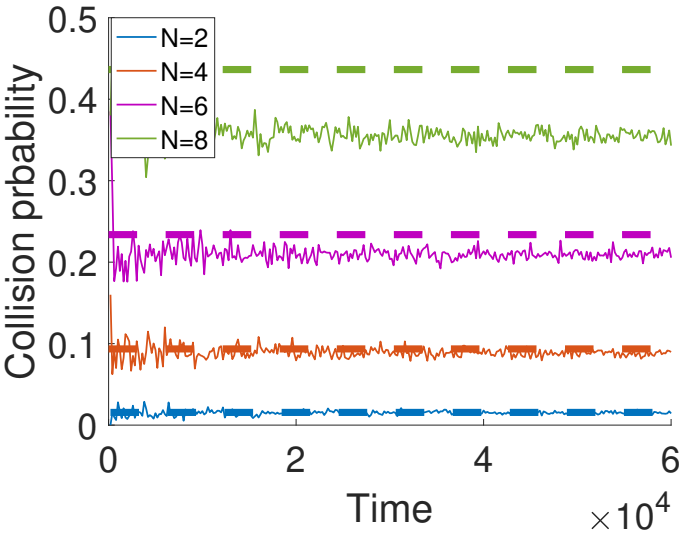


(b)

Figure 8.8. The integral regions (shaded regions) of (a) S_{1C}, and (b) S_1.

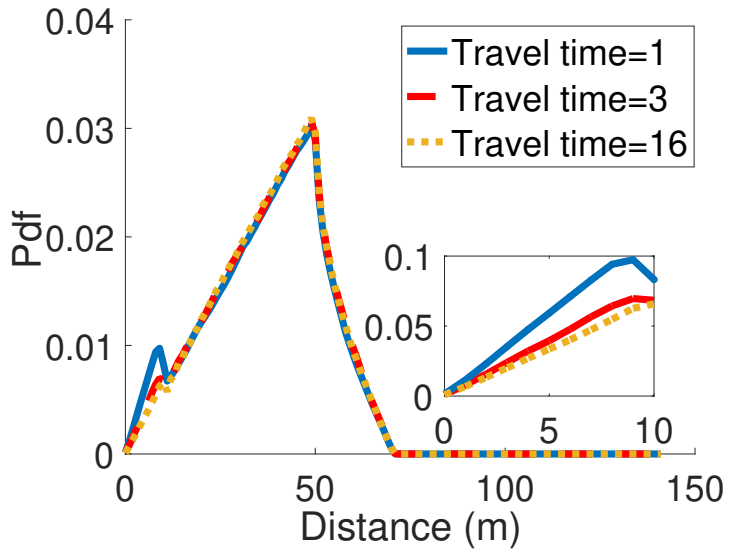


(a)

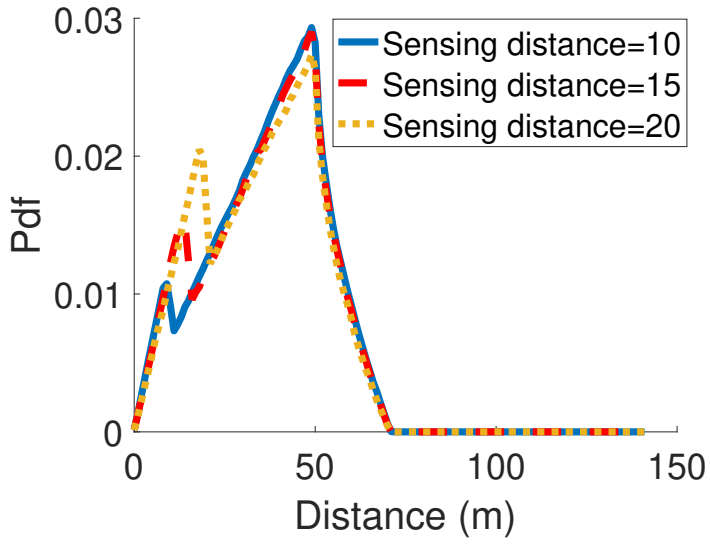


(b)

Figure 8.9. Collision probabilities among UAVs that follow (a) the independent RD RMM, and (b) the RD RMM equipped with S&S protocol. The solid lines are simulated collision probabilities. The dotted lines in (a) are theoretical values, and in (b) are theoretical upper bounds.

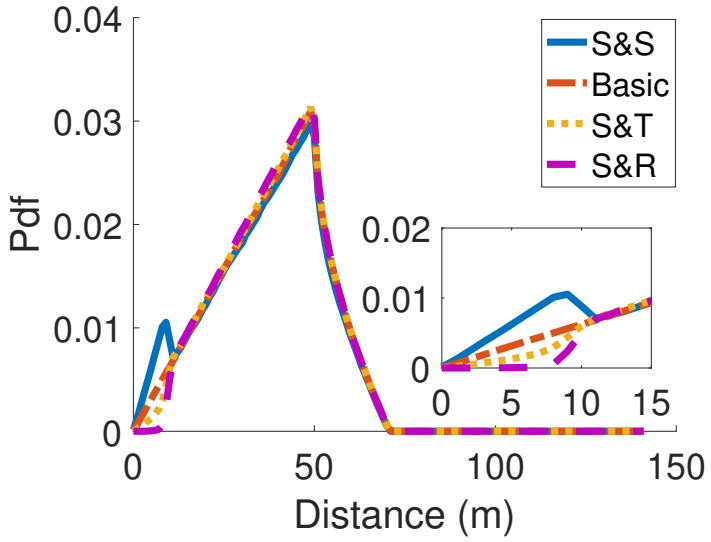


(a)

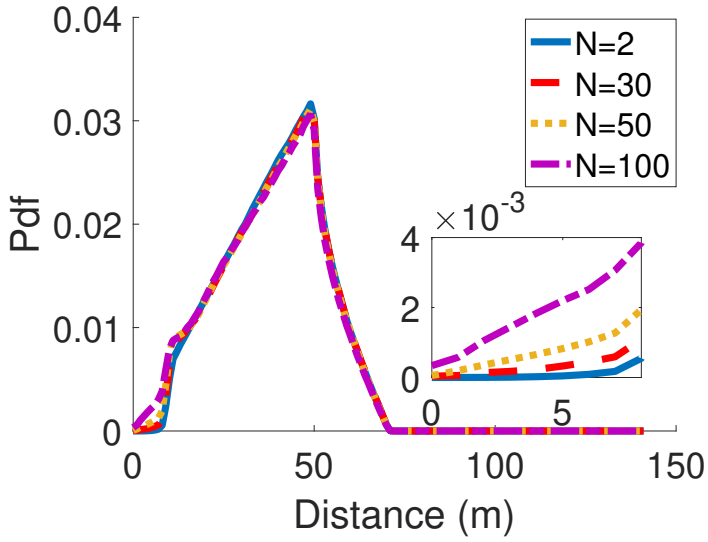


(b)

Figure 8.10. Inter-vehicle distance distribution with different (a) travel time and (b) sensing distances.

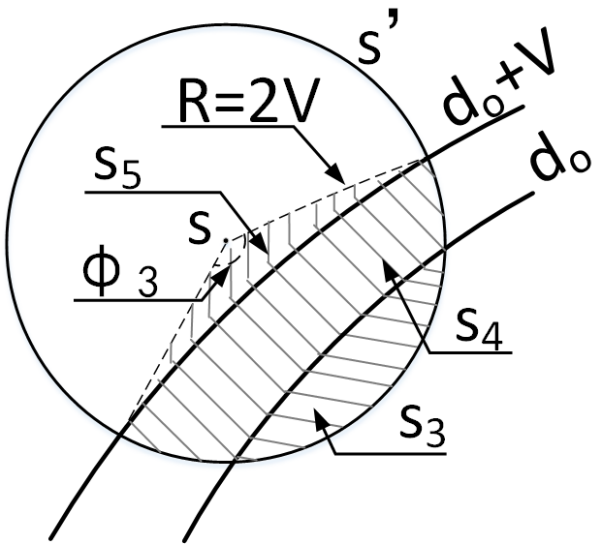


(a)

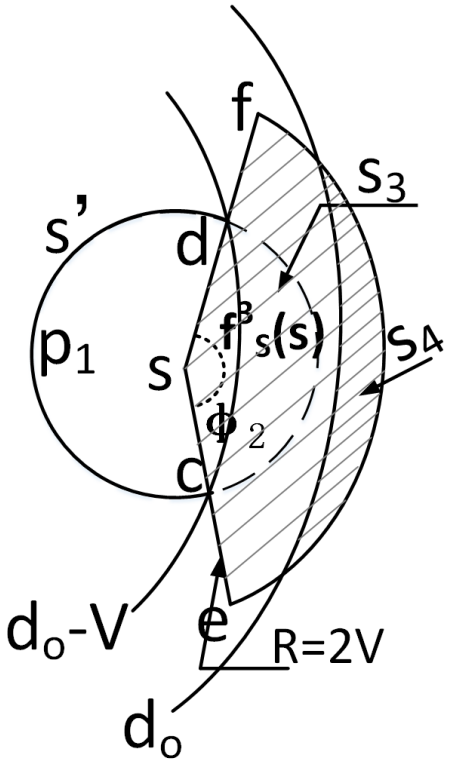


(b)

Figure 8.11. Inter-vehicle distance distribution when UAVs (a) follow different S&A protocols, and (b) S&R with different number of N.



(a)



(b)

Figure 8.12. (a) Illustration of Steps 3 and 4. (b) Illustration of Step 5.

CHAPTER 9

BAYESIAN ESTIMATION OF DEFECT PATTERNS IN COMPOSITE MATERIALS USING THROUGH-THICKNESS DIELECTRIC MEASUREMENTS

9.1 INTRODUCTION

Composite materials have been widely used in multifunctional applications, including biomedical (e.g. prostheses and devices), structural (e.g., vehicles and urban infrastructure), energy (e.g. conversion and storage), and communications (e.g. semiconductors and circuit boards) [188]. An effective approach that measures the defect patterns of materials and predicts the initiation of failures becomes extremely crucial to avoid "critical events" such as structural or functional failures. However, the diagnosis of the damage pattern in composite materials is challenging, since the interactions of the damage modes (e.g., matrix cracking, defects, delamination, fiber fracture etc.) are complex and may lead to critical fracture paths and in turn final failures. The relationship between damage development and material properties was studied in paper [3] and shown in Figure 9.1.

The current non-destructive evaluation (NDE) methods (e.g., Ultrasonic Testing (UT) and Acoustic Emission (AE)) have been used to find the area and size of damage in composite materials [189–192]. However, most of the NDE methods are point-to-point, and are not capable of diagnosing the overall state of the whole material. Detecting defects using dielectric responses has gained significant interests during the latest years [193–200], since it can find the overall state of material during its service life by measuring global dielectric properties. The most widely used indicator to test defects in composite materials is the dielectric resistivity [193–195]. The

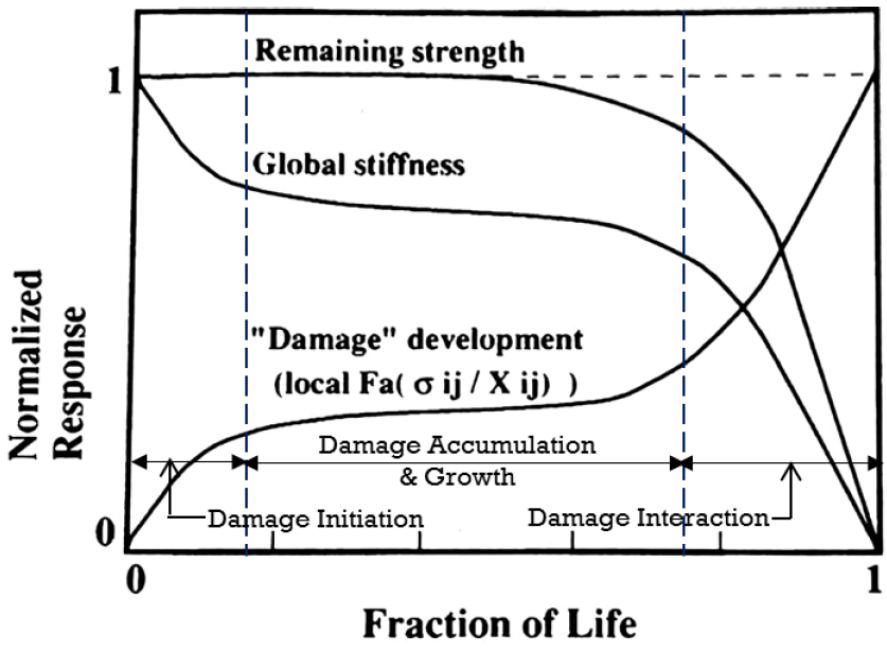


Figure 9.1. Relationship between damage development and material properties [3]..

electric resistance methods are capable of characterizing the damage states during the service life of composite material, however, they do not provide quantitative information about the specific damage modes such as the volumes of defects, orientations of the flaws, and the density of defects [201]. Broadband dielectric spectroscopy (BbDS) is the interaction of electromagnetic waves with matter in the frequency range from 10^{-6}Hz to -10^{12} Hz [197, 202–204]. BbDS has been used to detect the damage development in composite materials, since it can extract the material-level information, including the generation of micro-defects, and the volumes and orientations of those defects [197]. The BbDS method works as follows. A vector electric field is applied through the thickness of the specimen, and the dielectric properties are then measured to find out the current material state. Reifsnider showed that the shape, volume, and orientation of defects in materials can be found from the measured through-thickness dielectric permittivity [196–200]. These papers did not consider the interactions of

different defects, which can be important indicators of critical fracture paths. Per knowledge of the authors, there is no research work estimating the relative positions of defects from the global dielectric properties that considers the interactions between defects.

In this chapter, we find the relationship between the dielectric permittivity and the relative positions of defects using COMSOL®, and develop a Bayesian estimation method to estimate the relative positions of the defects from the global dielectric permittivity. The organization structure of this chapter is as follows. Section 9.2 presents the fundamental dielectric principle that leads to the detection and estimation method. Section 9.3 finds the relation between the relative positions of defects and the dielectric properties. Section 9.4 develops a Bayesian-based estimation method to estimate the current material state from the measured global permittivity.

9.2 The dielectric Principle and Modeling Framework

We first provide the principle of electromagnetic phenomena as the foundation for theanalysis in this chapter. A capacitor model is built to analyze the dielectric responses.

9.2.1 Principle of electromagnetic phenomena

In general, electromagnetic phenomena can be described by the Maxwell equations [198].

$$\nabla \cdot \vec{D} = \rho \tag{9.1}$$

$$\nabla \times \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t} \tag{9.2}$$

$$\nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0 \tag{9.3}$$

$$\nabla \cdot \vec{B} = 0 \tag{9.4}$$

where $\vec{D}$ is the dielectric displacement, ρ is the charge density, $\vec{H}$ is the magnetic field, $\vec{E}$ is the electric field, $\vec{B}$ is the magnetic induction, and $\vec{J}$ is the ohmic current density.

For linear materials, the interrelation between $\vec{D}$ and $\vec{E}$ is described as

$$\vec{D} = \varepsilon_0 \vec{E} + \vec{P} \tag{9.5}$$

where $\vec{P}$ is polarization determined by the charge density when there is no external source applied.

$$\nabla \cdot \vec{P} = -\rho \tag{9.6}$$

In addition, the relation between the dielectric displacement and electric field satisfies

$$\vec{D} = \varepsilon \varepsilon_0 \vec{E} \tag{9.7}$$

where ε_0 is the permittivity of vacuum, and ε is the relative permittivity of the material. Note that ε is a function of electric field frequency ω, and characterizes the material's dielectric behavior. In particular, the real part of ε represents the material's conductivity, and the imaginary part measures the material's dielectric loss.

Combining Equations (9.5) and (9.7), the relation between polarization and electric field is described as

$$\vec{P} = \varepsilon_0(\varepsilon - 1)\vec{E} = \chi \varepsilon_0 \vec{E} \tag{9.8}$$

where χ is the dielectric susceptibility determined uniquely by the material's relative permittivity ε.

We then describe the two important electromagnetic phenomena: polarization and interfacial polarization. In general, polarization refers to the phenomenon that electric charges accumulate at the material's interface when it is immersed in an electric field, as shown in Figure 9.2. In particular, the positive charges move along

the direction of the electric field and accumulate at one surface of the material, and the negative charges move in the opposite direction and accumulate at the opposite surface. The cumulative electric charges then generate a built-in electric field, which has the opposite direction to the external electric field shown in Figure 9.2. Note that the strength of the built-in electric field is determined by the real part of the material's permitivity. Larger ε leads to a stronger built-in electric field.

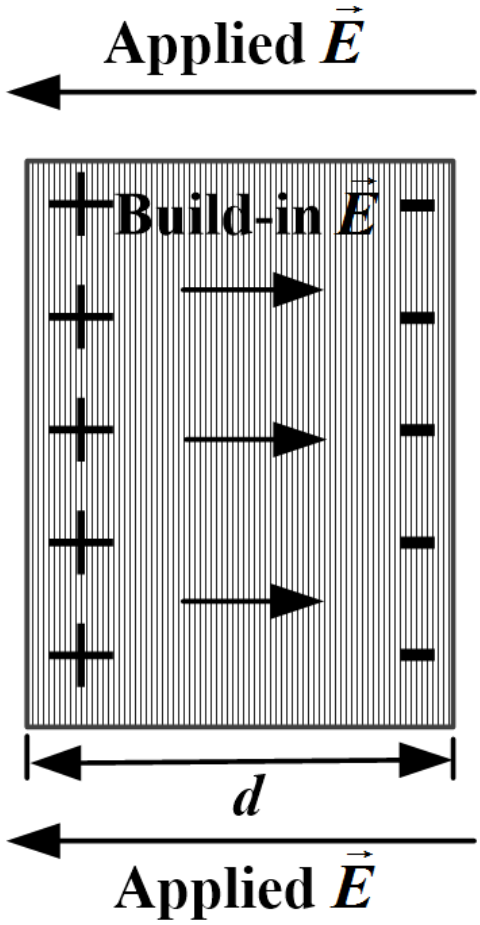


Figure 9.2. Illustration of interfacial polarization..

Similarly, when heterogeneous dielectric materials are immersed in an electric field, electric charges accumulate at the interfaces between the different constituents,

and this effect is called interfacial polarization [199]. As such, when there are defects in a material, a different "phases" with different conductivity and permittivity is introduced. At the boundary of these newly introduced defects, interfacial polarization occurs, and thus the global permittivity of the material changes.

9.2.2 The material modeling framework

A parallel plate capacitor model is designed to study the relation between a material's global dielectric permittivity and the relative positions of defects. In particular, a cylinder-shaped dielectric material is placed between two parallel metal plates, and the dielectric material and two metal plates compose a parallel plate capacitor as shown in Figure 9.3. The material's permittivity can be obtained by measuring the capacitance of the capacitor, when a direct voltage is applied on the two metal plates. The relation between the capacitance and the permittivity is described as

$$\varepsilon = \frac{C \cdot d \cdot 4\pi k}{\varepsilon_0 \cdot A} \tag{9.9}$$

where C is the capacitance, d and A are height and round area of the cylinder-shaped material respectively, and k is Coulomb's constant ($k \approx 9 \times 10^9 Nm^2C^2$). Here we use glass, which has the relative permittivity $\varepsilon = 6$, as the dielectric material. We use conductive copper for the metal plates.

The defects distributed in the dielectric material are modeled as ellipsoids filled with air. The model is shown in Figure 9.2.2. Here we introduce two defects in our capacitor model with the same size and orientation. We aim to study the impact of the two defects' relative position to the material's global permittivity. In particular, two groups of simulations are designed to test the impacts of the relative positions along the X and Z axes. In group 1, the two defects are distributed along the Z axis, i.e., they have the same X and Y coordinates, and differ in their Z coordinates. In

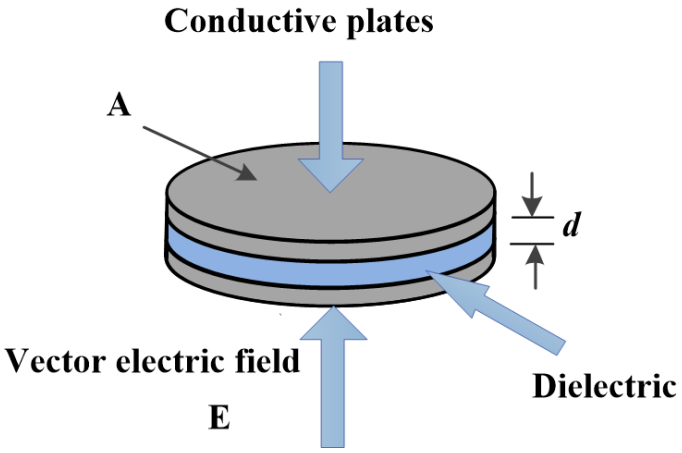


Figure 9.3. Capacitor model..

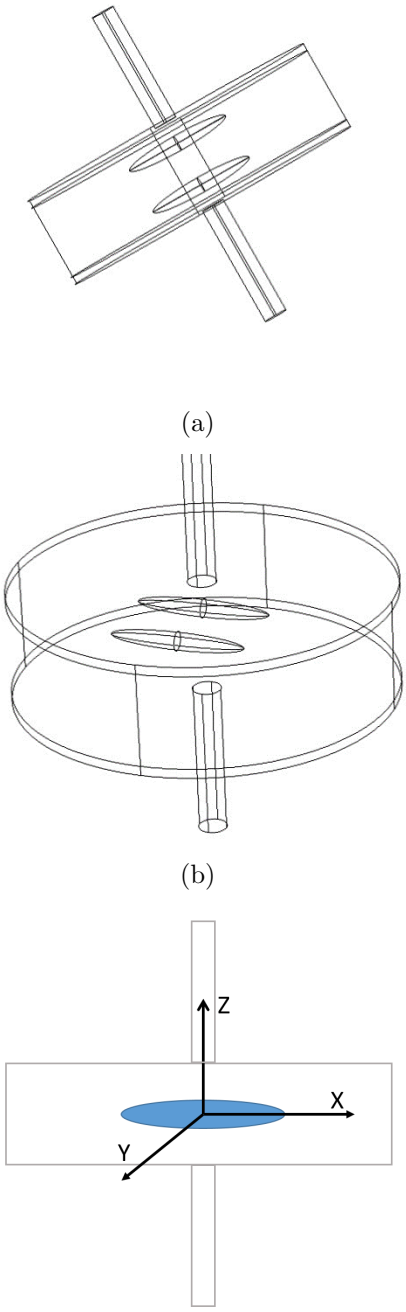
group 2, the two defects have the same X and Z coordinates, and differ in their Y coordinates. The three-dimensional coordinate system is shown in Figure 9.4(c).

9.3 Simulations in COMSOL Multiphysics®

In this section we study the impact of the defects' relative positions to materials' global permittivity. In particular, we use COMSOL Multiphysics® to simulate the capacitor models developed in Section 9.2.2. The material permittivity is then calculated from the measured capacitance.

9.3.1 Simulation setup

Here we use AC/DC Electrostatics module in COMSOL Multiphysics® to study the stationary circuit performance with a direct current. The size of the cylinder, which includes the dielectric material and two metal plates, is set as $10cm$ tall and $20cm$ wide, (i.e., with the radius of $10cm$). The defects are modeled as ellipsoids with the principal semi-axes $0.5cm$, $0.5cm$, and $3.5cm$ respectively. Simulations are



(a)

(b)

(c)

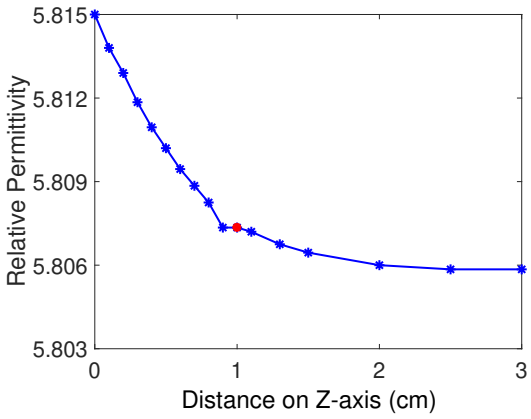
Figure 9.4. Capacitor model with two defects distributed along (a) the Z axis and (b) the Y axis. (c) Illustration of the coordinates..

conducted for the two groups of models described in Figure 9.4(a) and 9.4(b) respectively. In group 1, one defect is placed in the middle of the material with the coordinate $(0, 0, 0)$, and the other defect is placed at $(0, 0, 0)$, $(0, 0, 0.2)$, $(0, 0, 0.4)$, $(0, 0, 0.6)$, $(0, 0, 0.8)$, $(0, 0, 1)$, $(0, 0, 1.2)$, $(0, 0, 1.4)$, $(0, 0, 1.6)$, $(0, 0, 2)$, $(0, 0, 2.5)$, and $(0, 0, 3)$ respectively. In group 2, one defect is placed in the middle with the coordinate $(0, 0, 0)$, and the other defect is placed at $(0, 0, 0)$, $(0, 0.2, 0)$, $(0, 0.4, 0)$, $(0, 0.6, 0)$, $(0, 0.8, 0)$, $(0, 1, 0)$, $(0, 1.2, 0)$, $(0, 1.4, 0)$, $(0, 1.6, 0)$, $(0, 1.8, 0)$, $(0, 2, 0)$, $(0, 3, 0)$, and $(0, 4, 0)$ respectively.

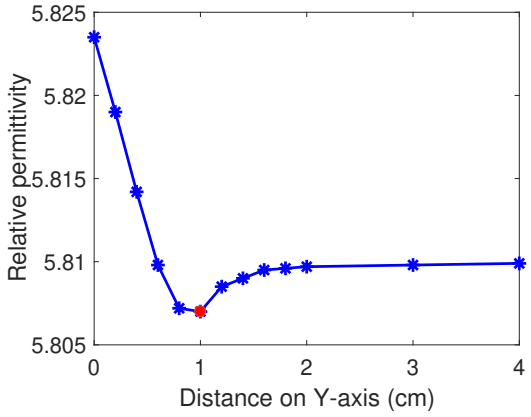
9.3.2 Simulation and analysis

Figures 9.5(a) and 9.5(b) show the relationships between the material's global permittivity and the inter-defect distances along the Z and Y axes respectively. Note that distance being $0cm$ means that the two defects coincide with their locations, which is equivalent to the single ellipsoidal defect case. Since the defect's radius is set as $0.5cm$, when the inter-defect distance is less than $1cm$, the two defects have a coinsident portion. The distances of $1cm$ are labeled with red dots in Figures 9.5(a) and 9.5(b). Figure 9.5(c) is the contour plot of the relationships. Note that the plots in the first quadrant (i.e., $y \in [0, 4]$, and $z \in [0, 2]$) is derived from the simulation in Comosl, and the plots in other quadrants are derived from the symmetrical characteristic, i.e., the two defects located at $(0, 0, 2)$ and $(0, 0, -2)$ have the same materials' permittivity.

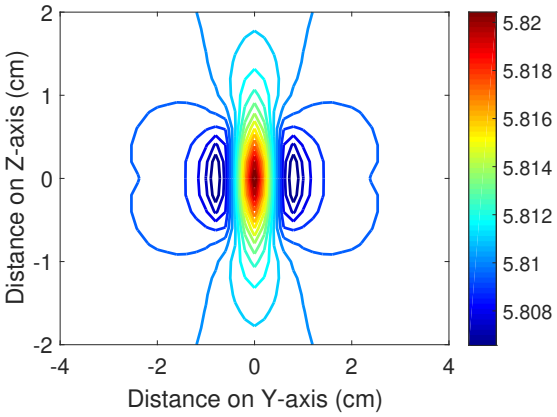
We can draw two observations from the avoe figures. First, when the inter-defect distance is less than $1cm$, the global permittivity decreases with the increase of distances along both the Y and Z axes. When the distance is less than $1cm$, the change of permittivity is dominated by the change of defects' volume. As such, it can be concluded that the increase of the defects' volume leads to the decrease of the



(a)



(b)



(c)

Figure 9.5. Relation between the material's permittivity and the inter-defect distance along the (a) Z axis, (b) Y axis, and (c) Y and Z axes of a contour plot.

global permittivity. Second, when the inter-defect distance is greater than $1cm$, in which case the defects' volume remains a constant, the material's global permittivity decreases with the increase of distance along the Z axis, or the decrease of the inter-defect distance along the Y axis.

The above two observations can be explained from the dielectric material properties and the principle of interfacial polarization. The first observation is straightforward since the relative permittivity of air is much less than that of the glass. Therefore, the increase of defect volume indicates more air in the material, and leads to less global permittivity. To explain the second observation, let us analyze the interfacial polarization phenomenon in the proposed capacitor models respectively. Denote the models described in Figures 9.4(a) and 9.4(b) as model 1 and 2 respectively, then when a direct voltage is added on the metal plates of the capacitor, an electric field, which is vertical to the metal plate, is produced as shown in Figure 9.6. With this electric field, the polarization and interfacial polarization are triggered, and charges accumulate at the surfaces of both the material and the defects. Note that the two defects have interations with each other because of the cumulative charges at their surfaces. In particular, for model 1, the polarization of one defect prompts the polarization of the other defect, as shown in Figure 9.6(a). As such, the increase of the inter-defect distance in model 1 weakens the interfacial polarization of the two defects, and thus results in smaller global permittivity. Similarly, for model 2, the polarization of one defect weakens the polarization of the other, as shown in Figure 9.6(b). As such, the increase of the inter-defect distance in model 2 enhances the global interfacial polarization, and thus leads to increased global permittivity.

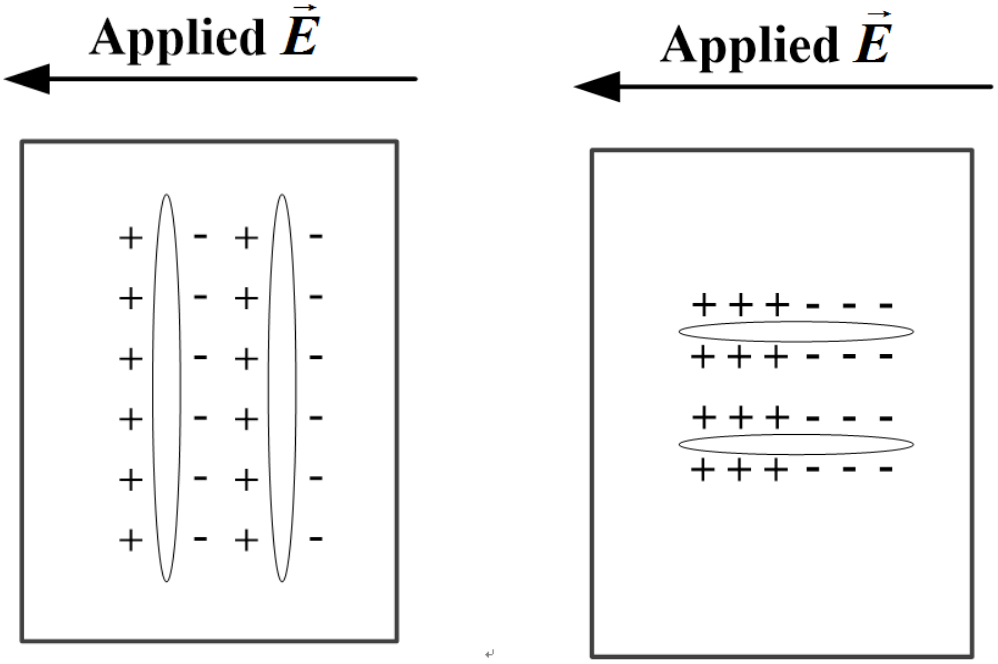


Figure 9.6. Illustration of interfacial polarization in (a) model 1, and (b) model 2.

9.4 Relative position estimation

In this section we first develop a Bayesian estimation based method to estimate the relative positions of the two defects given the permittivity measurements. Two numerical examples are then provided to illustrate the estimation method.

9.4.1 Bayesian estimation

Assume that the two defects are initially distributed uniformly in the material. The initial probability density function (pdf) of the defects' positions along the Y and Z axes are

$$f_Y(y) = \begin{cases} \frac{1}{D_1} & y \leqslant D_1 \\ 0 & \text{otherwise} \end{cases} \tag{9.10}$$

$$f_Z(z) = \begin{cases} \frac{1}{D_2} & z \leqslant D_2 \\ 0 & \text{otherwise} \end{cases} \tag{9.11}$$

respectively, where D_1 and D_2 are the width and height of the material respectively. Then the inter-defect distance distributions along the Y and Z axes are [205]

$$f_{|\Delta Y|}(\Delta y) = \begin{cases} \frac{2(D_1-\Delta y)}{D_1^2} & \Delta y \leqslant D_1 \\ 0 & \text{otherwise} \end{cases} \tag{9.12}$$

$$f_{|\Delta Z|}(\Delta z) = \begin{cases} \frac{2(D_2-\Delta z)}{D_2^2} & \Delta z \leqslant D_2 \\ 0 & \text{otherwise} \end{cases} \tag{9.13}$$

respectively.

Consider the measurement model described as

$$\hat{\epsilon} = g(\theta) + n \tag{9.14}$$

where $\hat{\epsilon}$ is the measured global permittivity, $\theta = [\Delta y, \Delta z]^T$ is the inter-defect relative position vector to be estimated. g is the relation between permittivity and the inter-defect distance, and n is the Gaussian noise with zero mean and variance σ^2, $n \backsim N(0, \sigma^2)$. Using the Bayesian estimation method, given the measured permittivity, the conditional pdf of the inter-defect distance can be derived as

$$f(\theta|\hat{\epsilon} = \epsilon_1) = \frac{f(\hat{\epsilon} = \epsilon_1|\theta) \cdot f(\theta)}{f(\hat{\epsilon} = \epsilon_1)} \tag{9.15}$$

Since $n \backsim N(0, \sigma^2)$, the conditional probability $f(\hat{\epsilon} = \epsilon_1|\theta)$ can be derived as

$$f(\hat{\epsilon} = \epsilon_1|\theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-\frac{1}{2\sigma^2}(\epsilon_1 - g(\theta))^2)} \tag{9.16}$$

As such, $f(\hat{\epsilon} = \epsilon_1)$ can be obtained from the integration of the conditional probability as

$$\begin{aligned} f(\hat{\epsilon} = \epsilon_1) &= \int f(\hat{\epsilon} = \epsilon_1|\theta) f(\theta) d\theta \\ &= \int_0^{D_2} \int_0^{D_1} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-\frac{1}{2\sigma^2}(\epsilon_1 - g(\theta))^2)} \cdot f(\theta) d\Delta y d\Delta z \end{aligned} \tag{9.17}$$

Combining Equations 9.15-9.17, the probability $f(\theta|\hat{\epsilon} = \epsilon_1)$ can be derived as:

$$f(\theta|\hat{\epsilon} = \epsilon_1) = \frac{\frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-\frac{1}{2\sigma^2}(\epsilon_1 - g(\theta))^2)} f(\theta)}{\int_0^{D_2} \int_0^{D_1} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-\frac{1}{2\sigma^2}(\epsilon_1 - g(\theta))^2)} f(\theta) d\Delta y d\Delta z} \tag{9.18}$$

As such, given the measured global permittivity $\hat{\epsilon}$, the distribution of the relative positions of the defects can be estimated using Equation (9.18).

9.4.2 Numerical examples

We use numerical examples to illustrate the proposed estimation method. Consider two dielectric material specimens, both of which are composed of glass and two ellipsoid defects filled with air. The two defects' relative positions in the two specimens are $(\Delta y_1, \Delta z_1) = (1, 0)$ and $(\Delta y_2, \Delta z_2) = (0, 1)$ respectively. Initially, we do not have any information concerning the defects' locations, and as such, we assume the two defects are distributed uniformly in the material as described in Equations (9.10) and (9.11). The initial distributions of the two defects' relative positions along the Y and Z axes (described in Equations (9.12) and (9.13)) are shown in Figure 9.4.2. Note that the "red" grid represents a "large" probability, and the "blue" grid means a "small" probability. Using the capacitor models described in Section 9.3, the permittivity of the two specimens are measured as $\hat{\epsilon}_1 = 5.8073$ and $\hat{\epsilon}_2 = 5.8126$ respectively. Then with the developed Bayesian estimation method described in equation (9.18), the probability distributions of the two defects' relative positions can be estimated from the measured permittivity as shown in Figures 9.8 and 9.9 for specimens 1 and 2 respectively.

It can be seen from the figures that 1) in specimen 1, the two defects' relative positions are "most probably" located in $(\Delta\hat{y}_1, \Delta\hat{z}_1) = (0.6, 0)$, where $\Delta\hat{y}$ and $\Delta\hat{z}$ are the estimated relative positions with the largest probability along the Y and Z axes respectively; 2) in specimen 2, the two defects' relative positions are "most probably" located in $(\Delta\hat{y}_1, \Delta\hat{z}_1) = (0, 0.8)$. As such, for both specimens, the estimation error is 0% along one axis, and within 10% for the other. This simulation study validates the effectiveness of the proposed estimation method. We leave the experimental validation to the future work.

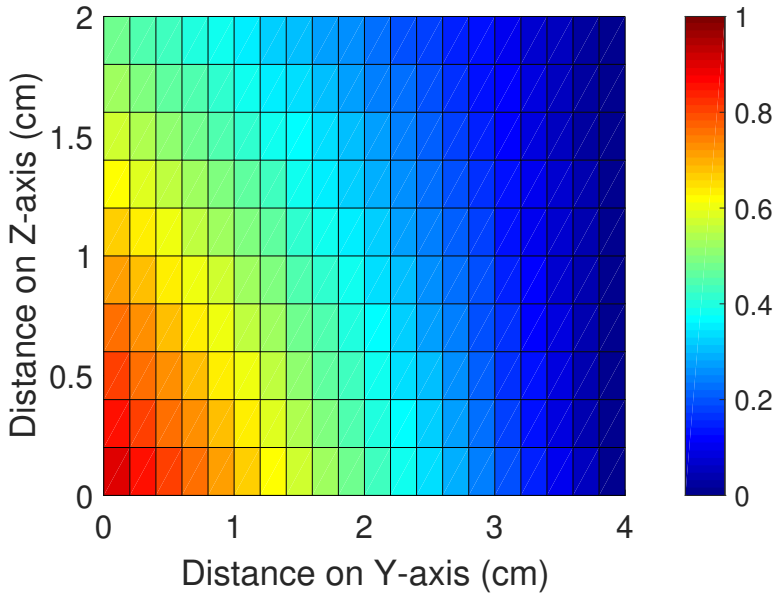


Figure 9.7. Initial distribution of the defects' relative positions.

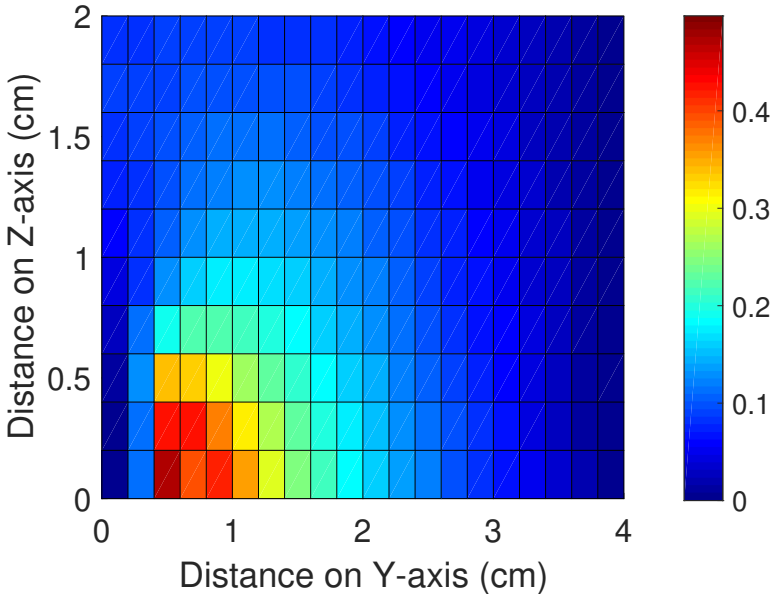


Figure 9.8. Estimated probability distribution of defects' relative positions in specimen 1.

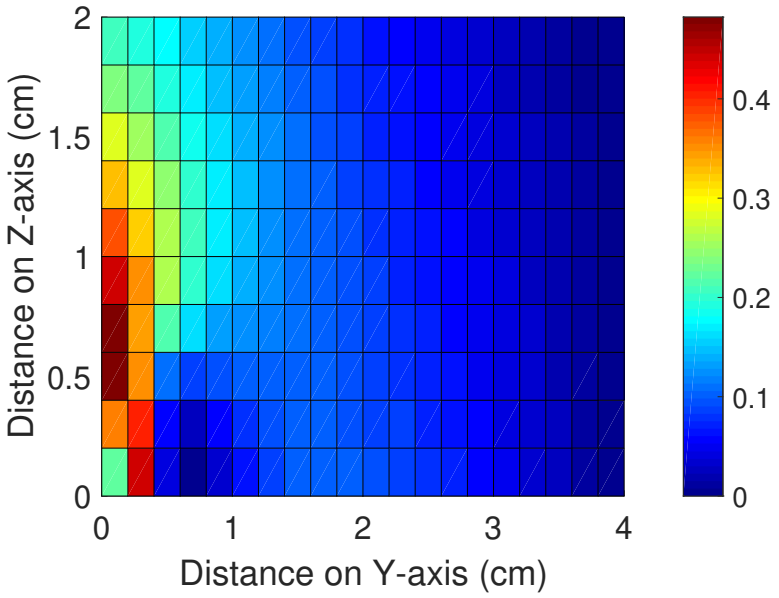


Figure 9.9. Estimated probability distribution of defects' relative positions in specimen 2.

CHAPTER 10

CONCLUSION AND FUTURE WORK

This dissertation studies distributed optimal policies for multi-agent systems under uncertainties. Both theoretical developments and their practical applications are investigated. Conclusions and future works are summarized as follows.

10.1 Theoretical Contributions

In Chapter 3, we develop optimal decision-making solutions for multi-agent random switching systems. An optimal controller and a practical state estimator, developed based on RMM, UKF, MPCM and RL constructs, are designed respectively. This work provides frameworks to solve controls and estimations for multiple agents moving with uncertain intentions or general highly flexible and uncertain movement patterns. Efficiency and accuracy of the proposed solutions are analyzed respectively.

In Chapter 4, to explore the optimal decisions for interacting agents in uncertain environments, we propose two novel stochastic differential games, where the system dynamics are modulated by randomly time-varying parameters. The optimal control policies for the two differential games, i.e., two-player zero-sum and multi-player nonzero-sum games, are obtained from the corresponding Hamiltonian functions. The system properties, including stability and Nash equilibrium, are analyzed. In addition, we develop IRL-based online learning algorithms for each game to find the optimal control solutions in real time. To evaluate the value functions with multi-dimensional 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.

In Chapter 5, we study multi-agent differential graphical games. We point out that the best response policies and minmax strategies for existing differential graphical games generally permit no global Nash equilibrium solution, where each agent only uses the state information of its own and its neighbors. To address this problem, we develop a novel graphical game formulation, with extra terms in the cost function to decouple the the HJ equation and thus guarantee the existence of a distributed value function. System properties, including stability and Nash equilibrium, are proven for this novel graphical game.

In Chapter 6, we study the stability margins of the graph-connected cooperative tracking systems. Unlike the single-agent system, where the phase and gain margins are constants, the stability margins of the cooperative tracking systems depend on the communication graph topology. In particular, both phase and gain margins are functions of $\underline{\lambda_R}$, which is the minimum real part of the eigenvalues of $L+G$. Motivated by this connection, we further study the value ranges of $\underline{\lambda_R}$ for general communication graphs. We find that $0 < \underline{\lambda_R} \leq 1$ holds for any possible communication graphs, and $\underline{\lambda_R} = 1$ if the communication graph is a directed tree. Linking the robustness analysis and the graph topology analysis, the limits of the phase and gain margins are then analyzed. In particular, we show that the directed tree graph promises the best phase and gain margins among all other possible communication graphs, and the performances are as good as the single-agent LQR system.

10.2 Application Contributions

In Chapter 7, we apply our developed decision-making solutions to antenna controls in the ACDA system, which aims to establish a robust long-distance air-to-air communication channel using pure directional antennas. In particular, to capture the uncertain intentions of UAVs executing surveillance-like missions for better tracking, we adopt a UAV ST RMM. To account for an unstable GPS environment, we apply the developed RL-based stochastic optimal control solution, which features a learning of communication RSSI models to provide an additional measurement that compensates GPS signals. This solution also features an integration of RL and MPCM to learn the environment-specific RSSI model and to provide online optimal control solutions. With the learned RSSI model, the optimal solutions in both GPS-available and GPS-denied environments are developed, respectively.

In Chapter 8, we propose a modeling framework of equipping RMMs with S&A protocols to quantitatively describe the highly random movement patterns of UAVs with safety constraints. We propose the RD RMM with the S&S protocol, and showe that stationary node distribution remains uniform, however the inter-vehicle distance distribution is not uniform any more. Based on the Markov analysis, we provide theoretical bounds for the stationary inter-vehicle distance distribution. We further define collisions between a pair of UAVs and among multiple UAVs based on the inter-vehicle distance, and found the stationary collision probabilities for both the independent RD RMM and the RD RMM equipped with the S&S protocol. We further define airspace capacity and derived it based on the stationary collision probabilities. Finally, we analyze the impact of model configurations, including travel time, sensing distance, and collision distance based on the proposed analytical framework. This analysis links local autonomy with global capacity, and provides insights on airspace capacity under highly flexible, variable, and uncertain mobility patterns of UAVs.

We found that the S&S protocol is not effective for UAVs of highly variable flight patterns. Compared to the independent RD RMM, it increases collision probability and reduces airspace capacity. The S&R and S&T protocols are more effective in increasing airspace capacity. The S&R performs the best among the three, however its performance is reduced with the increase of UAVs in the airspace.

In Chapter 9, we study the estimation of damage patterns, i.e., the relative positions of defects, in composite materials from the global dielectric response measurements. The interactions of defects in materials are important, as they can be indicators to the development of critical fracture paths. In particular, we first explain the fundamental dielectric principle that leads to the detection of defect patterns. Capacitor models are then built to measure the material permittivity, and the relationship between the dielectric permittivity and relative positions are found using COMSOL Multiphysics®. To estimate the relative positions of defects from the measured global permittivity, a Bayesian based estimation method is developed. It shows that the relative positions of defects can be estimated well from dielectric response measurements. This study lays the foundation for the early diagnosis and smart control of material systems to avoid potential structural or functional failures.

10.3 Future works

In the future work, we will enhance our current results to address more complex systems, e.g., systems with heterogeneous and nonlinear dynamics. In addition, we will continue to apply our developed solutions to real-world applications, e.g., autonomous driving systems and multi-UAV communication networks, and to further technology transfer to benefit our society.

REFERENCES

[1] Y. Wan, K. Namuduri, Y. Zhou, and S. Fu, "A smooth-turn mobility model for airborne networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3359–3370, 2013.

[2] S. Li, C. He, M. Liu, Y. Wan, Y. Gu, J. Xie, S. Fu, and K. Lu, "The design and implementation of aerial communication using directional antennas: Learning control in unknown communication environments," *IET Control Theory & Applications*, vol. 13, no. 17, pp. 2906–2916, 2019.

[3] K. Reifsnider, *Damage in composite materials: basic mechanisms, accumulation, tolerance, and characterization*, 1982, vol. 775.

[4] M. Liu, Y. Wan, and F. L. Lewis, "Adaptive optimal decision in multi-agent random switching systems," *IEEE Control Systems Letters*, vol. 4, pp. 265–270, 2019.

[5] M. Liu, Y. Wan, S. Li, and F. Lewis, "Learning and uncertainty-exploited directional antenna control for robust long-distance and broad-band aerial communication," *IEEE Transactions on Vehicle Technology*, vol. 69, pp. 593–606, 2020.

[6] M. Liu, Y. Wan, L. F. Lewis, and V. Lopez, "Adaptive optimal control for stochastic multi-agent differential games using on-policy and off-policy reinforcement learning," *accepted by IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[7] M. Liu, Y. Wan, V. Lopez, and F. L. Lewis, "Differential graphical game with distributed global nash solution," *submitted to IEEE Transactions on Control of Network Systems*, 2020.

[8] M. Liu, Y. Wan, Z. Lin, F. L. Lewis, J. Xie, and B. Jalaian, "Computational intelligence in uncertainty quantification for learning control and differential games," *Handbook on RL and Control*, 2020.

[9] M. Liu, Y. Wan, V. Lopez, and F. L. Lewis, "On the robustness of networked cooperative tracking systems," *submitted to IEEE Transactions on Automatic Control*, 2020.

[10] M. Liu, Y. Wan, F. L. Lewis, and E. Atkins, "Statistical properties of unmanned aerial vehicle networks subject to sense-and-avoid safety protocols," *submitted to IEEE Transactions on Intelligent Transportation Systems*, 2020.

[11] M. Liu and Y. Wan, "Analysis of random mobility model with sense and avoid protocols for uav traffic management," in *Proceedings of AIAA Information Systems-AIAA Infotech@ Aerospace*, Kissimmee, FL, 2018.

[12] 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, United Kingdom, 2019.

[13] M. Liu, Y. Wan, and F. L. Lewis, "Analysis of the random direction mobility model with a sense-and-avoid protocol," in *Proceedings of IEEE Globecom Workshops (GC Wkshps)*, Singapore, 2017.

[14] M. Liu, Y. Wan, V. Vadlamudi, F. L. Lewis, K. Reifsnider, and H. F. Wu, "Bayesian estimation of defect patterns in composite materials using through-thickness dielectric measurements," in *Proceedings of SPIE on Nondestructive*

Characterization and Monitoring of Advanced Materials, Aerospace, and Civil Infrastructure, Denver, CO, 2019.

[15] M. Liu, Y. Wan, S. Li, and F. L. Lewis, "A learning and uncertainty-exploited directional antenna control solution for robust aerial networking," in *Proceedings of IEEE Vehicle Technology Conference*, Honolulu, HI, 2019.

[16] S. Li, Y. Wan, S. Fu, M. Liu, and H. F. Wu, "Design and implementation of a remote uav-based mobile health monitoring system," in *Proceedings of SPIE on Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, and Civil Infrastructure*, Portland, OR, 2017.

[17] 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.

[18] B. Lian, Y. Wan, y. Zhang, M. Liu, and F. L. Lewis, "Distributed consensus-based kalman filtering for estimation with multiple moving targets," in *Proceedings of IEEE Conference on Decision and Control (CDC)*, Nice, France, 2019.

[19] V. G. Lopez, F. L. Lewis, Y. Wan, M. Liu, G. Hewer, and K. Estabridis, "Stability and robustness analysis of minmax solutions for differential graphical games," *Accepted by Automatica*, 2020.

[20] V. Lopez, F. L. Lewis, M. Liu, and Y. Wan, "Beyond nash solutions for differential graphical games," *submitted to IEEE Transactions on Automatic Control*, 2020.

[21] T. Li, Y. Wan, M. Liu, and F. L. Lewis, "Estimation of random mobility models using the expectation-maximization method," in *Proceedings of IEEE 14th International Conference on Control and Automation (ICCA)*, Anchorage, AK, 2018.

[22] P. An, , M. Liu, Y. Wan, and F. L. Lewis, "Multi-player h-infinity differential games using on-policy and off-policy reinforcement learning," in *submitted to IEEE International Conference on Control Automation (ICCA)*, Sapporo, Hokkaido, 2020.

[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] S. J. Julier and J. K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford, Tech. Rep., 1996.

[25] ——, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[26] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proceedings of IEEE American Control Conference*. Seattle, WA, 1995.

[27] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *Proceedings of Signal processing, sensor fusion, and target recognition VI*, 1997.

[28] E. A. Wan and R. Van Der Merwe, "The unscented kalman filter," *Kalman filtering and neural networks*, pp. 221–280, 2001.

[29] J. Xie, Y. Wan, J. H. Kim, S. Fu, and K. Namuduri, "A survey and analysis of mobility models for airborne networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1221–1238, 2014.

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

[31] 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.

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

[33] D. P. Bertsekas, "Value and policy iterations in optimal control and adaptive dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 500–509, 2017.

[34] F. Tatari, K. Vamvoudakis, and M. Mazouchi, "Optimal distributed learning for disturbance rejection in networked nonlinear games under unknown dynamics," *IET Control Theory & Applications*, 2018.

[35] K. Xiong, H. Zhang, and C. Chan, "Performance evaluation of ukf-based nonlinear filtering," *Automatica*, vol. 42, no. 2, pp. 261–270, 2006.

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

[37] M. J. Osborne *et al.*, *An introduction to game theory*. Oxford University Press, 2004.

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

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

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

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

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

[59] 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," *Journal of Operation and Automation in Power Engineering*, vol. 3, no. 1, pp. 83–93, 2015.

[60] Y. Matsuno, T. Tsuchiya, J. Wei, I. Hwang, and N. Matayoshi, "Stochastic optimal control for aircraft conflict resolution under wind uncertainty," *Aerospace Science and Technology*, vol. 43, pp. 77–88, 2015.

[61] 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.

[62] 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.

[63] 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.

[64] 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, 1979.

[65] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," *Management Science*, vol. 35, no. 11, pp. 1367–1392, 1989.

[66] S. Heinrich, "Multilevel monte carlo methods," in *Proceedings of International Conference on Large-Scale Scientific Computing*, Sozopol, Bulgaria, 2001.

[67] 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: Mathematical Modelling and Numerical Analysis*, vol. 48, no. 1, pp. 259–283, 2014.

[68] T. Hachisuka, W. Jarosz, R. P. Weistroffer, K. Dale, G. Humphreys, M. Zwicker, and H. W. Jensen, "Multidimensional adaptive sampling and reconstruction for ray tracing," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, p. 33, 2008.

[69] 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.

[70] 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, pp. 98–103, 2017.

[71] 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.

[72] 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.

[73] 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.

[74] F. Kozin, "A survey of stability of stochastic systems," *Automatica*, vol. 5, no. 1, pp. 95–112, 1969.

[75] 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.

[76] B. M. Chen, Z. Lin, and Y. Shamash, *Linear systems theory: a structural decomposition approach*. Springer Science & Business Media, 2004.

[77] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proceedings of IEEE American Control Conference (ACC)*, Portland, OR, 2005.

[78] W. Ren and R. W. Beard, *Distributed consensus in multi-vehicle cooperative control*. Springer, 2008.

[79] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[80] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: algorithms and theory," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 401–420, 2006.

[81] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[82] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[83] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 781–786, 2013.

[84] Q. Zhu and T. Basar, "Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 46–65, 2015.

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

[86] J. Li, H. Modares, T. Chai, F. L. Lewis, and L. Xie, "Off-policy reinforcement learning for synchronization in multiagent graphical games," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2434–2445, 2017.

[87] M. Mazouchi, M. B. Naghibi-Sistani, and S. K. H. Sani, "A novel distributed optimal adaptive control algorithm for nonlinear multi-agent differential graphical games," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 331–341, 2017.

[88] R. Kamalapurkar, J. R. Klotz, P. Walters, and W. E. Dixon, "Model-based reinforcement learning in differential graphical games," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 423–433, 2016.

[89] H. Cao, E. Ertin, and A. Arora, "Minimax equilibrium of networked differential games," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 3, no. 4, p. 14, 2008.

[90] Z. Qu and M. A. Simaan, "A design of distributed game strategies for networked agents," *IFAC Proceedings Volumes*, vol. 42, no. 20, pp. 270–275, 2009.

[91] V. G. Lopez, F. L. Lewis, Y. Wan, E. N. Sanchez, and L. Fan, "Solutions for multiagent pursuit-evasion games on communication graphs: Finite-time capture and asymptotic behavior," *IEEE Transactions on Automatic Control*, 2019.

[92] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative control of multi-agent systems: optimal and adaptive design approaches*. Springer Science & Business Media, 2013.

[93] 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.

[94] G. F. Franklin, J. D. Powell, A. Emami-Naeini, and J. D. Powell, *Feedback control of dynamic systems*. Addison-Wesley Reading, MA, 1994, vol. 3.

[95] H. W. Bode *et al.*, "Network analysis and feedback amplifier design," 1945.

[96] I. M. Horowitz, *Synthesis of feedback systems*. Elsevier, 2013.

[97] H. Rosenbrock, "Design of multivariable control systems using the inverse nyquist array," in *Proceedings of the Institution of Electrical Engineers*, vol. 116, no. 11, 1969, pp. 1929–1936.

[98] P. McMorran, "Extension of the inverse nyquist method," *Electronics Letters*, vol. 6, no. 25, pp. 800–801, 1970.

[99] A. MacFarlane, "A survey of some recent results in linear multivariable feedback theory," *Automatica*, vol. 8, no. 4, pp. 455–492, 1972.

[100] A. MacFarlane and J. Belletrutti, "The characteristic locus design method," *Automatica*, vol. 9, no. 5, pp. 575–588, 1973.

[101] M. Safonov and M. Athans, "Gain and phase margin for multiloop LQG regulators," *IEEE Transactions on Automatic Control*, vol. 22, no. 2, pp. 173–179, 1977.

[102] J. Doyle, "Robustness of multiloop linear feedback systems," in *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, 1979, pp. 12–18.

[103] M. G. Safonov, "Stability margins of diagonally perturbed multivariable feedback systems," in *IEE Proceedings D (Control Theory and Applications)*, vol. 129, no. 6, 1982, pp. 251–256.

[104] T. T. Tsao, F. C. Lee, and D. Augenstein, "Relationship between robustness /*mu*-analysis and classical stability margins," in *IEEE Aerospace Conference Proceedings*, vol. 4, 1998, pp. 481–486.

[105] K. Zhou, J. C. Doyle, K. Glover, *et al.*, *Robust and optimal control*. Prentice hall New Jersey, 1996, vol. 40.

[106] V. Balakrishnan, "Lyapunov functionals in complex/spl mu/analysis," *IEEE Transactions on Automatic Control*, vol. 47, no. 9, pp. 1466–1479, 2002.

[107] C. T. Lawrence, A. L. Tits, and P. Van Dooren, "A fast algorithm for the computation of an upper bound on the μ-norm," *Automatica*, vol. 36, no. 3, pp. 449–456, 2000.

[108] M. Halton, P. Iordanov, and J. Mooney, "Robust analysis and synthesis design tools for digitally controlled power converters," in *IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2015, pp. 317–322.

[109] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proceedings of IEEE American Control Conference*, 2005, pp. 1859–1864.

[110] J. G. Bender, "An overview of systems studies of automated highway systems," *IEEE Transactions on vehicular technology*, vol. 40, no. 1, pp. 82–99, 1991.

[111] J. Russell Carpenter, "Decentralized control of satellite formations," *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, vol. 12, no. 2-3, pp. 141–161, 2002.

[112] R. W. Beard, T. W. McLain, and M. Goodrich, "Coordinated target assignment and intercept for unmanned air vehicles," in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 3, 2002, pp. 2581–2586.

[113] K. H. Movric and F. L. Lewis, "Cooperative optimal control for multi-agent systems on directed graph topologies," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 769–774, 2013.

[114] H. Zhang, T. Feng, G.-H. Yang, and H. Liang, "Distributed cooperative optimal control for multiagent systems on directed graphs: An inverse optimal approach," *IEEE Transactions on Cybernetics*, vol. 45, no. 7, pp. 1315–1326, 2014.

[115] H. Zhang, F. L. Lewis, and Z. Qu, "Lyapunov, adaptive, and optimal design techniques for cooperative systems on directed communication graphs," *IEEE transactions on industrial electronics*, vol. 59, no. 7, pp. 3026–3041, 2011.

[116] W. Ren, R. W. Beard, and E. M. Atkins, "Information consensus in multivehicle cooperative control," *IEEE Control systems magazine*, vol. 27, no. 2, pp. 71–82, 2007.

[117] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[118] J. A. Fax and R. M. Murray, "Information flow and cooperative control of vehicle formations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1465–1476, 2004.

[119] H. Zhang, F. L. Lewis, and A. Das, "Optimal design for synchronization of cooperative systems: state feedback, observer and output feedback," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1948–1952, 2011.

[120] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative control of multi-agent systems: optimal and adaptive design approaches*. Springer Science & Business Media, 2013.

[121] H. K. Khalil, *Nonlinear systems*. Upper Saddle River, 2002.

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

[123] C. A. Desoer and M. Vidyasagar, *Feedback systems: input-output properties*. Siam, 1975, vol. 55.

[124] M. Fiedler and V. Ptak, "On matrices with non-positive off-diagonal elements and positive principal minors," *Czechoslovak Mathematical Journal*, vol. 12, no. 3, pp. 382–400, 1962.

[125] S. Winkler, S. Zeadally, and K. Evans, "Privacy and civilian drone use: The need for further regulation," *IEEE Security & Privacy*, vol. 16, no. 5, pp. 72–80, 2018.

[126] O. S. Oubbati, N. Chaib, A. Lakas, P. Lorenz, and A. Rachedi, "Uav-assisted supporting services connectivity in urban vanets," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3944–3951, 2019.

[127] S. Hu, "On ergodic capacity and optimal number of tiers in uav-assisted communication systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2814–2824, 2019.

[128] Y. Wan and S. Fu, "Communicating in remote areas or disaster situations using unmanned aerial vehicles," *Homeland Security Today Magazine*, pp. 32–35, 2015.

[129] K. Li, R. C. Voicu, S. S. Kanhere, W. Ni, and E. Tovar, "Energy efficient legitimate wireless surveillance of uav communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2283–2293, 2019.

[130] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2624–2661, 2016.

[131] B.-N. Cheng, A. Coyle, S. McGarry, I. Pedan, L. Veytser, and J. Wheeler, "Characterizing routing with radio-to-router information in a heterogeneous airborne network," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4183–4195, 2013.

[132] A. I. Alshbatat and L. Dong, "Performance analysis of mobile ad hoc unmanned aerial vehicle communication networks with directional antennas," *International Journal of Aerospace Engineering*, vol. 2010, 2010.

[133] J. Chen, J. Xie, Y. Gu, S. Li, S. Fu, Y. Wan, and K. Lu, "Long-range and broadband aerial communication using directional antennas (acda): design and implementation," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 10 793–10 805, 2017.

[134] Y. Gu, M. Zhou, S. Fu, and Y. Wan, "Airborne wifi networks through directional antennae: An experimental study," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, New Orleans, LA, 2015.

[135] J. Xie, F. AI-Emrani, Y. Gu, Y. Wan, and S. Fu, "Uav-carried long-distance wifi communication infrastructure," in *Proceedings of AIAA Infotech@ Aerospace*, San Diego, CA, 2016.

[136] E. Yanmaz, R. Kuschnig, and C. Bettstetter, "Achieving air-ground communications in 802.11 networks with three-dimensional aerial mobility," in *Proceedings of IEEE INFOCOM*, Turin, Italy, 2013.

[137] C. Danilov, T. R. Henderson, T. Goff, J. H. Kim, J. Macker, J. Weston, N. Neogi, A. Ortiz, and D. Uhlig, "Experiment and field demonstration of a 802.11-based ground-uav mobile ad-hoc network," in *Proceedings of IEEE Military Communications Conference (MILCOM)*, Boston, MA, 2009.

[138] K. Wakita, J. Huang, P. Di, K. Sekiyama, and T. Fukuda, "Human-walking-intention-based motion control of an omnidirectional-type cane robot," *IEEE/ASME Transactions On Mechatronics*, vol. 18, no. 1, pp. 285–296, 2013.

[139] Y. Li and S. S. Ge, "Human-robot collaboration based on motion intention estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 1, pp. 1007–1014, 2014.

[140] H. Modares, I. Ranatunga, F. L. Lewis, and D. O. Popa, "Optimized assistive human–robot interaction using reinforcement learning," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 655–667, 2016.

[141] T. Takeda, Y. Hirata, and K. Kosuge, "Dance step estimation method based on hmm for dance partner robot," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 2, pp. 699–706, 2007.

[142] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán, "Exploiting map information for driver intention estimation at road intersections," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, Baden-Baden, Germany, 2011.

[143] J. Xie, Y. Wan, K. Namuduri, S. Fu, G. L. Peterson, and J. F. Raquet, "Estimation and validation of the 3d smooth-turn mobility model for airborne networks," in *Proceedings of IEEE Military Communications Conference (MILCOM)*, San Diego, CA, 2013.

[144] J. Yan, Y. Wan, S. Fu, X. J., S. Li, and K. Lu, "Rssi-based decentralized control for robust long-distance aerial networks using directional antennas," *IET Control Theory and Applications*, vol. 11, no. 11, pp. 1838–1847, 2016.

[145] M. K. Haider and E. W. Knightly, "Mobility resilience and overhead constrained adaptation in directional 60 ghz wlans: protocol design and system implementation," in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Paderborn, Germany, 2016.

[146] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems*, vol. 32, no. 6, pp. 76–105, 2012.

[147] 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 11th Asian Control Conference (ASCC)*, Gold Coast, QLD, Australia, 2017.

[148] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on uav cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Communications Surveys & Tutorials*, 2019.

[149] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, 2016.

[150] B. Galkin, J. Kibilda, and L. A. DaSilva, "Coverage analysis for low-altitude uav networks in urban environments," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, Singapore, 2017.

[151] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Drone small cells in the clouds: Design, deployment and performance analysis," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015.

[152] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[153] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*. PTR New Jersey, 1996.

[154] L. Josefsson and P. Persson, *Conformal array antenna theory and design*. John Wiley & Sons, 2006.

[155] "Ubiquity nanostation loco m5," in *https://www.ubnt.com/airmax/nanostationm/*.

[156] J. D. Kraus, *Antennas*. McGraw-Hill Education, 1988.

[157] A. Gil, J. Segura, and N. M. Temme, *Numerical methods for special functions*. Siam, 2007.

[158] W. Y. Chung and E. E. L. Lau, "Enhanced rssi-based real-time user location tracking system for indoor and outdoor environments," in *Proceedings of International Conference on Convergence Information Technology (ICCIT)*, Gyeongju, South Korea, 2007.

[159] Z. Fang, Z. Zhao, D. Geng, Y. Xuan, L. Du, and X. Cui, "Rssi variability characterization and calibration method in wireless sensor network," in *Proceedings of IEEE International Conference on Information and Automation*, Harbin, China, 2010.

[160] A. P. Subramanian, H. Lundgren, T. Salonidis, and D. Towsley, "Topology control protocol using sectorized antennas in dense 802.11 wireless networks," in *Proceedings of 17th IEEE International Conference on Network Protocols*. Princeton, NJ, 2009.

[161] J. Xie, Y. Wan, B. Wang, S. Fu, K. Lu, and J. H. Kim, "A comprehensive 3-dimensional random mobility modeling framework for airborne networks," *IEEE Access*, vol. 6, pp. 22 849–22 862, 2018.

[162] N. E. Nahi, *Estimation theory and applications*. Wiley New York, 1969.

[163] X. Yuan, Z. Feng, W. Xu, W. Ni, A. Zhang, Z. Wei, and R. P. Liu, "Capacity analysis of uav communications: Cases of random trajectories," *IEEE Transactions on Vehicular Technology*, 2018.

[164] I. Maza, K. Kondak, M. Bernard, and A. Ollero, "Multi-uav cooperation and control for load transportation and deployment," in *Proceedings of the 2nd International Symposium on UAVs, Springer*. Reno, NV, 2009.

[165] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Communications Magazine*, pp. 36–42, 2016.

[166] Y. Wan, S. Fu, J. Zander, and P. Mosterman, "Transforming on-demand communications with drones: the needs, analyses, and solutions," *Homeland Security Today Magazine*, pp. 32–35, 2015.

[167] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-uav," *IEEE Communications Magazine*, pp. 82–89, 2018.

[168] "Code of federal regulations," https://www.law.cornell.edu/cfr/text/14/91.113.

[169] J. Xie, Y. Wan, J. H. Kim, S. Fu, and K. Namuduri, "A survey and analysis of mobility models for airborne networks," *IEEE Communications Surveys & Tutorials*, pp. 1221–1238, 2014.

[170] P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *Proceedings of 24th Annual Joint Conference of the IEEE Computer and Communications Societies*. Miami, FL, 2005.

[171] Y. Wan, K. Namuduri, Y. Zhou, and S. Fu, "A smooth-turn mobility model for airborne networks," *IEEE Transactions on Vehicular Technology*, pp. 3359–3370, 2013.

[172] J. Xie, Y. Wan, B. Wang, S. Fu, K. Lu, and J. H. Kim, "A comprehensive 3-dimensional random mobility modeling framework for airborne networks," *IEEE Access*, pp. 22 849–22 862, 2018.

[173] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, no. 5, pp. 483–502, 2002.

[174] C. Cho, S.-m. Jun, E. Paik, and K. Park, "Rate control for streaming services based on mobility prediction in wireless mobile networks," in *Proceedings of IEEE Wireless Communications and Networking Conference*. New Orleans, LA, 2005.

[175] Z. Cheng and W. B. Heinzelman, "Exploring long lifetime routing (llr) in ad hoc networks," in *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. Venice, Italy, 2004.

[176] G. Lim, K. Shin, S. Lee, H. Yoon, and J. S. Ma, "Link stability and route lifetime in ad-hoc wireless networks," in *Proceedings of International Conference on Parallel Processing Workshops*. Vancouver, BC, Canada, 2002.

[177] Y. Wan, C. Taylor, S. Roy, C. Wanke, and Y. Zhou, "Dynamic queuing network model for flow contingency management," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1380–1392, 2013.

[178] Y. Zhou, Y. Wan, S. Roy, C. Taylor, C. Wanke, D. Ramamurthy, and J. Xie, "Multivariate probabilistic collocation method for effective uncertainty evalu-

ation with application to air traffic flow management," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1347–1363, 2014.

[179] J. Krozel, M. Peters, and K. Bilimoria, "A decentralized control strategy for distributed air/ground traffic separation," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*. Denver, CO, 2000.

[180] V. Bulusu and V. Polishchuk, "A threshold based airspace capacity estimation method for uas traffic management system," in *Proceedings of IEEE Systems Conference*. Montreal, Quebec, Canada, 2017.

[181] Z. J. Haas, "A new routing protocol for the reconfigurable wireless networks," in *Proceedings of IEEE 6th International Universal Personal Communications*. San Diego, CA, 1997.

[182] B. Cook, K. Cohen, and E. H. Kivelevitch, "A fuzzy logic approach for low altitude uas traffic management (utm)," in *Proceedings of AIAA Infotech@ Aerospace*. San Diego, CA, 2016.

[183] R. Breil, D. Delahaye, L. Lapasset, and E. Féron, "Multi-agent systems to help managing air traffic structure," *Journal of Aerospace Operations*, pp. 1–30, 2017.

[184] L. Sedov and V. Polishchuk, "Centralized and distributed utm in layered airspace."

[185] J. Stachurski and V. Martin, "Computing the distributions of economic models via simulation," *Econometrica*, pp. 443–450, 2008.

[186] S. G. Henderson and P. W. Glynn, "Computing densities for markov chains via simulation," *Mathematics of Operations Research*, pp. 375–400, 2001.

[187] 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*, pp. 545–563, 2015.

[188] R. F. Gibson, *Principles of composite material mechanics*. CRC Press, 2011.

[189] T. Prasse, F. Michel, G. Mook, K. Schulte, and W. Bauhofer, "A comparative investigation of electrical resistance and acoustic emission during cyclic loading of cfrp laminates," *Composites Science and Technology*, vol. 61, pp. 831–835, 2001.

[190] L. Rippert, M. Wevers, and S. Van Huffel, "Optical and acoustic damage detection in laminated cfrp composite materials," *Composites Science and Technology*, vol. 60, pp. 2713–2724, 2000.

[191] K. Maslov, R. Y. Kim, V. K. Kinra, and N. J. Pagano, "A new technique for the ultrasonic detection of internal transverse cracks in carbon-fibre/bismaleimide composite laminates," *Composites Science and Technology*, vol. 60, pp. 2185–2190, 2000.

[192] K. V. Steiner, R. F. Eduljee, X. Huang, and J. W. Gillespie Jr, "Ultrasonic nde techniques for the evaluation of matrix cracking in composite laminates," *Composites Science and Technology*, vol. 53, pp. 193–198, 1995.

[193] A. Vavouliotis, A. Paipetis, and V. Kostopoulos, "On the fatigue life prediction of cfrp laminates using the electrical resistance change method," *Composites Science and Technology*, vol. 71, pp. 630–642, 2011.

[194] A. Todoroki, K. Omagari, Y. Shimamura, and H. Kobayashi, "Matrix crack detection of cfrp using electrical resistance change with integrated surface probes," *Composites Science and Technology*, vol. 66, pp. 1539–1545, 2006.

[195] P. Irving and C. Thiagarajan, "Fatigue damage characterization in carbon fibre composite materials using an electrical potential technique," *Smart Materials and Structures*, vol. 7, p. 456, 1998.

[196] R. Raihan, K. Reifsnider, D. Cacuci, and Q. Liu, "Dielectric signatures and interpretive analysis for changes of state in composite materials," *ZAMM-Journal*

of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, vol. 95, pp. 1037–1045, 2015.

[197] R. Raihan, J.-M. Adkins, J. Baker, F. Rabbi, and K. Reifsnider, "Relationship of dielectric property change to composite material state degradation," *Composites Science and Technology*, vol. 105, pp. 160–165, 2014.

[198] K. Reifsnider, M. R. Raihan, and V. Vadlamudi, "Heterogeneous fracture mechanics for multi-defect analysis," *Composite Structures*, vol. 156, pp. 20–28, 2016.

[199] Q. Liu and K. L. Reifsnider, "Heterogeneous mixtures of elliptical particles: directly resolving local and global properties and responses," *Journal of Computational Physics*, vol. 235, pp. 161–181, 2013.

[200] J. Baker, J. M. Adkins, F. Rabbi, Q. Liu, K. Reifsnider, and R. Raihan, "Meso-design of heterogeneous dielectric material systems: Structure property relationships," *Journal of Advanced Dielectrics*, vol. 4, pp. 1 450 008–1 450 017, 2014.

[201] M. R. P. Elenchezhian, V. Vadlamudi, R. Md Raihan, and K. Reifsnider, "Damage precursor identification in composite laminates using data driven approach," in *Proceedings of AIAA Scitech Forum*, San Diego, CA, 2019.

[202] P. D. Fazzino, K. L. Reifsnider, and P. Majumdar, "Impedance spectroscopy for progressive damage analysis in woven composites," *Composites Science and Technology*, vol. 69, pp. 2008–2014, 2009.

[203] D. Bekas and A. Paipetis, "Damage monitoring in nanoenhanced composites using impedance spectroscopy," *Composites Science and Technology*, vol. 134, pp. 96–105, 2016.

[204] P. Majumdar, Y. Bhuiyan, J. Clifford, F. Haider, and K. Reifsnider, "Multi-physical description of material state change in composite materials," in *Pro-*

ceedings of the Society for the Advancement of Material and Process Engineering, Baltimore, MD, 2015.

[205] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010.

BIOGRAPHICAL STATEMENT

Mushuang Liu was born in Shandong, China, in 1994. She received her B.S. degree in Electrical Engineering from University of Electronic Science and Technology of China, Chendu, China in 2016. She is now working toward her Ph.D. degree in the department of Electrical Engineering in University of Texas at Arlington. Her research interests include distributed decisions for multi-agent systems, uncertain systems, multi-player games, graphical games, reinforcement learning, and their applications to UAV networking and UAV traffic management.