SAFE CONTROL DESIGN FOR UNCERTAIN SYSTEMS

By

Zahra Marvi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical Engineering- Doctor of Philosophy

2021

ABSTRACT

SAFE CONTROL DESIGN FOR UNCERTAIN SYSTEMS

By

Zahra Marvi

This dissertation investigates the problem of safe control design for systems under model and environmental uncertainty. Reinforcement learning (RL) provides an interactive learning framework in which the optimal controller is sequentially derived based on instantaneous reward. Although powerful, safety consideration is a barrier to the wide deployment of RL algorithms in practice. To overcome this problem, we proposed an iterative safe off-policy RL algorithm. The cost function that encodes the designer's objectives is augmented with a control barrier function (CBF) to ensure safety and optimality. The proposed formulation provides a look-ahead and proactive safety planning, in which the safety is planned and optimized along with the performance to minimize the intervention with the optimal controller. Extensive safety and stability analysis is provided and the proposed method is implemented using the off-policy algorithm without requiring complete knowledge about the system dynamics. This line of research is then further extended to have a safety and stability guarantee even during the data collection and exploration phases in which random noisy inputs are applied to the system. However, satisfying the safety of actions when little is known about the system dynamics is a daunting challenge. We present a novel RL scheme that ensures the safety and stability of the linear systems during the exploration and exploitation phases. This is obtained by having a concurrent model learning and control, in which an efficient learning scheme is employed to prescribe the learning behavior. This characteristic is then employed to apply only safe and stabilizing controllers to the system. First, the prescribed errors are employed in a novel adaptive robustified control barrier function (AR-CBF) which guarantees that the states of the system remain in the safe set even when the learning is incomplete. Therefore, the noisy input in the exploratory data collection phase and the optimal controller in the exploitation phase are minimally altered such that the AR-CBF criterion

is satisfied and, therefore, safety is guaranteed in both phases. It is shown that under the proposed prescribed RL framework, the model learning error is a vanishing perturbation to the original system. Therefore, a stability guarantee is also provided even in the exploration when noisy random inputs are applied to the system. A learning-enabled barrier-certified safe controllers for systems that operate in a shared and uncertain environment is then presented. A safety-aware loss function is defined and minimized to learn the uncertain and unknown behavior of external agents that affect the safety of the system. The loss function is defined based on safe set error, instead of the system model error, and is minimized for both current samples as well as past samples stored in the memory to assure a fast and generalizable learning algorithm for approximating the safe set. The proposed model learning and CBF are then integrated together to form a learning-enabled zeroing CBF (L-ZCBF), which employs the approximated trajectory information of the external agents provided by the learned model but shrinks the safety boundary in case of an imminent safety violation using instantaneous sensory observations. It is shown that the proposed L-ZCBF assures the safety guarantees during learning and even in the face of inaccurate or simplified approximation of external agents, which is crucial in highly interactive environments. Finally, the cooperative capability of agents in a multi-agent environment is investigated for the sake of safety guarantee. CBFs and information-gap theory are integrated to have robust safe controllers for multi-agent systems with different levels of measurement accuracy. A cooperative framework for the construction of CBFs for every two agents is employed to maximize the horizon of uncertainty under which the safety of the overall system is satisfied. The information-gap theory is leveraged to determine the contribution and share of each agent in the construction of CBFs. This results in the highest possible robustness against measurement uncertainty. By employing the proposed approach in constructing CBF, a higher horizon of uncertainty can be safely tolerated and even the failure of one agent in gathering accurate local data can be compensated by cooperation between agents. The effectiveness of the proposed methods is extensively examined in simulation results.

To researchers who work hard and honest to make a positive impact to the world

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Bahare Kiumarsi, for her continuous support, time and superb guidance throughout my Ph.D. studies. I would like to express my appreciation to the members of my Ph.D. committee, Dr. Hayder Radha, Dr. Hamidreza Modares and Dr. Xiaboo Tan for their time, support and valuable suggestions. I would also like to thank all faculty and staff with Michigan State University; specially, I would like to thank Dr. Tim Hogan, Dr. Andrew Mason, Dr. Katy Luchini Colbry, and Dr. John Papapolymerou for their support. A special appreciation to my family for their love and support throughout my life, to my beloved mother for empowering me and her consistent emotional support, to my dear father and my dear brother. A special thanks to my beloved husband, Ehsan for his support, encouragement and standing by my side. I praise God for all the blessings.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **x**

LIST OF FIGURES . . . . . **xi**

LIST OF ALGORITHMS . . . . . **xiii**

Chapter1
Introduction and Literature Review **1**
1.1 Organization of the Dissertation 7

Chapter2
Safe Reinforcement Learning: A Control Barrier Function Optimization Approach **11**
2.1 Introduction 11
2.1.1 Notations 12
2.1.2 Organization of the Chapter 12
2.2 Preliminaries 13
2.2.1 Problem Statement 13
2.2.2 Barrier Function 15
2.3 Safe Optimal Control Approach 16
2.3.1 Safe Modified Formulation 16
2.3.2 Safety and Performance Analysis 20
2.3.2.1 Safety Analysis 21
2.3.2.2 Stability and Optimality Analysis 25
2.4 Algorithm for Safe Reinforcement Learning 28
2.4.1 Safe Off-policy Reinforcement Learning Algorithm 28
2.4.2 Neural Network Approximation of Safe RL Algorithm 31
2.5 Simulation Results 34
2.6 Conclusion 35

Chapter3
Reinforcement Learning based Control Design with Safety and Stability Guarantees During Exploration **39**
3.1 Introduction 39
3.1.1 Organization of the chapter 41
3.2 Problem Statement 41
3.3 Background 43
3.3.1 Control Barrier Functions 43
3.3.2 Adaptive Optimal Control Design 44
3.4 Robustified Safety and Stability using Experience Replay Learning 46

3.4.1 Experience Replay System Approximation . . . 47
3.4.2 Stability Analysis . . . 50
3.4.3 Adaptive Robustified CBF . . . 52
3.4.4 Safe and Stable Controller . . . 55
3.5 Barrier-certified Off-Policy Algorithm . . . 55
3.6 Simulation . . . 59
3.6.1 Simulation Setup . . . 59
3.6.2 Simulation Results and Discussion . . . 60
3.7 Conclusion and Future Work . . . 61

Chapter4
Barrier-certified Learning-enabled Safe Control Design for Systems Operating in Uncertain Environments . . . 65
4.1 Introduction . . . 65
4.1.1 Organization of the Chapter . . . 67
4.2 Problem Statement and Background . . . 67
4.2.1 Problem Statement . . . 68
4.2.2 Control Barrier Functions . . . 70
4.3 Learning-enabled ZCBF with Uncertain Sets . . . 72
4.3.1 Learning Safe Set Despite Uncertain Behaviors of External Agents . . . 73
4.3.2 External Dynamics Identifier . . . 80
4.4 Control Framework . . . 88
4.5 Case Study . . . 89
4.5.1 Control Scenario . . . 91
4.5.2 Mathematical Representation . . . 91
4.6 Simulation Results . . . 94
4.6.1 Zero Modeling Error Scenario . . . 95
4.6.2 First Non-zero Modeling Error Scenario . . . 96
4.6.3 Second Non-zero Modeling Error Scenario . . . 97
4.6.4 Discussion . . . 98
4.7 Conclusion and Future Work . . . 100

Chapter5
Robust Satisficing Cooperative Control Barrier Functions for Multi-Robots Systems using Information-Gap Theory . . . 102
5.1 Introduction . . . 102
5.1.1 Organization of the Chapter . . . 106
5.2 Problem Overview and Background . . . 106
5.2.1 Problem Overview . . . 106
5.2.2 Background . . . 107
5.2.2.1 Control Barrier Functions . . . 107
5.2.2.2 Information-Gap Theory . . . 109
5.3 Robust-Satisficing Control Barrier Function . . . 110
5.3.1 Problem Formulation . . . 112
5.3.2 Robust-satisficing Distributed CBF . . . 114

5.3.2.1 Distributed ZCBF 114
5.3.2.2 Robust-satisficing distributed ZCBF 115
5.3.2.3 Discussion 121
5.3.3 Controller Design 122
5.4 Simulation 124
5.5 Conclusion 127

Chapter6
Conclusion 131

BIBLIOGRAPHY 134

LIST OF TABLES

Table 2.1: Simulation Parameters 36

Table 3.1: Simulation Parameters 61

Table 4.1: Simulation Parameters 94

LIST OF FIGURES

Figure 2.1: Lateral displacement with and without CBF 36

Figure 2.2: The states of the system 36

Figure 2.3: Actor Weights 37

Figure 2.4: Critic Weights 37

Figure 3.1: Overview of the proposed approach 41

Figure 3.2: States of the system under the proposed framework 62

Figure 3.3: States of the system with plain off-policy with manual reduced noise . . . 62

Figure 3.4: NN Weight error 63

Figure 3.5: Convergence of P_k and K_k 63

Figure 4.1: (a): $\hat{\mathscr{C}}$ invariant, $\partial\mathscr{C}$ violated (b): $\mathscr{C}_c$ invariant (c): $\hat{\mathscr{C}}$ converges to $\mathscr{C}$ . . 74

Figure 4.2: Control scheme 89

Figure 4.3: Control scenario 95

Figure 4.4: Position of vehicles in 'y' coordinate (Scenario1) 96

Figure 4.5: (a) NN weights (Scenario 1), (b) Optimal solution without CBF 97

Figure 4.6: NN Weights (Scenario 2) 98

Figure 4.7: NN Weights with and without Experience replay 98

Figure 4.8: Position of vehicles in 'y' coordinate (Scenario2) 99

Figure 4.9: NN Weight (Scenario 3) 99

Figure 4.10: Position of agents in 'y' coordinate (Scenario 3) 100

Figure 5.1: Graph Topology 124

Figure 5.2: (a) Agents' trajectories, no measurement error (b) Corresponding $||\Delta\mathbf{p}_{ij}||$ 127

Figure 5.3: (a) Trajectories, measurement error without IG (b) Corresponding $||\Delta\mathbf{p}_{ij}||$ 127

Figure 5.4: (a) Trajectories, measurement error with IG (b) Corresponding $||\Delta\mathbf{p}_{ij}||$ . . 128

Figure 5.5: Pairwise distances between agents for different values of safety distance D_s 128

Figure 5.6: Time Lapse of agents' position with measurement error using IG method 129

LIST OF ALGORITHMS

Algorithm 1 Safe Off-policy RL 34

Algorithm 2 Safe and Stable Off-policy RL 58

Algorithm 3 Barrier-certified Learning-enabled Controller 90

Algorithm 4 Safe and Robust Control Design for each Agent i 123

Chapter1

Introduction and Literature Review

Safety-critical systems are the systems whose failure or malfunction can result in injury to people, damage to the equipment or harm to the environment [1]. Being that said, most control systems face instrumental or environmental limitations and thus are considered as safety-critical systems. The limitations of the system itself include states constraints, such as saturation of actuators, limited range of motion in a joint of a robotic arm, maximum allowable speed of a vehicle, the relative portion of materials in a chemical process, and so on. The environment in which the system is operating also imposes different types of safety constraints on the system. For example, when the operating environment is shared between different agents such as pick-and-place robotic arms in a factory, a multi-robot system and autonomous driving, the collision should be avoided between the nearby agents. In addition, the safety of human operators and nearby facilities must be guaranteed as well. All these safety constraints need to be satisfied for a safe and reliable operation. The set of states in which these safety constraints are satisfied are considered as the safe set. The controller need to be designed accordingly to get the desired performance within the safe set of the system to avoid safety violation. Moreover, conflicts can always arise between safety and performance requirements, and, in a conflicting situation, safety objectives must always be prioritized to the performance. For example, in the adaptive cruise control system, the system's performance level that can be achieved without safety violation in terms of reaching the desired speed depends on the traffic situation and assuring a safe maneuver (maintaining

a safe distance from the vehicle ahead) must be prioritized to the performance.

Specially, with the emerge of robotics and autonomous systems, which have a high level of interactions with humans, and typically operate in a cluttered and uncertain environment; it is crucial to design safe and smart controllers in the face of the model and environmental uncertainty, which is the goal of this dissertation.

Reinforcement learning (RL) is an emerging framework in control systems that learns the optimal controller for uncertain systems online in real time [2, 3, 4]. Although powerful, assuring its safety is one of the main challenges to pave the way to widespread deployment of RL in practice. RL algorithms typically consist of two phases of operation: exploration and exploitation. In the former phase, random noisy inputs are applied to the system to collect rich data. The collected data is used to learn improved control policies, followed in the exploitation phase to gain more rewards. However, under uncertainty, little or no knowledge of the system dynamics might be available, and therefore, RL agent faces the risk of stability or safety violation. Satisfaction of these properties is very challenging since, on one hand, noisy exploratory inputs must be applied to the system, and, on the other hand, their consequences cannot be fully predicted because the complete knowledge of the system dynamics is not known priori.

Different approaches are proposed in the literature to address the safe RL problem. Safety in RL framework has been addressed in two general ways; one takes into account the uncertainty of the reward and the other one deals with possible risks in the exploration process [5]. In the former case, stochastic cost-to-go functions are considered and appropriate risk measure functions are applied [6], while in the latter one, the learning agent is typically provided with some external knowledge or advice for safe exploration [5],[7], [8]. While applying risk measures on stochastic functions is a strong tool to deal with uncertainty, it does not take into account the constraints on the system's state and the control input. Economists have studied this risk-aware approach because the goal is to obtain the highest profit while the risk is the chance of loss which is inherent to the concept of profit. However,

for many control systems that risk arises from state or input constraints such as collision avoidance in multi-agent systems, forbidden states of a robotic arm, and safe autonomous vehicles, this approach cannot be directly applied. Risk in the exploration process has been addressed through learning simulators, using external advice and prior knowledge [9, 10]. However, all of these approaches need some prior knowledge about the risk or distance to the risk. These approaches are applicable for cases such as the risky height of flight for an airplane; but they are not constructive for applications that the information about dangerous occasions is not available which is somehow inherent to the concept of risk.

[11] employs expert demonstration in a surgical robotic system which provides an area with a high probability of safe task completion. The forward RL is then solved in this region in conjunction with an area with a return route to this region based on the task completion cost. [8] uses the idea of the escape route and backup, and therefore, in the face of a safety crisis, a backup safe path is taken. To reduce the need for prior knowledge, which might not be available, learning from data can be leveraged and combined with prior knowledge. For example, in [12], two stages of learning are considered: in the first one, a rule-based safeguard is employed. As more data become available, the rule-based safeguard is replaced with a data-driven counterpart in the second stage. [13] identifies undesirable actions in a set of previously learned tasks and uses transfer learning. In [14], a recovery RL algorithm is proposed which employs the offline data to learn about unsafe zones. Then, a recovery policy is employed which acts as a backup policy in the face of imminent risk. However, safe offline data collection demands human supervision. In addition, in a hostile environment, frequent alteration of policies might prevent reaching performance objectives.

A broad class of methods in safe RL are model-based and rely on information about the system/environment or prediction of the risk. This includes shielding approaches based on reachability [15], safety modules [16], or safety layers that adjust the policy to prevent violation of safety. [17] employs risk state estimation module, which activates the safe policy search module in the face of risk. Employing constrained Markov decision process, [18,

19] and safe region of attraction calculation [20] are other methods to tackle the safe RL problem.

Reachability analysis has also been widely used to handle safety in the exploration process by finding the set of initial states for which there is a control input that keeps the state of the system within the safe feasible set despite disturbance [21]. Moreover, in the boundary of the safe set, it needs to switch to a controller to push the state back into the safe region, which can cause chattering. In [22], the Gaussian regression is used to learn the disturbance and, also a term is added to the cost function to incorporate the risk within the learning scheme. More relevant work in safety within the context of reachability can be found in [23], [24]. [23] presented a safe control framework based on the Hamilton-Jacobi reachability method for partially unknown systems. The safety problem is then defined as a differential game between controller and disturbance. A Gaussian process is leveraged to learn about the disturbance, and Bayesian analysis computes its bound. In [25], a safeguard layer is incorporated using trajectory-parametrized reachable set analysis which is computed offline. Although elegant, reachability-based approaches are computationally demanding. In addition, these methods are still model-based, and offline model-learning results in dependency of safety to the accuracy of learning. In addition, by any change in the operation regime of the system, offline learning needs to be re-initiated.

Control barrier function (CBF) is another widely used method to guarantee the safety of the control systems [26, 27, 28, 29, 30, 31]. This includes adaptive cruise control problem in [27, 28], safe control of robots [29, 32] and collision-free multi-agents systems [30, 33]. These methods generally integrate CBFs and control Lyapunov functions and solve a point-wise quadratic programming optimization problem to certify the safety and stability of a nominal controller. CBFs are conceptually similar to Lyapunov functions and are used to ensure forward invariance of a specific set. However, these methods require complete knowledge of the system dynamics as well as the feasible set. therefore, it is not straightforward to integrate it with an RL framework for which the knowledge of the system model is not required. In

addition, for the systems that operate in uncertain environments, the safe or feasible set is uncertain: safety criteria are affected by some external factors with possibly uncertain or unknown behaviors which are not known a priori. For example, in autonomous vehicles, the operation platform of vehicles is highly complicated and shared between autonomous, semi-autonomous, and human driving vehicles and pedestrians. Therefore, it is necessary to design a controller that can ensure the safety of the system despite the uncertainty in the feasible set due to the existence of unknown external agents while reaching as much performance as possible.

To account for uncertainties in designing safe controllers, several robust and adaptive approaches are presented. In [34], the robustness of zeroing CBFs (ZCBFs) under model perturbation is investigated. It is shown that the existence of ZCBF ensures the input-to-state stability of the safe set under perturbations. However, external agents that affect the safety of the ego system cannot be modeled as a perturbation. In [35], an adaptive CBF (aCBF) is proposed to ensure safety despite parametric uncertainty. To reduce conservatism, [36] proposed a robust aCBF (RaCBF), which guarantees forward invariance of a tightened set within the safe set. However, in both approaches, the invariance criterion needs to be satisfied for all values within the uncertain parameters that are not always known ahead of time. In addition, the effect of external dynamics in the environment shared with the ego-system can be completely modeled as neither parametric uncertainty nor disturbance. In [37], uncertainties impacting CBFs are learned to design a safe controller for a wider class of uncertainties. However, it is assumed that the CBF for the nominal system is a CBF for the uncertain system, which is not always applicable. To partially compensate for the need for full knowledge on dynamics, [32, 38] have proposed data-driven methods which use the Gaussian process to learn about disturbance. [32] Uses learning to explore uncertain states to expand and maximize the barrier-certified safe region by updating probabilistic parameters and decreasing the variance of the disturbance Gaussian Model. Then, the least square method is used to find the closest control input to the nominal input, which ensures safety.

[38] uses a similar method to form the CBF by learning about the disturbance. It also takes the optimality into account by finding the optimal control input by policy-gradient RL, which is then combined with the control input obtained from CBF, which ensures safety. In both of these works, a nominal model is needed to form the CBF constraint and the disturbance is modeled by the Gaussian process, which is not always applicable. To ensure the convergence on the original goal and to avoid the conflict between safety and performance, [39] uses an iterative search algorithm using the sum-of-squares method to find the maximum region in which safety and stabilization are compatible. In [29], sparse optimization is used to extract the dynamical structure. The model and long-term reward are adaptively estimated, and the learned model is used at each instant to provide required information on the dynamics to ensure safety using the CBF method for non-stationary discrete-time control systems. RL method for handling constrained states is proposed in [40], in which a non-quadratic function is incorporated in the performance functional that becomes dominant in case of constraint violation. In this method, safety is considered as a soft constraint. [41] incorporates state and input constraints in RL framework using penalty function and barrier function (BF)-based state transformation; however, the possible conflict between safety and stability is not considered.

Model predictive control (MPC) is another suitable framework for handling state and input constraints. In some MPC approaches, a barrier function (BF) is incorporated into the cost function to convert a constrained optimization problem into an unconstrained optimization problem, which provides a smooth transition of states within the feasible set [42], [43]. However, they only deal with state constraints imposing the condition that safe set must contain the origin, while in practical applications, safety and performance/stability might be in conflict, and safety must be prioritized. Our previous work [44] has extended those approaches to a general safe set, capable of handling even complicated and nonlinear safety criteria due to the interaction of different states. MPC-based approaches are mainly model based, and since they are short-sighted, it is hard to guarantee the stability and feasibility

of the solution in the presence of uncertainty.

In [45], path planning in uncertain and dense obstacles environment is investigated in which a reachability set estimator of dynamic obstacles is employed to predict its threat. The CBF-based method, in contrast, ensures safety without the need for finding the reachability set, which is typically computationally demanding. Inverse RL is used in [46] to learn about the reward function and, consequently, the behavior of the human agent in control of human-robot systems. However, this line of work assumes that the human operators or external agents choose their course of actions based on a perfectly rational framework that makes optimal decisions with respect to a reward function, which might not coincide with reality and is also computationally expensive.

1.1 Organization of the Dissertation

Based on the above elaborated problems, the brief contribution and organization of this dissertation is as follows.

1. Chapter 2 presents a learning-based barrier-certified method to learn safe optimal controllers that guarantee the operation of safety-critical systems within their safe regions while providing optimal performance. The cost function that encodes the designer's objectives is augmented with a CBF to ensure safety and optimality. A damping coefficient is incorporated into the CBF, which specifies the trade-off between safety and optimality. The proposed formulation provides a look-ahead and proactive safety planning and results in a smooth transition of states within the feasible set. That is, instead of applying an optimal controller and intervening with it only if the safety constraints are violated, the safety is planned and optimized along with the performance to minimize the intervention with the optimal controller. It is shown that the addition of the CBF into the cost function does not affect the stability and optimality of the designed controller within the safe region. This formulation enables us to find

the optimal safe solution iteratively. An off-policy RL algorithm is then employed to find a safe optimal policy without requiring the complete knowledge about the system dynamics while satisfying the safety constraints. The efficacy of the proposed safe RL control design approach is demonstrated on the lane keeping as an automotive control problem.

2. Satisfaction of safety and stability properties of RL algorithms has been a long-standing challenge. These properties must be satisfied even during learning, for which exploration is required to collect rich data. However, satisfying the safety of actions when little is known about the system dynamics is a daunting challenge. After all, predicting the consequence of RL actions requires knowing the system dynamics. Chapter 3 presents a novel RL scheme that ensures the safety and stability of the linear systems during the exploration and exploitation phases. First, the system model is learned for the sake of safety. That is, the update law is designed to assure that the actual model's safety properties are preserved by the learned model. Second, a fast and efficient learning scheme is presented to ensure that the model learning error remains in a prescribed bound with a desired convergence rate. This occurs because of the efficient deployment of data from past experiences in an off-policy RL framework. Then, the presented model and its prescribed errors are employed in a novel adaptive robustified control barrier function (AR-CBF) which guarantees that states of the system remain in the safe set even when the learning is incomplete. Therefore, the noisy input in the exploratory data collection phase and the optimal controller in the exploitation phase are minimally altered such that the AR-CBF criterion is satisfied, and therefore, safety is guaranteed in both phases. It is shown that under the proposed prescribed RL framework, the model learning error is a vanishing perturbation to the original system. Therefore, a stability guarantee is also provided even in the exploration when noisy random inputs are applied to the system.

3. Chapter 4 presents learning-enabled barrier-certified safe controllers for systems that operate in a shared environment for which multiple systems with uncertain dynamics and behaviors interact. That is, safety constraints are imposed by not only the ego system's own physical limitations but also other systems operating nearby. Since the model of the external agent is required to impose CBFs as safety constraints, a safety-aware loss function is defined and minimized to learn the uncertain and unknown behavior of external agents. More specifically, the loss function is defined based on barrier function error, instead of the system model error, and is minimized for both current samples as well as past samples stored in the memory to assure a fast and generalizable learning algorithm for approximating the safe set. The proposed model learning and CBF are then integrated together to form a learning-enabled zeroing CBF (L-ZCBF), which employs the approximated trajectory information of the external agents provided by the learned model but shrinks the safety boundary in case of an imminent safety violation using instantaneous sensory observations. It is shown that the proposed L-ZCBF assures safety guarantees during learning and even in the face of inaccurate or simplified approximation of external agents, which is crucial in safety-critical applications in highly interactive environments. The efficacy of the proposed method is examined in a simulation of safe maneuver control of a vehicle in an urban area.

4. Chapter 5 integrates the CBFs and information-gap theory to present robust safe controllers for collision avoidance problem in multi-agent systems with different levels of measurement accuracy. It is assumed that agents have uncertain and inaccurate measurements about the relative distance to neighboring agents. A cooperative framework for the construction of CBFs for every two agents is employed to avoid collision and ensure the safety of the overall system. To maximize the horizon of uncertainty under which the safety of the overall system is satisfied, the information-gap theory is leveraged to determine the contribution and share of each agent in the construction of CBFs.

This results in the highest possible robustness against measurement uncertainty. It is shown that the overall system can tolerate higher measurement uncertainty and safely operate if the agent that is more confident about its measurement contributes more to the construction of the CBF. By employing the proposed approach in constructing CBF, the possible failure of one agent in gathering accurate local data can be compensated by cooperation between agents. The effectiveness of the proposed method is demonstrated via performing simulations for multi-robot systems.

5. Chapter 6 summarizes and concludes the dissertation and provides future research directions.

The contributions of this dissertation are published in [44, 47, 48, 49, 50, 51, 52, 53].

Chapter2

Safe Reinforcement Learning: A Control Barrier Function Optimization Approach

Contents of this chapter first appeared as [50] and have been reformatted to fit the requirements of this dissertation.

2.1 Introduction

In this chapter, a safe RL scheme is proposed which is based on optimization of a cost function that is augmented with a CBF candidate. The proposed approach is capable of handling a pre-defined safe and feasible polytope set formed by state constraints and process risk. RL algorithm is used to learn the optimal control policy that minimizes this augmented cost function without requiring the complete knowledge about the system dynamics. It is shown that sequential improvement of the controller ensures safety and stability within the safe region. The main contribution is that the concept of the CBF is unified with an RL scheme to bring together the best of two worlds, i.e, to guarantee safety in a data-driven fashion. It also provides a look-ahead and proactive approach for safety planning for smooth handling of a sudden danger. Although the idea of using BFs in the cost function has been used in the context of MPC and dynamic programming, its main goal is to alter a constrained optimization problem into an unconstrained optimization. However, the proposed approach

here differs in the following aspects:

1. It addresses possible conflict between safety and performance and the safe set does not necessarily contain the origin.

2. Off-policy RL is employed which allows to learn about an optimal safe policy that minimizes the augmented cost while applying a safe and possibly conservative policy to collect data during learning. This is because off-policy RL separates the target policy (policy we learn about) from the behavior policy (policy we apply to the system to collect data). Rigorous proofs are provided to show that sequential improvement of the control policy provides optimality and guarantees safety. That is, the safety of the optimal solution is verified.

3. To provide an optimal performance, instead of using a zeroing factor, a function is considered as a CBF that rapidly damps to zero within a specific distance to the safety boundary; this facilitates taking safety as a control objective not only as a constraint. A parameter is incorporated in CBF which determines the relative importance of the original control objectives to the safety.

2.1.1 Notations

The interior of set $\mathscr{C}$ is denoted as $int\,\mathscr{C}$ and $\partial\mathscr{C}$ stands for its boundary. Throughout the paper, $\|\cdot\|_M$ denotes the weighted Euclidean norm of a vector i.e. $\|x\|_M = \sqrt{x^T M x}$ in which M is a positive semi-definite matrix. $\mathscr{U}$ is the set of all admissible control inputs. C^1 denotes the set of continuously differentiable functions.

2.1.2 Organization of the Chapter

Background information, preliminaries and problem statement are given in Section 2. Safe optimal control approach with safety and stability proofs are provided in Section 3. Section

4 employs neural networks for estimation of optimal controller and value function using off-policy RL algorithm. Section 5 shows the efficiency of the proposed method by providing comprehensive simulation results and section 6 concludes the chapter.

2.2 Preliminaries

Consider a nonlinear system described by the following differential equation

$$\dot{x} = f(x) + g(x)u \tag{2.1}$$

where $x \in \mathscr{C} \subset \mathbb{R}^n$ and $u \in \mathscr{U} \subset \mathbb{R}^m$ are the state of the system and the control input, respectively. $\mathscr{C}$ represents the set of safe feasible states while $\mathscr{U}$ denotes the set of all admissible inputs. Moreover, $f(x) \in \mathbb{R}^n$ is the drift dynamics and $g(x) \in \mathbb{R}^{n\times m}$ is the input dynamics. $f(x)$ is C^1 and $f(0) = 0$. It is also assumed that the system is stabilizable. Before proceeding, the problem formulation and a short background are provided as follows.

2.2.1 Problem Statement

The goal is to design a safe optimal control policy for the system (2.1). To take into account optimality, an infinite horizon cost function is considered and is minimized along with the trajectories of the system (2.1) and within a safe set. That is, the safe optimal control problem is formulated as

$$\begin{aligned} &\min_{u\in\mathscr{U}} \; J(u,x) = \int_t^\infty r(x(\tau),u(\tau))d\tau \\ &\text{s.t. } (2.1),\; x(0) = x_0,\, x \in \mathscr{C}, \end{aligned} \tag{2.2}$$

where the utility function $r(x, u)$ is defined as

$$r(x, u) = Q(x) + u^T R u \tag{2.3}$$

where $Q(x)$ is a positive-definite function and R is a symmetric positive-definite matrix $R = R^T > 0$. The set $\mathscr{C}$ is called the safe set inside which the system's state must evolve to assure a safe operation. The safe set is formed by operational inequality constraints of the system such as actuator saturation of a robotic arm or unsafe region of exploration of a mobile robot and it is mathematically defined as

$$\mathscr{C} = \{x | h(x) \geq 0\} \tag{2.4}$$

where $h(x)$ is a continuously differentiable function of x. Note that $h(x) > 0$ represents the admissible state space that respects safety constraints. For example if $-1 < x < 1$, then $h(x) = [h_1, h_2]$ where $h_1 = 1 - x$ and $h_2 = x + 1$.

In the absence of safety constraints, using (2.3) in the cost function J in (2.2), the optimal value function is defined as [54]

$$V^*(x) = \min_u \int_t^\infty (Q(x) + u^T R u) d\tau \tag{2.5}$$

Denoting the minimizer policy by u^*, the Hamiltonian function is defined as

$$H(x, u^*, \nabla V) = r(x, u^*) + (\nabla V)^T (f(x) + g(x)u^*) \tag{2.6}$$

The right-hand side of (2.6) is the infinitesimal equivalent of (2.5) which is a nonlinear Lyapunov equation. $H = 0$ forms the continuous-time (CT) Bellman equation and is used for obtaining the optimal solution [54]. This framework, however, cannot guarantee safety.

One standard approach to design a safe control policy for system (2.1) utilizes the concept

of CBFs. We will discuss it briefly in the following subsection.

2.2.2 Barrier Function

A BF is a function which is positive within a set and reaches infinity at the boundary of this set. Moreover, the BF has a negative derivative in the vicinity of the boundary, and thus, it never reaches infinity. In other words, if the initial state is within a set, existence of the BF on that set guarantees its forward invariance. The BFs or barrier certificate functions (BCFs) are defined and used to certify safety of dynamical systems and control barrier functions (CBFs) is the terminology of the same concept for control systems. Under this approach, the control input is designed to satisfy the properties of a CBF candidate. The above mentioned properties of a CBF are formally defined as follows.

Definition 2.1. Class $\mathcal{K}$ Function.

A continuous function $\alpha : [0, a) \to [0, \infty)$ is a class $\mathcal{K}$ function if it is strictly increasing and $\alpha(0) = 0$ [55].

Definition 2.2. CBF Properties.

For the control system (2.1), the C^1 function $B : \mathscr{C} \to \mathbb{R}$ is a CBF for the set (2.4), if there exist locally Lipschitz class $\mathcal{K}$ functions α_1, α_2 and α_3 such that [27]

$$\frac{1}{\alpha_1(h(x))} \leq B(x) \leq \frac{1}{\alpha_2(h(x))}, \forall x \in int\mathscr{C} \tag{2.7}$$

$$\dot{B}(x) \leq \alpha_3(h(x)), \forall x \in int\mathscr{C} \tag{2.8}$$

Remark 2.1. The condition $\dot{B} < 0$ also can be used instead of the CBF derivative condition (2.8) in Definition 2.2. However, compared to (2.8), it could unnecessary shrink the sub-levels even if they are within the desired set [26]. The condition (2.8) let $\dot{B}$ increase when it is far from the boundary and makes it negative only in the vicinity of the boundary.

Remark 2.2. The control input is designed by choosing a CBF candidate that satisfies (2.7) and, then, (2.8) is imposed as an inequality constraint to the control problem. While elegant, this framework does not consider the optimality of the solution and the complete knowledge of the system dynamics is required to check if the condition (2.8) is satisfied, because trajectory information $\dot{x}$ appears in $\dot{B} = \frac{\partial B}{\partial x}\dot{x}$. To obviate these requirements and design an optimal safe control policy, RL will be integrated with the CBF concept in the subsequent sections.

2.3 Safe Optimal Control Approach

We present a new formulation for designing a safe and optimal control input by integration of CBF into performance (2.2). The proposed approach guarantees safety in case it has a conflict with other control objectives, and in a safe condition, it guarantees an optimal performance. This formulation enables us to learn an optimal safe policy in a data-driven fashion using off-policy RL algorithm.

2.3.1 Safe Modified Formulation

To ensure safety, the cost-to-go function is augmented with a CBF term $B_\gamma(x)$ and the performance defined in (2.2) is modified to

$$\begin{aligned} &\min_{u\in\mathscr{U}} \; J(x,u) = \int_t^\infty (Q(x) + u^T R(x) u + B_\gamma(x)) d\tau \\ &\text{s.t. } (2.1),\; x(0) = x_0 \end{aligned} \tag{2.9}$$

where $B_\gamma(x) : \mathscr{C} \to \mathbb{R}$ has the following properties.

Assumption 2.1. CBF Properties.

B_γ in (2.9) is a function with the following properties,

1. $B_\gamma(x) > 0 \; \forall x \in \mathscr{C}$

2. $B_\gamma(x) \to \infty \; \forall x \in \partial\mathscr{C}$ with $\partial\mathscr{C}$ as the boundary of the safe set $\mathscr{C}$

3. $B_\gamma(x)$ is monotonically decreasing $\forall x \in \mathscr{C}$.

A coefficient γ is included in the CBF to specify the relative dominancy of the CBF to the utility function. While any CBF function that satisfies Assumption 2.1 can be used, a possible candidate is used in this chapter as follows

$$B_\gamma(x) = -log(\frac{\gamma h(x)}{\gamma h(x) + 1}) \tag{2.10}$$

The parameter γ determines how rapidly $B_\gamma(x)$ damps as it gets further away from the safety boundary. In other words, the coefficient γ trades-off between safety and optimality by specifying the margin that safety dominates other control objectives.

Compared to (2.3) and (2.5), the augmented utility function and the augmented value function are defined, respectively, as

$$r_a(x, u) = Q(x) + u^T R u + B_\gamma(x) \tag{2.11}$$

and

$$V^*{}_{aug}(x) = \min_u \int_t^\infty (Q(x) + u^T R u + B_\gamma(x)) d\tau \tag{2.12}$$

Remark 2.3. In contrast to the condition (2.8), the new formulation does not impose any conditions on the derivative of $B_\gamma(x)$; the reason is that $B_\gamma(x)$ is incorporated into the cost function and in the vicinity of the safety boundary, $B_\gamma(x)$ becomes the dominant term in (2.12) and the optimal controller acts in a descent direction of $B_\gamma(x)$. In other words, $\dot{B}_\gamma$ implicitly becomes negative near the boundary in an optimal manner without imposing any inequality constraints. Moreover, numerical methods for solving unconstrained optimization problem are applicable. Finally, safety satisfaction over a long horizon plays an important

role in performing anticipatory safe planning, and avoiding excessive intervention with the optimal solution.

Before proceeding to the next section, some definitions and assumptions are given.

Definition 2.3. The set of safe inputs.

The set of safe inputs for the current state x is defined as

$$\mathscr{U}_c = \{u \in \mathbb{R}^m | x^u \in int\mathscr{C}\} \tag{2.13}$$

where $int\mathscr{C}$ is the interior of the set defined in (2.4) and x^u is the state of the system evolved by the input u.

Definition 2.4. Admissible policy.

A control policy is said to be admissible for an optimal control problem if it stabilizes the system (2.1) and its associated cost is bounded.

The following proposition shows that every admissible policy for the original optimization problem (2.2) that satisfies the safety and state constraints, is an admissible policy for the modified formulation (2.9).

Proposition 2.1. *A control policy is admissible for the modified optimal control problem* (2.9)*, if and only if,*

$$u \in \mathscr{U} \cap \mathscr{U}_c$$

where $\mathscr{U}$ is the admissible control policy for the optimal control problem (2.2) and $\mathscr{U}_c$ is defined in (2.13).

Proof. The cost function in (2.9) augments a utility function with a CBF. Therefore, to have an admissible policy, in addition to $r(x, u)$, $B_\gamma(x)$ should also remain bounded. Since $r(x, u)$

is bounded for $u \in \mathscr{U}$, and B_γ remains within the safe set and is bounded for $u \in \mathscr{U}_c$, as a result, for the modified formulation, a policy results in a bounded cost function for (2.9) if

$$u \in \mathscr{U} \cap \mathscr{U}_c \tag{2.14}$$

On the other hand, since $u \in \mathscr{U}$ stabilizes the system (2.1) by definition, $u \in \mathscr{U} \cap \mathscr{U}_c$ also stabilizes the system (2.1). This completes the proof. □

The set of admissible inputs for (2.9) is now defined as

$$\mathscr{U}_a = \mathscr{U} \cap \mathscr{U}_c$$

Assumption 2.2. Strict interiority of the initial condition.

The initial condition of (2.1) belongs to the interior of $\mathscr{C}$. That is,

$$x_0 \in int\mathscr{C}$$

Assumption 2.3. Existence of an admissible control input.

We assume the set of admissible inputs for (2.9) is non-empty, i.e., $\mathscr{U} \cap \mathscr{U}_c \neq \emptyset$ and for any initial condition x_0 satisfying Assumption 2.2, there exists a control policy $u(x_0) \in \mathscr{U}_a$.

Remark 2.4. Assumptions 2.1, 2.2, 2.3 are standard assumptions in safe control design. More specifically, the function $B_\gamma(x)$ in (2.10) actually satisfies Assumption 2.1. However, besides the CBF in (2.10), any other CBF that satisfies this assumption would also be acceptable. Assumptions 2.2 and 2.3 imply that the system must start from a safe initial condition and that a feasible control input exists to keep the system in its safe set. If these assumptions are not satisfied, then there is no hope to maintain the system safety using any control strategy, and thus the system itself is ill-posed. Other assumptions are also

made throughout the chapter such as Lipschitz continuity of the system or existence of value functions, which are standard in optimal control literature, for example see [54].

The Hamiltonian function H_j (2.6) for the augmented utility function (2.11) and the value function W_j is given as

$$H_j(x, u_j, \nabla W_j) = r_a(x, u_j) + (\nabla W_j)^T (f + g u_j) \tag{2.15}$$

Then, H_{min_j}, i.e., the minimizer of H_j, is obtained by the control input

$$u_j{}^* = -0.5 R^{-1} g^T(x) \nabla W_j \tag{2.16}$$

and is given by

$$H_{min_j} = H_j(x, u_j{}^*, \nabla W_j) \tag{2.17}$$

In the following subsections, RL is employed to solve the modified safe optimal formulation (2.9), which iteratively estimates the value function and sequentially improves the control input toward the optimal minimizer while not violating safety constraints.

2.3.2 Safety and Performance Analysis

We now present how the formulation (2.9) trades-off between safety and performance. In this approach, safety is ensured while a desired performance is maintained within the safe region. In addition, to improve safety robustness and avoid taking myopic safe actions, the CBF acts as a safety measure along other control objectives to be optimized over time. As a result, it provides a platform for safety planning and to specify the importance of safety compared to other objectives. All of these goals should be achieved in an iterative method while a closed-form solution to the value function is not available.

To prove the claims, a couple of theorems are presented. First, it is proved that the

proposed approach guarantees the safety of the system. Second, the concept of safe region is introduced. Finally, stability and optimality of the solution in the safe region are shown.

2.3.2.1 Safety Analysis

First, the existence of the value function is shown and, inspired by [56] the boundedness of the CBF is demonstrated through sequential improvement of the controller, and, finally, based on these results, the main theorem on guaranteeing safety is provided.

Lemma 2.1. *Consider an admissible feedback control policy $u_1 \in \mathscr{U}_a$. If a time invariant positive-definite function $W \in C^1$ exists such that*

$$\frac{\partial W^T}{\partial x}(f(x) + g(x)u_1) + Q(x) + B_\gamma(x) + u_1^T R u_1 = 0 \tag{2.18}$$

$$W(x_0, u_1) = J(x_0, u_1) \tag{2.19}$$

then, W is the value function of the system for all $t \in [0, \infty)$, i.e.,

$$W(x, u) = J(x, u)$$

Proof. Assume $W(x, u_1) > 0$ exists; since it is a continuously differentiable function, one has

$$\begin{aligned} W(x(t), u_1) - W(x_0, u_1) &= \int_0^t \dot{W}(x(\tau), u_1) d\tau \\ &= \int_0^t \frac{\partial W}{\partial x}(f + g u_1) d\tau \end{aligned} \tag{2.20}$$

Considering (2.9) and (2.11), one has

$$J(x(t), u_1) - J(x_0, u_1) = -\int_0^t r_a(x(\tau), u_1) d\tau \tag{2.21}$$

Subtracting both sides of (2.21) from (2.20) yields

$$J(x(t),u_1)-W(x(t),u_1)=\int_0^t(-\frac{\partial W}{\partial x}(f+gu_1)-r_a(x(\tau),u_1))\,d\tau+J(x_0,u_1)-W(x_0,u_1) \tag{2.22}$$

Considering (2.18) and (2.19) in (2.22) gives

$$J(x(t),u_1)-W(x(t),u_1)=\int_0^t r_a(x(\tau),u_1)-r_a(x(\tau),u_1)d\tau=0$$

Therefore, one has

$$J(x(t),u_1)=W(x(t),u_1)$$

which completes the proof. □

Lemma 2.2. *Consider positive-definite value functions* $W(x,t,u_1),W(x,t,u_2),...,W(x,t,u_i)$ *abbreviated by* $W_1,W_2,...,W_i$ *which are associated with the sequence of admissible inputs* $u_1(x,t)$, $u_2(x,t)$, ..., $u_i(x,t)\in\mathscr{U}_a$. *If corresponding minimized Hamiltonian values defined in* (2.17) *satisfy*

$$H_{min_1}\leq H_{min_2}\leq ...\leq H_{min_i} \tag{2.23}$$

then, the CBF candidate B_γ^j, $1\leq j\leq i$ *at each step of the sequence is bounded.*

Proof. For any j and k such that $0\leq j\leq k\leq i$, assume $H_{min_j}\leq H_{min_k}$; consider

$$W_k=W_j+W_d \tag{2.24}$$

where

$$W_d \triangleq W_d(x(t), u_j)$$

then, by applying $u_k^* = -0.5R^{-1}g^T\nabla W_k$, one has

$$H_{min_k} = Q(x) + B_\gamma(x) + \frac{1}{4}\nabla W_k^T g R^{-1} g^T \nabla W_k + \nabla W_k^T (f + g(-0.5R^{-1}g^T\nabla W_k))$$

Considering $L(x) = Q(x) + B_\gamma(x)$, using (2.24), and doing some manipulations yield

$$\begin{aligned} H_{min_k} = L(x) + \nabla W_j^T f - \frac{1}{4}\nabla W_j^T g R^{-1} g^T \nabla W_j + \nabla W_d^T f \\ - \frac{1}{4}\nabla W_d^T g R^{-1} g^T \nabla W_d - \frac{1}{2}\nabla W_d^T g R^{-1} g^T \nabla W_j \end{aligned}$$

or equivalently

$$H_{min_k} = H_{min_j} + \nabla W_d^T (f + g{u^*}_j) - ({u^*}_d^T R {u^*}_d)$$

Since $H_{min_k} - H_{min_j} + {u^*}_d^T R {u^*}_d \geq 0$, one has

$$\frac{dW_d(x, u_j)}{dt} \geq 0$$

In addition, $\lim_{t\to\infty} W_d(x(t)) = 0$. Therefore,

$$W_d \leq 0$$

As a result,

$$W_k \leq W_j$$

Considering the sequence in (2.23) results in

$$W(x,t,u_1) > W(x,t,u_2) > ... > W(x,t,u_i) \tag{2.25}$$

In other words,

$$W_j < W_1 \ \ \forall 1 \leq j \leq i$$

From Lemma 2.1,

$$J(x(t),u_j) < J(x(t),u_1) \ \ \forall 1 \leq j \leq i$$

Since $J(x(t),u_j) = \int_t^\infty r_a(x(\tau,u_j))d\tau$ is bounded and r_a is positive definite, then, r_a, and as a result B_γ^j are bounded. This completes the proof that the CBF is bounded at each sequence. □

Theorem 2.1. *Consider the optimization problem defined in* (2.9) *and let Assumptions 2.2 and 2.3 be satisfied. Then, the states of the system evolving through sequential improvement of the control input* (2.16) *stay within the safe set and safety of the system is ensured for all* $t > 0$.

Proof. Lemma 2.2 shows that the performance function $J(x,u_j)$ and consequently the barrier function ${B_\gamma}^j$ remain bounded after each policy improvement step (2.16). On the other hand, based on Assumption 2.1, the value of the CBF function ${B_\gamma}^j$ becomes infinity only at the boundary of the safe set . Therefore, since the barrier function remains bounded after every iteration, it guarantees that the system states never reach the boundary of the safe set. This in turn guarantees safety. □

Remark 2.5. By using Lemma 2.1, Lemma 2.2 and subsequently Theorem 2.1, it is proved that the safety of the control system is ensured for all $t > 0$ and $0 < \gamma < \infty$.

2.3.2.2 Stability and Optimality Analysis

Although safety is assured in Theorem 2.1, since a term is added into the cost function, the stability and performance of the system also need to be investigated. A desired safe controller should prioritize safety in case of a conflict with the desired performance. However, it still needs to ensure stability and demonstrate a good performance within the safe region. Feasible set and safe region are defined as follows and stability and optimality proofs are then given.

Definition 2.5. Feasible set.

The $int\mathscr{C}$ defined in (2.4) is considered as the feasible set.

Definition 2.6. Safe region.

The safe region is defined based on the feasible set as

$$D = \{x | x \in int\mathscr{C} - \beta(x_h^*, r_0)\}, \;\; x^* \in D$$

where $x_h^* = \{x | h(x) = 0\}$ and β is the ball around the boundary with radius of r_0 and x^* is the equilibrium point of the system assumed to be the origin.

The damping factor γ is chosen such that

$$\frac{B_\gamma(x)}{B_\gamma(x) + Q(x)} \leq 0.5 \;\; \forall x \in D$$

Therefore, within the safe region, $Q(x)$ is the dominant term in the optimization problem.

Remark 2.6. The safe region is the set containing the origin such that the CBF is not dominant compared to $Q(x)$.

Remark 2.7. The safe set might or might not contain the origin. However, as it is shown in the previous section, safety is guaranteed either way. Here, the safe region is defined for the condition that safety is not in conflict with the performance. Then, it is demonstrated that under this condition, optimality is achieved and uniform stability is also guaranteed.

Lemma 2.3. *Assume that $x = 0$ is the equilibrium point of the system (2.1), and $D \subset \mathbb{R}$ contains the origin. Let $M : [0, \infty) \times D \to \mathbb{R}$ be a continuously differentiable function such that*

$$\Lambda_1(x) \leq M(t, x) \leq \Lambda_2(x), \tag{2.26}$$

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial x}(f(x) + gu) \leq 0, \forall t > 0, \, \forall x \in D \tag{2.27}$$

where Λ_1 and Λ_2 are continuous positive-definite functions on D. Then, the origin is uniformly stable.

Proof. See [55] Theorem 4.8 page 151. □

Theorem 2.2. *The sequence of control inputs $u_j{}^*$ obtained by optimization over Hamiltonian functions (2.15) associated with positive-definite value functions W_j and the augmented utility function r_a, uniformly stabilize the system within the safe region D.*

Proof. Lemmas 2.1 and 2.2 prove that $W_j(t, x)$ is positive definite and

$$0 < W_j(t, x) < W_1(t, x), \; \forall 1 \leq j \leq i \tag{2.28}$$

where $W_1(t, x)$ is bounded and one can define positive-definite function Λ as

$$\Lambda(x) = \max_t W_1(t, x) \tag{2.29}$$

Thus, condition (2.26) is satisfied. Moreover, (2.25) proves that W_j is decreasing at each sequence. Therefore, using results of Lemma 2.3, the control system (2.1) is uniformly stable. □

Remark 2.8. The CBF in (2.9) is included as a safety objective to ensure safety for any value of $0 < \gamma < \infty$. Meanwhile, the trade-off bewteen $Q(x)$ and $B_\gamma(x)$ within the safe region is specified by the coefficient γ; larger values of γ could speed up damping of $B_\gamma(x)$ when

it goes further away from the safety boundary and retaining the original utility function $r = Q(x) + u^T Ru$, while smaller values of γ lead to more emphasis on safety and a more conservative control design. The CBF candidate $B_\gamma(x) = -log(\frac{\gamma h(x)}{\gamma h(x)+1})$ rapidly goes to zero, for example for $h(x) = 1$ and $\gamma = 5$, $B_\gamma(x) = 0.08$. In other words, one may design γ in the safe region such that $B_\gamma(0)$ gets arbitrary close to zero, which means depending on the application, optimality of the controller is achievable. This is proved in the following theorem.

Theorem 2.3. *Assume that the equilibrium point of the system is located at the origin. Assume* (2.9) *has a minimizer in the safe region denoted by* u^**. Then, by proper selection of* γ *within the safe region, the minimum can get arbitrarily close to zero and,*

$$\lim_{\gamma h(x)\to\infty} r_a(u^*) = 0$$

Proof. For an arbitrary small value ϵ, define $\gamma_1 = \frac{e^\epsilon}{h\cdot(1-e^\epsilon)}$. Then, for any $\gamma \geq \gamma_1$, one has

$$\begin{aligned} 0 \leq r(u^*) + B_\gamma(x^{u*}) &\leq r(u^*) + B_{\gamma_1}(x^{u*}) \\ &\leq r(u^*) + \epsilon \end{aligned}$$

In other words,

$$\forall \epsilon > 0 \;\; \exists \gamma \;\; s.t. \;\; B_\gamma(x) \leq \epsilon$$

For $\epsilon = 0$ and a finite value of $h(x)$, $\gamma_1 \to \infty$. Therefore, $\gamma h(x) \to \infty$; which completes the proof showing that the minimum of augmented utility function converges to zero. □

Remark 2.9. Based on Theorem 2.3, the optimal solution is feasible if $h(x)$ has a finite value and γ is selected properly.

Remark 2.10. From the theoretical perspective, the convergence of the proposed approach to the origin within the safe region is guaranteed if γ is large enough. However, in practice,

the value of γ depends on the physical system and we can achieve convergence with even small values of γ for some systems. For example, in the lane changing problem in Section 5, the states of the system have reached the origin with $\gamma_1 = 0.95$ and $\gamma_2 = 2$.

2.4 Algorithm for Safe Reinforcement Learning

In this section, an off-policy RL algorithm is presented to find a safe solution to the optimization problem (2.9). First, the off-policy RL algorithm is presented and then, neural networks (NNs) are used to approximate its solution for systems with the lack of knowledge about their dynamics.

2.4.1 Safe Off-policy Reinforcement Learning Algorithm

Off-policy RL is a policy iteration algorithm to find an optimal controller without requiring the knowledge on the system dynamics [57, 58, 59]. This method uses two different policies, called behavior policy and target policy. The behavior policy is a safe policy that is applied to the system for gathering data and the target policy is a policy that is updated toward the optimal policy using the collected data. Any available prior knowledge about the system dynamics can be used to find a safe but possibly conservative behavior policy to ensure safety during learning. The safety of optimal policy found by iterating on the target policy is also guaranteed based on Theorem 2.1. The optimal safe policy is applied to the system once the learning is finished. The details of the proposed off-policy RL algorithm is provided in the following.

Considering the augmented cost function (2.9), its infinitesimal version is the Bellman equation

$$0 = J_x^T \dot{x} + r_a + B_\gamma \tag{2.30}$$

where

$$\dot{J} = \frac{\partial J}{\partial x}\frac{\partial x}{\partial t} = J_x^T \dot{x} \tag{2.31}$$

In the off-policy approach, the dynamics (2.1) is rewritten to separate the behavior policy and the target policy. This yields

$$\dot{x} = f(x) + g(x)u^i + g(x)(u - u^i) \tag{2.32}$$

where u^i is the target policy which is updated in the algorithm but not applied to the system; while u is the behavior policy which is applied to the system to generate data for learning.

Integrating from both sides of (2.31) and considering (2.30) and (2.32) yield

$$J^i(x(t)) - J^i(x(t-T)) = -\int_{t-T}^{t} (Q(x) + B_\gamma(x))d\tau - \int_{t-T}^{t} u^{iT} R u^i d\tau + \int_{t-T}^{t} (J_x^{iT} g(x)(u - u^i))d\tau \tag{2.33}$$

The control input u^i is updated by optimizing over the Hamiltonian function

$$u^{i+1} = -0.5 R^{-1} g^T J_x^i \tag{2.34}$$

Substituting $g^T J_x^i$ term in (2.33) using (2.34) yields the off-policy Bellman equation

$$\begin{aligned} J^i(x(t)) - J^i(x(t-T)) = &-\int_{t-T}^{t} (Q(x) + B_\gamma(x))d\tau - \int_{t-T}^{t} u^{iT} R u^i d\tau \\ &-2\int_{t-T}^{t} (u^{(i+1)T} R(u - u^i))d\tau \end{aligned} \tag{2.35}$$

In the off-policy Bellman equation (2.35), both control policy (i.e. u^{i+1}) and value function (i.e. J^i) are updated simultaneously for a given target policy u^i using collected data by applying the behavior policy u.

Remark 2.11. Compared to on-policy method that improves the same policy that is applied to the system, the off-policy RL algorithm is a data-efficient method in which the learning agent evaluates as many policies as required without even applying them to the system using only a set of collected data. Being able to evaluate possibly unsafe policies without even applying them to the system is of vital importance for safety-critical systems.

Lemma 2.4. *Off-policy Bellman equation* (2.35) *is equivalent to the Bellman equation* (2.30) *and both have the same update law* (2.34).

Proof. Equations (2.30) - (2.35) demonstrate that off-policy Bellman equation is obtained by manipulating Bellman equation (2.30) and update law (2.34). Interchangeably, the Bellman equation can be obtained using (2.35). By dividing both sides of (2.35) by T and taking the limit from both sides, one has

$$\lim_{T\to 0} \frac{J^i(x(t)) - J^i(x(t-T))}{T} \\ - \lim_{T\to 0} \frac{-\int_{t-T}^{t} (Q(x) + B_\gamma(x))d\tau - \int_{t-T}^{t} u^{iT} R u^i d\tau - 2\int_{t-T}^{t} (u^{(i+1)T} R(u-u^i))d\tau}{T} = 0$$

By using L'Hopital's rule, one has

$$\dot{J}^i(x(t)) + (Q(x) + B_\gamma(x) + u^{iT} R u^i + 2u^{i+1} R(u-u^i)) = 0$$

using (2.31) and (2.32), one has

$${J_x}^{iT}(f(x) + g(x)u^i + g(x)(u-u^i)) + Q(x) + B_\gamma(x) + u^{iT} R u^i + 2u^{i+1} R(u-u^i) = 0$$

Then, using (2.34), one has

$${J_x}^{iT}(f(x) + g(x)u^i + g(x)(u-u^i)) + Q(x) + B_\gamma(x) + u^{iT} R u^i - {J_x}^{iT} g(x)(u-u^i) = 0$$

which is equivalent to the Bellman equation (2.30). This completes the proof. □

2.4.2 Neural Network Approximation of Safe RL Algorithm

In this section, the solution to the off-policy RL algorithm is learned using an actor-critic structure which does not require knowledge of the system dynamics. The critic network estimates the value function J^i and the actor network represents the control input u^{i+1} as follows

$$\hat{J}^i(x) = \hat{W}^i \Phi(x) \tag{2.36}$$

$$\hat{u}^{i+1}(x) = \hat{P}^i \Psi(x) \tag{2.37}$$

where $\Phi = [\Phi_1 \ \ \Phi_2 \ \ ... \ \ \Phi_{l_\Phi}] \in \mathbb{R}^{l_\Phi}$ and $\Psi = [\Psi_1 \ \ \Psi_2 \ \ ... \ \ \Psi_{l_\Psi}] \in \mathbb{R}^{l_\Psi}$ are the suitable activation functions for critic and actor networks with l_Φ and l_Ψ neurons, respectively; in addition, $\hat{W}^i \in \mathbb{R}^{l_\Phi}$ and $\hat{P}^i \in \mathbb{R}^{m \times l_\Psi}$ are the weight vectors. Note that u^{i+1} in (2.34) is estimated by a NN as (2.37) and no knowledge about the system dynamics is required. We define $v^i = [v_1^i \ \ ..., \ \ v_m^i] = u - u^i$ and define the error on off-policy Bellman equation (2.35) using (2.36) and (2.37) [58],

$$e^i(t) = \hat{W}^{iT}(\Phi(x(t))) - \hat{W}^{iT}(\Phi(x(t-T))) - \int_{t-T}^{t} (-Q(x) - B_\gamma(x) - u^{iT} R u^i) d\tau \\ + 2 \sum_{j=1}^{m} \rho_j \int_{t-T}^{t} (\hat{P}_j^{iT} \Psi(x(t)) v_j^i) d\tau \tag{2.38}$$

where ρ_j is the j^{th} diagonal element of R and $\hat{P}_j^{iT}$ is the j^{th} column of $\hat{P}^{iT}$. The least squares method is used to obtain the minimum of Bellman approximation error (2.38). In doing so, (2.38) is rewritten in regression form as

$$y^i(t) + e^i(t) = \hat{W}_t^{\ iT} h(t) \tag{2.39}$$

in which $\hat{W}_t^{\ i}$ is a matrix composed of weight vectors as

$$\hat{W}_t^{\ iT} = [\hat{W}^{iT},\ \hat{P}_1^{iT}, \ldots, \hat{P}_m^{iT}]$$

and $h^i(t)$ is

$$h^i(t) = \begin{bmatrix} \Phi(x(t)) - \Phi(x(t-T)) \\ 2\rho_1 \int_{t-T}^{t} (\Psi(x(t))v_1^i)d\tau \\ \vdots \\ 2\rho_m \int_{t-T}^{t} (\Psi(x(t))v_m^i)d\tau \end{bmatrix} \tag{2.40}$$

and $y^i(t)$ is

$$y^i(t) = \int_{t-T}^{t} (-Q(x) - B_\gamma(x) - u^{iT}Ru^i)d\tau \tag{2.41}$$

We collect the state and input data at N points at the time interval T to solve (2.39) for $\hat{W}_t^{\ iT}$. Let the collected information be saved in matrices H^i and Y^i as

$$H^i = [h^i(t_1)\,, \ldots,\, h^i(t_N)]$$
$$Y^i = [y^i(t_1)\,, \ldots,\, y^i(t_N)]^T$$

Therefore,the least-square equation is

$$\hat{W}_t^{\ iT} H^i = Y^i \tag{2.42}$$

and its solution is

$$\hat{W}_t^{\ iT} = (H^i H^{iT})^{-1} H^i Y^i \tag{2.43}$$

Equation (2.43) has a solution if

$$N > l_1 + ml_2 \tag{2.44}$$

Remark 2.12. The CBF is the dominant term in the vicinity of the risky area, while it rapidly damps as gets further away from the safety boundary. As a result, for having a reliable training, one needs to collect samples from both the safe region and the region in vicinity of the safety boundary in which the CBF comes to play.

Remark 2.13. The off-policy RL algorithm provides an optimal and safe solution to the optimization problem defined in (2.9). This is because, the off-policy RL algorithm applies a safe (possibly conservative) policy to the system while learning about an optimal and safe policy. Only the behavior policy requires partial knowledge of the dynamics and the learning process is model free.

Theorem 2.4. *Algorithm 1 converges to a safe optimal solution.*

Proof. Algorithm 1 iterates on the off-policy Bellman equation (2.35) with update law (2.34). According to Lemma 2.4, the off-policy Bellman (2.35) is equivalent to Bellman equation (2.30) with the same update law (2.34). On the other hand, it is shown in Lemma 2.2 that the value function obtained by iterating on the Bellman equation is monotonically decreasing and bounded. Therefore, Algorithm 1 converges to the optimal solution. □

Remark 2.14. Note that the behavior policy is assumed to be an exploratory policy and provides rich data for learning, i.e., the collected data guarantees that the least-square equation (2.42) has a feasible solution. Under this condition, similar to [59], it can be shown that at each iteration of Algorithm 1, the controller's weights converge to their desired values, which in turns results in an admissible policy that makes the system stable and the value function bounded. Boundedness of the value function guarantees safety at each iteration.

Algorithm 1 Safe Off-policy RL

1: Initialize actor and critic networks (2.36), (2.37).
2: **procedure** DATA COLLECTION
3: Employ the initial noisy stabilizing control policy $u \in \mathscr{U}_a$ as (2.14) until (2.44) is satisfied. This input must bring the system in vicinity of the risky area as well for reliable learning.
4: **end procedure**
5: **procedure** FIND AN OPTIMAL SOLUTION BY REUSING THE COLLECTED DATA
6: For all $t = t_1, ..., t_N$, given u^i, obtain matrices $h^i(t)$ and $y^i(t)$ as (2.40), (2.41).
7: Find NNs weights using (2.43), update J^i and u^{j+1} in (2.36), (2.37).
8: Stop if a stopping criterion is met, otherwise set $i = i + 1$ and go to 5.
9: **end procedure**

2.5 Simulation Results

The efficiency of the proposed method is examined in lane keeping problem for an autonomous vehicle. This problem aims to keep the car centered in the lane in spite of possible curvature of the road. In addition to this regulation objective, there is a safety objective which specifies the maximum allowable lateral displacement of the car according to the width of the road. The linear tire-force model and constant longitudinal speed is considered [28]. More details on system model and its formulation can be found in [60].

The state model of the system is given as

$$\begin{bmatrix} \dot{y} \\ \dot{v} \\ \dot{\phi} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 1 & v_{l0} & 0 \\ 0 & -\frac{C_f+C_r}{Mv_{l0}} & 0 & \frac{bC_r-aC_f}{Mv_{l0}} - v_{l0} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{bC_r-aC_f}{I_z v_{l0}} & 0 & 0 \end{bmatrix} \begin{bmatrix} y \\ v \\ \phi \\ \psi \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{C_f}{M} \\ 0 \\ a\frac{C_f}{I_z} \end{bmatrix} u + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} d$$

where y and v are the lateral displacement and its velocity, respectively, while y_{max} and y_{min} show the maximum and the minimum allowable displacement from the center of the road. ϕ is the error yaw angle and ψ is its derivative. u is the steering angle, while d is the desired yaw rate obtained from the curvature of the road as $d = \frac{v_{l0}}{R_r}$; v_{l0} is the longitudinal speed and R_r is the road radius of curvature. M is the total mass of the car and I_z is its moment

of inertia with respect to the center of the mass. C_r and C_f are stiffness parameters of tire. Finally, a and b show the distance of front and rear tires to the center of the mass. The value of parameters used in simulation are given in Table 4.1. To have a unified notation, the states of the system are denoted by $x = [x_1,\ x_2,\ x_3,\ x_4]^T = [y,\ v,\ \phi,\ \psi]^T$

The modified formulation (2.9) is employed with the following utility function,

$$r_a(x,u) = x^T Q x + u^T R u - log(\frac{\gamma_1(x_1 - y_{min})}{\gamma_1(x_1 - y_{min}) + 1}) - log(\frac{\gamma_2(-x_1 + y_{max})}{\gamma_2(-x_1 + y_{max}) + 1})$$

where Q, R, γ_1, γ_2 are design parameters. The activation functions for critic and actor networks are considered respectively, as

$$\Phi(x) = [x_1^2\ x_2^2\ x_3^2\ x_4^2\ x_1x_2\ x_1x_3,\ x_1x_4\ x_2x_3\ x_2x_4\ x_3x_4\ (x_1 - y_{max})\ {x_1}^4]$$
$$\Psi(x) = [x_1\ x_2\ x_3\ x_4]^T$$

Then, these networks are trained using off-policy Algorithm 1.

Critic and actor networks should be trained in the safe states as well as close to the risky states. So, the learned network is reliable in recognizing risk. After six iterations, the learning process is completed. The lateral displacement of the car is shown in Figure 2.1 with and without incorporation of the CBF. As it can be seen after learning, the states of the system have stayed within the safe region and have not exceeded the limits. Trajectories of other states of the system are given in Figure 2.2, which are converged to the origin. The actor weights and critic weights are shown in Figures 2.3 and 2.4, respectively. The graphs with different ranges have been separated.

2.6 Conclusion

In this chapter, a safe off-policy RL scheme is proposed which trades-off between safety and performance. This method guarantees and plans for the safety by incorporation of a

Table 2.1: Simulation Parameters

Parameter	Value	Parameter	Value
M	1650 Kg	$\|R_r\|$	$0 \rightarrow 0.1$
I_z	2315.3 $m^2.Kg$	y_{max}, y_{min}	0.45, -0.45 m
v_0	27.7 m/s	Q	$2 \times I_{n \times n}$
C_f	133000 N/rad	R	1
C_r	98800 N/rad	γ_1, γ_2	0.95, 2
a	1.11 m	b	1.59 m

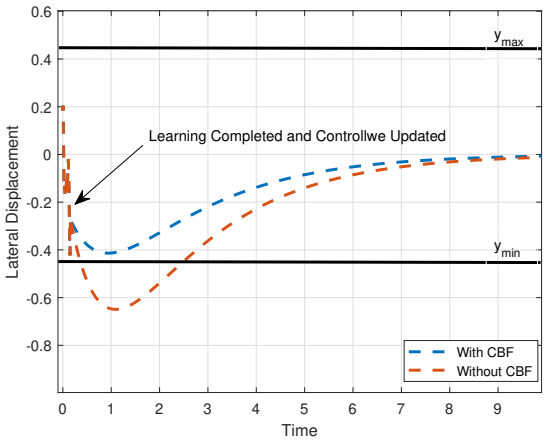

Figure 2.1: Lateral displacement with and without CBF

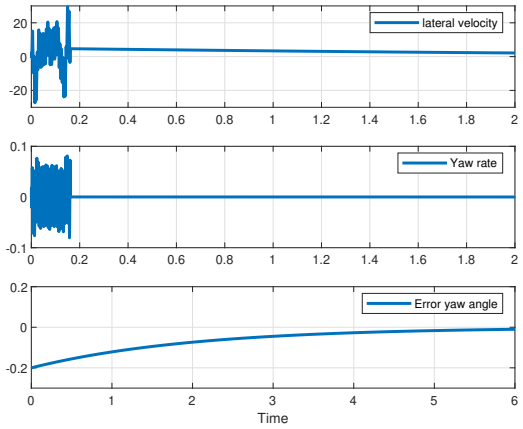

Figure 2.2: The states of the system

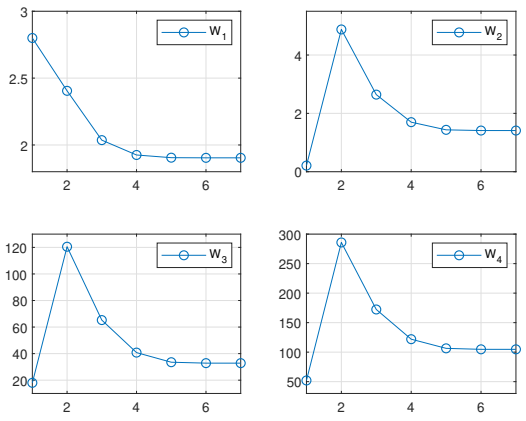

Figure 2.3: Actor Weights

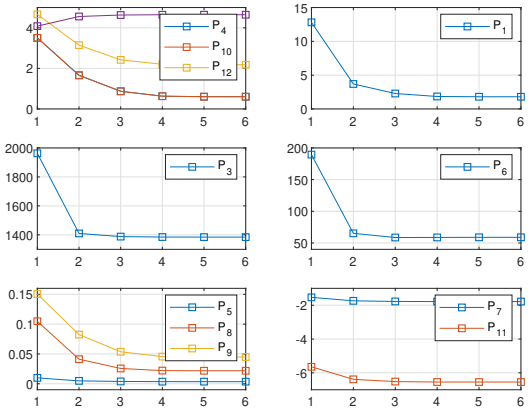

Figure 2.4: Critic Weights

CBF term into the cost function and forming an augmented value function. Using iterative approximation of the augmented value function, the application of CBF is extended to a data-driven approach. Rigorous proof of safety is presented. The notion of safety region is introduced for the case of no conflict between safety and performance and proof of stability and optimality in the safe region is derived accordingly.

Chapter3

Reinforcement Learning based Control Design with Safety and Stability Guarantees During Exploration

Contents of this chapter first appeared as [53] and have been reformatted to fit the requirements of this dissertation.

3.1 Introduction

Proper learning-based algorithms require satisfaction of the persistence of excitation (PE) condition. The PE condition is typically satisfied by applying noisy inputs to the system to excite all its dynamical modes. Since this noise is random and arbitrary, it might result in violation of safety. All methods mentioned in Chapter 1 need information about the system dynamics, environment or human supervision for safe exploratory data collection.

This chapter proposes a novel off-policy RL algorithm with prescribed learning performance with safety and stability guarantees during exploration and exploitation phases. To the best of our knowledge, it is the first time that safety and stability guarantees of the system during the excitation of the system in the presence of noisy input is ensured without any external knowledge about the risk, dynamics or environment. The schematic of the presented idea is depicted in Figure 3.1 as two main interconnected modules: i) a prescribed learning method with verifiable PE condition. ii) a robustified safe control design. In the first

module, experience replay-based safe model learning along with an off-policy RL algorithm are employed to present a framework to specify conditions under which the learning can be prescribed and how the data quality affects it. This method is capable of guaranteeing the exponential convergence of the learning error to zero with a prescribed bound that can be considered as a vanishing perturbation term to the nominal system, enabling stabilizing controller design. The outcome of the first module is then employed in designing a novel adaptive robustified control barrier function (AR-CBF). AR-CBF benefits from learning to compensate for uncertainties without being overly conservative and accounts for estimation error to guarantee safety despite learning inaccuracy. Any policy that satisfies this criterion assures the safe performance of the system. Since AR-CBF criterion is built based upon the current approximation of the dynamics, it can ensure safety during learning. The safe and stabilizing input obtained in the robustified design module is employed to collect more safe exploratory data. This collected data is then repetitively used to update the approximation of the dynamics and find the optimal target policy. The relationship between these two modules is reciprocal. The proposed learning approach provides a better description about the behavior of model learning error and its bound without excess conservatism, and the robustified design module enables deriving a noisy random and yet safe and stabilizing controller for further data collection. As the learning improves, the AR-CBF converges to the nominal CBF exponentially fast and provides more room for taking safe actions. When the optimal target policy is found, it is minimally altered to respect AR-CBF and safely applied to the system. Therefore, even if the system model is not perfectly approximated, the safe and optimal target policy can be successfully found and be applied to the system.

In a nutshell, the contributions of the chapter are as follows.

1. Proposing a learning-enabled safe model-free RL framework with safety and stability guarantee during data collection, exploration, and exploitation without external advice.

2. Integrating efficient RL with prescribed learning and verifiable PE condition in conjunction with a robustified formulation.

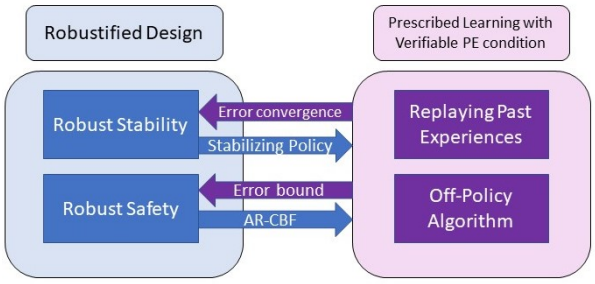

Figure 3.1: Overview of the proposed approach

3. Employing prescribed performance in the stability analysis based on perturbation theory.

4. Presenting a novel AR-CBF for safe control of uncertain systems with safety verification during learning.

3.1.1 Organization of the chapter

Section II is allocated for problem statement. Background information on CBFs and RL techniques is presented in Section III. The robustified safety and stability design using experience replay method is given in Section IV. Section V represents the proposed barrier-certified off-policy RL algorithm. Section IV represents the simulation results and Section V concludes the chapter.

3.2 Problem Statement

Consider a continuous-time linear system as

$$\dot{x} = Ax + Bu \tag{3.1}$$

where $x \in \mathbb{R}^n$ is the system state and $u \in \mathbb{R}^m$ is the control input. It is assumed that the system is stabilizable.

Assumption 3.1. The dynamics and input matrices $A \in R^{n\times n}$ and $B \in R^{n\times m}$ are unknown. Moreover, their initial approximations $\hat{A}_0$ and $\hat{B}_0$ can be chosen arbitrarily within the set $\{(\hat{A}_0, \hat{B}_0)|(\hat{A}_0, \hat{B}_0) is\ stabilizable\}$.

The control objective is to design u to optimize a performance function while assuring satisfaction of safety specifications. The safety objective is to ensure that, as the system's states evolve according to (3.1), they never leave a safe set $\mathscr{C}$, i.e.,

$$x(t) \in \mathscr{C}, \forall t \geq 0$$

where the safe set is formed using a safety criterion as

$$\mathscr{C} = \{x|h(x) \geq 0\} \tag{3.2}$$

where $h(x) : \mathbb{R}^n \to \mathbb{R}$ is a smooth function.

The performance objective encodes the quality of the control solution in achieving a goal. For the optimal stabilizing problem, the long-term cost function is typically chosen as

$$J = \int_0^\infty (x^T Q x + u^T R u) d\tau \tag{3.3}$$

where $Q = Q^T$ is a positive semi-definite matrix, while $R = R^T$ is a positive definite matrix. It is assumed that $(A, Q^{\frac{1}{2}})$ is observable.

Remark 3.1. Safety and performance can be in conflict and the performance level that can be achieved safely depends on the uncertainty level. Therefore, possible conflicts between safety and performance is considered in the proposed framework. When conflicts arise, the safety satisfaction is prioritized by imposing it as a hard constraint while the performance is considered as a soft constraint.

Therefore, the controller is in the form of

$$u = -Kx + \delta \tag{3.4}$$

where $u^* = -Kx$ is the optimal controller obtained by minimizing (3.3) without considering safety constraints, while δ is the safety modifier added to the optimal feedback policy to certify the safety of the system while minimally altering its actions. In case of no conflict between safety and performance, $\delta = 0$.

Finding the optimal control policy for uncertain systems is not directly possible and demands iterative approaches to approximate the optimal controller and the value function using neural networks (NNs). This, however, does not account for the safety of the system. Safety and stability guarantees are especially challenging at the beginning, as the collection of rich data for training NNs is required. This chapter presents a method with safety and stability guarantee in data collection, exploration, and exploitation phases.

3.3 Background

In this section, the background on CBFs and off-policy RL algorithm are briefly reviewed.

3.3.1 Control Barrier Functions

CBFs provide conditions for the control input that restricts the trajectories of the system to evolve in a pre-defined safe set by ensuring forward invariance of the set. Thus, by starting initially within the safe set and designing the controller to respect the CBF conditions, the safety of the system is guaranteed. Zeroing CBF as one major form of CBFs is formally defined as follows.

Definition 3.1. A continuous function $\alpha : (-b, a) \to (-\infty, \infty)$ with $a, b > 0$ is an extended class $\mathscr{K}$ function, if it is strictly increasing and $\alpha(0) = 0$ [55, 34]. ■

Definition 3.2. Considering the dynamical system (3.1) and the set $\mathscr{C} \subset \mathbb{R}^n$ (3.2) defined using a C^1 function $h(x)$, if there exists a locally Lipschitz extended class $\mathscr{K}$ function α such that

$$\sup_{u \in \mathscr{U}} [\frac{\partial h}{\partial x} A + \frac{\partial h}{\partial x} Bu + \alpha(h(x))] \geq 0, \quad \forall x \in \mathscr{D} \tag{3.5}$$

then, the function $h(x)$ is a ZCBF on $\mathscr{D}$ with $\mathscr{C} \subseteq \mathscr{D} \subset \mathbb{R}^n$ [28]. ■

The set of safe control inputs for $h(x)$ is formed accordingly as

$$\mathscr{U}_m(x) = \{u \in \mathscr{U} | \frac{\partial h}{\partial x} A + \frac{\partial h}{\partial x} Bu + \alpha(h(x)) \geq 0\}$$

Ensuring the forward invarinace of a set using ZCBFs is the result of the following theorem.

Theorem 3.1. *Given dynamical system* (3.1) *and the set* $\mathscr{C} \subseteq \mathscr{D}$ (3.2) *defined for a* C^1 *function* $h(x)$*, if* h *is a ZCBF on* $\mathscr{D}$*, any Lipschitz continuous controller* $\{u : \mathscr{D} \to \mathbb{R} | u \in \mathscr{U}_m(x)\}$ *renders the set* $\mathscr{C}$ *forward invariant.*

Proof. See [28]. □

Remark 3.2. Note that complete knowledge of the system dynamics, i.e., A and B matrices are required to guarantee (3.5). To obviate these requirements, a novel robustified CBF is proposed, which accounts for a non-conservative bound of error as well.

3.3.2 Adaptive Optimal Control Design

Having A and B known for the system (3.1), the optimal value function for the objective function (3.3) in the form of [54]

$$V(x) = x^T P x \tag{3.6}$$

where P is the solution of well-known algebraic Riccati equation (ARE)

$$A^T P + PA + Q - PBR^{-1}B^T P = 0 \tag{3.7}$$

which is quadratic in P. To sidestep the difficulty of solving quadratic equations, the Bellman equation is iteratively solved. Considering (3.3) and (3.6), the Bellman equation is formed as

$$x(t+\delta t)^T P x(t+\delta t) - x(t)^T P x(t) = \int_t^{t+\delta t} (x^T Q x + u^T R u) d\tau \tag{3.8}$$

To iteratively solve the Bellman equation, and by having $K_0 \in \mathbb{R}^{m\times n}$ as a stabilizing feedback gain matrix, the Lyapunov equation is formed

$$(A - BK_k)^T P_k + P_k(A - BK_k) + Q + {K_k}^T R K_k = 0 \tag{3.9}$$

where $P_k = {P_k}^T$ is the solution of (3.9) and is positive definite. Then, K_k is recursively defined as

$$K_k = R^{-1}B^T P_{k-1},\ k = 1, 2, ... \tag{3.10}$$

Then, one achieves the following properties:

1) $A - BK_k$ is Hurwitz

2)$P^* \leq P_{k+1} \leq P_k$

3)$\lim_{k\to\infty} K_k = K^*$, $\lim_{k\to\infty} P_k = P^*$

where P^* is the solution of ARE (3.7) and K^* is the optimal feedback gain. Therefore, the solution of ARE is approximated by iteratively solving (3.9) which is linear with respect to P_k.

However, A, B are needed in (3.10). To overcome this issue, [57] proposed online strategy

to solve (3.9) when the system is fully unknown. The system (3.1) is re-written as

$$\dot{x} = A_k x + B(K_k x + u) \tag{3.11}$$

where $A_k = A - BK_k$. Then, using (3.9), (3.10) and (3.11), the off-policy Bellman equation is formed

$$\begin{aligned} & x(t+\delta t)^T P_k x(t+\delta t) - x(t)^T P_k x(t) \\ &= \int_t^{t+\delta t} [x^T ({A_k}^T P_k + P_k A_k) x + 2(u + K_k x)^T B^T P_k x] d\tau \\ &= -\int_t^{t+\delta t} x^T Q_k x d\tau + 2\int_t^{t+\delta t} (u + K_k x)^T R K_{k+1} x d\tau \end{aligned} \tag{3.12}$$

where $Q_k = Q + {K_k}^T R K_k$. (3.12) is equivalent to the on-policy Bellman equation (3.8). However, this method does not consider safety of the system. In this chapter, a novel method to certify the safety of this algorithm is proposed.

3.4 Robustified Safety and Stability using Experience Replay Learning

In an off-policy algorithm, the behavior policy is applied to the system to collect data. A NN is assigned to learn about the dynamics of the system which its weights are updated by means of replaying the past experiences. After applying a few initial policies, a mild rank condition is satisfied which ensures in continuation of dynamics approximation, the learning error exponentially fast converges to zero with a predefined rate. This prescribed behavior of the learning error along the current rough approximation of the system is employed in design of robustified safe and stabilizing controller which is then integrated to the off-policy learner for safe data acquisition.

In this section, the experience replay approximation and the prescribed behavior of the

learning error is presented. It is shown that using this learning platform, the learning error is a vanishing perturbation to the system and condition for having a stabilizing controller is derived. Finally, a novel non-conservative robustified CBF is presented which ensures safety during learning.

3.4.1 Experience Replay System Approximation

The system dynamics (3.1) can be written in the form of

$$\dot{x} = W\phi(x, u) \tag{3.13}$$

where $W = [A, B] \in \mathbb{R}^{n\times(n+m)}$ and $\phi(x, u) = [x, u]^T \in \mathbb{R}^{(n+m)\times 1}$. The system dynamics (3.13) is written as a compact linear form

$$\dot{x} = G(t)\psi \tag{3.14}$$

where $G(t) \triangleq \phi(x, u)^T \otimes I_n \in \mathbb{R}^{(n)\times(mn+n^2)}$ and $\psi = vec(W) \in \mathbb{R}^{((n^2+nm)\times 1)}$. Let $\hat{\psi}$ be a rough estimation of ψ and $\tilde{\psi} = \psi - \hat{\psi}$ be the estimation error. The following filters are applied to $\dot{x}$, $G(t)$ in (3.14) and $\phi(x, u)$ in (3.13) in terms of σ, Ω and x_s, respectively as

$$\dot{\sigma}(t) = -\beta\sigma(t) + \dot{x} \tag{3.15}$$

$$\dot{\Omega}(t) = -\beta\Omega(t) + G(t) \tag{3.16}$$

$$\dot{x}_s(t) = -\beta x_s(t) + \phi(x, u) \tag{3.17}$$

where $\beta > 0$ is a design gain and $\Omega(0) = 0$, $x_s(0) = 0$. The filtered signal Ω in (3.16) can be written using x_s in (3.17) as

$$\Omega(t) = {x_s}^T \otimes I_n$$

The solution of (3.15), (3.16) and (3.17) are given, respectively as

$$\sigma(t) = e^{-\beta t} \int_0^t e^{\beta \tau} \dot{x}(\tau) d\tau \tag{3.18}$$

$$\Omega(t) = e^{-\beta t} \int_0^t e^{\beta \tau} G(\tau) d\tau \tag{3.19}$$

$$x_s(t) = e^{-\beta t} \int_0^t e^{\beta \tau} \phi(x, u) d\tau \tag{3.20}$$

where $\sigma \in \mathbb{R}^n$, $\Omega \in \mathbb{R}^{n \times (mn+n^2)}$, and $x_s \in \mathbb{R}^{(m+n)}$. The system dynamics (3.13) can be written using filtered signals as

$$\sigma(t) = W x_s \tag{3.21}$$

Using (3.14), (3.18) and (3.19), one has

$$\sigma(t) = \Omega(t) \psi \tag{3.22}$$

From (3.18) and using integration by part, σ can be expressed in terms of known variables $x(t)$ and $x_s(t)$ as

$$\sigma(t) = x(t) - e^{\beta t} x(0) - \beta x_s(t) \tag{3.23}$$

According to (3.22) and (3.23), the prediction error is defined as

$$e(t) = \sigma(t) - \Omega(t) \hat{\psi}(t) \tag{3.24}$$

where $\hat{\psi}$ is an estimation of ψ, and $\hat{\psi} = vec(\hat{W})$. In order to store and use the past data in the update law, two memory stacks $\{\sigma_i\}_{i=1:p}$, $\{\Omega_i\}_{i=1:p}$ are employed, which store the values of $\sigma(t_i)$ and $\Omega(t_i)$, respectively at each time instance t_i. The prediction error at time

constant t_i is defined accordingly as

$$e_i(t) = \sigma_i - \Omega_i \hat{\psi}(t) \tag{3.25}$$

The following update law using the past stored data is then employed

$$\dot{\hat{\psi}} = \beta_{\psi 1} \Omega^T(t) e(t) + \beta_{\psi 2} \sum_{i=1}^{p} {\Omega^T}_i e_i(t) \tag{3.26}$$

where $\beta_{\psi 1}$ and $\beta_{\psi 2}$ are positive scalar gains. This update law ensures exponential convergence of $\hat{\psi}$ to ψ under a rank condition and in the presence of enough stored data. This result if formally represented as follows.

Lemma 3.1. *[61] Considering the dynamics* (3.26)*, if there exists* p^* *such that for all* $p \geq p^*$*, for any sequence* $t_1 <_2< .... < t_p$*,*

$$rank([{\Omega_1}^T, {\Omega_2}^T, ..., {\Omega_p}^T]) = mn + n^2 \tag{3.27}$$

Then, using the update law (3.26) $\hat{\psi}$ *converges to* ψ *exponentially fast with employing the Lyapaunov function* $V_\psi = 0.5\tilde{\psi}^T\tilde{\psi}$ *and there exists a positive gain* $\beta_{\psi 12}$ *such that*

$$\dot{V}_x \leq -2(\beta_{\psi 12}) V_x \tag{3.28}$$

Remark 3.3. In a nutshell, in experience-replay dynamics approximation, the regressor form of the dynamics is derived and an update law which incorporates past stored data is employed. By satisfaction of a rank condition (3.27), and considering (3.28) the learning error exponentially fast converges to zero. In other words, after $mn + n^2$ number of samples are collected fast convergence of error to zero is ensured with a prescribed rate.

Remark 3.4. The significance of this method in safe RL exploration is twofold. First,

since the learning error exponentially converges to zero, the error term can be taken as a vanishing perturbation to the approximated dynamics and therefore perturbation theory can be employed to design a controller based on the approximated dynamics which guarantees stability for the true dynamics. Second, an accurate bound of learning error can be derived. This bound can be taken as a non-conservative worst case in formation of the novel robustified CBF.

3.4.2 Stability Analysis

Stability analysis is performed based on control Lyapunov function (CLF). However, due to uncertainty in the model, CLF is built based on the available approximated model and its validity for the original system needs to be investigated. Having the experience replay model learning enables designing stabilizing controllers for the true system based on the approximated dynamics.

Theorem 3.2. *Let $x = 0$ be an exponentially stable equilibrium point of the following closed loop approximated system with stabilizing feedback gain k*

$$\dot{x} = \hat{A}x - \hat{B}kx \tag{3.29}$$

Let $V(x) = x^T Px$ be the Lyapunov function for (3.29)*, where P is the solution to the Lyapunov equation. Suppose that update law* (3.26) *is employed and* (3.27) *is satisfied. Then the origin is exponentially stable for the original system* (3.1)*.*

Proof. By considering (3.14) in closed loop format and taking $\hat{A}_c = A - Bk$ where k is the feedback gain, (3.1) can be written as

$$\dot{x} = \hat{A}_c x + G\tilde{\psi} \tag{3.30}$$

Model learning error $G\tilde{\psi} = \tilde{W}[x, -kx]$ which is equal to zero at the origin. Therefore, error

term vanishes at the origin. Furthermore, satisfaction of (3.27) results in satisfaction of (3.33). Therefore, one has

$$G(t)\tilde{\psi} \leq G(t)\tilde{\psi}(0)$$

Thus,

$$G(t)\tilde{\psi} \leq \tilde{A}_c(0)x$$

Thus, the error term satisfies linear growth bound and there exists a coefficient γ such that

$$G(t)\tilde{\psi} \leq \gamma||x||$$

Therefore, the error term $G\tilde{\psi}$ is a vanishing perturbation to the approximated system $\dot{x} = \hat{A}_c x$.

Since $\hat{A}_c$ is Hurwitz and $P = P^T$ is the solution to the Lyapunov equation

$$P\hat{A}_c + \hat{A}_c^T P = -Q$$

Then, the quadratic Lyapunov function $V = x^T P x$ satisfies the following properties [55].

$$\begin{aligned}
&\lambda_{min}(P)||x||^2 \leq V(x) \leq \lambda_{max}(P)||x||^2 \\
&\frac{\partial V}{\partial x}\hat{A}_c x = -x^T Q x \leq -\lambda_{min}(Q)||x||^2 \\
&||\frac{\partial V}{\partial x}|| = ||2x^T P|| \leq 2\lambda_{max}(P)||x||
\end{aligned}$$

The derivative of $V(x)$ along the trajectory of the original system (3.30) becomes

$$\dot{V}(x) = \frac{\partial V}{\partial x}\hat{A}_c x + \frac{\partial V}{\partial x}G\tilde{\psi}$$

which satisfies

$$\dot{V}(x) \leq -\lambda_{min}(Q)||x||^2 + 2\lambda_{max}(P)\gamma||x||^2$$

Thus, if

$$\gamma \leq \frac{\lambda_{min}(Q)}{2\lambda_{max}(P)} \tag{3.31}$$

Then, the origin of the original system (3.1) is exponentially stable. This completes the proof. □

In other words, by employing experience-replay model learning, the modeling error can be taken as a vanishing perturbation to the approximated system, where the bound in (3.31) depends on choice of Q. Hence, stability analysis for the approximated system is valid in stability guarantee for the original system with proper design.

3.4.3 Adaptive Robustified CBF

The ARCBF condition is a stricter version of the CBF based on a rough estimation and worst-case model-learning's error which its satisfaction ensures the safety of the system.

Definition 3.3. Consider the dynamical system (3.13) and the set $\mathscr{C} \subset \mathbb{R}^n$ (3.2) defined using a C^1 function $h(x)$. Let there exist a locally Lipschitz extended class $\mathscr{K}$ function α such that

$$\sup_{u \in \mathscr{U}} [\frac{\partial h}{\partial x} G(t)\hat{\psi} - ||\frac{\partial h}{\partial x} G(t)||a + \alpha(h(x))] \geq 0, \quad \forall x \in \mathscr{D} \tag{3.32}$$

where a is the bound of estimation error as $||\tilde{\psi}|| \leq a$. Then, the function $h(x)$ is an ARCBF on $\mathscr{D}$ with $\mathscr{C} \subseteq \mathscr{D} \subset \mathbb{R}^n$. ■

The set of safe control inputs for $h(x)$ is formed accordingly as

$$\mathscr{U}_r(x) = \{u \in \mathscr{U} | \frac{\partial h}{\partial x} G(t)\hat{\psi} - ||\frac{\partial h}{\partial x} G(t)||a + \alpha(h(x)) \geq 0\}$$

Ensuring the forward invariance of a set using ARCBF is presented in the following theorem.

Theorem 3.3. *Consider dynamical system* (3.1)*, and its compact form* (3.14) *with estimation update law given by* (3.26) *and error bound of a as defined in Definition 3.3, and the set* $\mathscr{C} \subseteq \mathscr{D}$ (3.2) *defined for the* C^1 *function* $h(x)$*. If* h *is an ARCBF on* $\mathscr{D}$*, then any Lipschitz continuous controller* $\{u : \mathscr{D} \to \mathbb{R} | u \in \mathscr{U}_r(x)\}$ *renders the set* $\mathscr{C}$ *forward invariant.*

Proof. Considering (3.32), one has

$$\begin{aligned}
&\frac{\partial h}{\partial x} G(t)\hat{\psi} - ||\frac{\partial h}{\partial x} G(t)||a + \alpha(h(x)) \leq \\
&\frac{\partial h}{\partial x} G(t)\hat{\psi} - ||\frac{\partial h}{\partial x} G(t)\tilde{\psi}|| + \alpha(h(x)) \leq \\
&\frac{\partial h}{\partial x} G(t)\hat{\psi} + \frac{\partial h}{\partial x} G(t)\tilde{\psi} + \alpha(h(x))
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\partial h}{\partial x} G(t)\hat{\psi} - ||\frac{\partial h}{\partial x} G(t)||a + \alpha(h(x)) \leq \\
&\frac{\partial h}{\partial x} G(t)\psi + \alpha(h(x))
\end{aligned}$$

Since the left-hand side of above equation is positive, it ensures positiveness of its right-hand side and therefore the original CBF.

$$\frac{\partial h}{\partial x}(Ax + Bu) + \alpha(h(x)) \geq 0$$

Considering Theorem 3.1, the safety of the system is ensured. This completes the proof. □

As seen above, in addition to the estimation of the dynamics, the bound of modeling

error is employed in formation of ARCBF. Improper model learning and inaccurate worst-case value of error result in conservatism of the controller. This issue is obviated using the results of Lemma 3.1.

Considering (3.28) and comparison lemma, one has

$$V_x \leq V_x(t_0)e^{-2(\beta_{\psi 12})(t-t_0)}$$

Therefore,

$$||\tilde{\psi}|| \leq ||\tilde{\psi}(t_0)||e^{-(\beta_{\psi 12})(t-t_0)} \tag{3.33}$$

This gives an accurate bound of approximation error. By employing (3.33) in (3.32), the ARCBF criterion becomes

$$\begin{aligned}[\frac{\partial h}{\partial x}G(t)\hat{\psi} - ||\frac{\partial h}{\partial x}G(t)||||\tilde{\psi}(t_0)||e^{-(\beta_{\psi 12})(t-t_0)} \\ + \alpha(h(x))] \geq 0, \quad \forall x \in \mathscr{D}\end{aligned} \tag{3.34}$$

From Theorem 3.3, if the control policy satisfies (3.34), then the safety of the system is ensured.

Remark 3.5. Note that the ARCBF provides an invariance safety criterion for the worst-case uncertainty by incorporation of error bound, while experience-replay model learning quantifies the exponential convergence rate of the error to zero. In other words, an accurate bound of uncertainty is obtained that rapidly vanishes and results in convergence of ARCBF to the original CBF.

Remark 3.6. While (3.27) is satisfied by applying $mn + n^2$ safe initial actions, ARCBF is formed and the rest of policies which are needed for data acquisition are chosen such that (3.34) is satisfied. Therefore, off policy RL and model learning are safe without human intervention.

Remark 3.7. Since safety is certified after that rank condition (3.27) is satisfied, one might consider $||\tilde{\psi}(0)|| + \epsilon$ instead of $||\tilde{\psi}||(0)$ in (3.34), where $\epsilon > 0$. This gives a room of safe action for the initiation of model learning.

3.4.4 Safe and Stable Controller

Safety condition is encoded in (3.34), while quadratic CLF satisfying (3.31) ensures exponential stability of the system. Therefore, to have safety and stability, any policy is first minimally modified to satisfy these conditions through a quadratic programming optimization [27].

$$\begin{aligned} &\min_{u,\rho} \ \ ||u - u^*|| + ||\rho|| \\ &\text{s.t.} \ \ (3.34), \\ &\quad \dot{V} < -\lambda_{min}(Q)||x||^2 + \rho. \end{aligned} \tag{3.35}$$

where V is the control Lyapunov function that encodes the performance objective which is relaxed by factor ρ, while safety is applied as the hard constraint.

Remark 3.8. Quadratic programming formulation (3.35) is based on u; while considering $G(t) = [x, u]^T \otimes I_n$, $||u||$ appears in (3.34). Therefore, $||u|| = sgn(u)u$ is used, and thus ARCBF criterion and and Lyapunov derivative condition incorporated in (3.35) are linear with respect to control policy u and therefore it is a well-defined optimization problem to be solved.

3.5 Barrier-certified Off-Policy Algorithm

In the off-policy RL, two policies are defined. The behavior policy which is applied to the system to gather training data and target policy which is updated toward the optimal policy

using training data. This method is superior in safety critical applications for a couple of reasons. First, the target policy is updated without even being applied to the system. Second, it is efficient and repetitively uses the same data and therefore it demands application of less noisy inputs to the system. This method however faces safety risk at two stages. First, at the beginning, where no model about the system is available and noisy behavior policy is applied to the system and second when the learned target policy is applied to the system which is not necessarily safe. Therefore, it is desired to make sure the safety and stability of the system is preserved in the whole operation.

It is shown so far that by employing the experience replay dynamics approximation, any behavior policy that stabilizes the approximated dynamics and satisfies (3.34) is safe and stabilizing for the original system. The outcome of this robustified design is integrated into off-policy RL in order to have a safe and stable data acquisition, exploration and exploitation. Note that although the system is getting approximated for the sake of reduced conservatism of controller, the controller does not need to wait until the identification is complete; rather, the identification and off-policy controller are working at the same time.

For a given stabilizing K_k, (3.12) can be written in the matrix form [57]. To do so, $\hat{P} \in \mathbb{R}^{\frac{1}{2}n\times(n+1)}$ and $\bar{x} \in \mathbb{R}^{\frac{1}{2}n\times(n+1)}$ are defined based on $P \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, as

$$\hat{P} = [p_{11}, 2p_{12}, ..., 2p_{1n}, p_{22}, 2p_{23}, ...2p_{(n-1),n}, p_{nn}]^T$$
$$\bar{x} = [{x_1}^2, x_1x_2, ..., x_1x_n, {x_2}^2, x_2x_3, ..., x_{n-1}x_n, {x_n}^2]^T$$

The matrix form of (3.12) is

$$\Theta_k \begin{bmatrix} \hat{p}_k \\ vec(K_{k+1}) \end{bmatrix} = \Xi_k \tag{3.36}$$

where $\Theta_k \in \mathbb{R}^{l\times(\frac{1}{2}n(n+1)+mn)}$, $\Xi_k \in \mathbb{R}^l$ are defined as

$$\Theta_k = [\delta_{xx}, -2I_{xx}(I_n \otimes {K_k}^T R) - 2I_{xu}(I_n \otimes R)]$$
$$\Xi_k = -I_{xx} vec(Q_k)$$

In which for a positive integer l and time sequence $0 \leq t_0 < t_1 < ... < t_l$

$$\delta_{xx} = [\bar{x}(t_1) - \bar{x}(t_0), \bar{x}(t_2) - \bar{x}(t_1), ..., \bar{x}(t_l) - \bar{x}(t_{l-1})]^T,$$
$$I_{xx} = [\int_{t_0}^{t_1} x \otimes x d\tau, \int_{t_1}^{t_2} x \otimes x d\tau, ..., \int_{t_{l-1}}^{t_l} x \otimes x d\tau]^T$$
$$I_{xu} = [\int_{t_0}^{t_1} x \otimes u d\tau, \int_{t_1}^{t_2} x \otimes u d\tau, ..., \int_{t_{l-1}}^{t_l} x \otimes u d\tau]^T$$

where $\delta_{xx} \in \mathbb{R}^{l\times\frac{1}{2}n(n+1)}$, $I_{xx} \in \mathbb{R}^{l\times n^2}$ and $I_{xu} \in \mathbb{R}^{l\times mn}$. Note that if Θ_k is full rank, then (3.36) can be uniquely solved. This criterion is employed for rich data collection.

Safe off-policy algorithm is achieved in the following three phases.

Safe and Stable Data Collection: In this phase a few safe policies are applied to satisfy experience replay rank condition (3.27). The guaranteed prescribed behavior of the learning error is employed in deriving the condition on safe and stabilizing controller. The noisy input is modified accordingly and then is applied to the system for more safe data collection. The AR-CBF and the system approximation enhances at each iteration by collection of more data exponentially fast. Collection continues until it suffices for off-policy optimal controller calculation.

Optimal Policy Approximation: In this phase, the safe collected data is repetitively used at each target policy iteration toward the optimal controller.

Safe Target Policy Calculation: In this phase the optimal controller is minimally altered to satisfy AR-CBF condition and then is safely applied to the system. Experience replay approximation continues at each time instance by replaying the stored noisy input along the current response of the system until it converges to true dynamics which is equivalent with

convergence of AR-CBF to CBF.

More mathematical and detailed steps of the algorithm is represented in the following.

Algorithm 2 Safe and Stable Off-Policy RL

Initialization:
Initiate a NN for the dynamics (3.13) with stabilizing K_0, set numerator $k = 0$.

procedure Safe and Stable Data Collection
Apply p initial policies until (3.27) is satisfied.
Update dynamics NN weights based on (3.26).
Form the quadratic Lyapunov function such that (3.31) is satisfied.
Form AR-CBF based on (3.34).
Form the noisy input $u = \hat{K}x + e$ and modify it based on (3.35).
Go to the procedure "Optimal Policy Approximation" if Θ_k is full rank in (3.36). Otherwise, repeat the procedure until enough data for optimal policy approximation is collected.
end procedure

procedure Optimal Policy Approximation
Solve P_k and K_{k+1}.
If $||P_k - P_{k+1}|| < \epsilon$ then go to "Safe Target Policy Calculation", otherwise $k = k + 1$ and repeat the procedure.
end procedure

procedure Safe Target Policy Calculation
Minimally modify optimal controller u^* using (3.35) and apply it to the system.
If dynamics approximation is not converged, use the outcome of the system along the previous stored data to update dynamics NN.
Update AR-CBF based on the new approximation. Repeat procedure until control objectives are met.
end procedure

3.6 Simulation

3.6.1 Simulation Setup

Consider the following dynamical system

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} u \tag{3.37}$$

where $x = [x_1, x_2]$. The safety set in which the states of the system should belong to is defined as

$$\mathscr{C} = \{x | -a_1 \leq x_1 \leq a_1, -a_2 \leq x_2 \leq a_2\} \tag{3.38}$$

It is desired to ensure (3.38) is forward invariant. Thus, to certify the safety of the system the following CBFs are defined

$$h_1 = a_1 - x_1$$

$$h_2 = x_1 + a_1$$

$$h_3 = a_2 - x_2$$

$$h_4 = x_2 + a_2$$

To form AR-CBF condition (3.34), single layer NNs are used to learn the system dynamics.

$$\hat{\dot{x}} = \begin{bmatrix} \hat{w}_1 & \hat{w}_2 \\ \hat{w}_4 & \hat{w}_5 \end{bmatrix} x + \begin{bmatrix} \hat{w}_3 \\ \hat{w}_6 \end{bmatrix} u$$

Which is updated using experiance replay update law (3.26) and after applying a few safe initial condition until (3.27) is satisfied. The bound of learning error is obtained using (3.33)

and denoted by $\bar{w}_i$ for $1 \leq i \leq 6$.

ARCBF criteria is formed accordingly based on (3.34) as

$$\begin{aligned}
-\dot{\hat{x}}_1 + \alpha_1 h_1 - ||\bar{w}_1 x_1|| - ||\bar{w}_2 x_2|| - ||\bar{w}_3|| sgn(u)u \geq 0 \\
\dot{\hat{x}}_1 + \alpha_2 h_2 - ||\bar{w}_1 x_1|| - ||\bar{w}_2 x_2|| - ||\bar{w}_3|| sgn(u)u \geq 0 \\
-\dot{\hat{x}}_2 + \alpha_3 h_3 - ||\bar{w}_4 x_1|| - ||\bar{w}_5 x_2|| - ||\bar{w}_6|| sgn(u)u \geq 0 \\
\dot{\hat{x}}_2 + \alpha_4 h_4 - ||\bar{w}_4 x_1|| - ||\bar{w}_5 x_2|| - ||\bar{w}_6|| sgn(u)u \geq 0
\end{aligned} \tag{3.39}$$

The noisy input in the form of (3.4) is modified in the (3.35) with (3.39) as a hard inequality constraint and solution of linear quadratic programming (LQR) with $Q = I, R = 1$ which satisfies (3.31) as its soft equality constraint. The output is then applied to the system for further data collection. The collected data is iteratively used for optimal policy approximation which is then minimally modified using (3.35) for a safe and optimal operation as Algorithm 2. The numerical details of simulation setup are given is Table 3.1.

3.6.2 Simulation Results and Discussion

The states of the system under the proposed RL controller is depicted in Figure 3.2 where the safety boundary is shown with dashed red lines. To have a safe performance, the trajectory of the system must stay between two lines. As can be seen in this figure, the safety and stability of the system is preserved even at the beginning of the simulation where noisy input is applied to the system.

To demonstrate the advantage of the proposed method, plain off-policy under the same setup is applied to the system. With the same value of added noise, the system becomes unstable in the data collection phase. To avoid instability of the system the value of noise is manually reduced. Its result is shown in Figure 3.3. As can be seen in this figure, although the stability of the system is satisfied by manual modification of the noise, the safety of the system is violated.

Remark 3.9. Comparison of Figures 3.2 and 3.3 reveals two significant advantages of the proposed method. First of all, we ensure automatic stability guarantee during exploration. This obviates the need of manual adjustment of noise to avoid instability. Second, we ensure safety guarantee during the challenging phase of data collection which is not tractable to do it manually.

The weight errors and their exponential bound is shown in Figure 3.4. As can be seen in this figure, with replaying the past experiences, the behavior of the learning error is properly prescribed and exponentially fast has converged to zero.

Remark 3.10. Considering the time scale of Figures 3.4 and (3.2) reveals that at early stages of data collection, although the learning error is high, still, the safety of the system is satisfied. As mentioned earlier, system's dynamics is approximated along the operation of off-policy controller and off-policy controller does not need to wait until system approximation is finished. In other words, safety during learning is ensured.

The result of iterations toward optimal policy is shown in Figure 3.5. As can be seen in this figure, K_k and P_k are successfully converged to their optimal values by repetitive employment of safe collected data.

Table 3.1: Simulation Parameters

Parameter	Value	Parameter	Value
α_1, α_2	40	Q	$I_{2\times2}$
α_3, α_4	10	R	1
$\beta_{\psi1}$, $\beta_{\psi2}$	10	H	[1,0; 0, 10]
$\bar{w}_i, i = 1, 2, 4, 5$	$1.8e^{-0.5t}$	F	[1, -1]
$\bar{w}_3$	$e^{-0.5t}$	$\bar{w}_6$	$0.5e^{-0.5t}$
a_1	1	a_2	1.4

3.7 Conclusion and Future Work

A barrier-certified safe RL framework with safety and stability guarantee in exploration and exploitation phases is proposed. It is obtained my means of efficient learning with prescribed

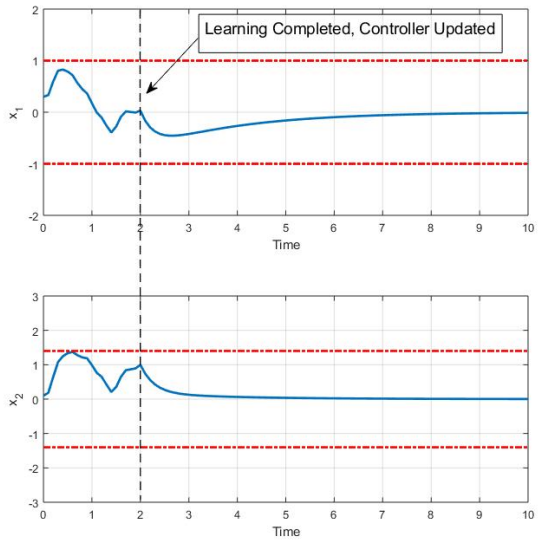

Figure 3.2: States of the system under the proposed framework

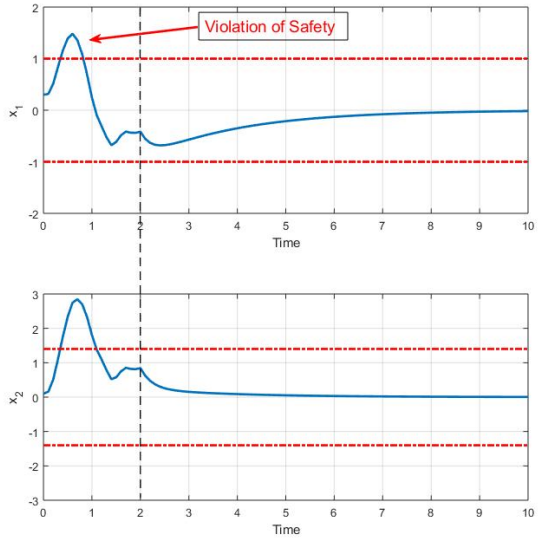

Figure 3.3: States of the system with plain off-policy with manual reduced noise

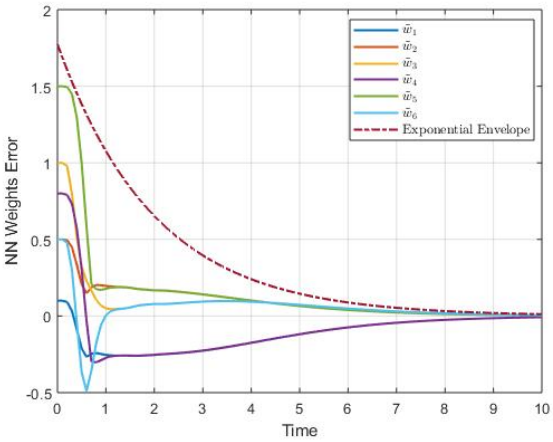

Figure 3.4: NN Weight error

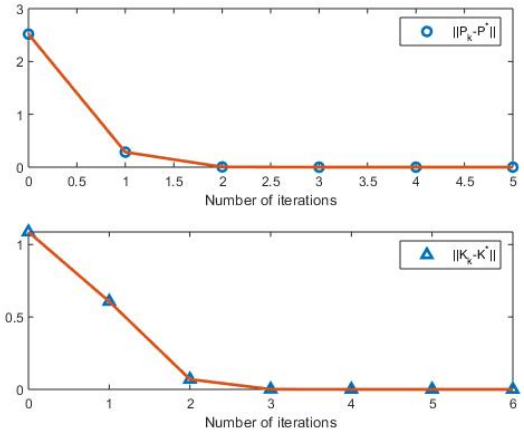

Figure 3.5: Convergence of P_k and K_k

performance along a robutified safe and stabilizable controller throughout the algorithm including the data collection phase. Experience replay-based model approximation is employed which ensures the exponential convergence of the learning error to zero after a mild rank condition is satisfied. This makes the learning error as a vanishing perturbation to the approximated model which facilitates designing stabilizing controller using the available rough knowledge of the system. The accurate bound of error is then employed in formation of a novel non-conservative AR-CBF which ensures safety during learning. AR-CBF and stabilizing controller are integrated through quadratic programming and is used for further data collection needed for off-policy iteration. The noisy input is modified accordingly to result in safe and stable action. After collecting safe rich data, the optimal policy is approximated and then again is certified using AR-CBF for safe exploitation. The efficacy of the proposed method is demonstrated in simulation. Extension to nonlinear dynamics, considering the effect of network reconstruction error are future directions of this line of research.

Chapter4

Barrier-certified Learning-enabled Safe Control Design for Systems Operating in Uncertain Environments

Contents of this chapter first appeared as [51] and have been reformatted to fit the requirements of this dissertation.

4.1 Introduction

This chapter presents a method for designing a learning-enabled safe controller for systems that must operate in environments that are shared with other agents with uncertain behaviors: The behaviors of surrounding agents affect the safe set and thus safe control design of the ego system, which are unknown and uncontrollable from the ego system's perspective. This is in sharp contrast with existing safe control methods requiring complete knowledge of the safe set. The uncertainty of the safe set caused by the uncertain behaviors of surrounding agents makes safe control design much more challenging. Fast and sample-efficient learning of uncertainties is of vital importance to avoid an overly conservative control design (which can also result in infeasibility) or unsafe behavior. A slow model-learning approach also avoids proactive safe control design, which can jeopardize the performance. Moreover, and even more importantly, a naive model learning approach based on minimizing the modeling error cannot account for safety even if the expected estimation error decreases over

time; This is because different models with the same modeling errors might have different characteristics in preserving the invariant behaviors of the actual system: Novel learning algorithms are required to avoid misrepresentation of the safe set as much as possible.

The interaction between agents is formulated using two sets of decoupled differential equations corresponding to the ego system and the risk-imposing external agent. A safety criterion is defined as a function of both subsystems' states. This is in sharp contrast with the existing works, which only consider partial uncertainty in the system dynamics and define the safety criterion solely based on the ego-system's states. The proposed framework is far more inclusive for safety-critical control scenarios where the agent operates in a cluttered uncertain environment shared with other agents. Since the trajectory information is required to form ZCBFs, the unknown external agent dynamics need to be learned. To make less conservative decisions and avoid misrepresentation of the safe set, a safety-aware model-learning approach that leverages safety-aware loss functions and the experience replay method is presented to learn uncertain and unknown behavior of the external agent. More specifically, the loss function is defined based on the barrier function error, instead of the system model error, and is minimized for both current samples and past samples stored in the memory to assure a fast and generalizable learning algorithm for approximating the safe set. Moreover, it provides an easy-to-verify metric on collected data to assure learning of the actual safe set, allowing to make more informative control decisions. Then, a learning-enabled ZCBF (L-ZCBF) is presented that integrates the proposed safety-aware model learning and a novel ZCBF to assure the safety of the ego system in the presence of uncertainty in the behavior of its surrounding agents. Since ensuring forward invariance of the approximated safe set does not necessarily ensure forward invariance of the actual safe set and strict safety might be violated, L-ZCBF employs the approximated trajectory of external agent and also the trajectory of ego-system, but shrinks the boundary of the safe set in case of an imminent risk that can be predicted using observations of the states of the external agents. These observations can be acquired using embedded sensors such as Light Detection and Ranging

(LIDAR). Guaranteeing forward invariance of the intersection of the approximated safe set and the actual safe set assures safety during learning and automatically shrinks the boundary of the safe set to the extent that safety of the overall system is guaranteed despite uncertainty. As learning enhances, this set expands to the actual safe set.

In a nutshell, the contributions of the chapter are as follows.

1. The problem of safe control design for systems operating in uncertain shared environments is formulated as two sets of decoupled dynamics with a safety criterion defined as a function of both ego and external agent's states to have a more inclusive scheme for safety-critical systems operating in the cluttered environment.

2. A novel learning-enabled ZCBF is proposed, which is capable of safety guarantee during learning of unknown dynamics.

3. The safety-aware model learning is proposed for rapid convergence of the approximated safe set to the exact one.

4.1.1 Organization of the Chapter

Section 2 provides the problem statement as well as preliminaries and background information. The main idea of learning-enabled ZCBFs for uncertain sets is presented in Section 3. Section 4 represents the overall control framework and the proposed algorithm. Case study, simulation results, and conclusion are given in Section 5, 6 and Section 7, respectively.

4.2 Problem Statement and Background

In this section, the safe control problem in the presence of external agents is stated, and some background information on ZCBFs and safe control design is provided.

4.2.1 Problem Statement

Consider the control system in the nonlinear affine form as

$$\dot{x} = f(x) + g(x)u \tag{4.1}$$

where $x \in \mathscr{X}$ and $u \in \mathscr{U}$ are the states of the controlled system and the control input, respectively. $f(x) \in \mathbb{R}^n$ is its drift dynamics and $g(x) \in \mathbb{R}^{n \times m}$ is its input dynamics. $f(x)$ is C^1 and $f(0) = 0$. It is assumed that the ego system is stabilizable and $\mathscr{U}$ is non-empty.

The goal is to ensure the safety of the control system (4.1) in a shared environment with external agents with uncertain and unknown behaviors. The dynamics of the external agents that affect the safety of (4.1) is given as

$$\dot{z} = f_2(z) \tag{4.2}$$

which is unknown and out of control of the ego system and $z = [z_1, ..., z_{p_2}]$ is the state vector of external agents which can be measured in real-time by the ego system (e.g., measuring the position of a leading vehicle using embedded sensors that measure the distance and relative steering) and $f_2 \in \mathbb{R}^{p_2}$ is assumed to be locally Lipschitz. Note that (4.2) does not need to capture the complete dynamical behavior of external agents in the surrounding environment, as it might require high-dimensional dynamics, which makes their learning computationally intractable; rather, it concerns simplified dynamics that best captures the effect of external agents on the safety of the ego system. For example, in urban driving, distance to other agents and obstacles and how they are approaching the ego vehicle matter most when safety is the main concern. The safety of the ego system is then formulated as a function of both x and z, which is uncertain due to unknown dynamics of z.

It is desired to satisfy uncertain safety criteria which are impacted by states of the system (4.1) and the external dynamics (4.2) and achieve stability and performance specifications

as long as it is safe.

Control Objectives: The following objectives must be achieved for the system 4.1:

1. Assuring safety by guaranteeing that the following safety conditions are satisfied all the time

$$l_i(x, z) \geq 0, \ \forall 1 \leq i \leq q$$

where $l_i(x, z) > 0$ is the i^{th} element of the safety criteria which is a smooth function describing a constraint on the system, and q is the total number of constraints.

2. Guaranteeing stability of the controlled system, i.e., $x \to 0$ as $t \to \infty$ in the case of no conflict with safety.

The safe set is formed as the intersection of all the sets, each satisfying a safety constraint. That is, the safe set is defined as

$$\mathscr{C} = \mathscr{C}_1 \cap \mathscr{C}_2 ... \cap \mathscr{C}_q \tag{4.3}$$

where

$$\mathscr{C}_i = \{x | l_i(x, z) \geq 0\}, \ \forall 1 \leq i \leq q \tag{4.4}$$

Safety imposes hard constraints on the control design, while performance is a soft constraint satisfied in the case of no conflict with safety.

Remark 4.1. Note that two sets of dynamics are considered in this framework in which (4.1) represents the first one and is known, and (4.2) represents the second one and is assumed to be uncertain and unknown. The safety set is represented as a function of both dynamics' states (4.4). Therefore, even when the dynamics of the ego system is partially available, this method is applicable since this unknown part is included in (4.2) which is learned.

Therefore, this covers not only disturbances that can be learned by collecting data but also a more general class of uncertainties in the environment and the ego-system dynamics.

4.2.2 Control Barrier Functions

Guaranteeing positive invariance of a set of states has broad applications in control system design, such as control of constrained systems and region of attraction maximization. For a dynamical system, positive invariance of a set means that inclusion of states in a specific set at any time ensures the inclusion of states in that set in the future time. Extension of this notion to control systems is called controlled positive invariance of a set, which guarantees forward invariance of the set by designing a proper control input. One of the widely referred theorems in the characterization of positive invariant sets is the Nagumo's theorem [62, 63, 64]. This theorem is presented using the concept of the tangent cone of a set [65, 64].

Theorem 4.1. ***Nagumo's Theorem.***

Given a dynamical system $\dot{x} = f(x)$ which has a globally unique solution for any initial condition $x_0 \in \mathscr{X}$, let $\mathscr{S} \subset \mathscr{X}$ be a closed set. Then, $\mathscr{S}$ is positively invariant if and only if

$$f(x) \in \mathscr{T}_{\mathscr{S}}(x), \quad \forall x \in \partial\mathscr{S} \tag{4.5}$$

where $\partial\mathscr{S}$ is the boundary of the set $\mathscr{S}$ and $\mathscr{T}_{\mathscr{S}}(x)$ is the tangent cone to $\mathscr{S}$.

Proof. See Theorem 3.1 in [63] and Theorem 4.7 in [64]. □

Remark 4.2. The Nagumo's theorem implies that to have a positive (forward) invariant set, $\dot{x}$ should point inside the set at the boundary, or in the worst case, it should be tangent to the boundary.

CBFs are used to ensure forward invariance of a specific set in a control system. ZCBF is a positive function within a set and zero at its boundary and thus, having a zeroing derivative

in the vicinity of the boundary prevents the states of the system from exceeding the limits. Theretofore, forward invariance of the set is ensured while handling unbounded functions are avoided [28]. Based on the definition of class $\mathscr{K}$ function in [55], extended class $\mathscr{K}$ function is defined as follows.

Definition 4.1. A continuous function $\alpha : (-b, a) \to (-\infty, \infty)$ with $a, b > 0$ is an extended class $\mathscr{K}$ function, if it is strictly increasing and $\alpha(0) = 0$ (Definition 1 in [34]).

■

Definition 4.2. ZCBF Properties.

For the control system (4.1) and a given set $\mathscr{M} \subseteq \mathscr{D} \subset \mathbb{R}^n$ defined as

$$\mathscr{M} = \{x | l(x) \geq 0\} \tag{4.6}$$

the C^1 function $l : \mathbb{R}^n \to \mathbb{R}$ is a ZCBF on the set $\mathscr{D}$, if there exists an extended class $\mathscr{K}$ function α such that

$$\sup_{u \in \mathscr{U}} [L_f l(x) + L_g l(x) u + \alpha(l(x))] \geq 0, \ \forall x \in \mathscr{D} \tag{4.7}$$

where L_f and L_g are Lie derivatives of $l(x)$ along f and g, respectively, and

$$\frac{dl}{dt} = \frac{\partial l}{\partial x} \dot{x} = L_f l(x) + L_g l(x) u$$

Then, the set of inputs that satisfy (4.7) is

$$\mathscr{V}_{zcbf} = \{u \in \mathscr{U} \, | [L_f l(x) + L_g l(x) u + \alpha(l(x))] \geq 0\}, \forall x \in \mathscr{D}$$

■

Lemma 4.1. *For the given set $\mathscr{M} \subseteq \mathscr{D} \subset \mathbb{R}^n$ with function l; if l is a ZCBF on $\mathscr{D}$, then, any Lipschitz continuous controller $u \in \mathscr{V}_{zcbf}$ for the system* (4.1) *renders the set $\mathscr{M}$ forward*

invariant.

Proof. See Proposition 1 in [28]. □

In safety-critical control systems, the safe set is presented by $\mathscr{M}$ with a safety criterion expressed by $l(x) \geq 0$ as (4.6). By starting from a safe initial condition $x_0 \in \mathscr{M}$ and selection of a control input that satisfies (4.7), the system never leaves $\mathscr{M}$ and thus guaranteeing safety. Despite the incredible power of ZCBFs in ensuring the safety of control systems, this method faces a couple of challenges. First of all, to ensure the satisfaction of (4.7), complete information about the dynamics of the system is needed. Second, the safety criteria and the safe set are assumed to be certain and known. However, in many real scenarios, the safe set is uncertain and affected by unknown external dynamics as described in (4.4). In the following section, the application of ZCBFs to guarantee safety under uncertain safety criteria in the presence of unknown external dynamics is investigated.

4.3 Learning-enabled ZCBF with Uncertain Sets

The system is considered to be operating in an environment that is shared with other agents. These external agents impose safety consideration on the system, while their dynamics are unknown and uncontrollable. This results in uncertainty in the environment and designing a safe controller. Therefore, in this section, the L-ZCBF platform is presented to ensure safety despite uncertainty in the behavior of external agents. The influential unknown dynamics of the external agents are learned, and consequently, an L-ZCBF is formed that assures the forward invariance of a set that is contained in the safe set, and its size becomes closer to the size of the actual safe set as the learning progresses.

4.3.1 Learning Safe Set Despite Uncertain Behaviors of External Agents

In order to design a safe controller for (4.1) in the presence of uncertain external agents in the environment, first, influential dynamics of external agents need to be approximated. Considering the Lipschitz continuity assumption on f_2 and the fact that any smooth function within a compact set can be approximated by an NN [66], (4.2) is approximated as

$$\dot{\hat{z}} = \hat{W}\Phi(\hat{z}) \tag{4.8}$$

where $\hat{W}$ is the estimated NN weights and Φ is its activation function. Then, considering (4.4), the approximated safe set is defined as formed by

$$\hat{\mathscr{C}} = \hat{\mathscr{C}}_1 \cap \hat{\mathscr{C}}_2 ... \cap \hat{\mathscr{C}}_q$$

where

$$\hat{\mathscr{C}}_i = \{x | l_i(x, \hat{z}) \geq 0\}, \ \ \forall 1 \leq i \leq q \tag{4.9}$$

where $\hat{z}$ is the state of the approximated external dynamics represented in (4.8). Fig. 4.1 shows an example with both the actual safe set and its approximated one for a specific time. As can be seen from Fig. 4.1(a), designing a controller based on the ZCBF (4.7) to ensure the forward invariance of this approximated set $\hat{\mathscr{C}}$ does not guarantee the forward invariance of the actual safe set and safety boundary might be violated. On the other hand, while the actual safe set can be formed based on the real-time measurement of the state of the external agent z, its forward invariance requires knowing the entire trajectory of the external agent, which is not available, and it is impossible to design (4.7) to make the actual safe set forward invariant. However, as shown in Fig. 4.1(b), if the control input is designed to assure the

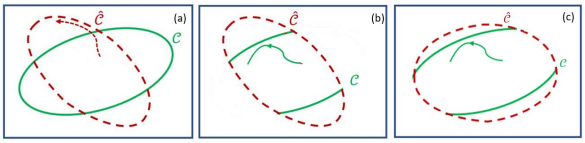

Figure 4.1: (a): $\hat{\mathscr{C}}$ invariant, $\partial\mathscr{C}$ violated (b): $\mathscr{C}_c$ invariant (c): $\hat{\mathscr{C}}$ converges to $\mathscr{C}$

forward invariance of the intersection of the actual safe set and its approximation, which is contained in the actual safe set, the safety of the system is guaranteed. This shrinks the boundary of the approximated safe set to assure that it is contained in the actual safe set. Note that this set can be made forward invariant using (4.7) since the approximate knowledge of the state trajectory of the external agent is available through (4.8). As learning progresses and the external dynamics becomes more accurate, as shown in Fig. 4.1(c), the approximated safe set becomes more accurate, and the system's maneuverability improves. The faster the external dynamics converges, the faster the intersection of the approximated safe set and actual safe set expands which provides more room of safe maneuver of the ego system.

In order to shrink the boundary of the approximated safe set and assure that it is contained in the actual safe set, the instantaneous sensory observations of the ego system from z are used to form the actual safe set, and the intersection of the safe set and its approximation is derived accordingly. $\mathscr{C}_c$ is defined as the intersection of $\mathscr{C}$ and $\hat{\mathscr{C}}$

$$\mathscr{C}_c = \hat{\mathscr{C}} \cap \mathscr{C} \tag{4.10}$$

Before presenting the proposed approach, the following assumptions are made.

Assumption 4.1. Strict interiority of the initial condition.

The initial condition of the system (4.1) belongs to the interior of the safe set $\mathscr{C}$. That is,

$$x_0 \in int\mathscr{C}$$

Assumption 4.2. The initial value of the approximated external dynamics $\hat{z}$ satisfies

$$l_i(x_0, \hat{z}_0) > 0, \ \forall 1 \leq i \leq q$$

Remark 4.3. Considering (4.3) and (4.9), Assumptions 4.1 and 4.2 imply that $\mathscr{C}_c = \hat{\mathscr{C}} \cap \mathscr{C}$ is non-empty.

Remark 4.4. Note that Assumptions 4.1 and 4.2 which state strict interiority of the initial condition and also its approximation, respectively, are mild and reasonable because if the initial condition is not safe, no controller can be designed to ensure safety in the future time.

Lemma 4.2. *Consider Assumptions 4.1, 4.2, and the set $\mathscr{C}_l$ defined as*

$$\mathscr{C}_l = \mathscr{C}_{l1} \cap \mathscr{C}_{l2} ... \cap \mathscr{C}_{lq}$$

where

$$\mathscr{C}_{li} = \{x | \min(l_i(x, z), l_i(x, \hat{z})) \geq 0\}, \ \forall 1 \leq i \leq q \tag{4.11}$$

Then, $\mathscr{C}_l = \mathscr{C}_c$ where $\mathscr{C}_c$ is defined in (4.10) *as the intersection of sets $\mathscr{C}$ and $\hat{\mathscr{C}}$.*

Proof. Given any $x \in \mathscr{X}$ and $1 \leq i \leq q$, if $x \in \mathscr{C}_l$, from (4.11), one has

$$l_i(x, z) \geq \min(l_i(x, z), l_i(x, \hat{z})) \geq 0 \Rightarrow x \in \mathscr{C}$$
$$l_i(x, \hat{z}) \geq \min(l_i(x, z), l_i(x, \hat{z})) \geq 0 \Rightarrow x \in \hat{\mathscr{C}}$$

Therefore,

$$\forall x \in \mathscr{C}_l \Rightarrow x \in \mathscr{C}_C$$

which means

$$\mathscr{C}_l \subset \mathscr{C}_c \tag{4.12}$$

On the other hands, if $x \in \mathscr{C}_c$

$$l_i(x, z) \geq 0, \;\; l_i(x, \hat{z}) \geq 0$$

and therefore,

$$\min(l_i(x, z), l_i(x, \hat{z})) \geq 0$$

which implies

$$\mathscr{C}_c \subset \mathscr{C}_l \tag{4.13}$$

From (4.12) and (4.13), one has

$$\mathscr{C}_c = \mathscr{C}_l$$

□

The boundary of $\mathscr{C}_c$ is defined as

$$\partial\mathscr{C}_c = \{x | \min(l_i(x, z), l_i(x, \hat{z})) = 0\}, \;\; \forall 1 \leq i \leq q$$

Definition 4.3. Given the control system (4.1), the smooth function $l = [l_1, ..., l_q] \in C^1$ is L-ZCBF for the set $\mathscr{C}_c$ if for each $1 \leq i \leq q$

$$\sup_{u \in \mathscr{U}} \{ \frac{\partial l_i}{\partial x} \dot{x} + \frac{\partial l_i}{\partial \hat{z}} \dot{\hat{z}} + \alpha(\min(l_i(x,z), l_i(x,\hat{z}))) \} \geq 0 \tag{4.14}$$

Moreover, the set of inputs that satisfy L-ZCBF condition is

$$\begin{aligned} &\mathscr{U}_{zcbf} = \{ u \in \mathscr{U} \mid \\ &\frac{\partial l_i}{\partial x} \dot{x} + \frac{\partial l_i}{\partial \hat{z}} \dot{\hat{z}} + \alpha(\min(l(x,z), l(x,\hat{z}))) \geq 0, \forall 1 \leq i \leq q \} \end{aligned} \tag{4.15}$$

where α is an extended class $\mathscr{K}$ funciton. ■

This definition is used to guarantee the safety of the system and forward invariance of $\mathscr{C}_c$ using tangent cone of practical sets and the Nagumo's theorem.

Definition 4.4. Practical Set (Definition 4.9 in [64])

Let $\mathscr{O}$ be an open set. Consider the set $\mathscr{S}_1 \subset \mathscr{O}$ defined by a set of inequalities in the form of

$$\mathscr{S}_1 = \{ x | l_i(x) \geq 0, i = 1, 2, ..., q \}$$

where l_i is continuously differentiable function in $\mathscr{O}$. The set $\mathscr{S}_1$ is said to be a practical set if

1) For all $x \in \mathscr{S}_1$, there exists y such that

$$l_i(x) + \nabla^T l_i(x) y > 0, \quad \forall i = 1, 2, ..., q \tag{4.16}$$

2) There exists a Lipschitz continuous vector field $\psi(x)$ such that for all $x \in \partial \mathscr{S}_1(x)$,

$$\nabla l_i(x) \psi(x) > 0 \tag{4.17}$$

■

For all $x \in \partial\mathscr{S}_1$, the tangent cone of the practical set is

$$\mathscr{T}_{\mathscr{S}_1}(x) = \{y | \nabla^T l_i(x) y \geq 0, \forall i \in \mathscr{S}_{1Act}(x)\} \tag{4.18}$$

where $\mathscr{S}_{1Act}(x)$ is the set of active constraints, which is defined as

$$\mathscr{S}_{1Act}(x) = \{x | l_i(x) = 0\}$$

For more details, see [64].

Theorem 4.2. *Given the control system* (4.1) *and the set* $\mathscr{C}_c$ (4.10)*, any Lipschitz controller* $u \in \mathscr{U}_{zcbf}$ *defined in* (4.15) *ensures safety criteria* $l_i(x, z) \geq 0$, $\forall 1 \leq i \leq q$.

Proof. For each $1 \leq i \leq q$, if $l_i(x, z) \geq l_i(x, \hat{z})$, the L-ZCBF condition (4.14) becomes

$$\frac{\partial l_i}{\partial x}\dot{x} + \frac{\partial l_i}{\partial \hat{z}}\dot{\hat{z}} + \alpha(l_i(x, \hat{z})) \geq 0$$

From direct result of Lemma 4.1, $l_i(x, \hat{z}) \geq 0$ and since $l_i(x, z) \geq l_i(x, \hat{z})$, then $l_i(x, z) \geq 0$. In addition, existence of $u \in \mathscr{U}_{zcbf}$ implies that for all $(x, \hat{z}) \in \mathscr{C}_c$

$$\begin{aligned}&\nabla^T l_i(x, \hat{z})[\dot{x}, \dot{\hat{z}}] + \alpha(l(x, \hat{z}))) \geq \\ &\nabla^T l_i(x, \hat{z})[\dot{x}, \dot{\hat{z}}] + \alpha(\min(l(x, \hat{z}), l(x, z)))) \geq 0\end{aligned}$$

Therefore, (4.16) is satisfied. If $l_i(x, z) < l_i(x, \hat{z})$, then L-ZCBF (4.14) turns to

$$\frac{\partial l_i}{\partial x}\dot{x} + \frac{\partial l_i}{\partial \hat{z}}\dot{\hat{z}} + \alpha(l_i(x, z)) \geq 0$$

Thus, at the boundary of $\mathscr{C}_c$ in which $l_i(x,z) \to 0$, one has

$$\frac{\partial l_i}{\partial x}\dot{x} + \frac{\partial l_i}{\partial \hat{z}}\dot{\hat{z}} \geq 0$$

In other words,

$$\nabla^T l_i(x)[\dot{x}, \dot{\hat{z}}] > 0$$

Therefore, (4.17) is satisfied for all $(x, \hat{z}) \in \partial\mathscr{C}_c$. Considering the definition of practical set and from (4.18), $[\dot{x}, \dot{\hat{z}}]$ is within the tangent cone of $\mathscr{C}_c$ (4.11) as

$$[\dot{x}, \dot{\hat{z}}] \in \mathscr{T}_{\mathscr{C}_c}(x,z)$$

This implies that if $l_i(x,z) \to 0$, then $(\dot{x}, \dot{\hat{z}})$ point inside the set at the boundary of $\mathscr{C}_c$ or in the worst case is tangent to the boundary. According to the Nagumo's Theorem, $\mathscr{C}_c$ is forward invariant and since $\mathscr{C}_c \subset \mathscr{C}$, therefore, the approximated trajectories do not exceed the boundary $l_i(x,z) = 0$ implying $l_i(x,z) \geq 0$ for all $t > 0$. Since this proof is valid for all $1 \leq i \leq q$, safety of the system is ensured. This completes the proof. □

Corollary 4.1. *Given the control system* (4.1)*, L-ZCBF introduced in* (4.14) *renders the intersection of the safe set and its approximation,* $\mathscr{C}_c$*, forward invariant.*

Proof. According to Theorem 4.2, boundary of the positive invariant set is shrunk to a more conservative value that provides a bigger margin to the safety boundary. According to Lemma 4.2, this forms the intersection of the safe set and its approximation. Therefore, the introduced L-ZCBF renders $\mathscr{C}_c$ invariant indeed.

□

Remark 4.5. The proposed L-ZCBF assures that at least a conservative safe set remains forward invariant, which guarantees safety. The conservativeness will be reduced next by presenting a fast and data-efficient learning approach for modeling the external agent.

Remark 4.6. It is shown that the external agent dynamics and, consequently, the unknown safe set are approximated using NNs. However, these approximations alone cannot be relied upon to ensure the forward invariance of the safe set. The reason is that the approximation might not be perfect and lead to exceeding the safety limits, which is not acceptable for safety-critical systems. Therefore, to design a more realistic and practical controller, the system observations and the approximated external agents dynamics are also combined with ZCBF.

Although the safety of the system can be guaranteed with an inaccurate model of the external dynamics, as learning enhances, the intersection set $\mathscr{C}_c$ expands to the exact safe set $\mathscr{C}$ and the system would be able to take less conservative control actions. In other words, employing a proper learning approach that suits the application boosts the control system's performance. In the following subsection, the application of the experience replay method is demonstrated in this problem to identify the dynamical behaviors of external agents. This method provides a fast convergence of the network leading to the control system's fast response, which is crucial in safety-critical applications.

4.3.2 External Dynamics Identifier

The motivation behind learning about the dynamics of external agents is to provide the ego system with a larger set of feasible actions and reduce the conservatism of the controller. In other words, enhancing the approximation of the safe set has higher importance compared to learning about the external agent states, and inspired by [67], an experience replay-based method is proposed which updates the identifier weights to reduce the set approximation error rather than the external state estimation error. Experience replay method uses recorded and stored data in the update law and provides fast convergence and an easy-to-check and verifiable the persistence of excitation (PE) condition, which is necessary to guarantee the convergence of the identifier weights. In contrast, online checking of this condition is generally difficult and even infeasible [67, 68, 69, 70].

Considering (4.2) and (4.8), the external dynamics model is formulated into a filtered regressor form

$$\dot{z} = W\Phi(z) + \epsilon_f \tag{4.19}$$

where W and Φ are the weight matrix and the activation function, respectively. Also, ϵ_f is the model approximation error.

To convert the dynamics into the regressor form, let Az be added to the both sides of (4.19), where $A = aI_{j\times j}$, $a > 0$

$$\dot{z} = -Az + W\Phi(z) + Az + \epsilon_f \tag{4.20}$$

Assumption 4.3. There exists a constant $0 < {\epsilon_f}^* < \infty$ such that

$$\|\epsilon_f(t)\| \leq {\epsilon_f}^* \tag{4.21}$$

Note that ${\epsilon_f}^*$ is unknown and depends on the quality of selected basis functions. If the basis functions are chosen such that the unknown function dynamics is near the span of the basis functions, this error will be small. Note also that the boundedness of reconstruction error and its gradient are standard assumptions in neural network identification literature. Furthermore, using neural networks, the approximation guarantees are limited to a compact set. Since for safety-critical systems, the safe set is generally compact, and the system must not leave this set, therefore, approximation over a compact set is reasonable (Chapter 1 in [66]).

Lemma 4.3. *Considering* (4.20)*, Eq.* (4.19) *can be written as*

$$
\begin{aligned}
z &= Wh(z) + ad(z) + \epsilon \\
\dot{h}(z) &= -ah(z) + \Phi(z),\ h(0) = 0 \\
\dot{d}(z) &= -Ad(z) + z,\ d(0) = 0
\end{aligned}
\tag{4.22}
$$

where

$$
\begin{aligned}
h(z) &= \int_0^t e^{-a(t-\tau)} \Phi(z(\tau)) d\tau \\
d(z) &= \int_0^t e^{-A(t-\tau)} z(\tau) d\tau \\
\epsilon(t) &= e^{-At} z(0) + \int_0^t e^{-A(t-\tau)} \epsilon_f d\tau
\end{aligned}
$$

Proof. See Lemma 1 in [67]. □

Consider identifying weight estimator as

$$
\hat{z}(t) = \hat{W}(t) h(z) + ad(z) \tag{4.23}
$$

where $\hat{W}(t)$ is the estimated value of the weight matrix W at time t. The state estimation error is defined as

$$
e_z(t) = \hat{z}(t) - z(t) \tag{4.24}
$$

By considering (4.22), (4.23) and (4.24), one has the state estimation error as

$$
e_z(t) = \hat{W}(t) h(z) + ad(z) - Wh(z) - ad(z) - \epsilon
$$

which is simplified to

$$e_z(t) = \tilde{W}(t)h(z(t)) - \epsilon$$

where $\tilde{W}(t) = \hat{W}(t) - W$ is the weight estimation error. The approximation of external agents dynamics is needed to expand the approximated safe set to the exact one, which reduces conservatism and provides more room of safe maneuver for the ego system. To accelerate the convergence of the approximated safe set and its approximation, weights are updated in a way to decrease set approximation error rather than the state estimation error. Set approximation error is defined as

$$e(t) = l(x, \hat{z}) - l(x, z)$$

By using the Taylor expansion around (x, z) and some manipulations, one has

$$e(t) = e_z(t)K(x, z) \tag{4.25}$$

with

$$K(x, z) = \\ \frac{\partial l(x, z)}{\partial z} + \frac{\partial^2 l(x, z)}{2!\partial z^2})e_z(t) + ... + \frac{\partial^{q_1+1} l(x, z)}{(q_1+1)!\partial z^{q_1+1}})e_z(t)^{q_1}$$

where q_1 is the maximum degree of z in $l(x, z)$. The derivatives with the order of higher than $q_1 + 1$ are zero and eliminated from the Taylor expansion.

In experience replay method, recorded samples are used in the update law. Define the state estimation error using the k^{th} sample as

$$e_z(t_k) = \hat{z}(t, t_k) - z(t_k) \tag{4.26}$$

where

$$\hat{z}(t, t_k) = \hat{W}(t)h(z(t_k)) + ad(z(t_k)) \tag{4.27}$$

Using (4.22) and (4.27), the error defined in (4.26) becomes

$$\begin{aligned} &e_z(t_k) = \\ &\hat{W}(t)h(z(t_k)) + ad(z(t_k)) - Wh(z(t_k)) - ad(z(t_k)) - \epsilon(t_k) \end{aligned}$$

which further is simplified to

$$e_z(t_k) = \tilde{W}(t)h(z(t_k)) - \epsilon(t_k) \tag{4.28}$$

and the set estimation error at the k^{th} sample is defined accordingly as

$$e(t_k) = e_z(t_k)K(x(t), z(t))$$

The update law is then given as

$$\begin{aligned} \dot{\hat{W}}(t) = &- \Gamma e(t)(h(z(t))K(x(t), z(t)))^T \\ &- \Gamma \sum_{k=1}^{P} e(t_k)(h(z(t_k))K(x(t), z(t)))^T \end{aligned} \tag{4.29}$$

where P is the overall number of stored data and Γ is a positive-definite matrix which determines the learning rate. Let the matrix of stored data be

$$Z = [h(z(t_1), ..., h(z(t_P)] \tag{4.30}$$

Then, the persistence of excitation condition is defined as

$$If \;\; h \in \mathbb{R}^{o_1}, then \;\; rank(Z) = o_1 \tag{4.31}$$

Remark 4.7. Using the history of data in the experience replay approach makes learning the safe set fast and data-efficient. This is of vital importance for safety-critical systems operating in an uncertain environment since the learning phase and operation phase in these systems are not separated. Therefore, control approaches with fast convergence capability in the learning process make control of safety-critical systems more practical.

Remark 4.8. Adaptive optimal control schemes require a PE condition to ensure the sufficient exploration of the state space. An exploratory signal consisting of sinusoids of varying frequencies can be added to the control input to ensure PE qualitatively. Note that the requirement of rank satisfaction is much less restrictive than the standard PE condition requirement and is much easier to verify online. The exploratory noise can be removed as soon as the rank condition is satisfied, which can be easily certified.

Theorem 4.3. *Consider the model* (4.19)*, the update law* (4.29) *and assume full rank of the matrix* Z *in* (4.30) *1) If there is no reconstruction error, i.e.,* $\epsilon_f = 0$*, then the set approximation error* (4.28) *converges to zero exponentially fast. 2) If* $\epsilon_f \neq 0$*, then the set estimation error is uniformly ultimately bounded (UUB), and the ultimate bound can be made small by recording rich data in the history stack.*

Proof. Let the Lyapunov function on weight error be as

$$V_W = 0.5tr(\tilde{W}\Gamma^{-1}\tilde{W})$$

By differentiating along the trajectory of (4.29) and considering the fact that $\dot{\hat{W}}(t) = \dot{\tilde{W}}(t)$,

one has

$$\dot{V}_W = -tr(\tilde{W}(t)[K^T(t)h^T(z(t))h(z(t))K(t)) + \sum_{k=1}^{P}(K^T(t)h^T(z(t_k))h(z(t_k))K(t)]\tilde{W}^T(t)) + tr([\epsilon(t)K^T(t)h^T(z(t)) + \sum_{k=1}^{P}\epsilon(t_k)K^T(t)h^T(z(t_k))]\tilde{W}^T(t) \quad (4.32)$$

where $K(t)$ stands for $K(x(t), z(t))$. Eq. (4.32) is simplified as

$$\dot{V}_W = -tr(\tilde{W}(t)K^T(t)[h^T(z(t))h(z(t))) + \sum_{k=1}^{P}(h^T(z(t_k))h(z(t_k))]K(t)\tilde{W}^T(t)) + tr([\epsilon(t)K^T(t)h^T(z(t)) + \sum_{k=1}^{P}\epsilon(t_k)K^T(t)h^T(z(t_k))]\tilde{W}^T(t) \quad (4.33)$$

If the rank condition on Z holds, then

$$\sum_{k=1}^{P}(h^T(z(t_k))h(z(t_k)) > 0$$

Therefore, for case of no reconstruction error, $\dot{V}_W < 0$ means that $\tilde{W}$ exponentially converges to 0. This completes the first part of the proof. For the second part and under reconstruction error, assume

$$B = K^T(t)[h^T(z(t))h(z(t))) + \sum_{k=1}^{P}(h^T(z(t_k))h(z(t_k))]K(t) \quad (4.34)$$

$$\epsilon_n = \epsilon(t)K^T(t)h^T(z(t)) + \sum_{k=1}^{P}\epsilon(t_k)K^T(t)h^T(z(t_k)) \quad (4.35)$$

Using (4.33), one has

$$if \ \left\|\tilde{W}\right\| \geq \frac{\|\epsilon_n\|}{\lambda_{min}(B)} \Rightarrow \dot{V}_w < 0 \tag{4.36}$$

where λ_{min} is the smallest eigen value of B defined in (4.34). Therefore, if $\epsilon_f = o$, $\tilde{W}$ converges to zero exponentially fast and thus $e(t)$ in (4.25) converges to zero exponentially fast. According to (4.19), (4.21), (4.22), and (4.35), one has

$$\|\epsilon_n\| \leq \frac{P+1}{a}({\epsilon_f}^*)$$

where by proper selection of a as identifier design parameter, (4.36) is satisfied. For any value of a, $\dot{V}_w$ is negative outside of the following compact set

$$\omega = \{\tilde{W} | \left\|\tilde{W}\right\| \leq \frac{P+1}{a\lambda_{min}(B)}({\epsilon_f}^*)\} \tag{4.37}$$

Based on (4.25), $e(t)$ will also remain bounded, and this completes the proof of the second part of the theorem. □

Remark 4.9. Note that the proposed set identification method provides exponential convergence of the set approximation error to zero. This implies that there are times $t_1, t_2, ...$ during learning that set approximation error $e(t_k) = l(x(t_k), \hat{z}(t_k)) - l(x(t_k), z(t_k))$ has decreased i.e., $e(t_{k+1}) < e(t_k)$. Considering the approximated safe set at this sequence $\{x(t_k)|l(x(t_k), \hat{z}(t_k)) \geq 0\}$ which is equivalent to the set $\{x(t_k)|l(x(t_k), z(t_k)) \geq -e(t_k)\}$ reveals that by decreasing the error, the approximated set gets closer to the exact safety set $\mathscr{C}$. Utilizing experience replay technique has at least two advantages: 1) it significantly improves the decay rate of the approximation error and thus reduces the conservatism, and 2) it provides the ego system with an easy-to-verifiable metric to check if the approximated safe set converges to the actual safe set.

Remark 4.10. Fast convergence of the model enables the control system to act in a less

conservative manner leading to enhanced performance. Even in the case of an inaccurate model with a non-zero reconstruction error, this method provides an acceptable performance with a UUB weight estimation error.

4.4 Control Framework

The proposed control framework is demonstrated in Fig. 4.2. First, the control system gathers data, e.g., distance from other agents in the environment collected by camera or LIDAR sensor, by observing its surrounding environment. The observed data are labeled as risky and safe, respectively. The safe data coming from the external agents that do not impose any risk on the control system are removed from the collected data. Then, the risky data representing external agents that can impose risk on the control system are applied to identifier blocks that approximate the dynamics of risky external agents using the modified experience replay method. Next, the state of the system and the output of identifier modules are injected into the CBF block to form L-ZCBF constraints according to the strict safety criterion. Finally, these L-ZCBF constraints govern the performance of the controller block, and control action must satisfy L-ZCBF constraints. The combination of identifier networks and the CBF block is called the guardian block.

The quadratic programming [27, 28] is employed to design the controller for this platform. The performance objective is formulated as a soft inequality constraint on derivative of the Lyapunov function. This constraint on the Lyapunov function and ZCBF constraint are unified by imposing them as constraints of quadratic programming problem, which aims to minimize a cost function. The cost function is a combination of the control input u and the relaxation factor η which is considered in the performance objective to make it a soft constraint. As a result, the minimum value of the control input, which satisfies safety is obtained and the system gets close to desired performance as much as possible. ZCBF-based quadratic programming is formulated as

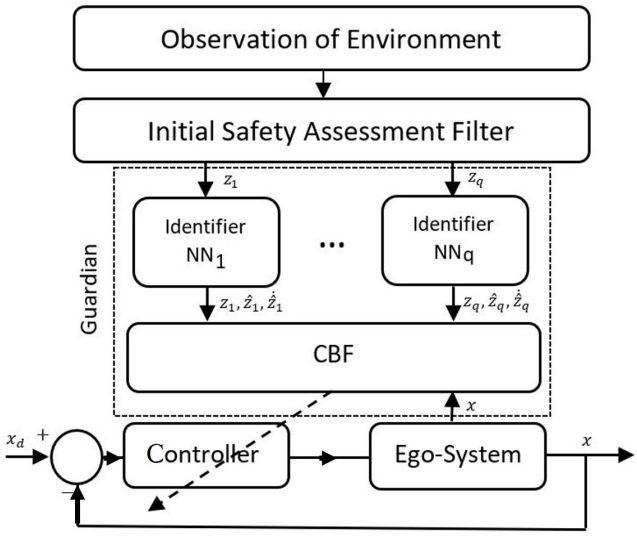

Figure 4.2: Control scheme

$$\begin{aligned} \min_{u,\eta} \quad & {u_t}^T H u_t + F u_t \\ \text{s.t.} \quad & (4.14),\ \dot{V} < \rho\eta. \end{aligned} \tag{4.38}$$

where $u_t = [u, \eta]$, H and F are the weight matrices, V is the Lyapunov function and ρ is the coefficient of the relaxation factor η.

Remark 4.11. If optimal controller u^* for performance objective is available, such as linear quadratic regulator (LQR) solution in a linear control system, then, the Lyapunov inequality in (4.38) can be replaced by equality $u = u^* + \rho\eta$.

The overall algorithm is given in Algorithm 3.

4.5 Case Study

The effectiveness of the proposed approach is verified here by designing a safe maneuver controller for an autonomous vehicle in the presence of other vehicles on the road.

Algorithm 3 Barrier-certified Learning-enabled Controller

Start with a safe initial condition $x_0 \in int\mathscr{C}$.
procedure
  procedure OBSERVATION
    Store states of previously observed agents $z_i, i \in 1, ..., n_i$ with n_i as the number of the previously observed agents (store null if agent vanished).
    Store states of new observed agents $z_{n_i+j}, j \in 1, ..., n_j$ with n_j as the number of the new observed agents .
    Go to "Initial Safety Assessment".
  end procedure
  procedure INITIAL SAFETY ASSESSMENT
    Check states of previously observed agents $z_i, i \in 1, ..., n_i$. Store them if they are still risky or null. Store null if they are safe.
    Check states of new observed agents $z_{n_i+j}, j \in 1, ..., n_j$. Store if they are risky. Discard if they are safe.
    Go to "Update".
  end procedure
  procedure UPDATE
    If stored data corresponds from previously observed agents, go to "Existing Agents".
    If stored data corresponds from new agents, go to "New Agents".
    procedure EXISTING AGENTS
      If stored data is null, discard corresponding identifier and L-ZCBF constraint.
      If stored data is not null, update weights using (4.29).
    end procedure
    procedure NEW AGENTS
      Initialize an identifier network $\hat{W}_{n_i+j_0}$ and safe initial approximation $\hat{z}_{n_i+j_0}$ for each new state $j \in 1, ..., n_j$.
      Form L-ZCBF constraint using (4.14) and incorporate it in quadratic programming (4.38).
    end procedure
    Go to "Quadratic Programming".
  end procedure
  procedure QUADRATIC PROGRAMMING
    Solve the quadratic programming problem (4.38).
    Apply the obtained controller to the system and update $n_i = n_i + n_j, n_j = 0$.
    Go to "Observation" and repeat until performance objectives are met.
  end procedure
end procedure

4.5.1 Control Scenario

Fig. 4.3 shows a safety-critical maneuver for autonomous vehicles in an urban area. The ego vehicle, specified by its position (x_e, y_e), is traveling in the road, and the control objective is to reach a pre-defined destination, which is marked in Fig. 4.3, in an optimal manner. However, the road is shared with other vehicles with uncertain behaviors, and their objectives might be in conflict with the ego vehicle desired objective. Vehicle (x_1, y_1) is traveling next to the ego vehicle; although it is very close, it looks safe. Vehicle (x_2, y_2) was not previously observable to the ego vehicle while it is now reaching the cross-section and might impose risk on the ego vehicle passing the crossroad. Vehicle (x_3, y_3) is farther but, it is moving in the same path as the ego vehicle and might impose risk on its maneuver in the future time. These types of maneuver scenarios are practically challenging but so common in everyday driving. This even becomes more challenging if instead of vehicles, bicycles or pedestrian are in the road which add more unpredictability and complexity to the control scenario. To elaborate on that, the effect of having an agent with a more complicated behavior instead of vehicle three is investigated as well. The following section mathematically formulates this scenario.

4.5.2 Mathematical Representation

The simple mass point model for vehicles is used [46].

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\psi} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \cdot \cos\psi \\ v \cdot \sin\psi \\ v \cdot u_s \\ u_a - \mu \cdot v \end{bmatrix}$$

where x, y are cartesian coordinate of the vehicle, v stands for vehicle's velocity and ψ is the heading angle of the vehicle. u_s is the steering angle and u_a is acceleration. μ is the friction

coefficient. For simplicity, the state vector is presented by $X = [x\ y\ \psi\ v]$. When moving in a straight line with zero friction coefficient, the dynamics of the ego vehicle and vehicles 1 and 3 are simplified as a double integrator

$$\begin{bmatrix} \dot{y}_i \\ \dot{v}_i \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_i \\ v_i \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{ai} \quad , i = e, 1, 3$$

and the dynamic of vehicle 2 is given as

$$\begin{bmatrix} \dot{x}_2 \\ \dot{v}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ v_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{a2}$$

Note that although the open-loop system is unstable, its controllability matrix is $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, which is full rank. Therefore, the system satisfies the stabilizability assumption, and there is a control input to make the closed-loop system stable. As explained in Algorithm 3, after gathering observation data, an initial safety assessment is required. In this scenario, using distance to assess safety is not functional since vehicle one is close to the ego vehicle, but it is safe. However, other vehicles are far from the ego vehicle, but they might impose risk on it. Therefore, the minimum distance of surrounding agents to the center of the road that the ego vehicle is moving along is considered for initial safety assessment and is named as the minimum safe lateral distance r_{min}. In this scenario, r_{min} is in x-coordinate defined as x_{min}, which is defined to be the lane width here and can be modified based on the application. As a result, if surrounding agents are within this range, they will be considered risky. Therefore, vehicle one is safe and will not be included in the loop as long as its lateral distance is in the safe range. However, other agents are considered risky, and headway safety criteria are

applied to the guardian block regarding them. Headway rule stated in [27, 71] is employed

$$D > v_e/2$$

where v_e is the ego vehicle speed, and D is the distance between two vehicles. Then, the safety criteria for vehicles 2 and 3 in this scenario would be

$$\begin{aligned} l_2(y_e, y_2) &= y_2 - y_e - v_e/2 > 0 \; when \; |x_2 - x_e| < x_{min} \\ l_3(y_e, y_3) &= y_3 - y_e - v_e/2 > 0 \end{aligned} \tag{4.39}$$

This formulation shows that if any vehicle gets very close to the lane that the ego vehicle is moving, then a minimum headway is required. Therefore, if the distance between the ego vehicle and any other vehicle gets shorter, the ego vehicle should decline its velocity to operate under these safety criteria.

The ego vehicle observes the identified external agents as black boxes in which only their current states are measurable. Thus, the identifier NNs for vehicles 2 and 3 are defined as

$$\begin{aligned} \dot{\hat{y}}_2 &= \hat{W}_2 \phi_3(y_2) \\ \dot{\hat{y}}_3 &= \hat{W}_3 \phi_3(y_3) \end{aligned}$$

The ego vehicle identifies the dynamics of vehicle 2 only in y coordinate because x is only needed for initial assessment, and it is not included in the headway criterion.

One of the advantages of the proposed approach is that safety is ensured even with inaccurate modeling of the external agents. Thus, to reduce computational cost and learning time, simple single layer perceptron with polynomial activation functions are employed as

$$\begin{aligned} \dot{\hat{y}}_2 &= \hat{W}_2 \cdot y_2 \\ \dot{\hat{y}}_3 &= \hat{W}_3 \cdot y_3 \end{aligned} \tag{4.40}$$

Remark 4.12. Since vehicle 2 is crossing the lane, the corresponding identifier is activated after it becomes observable for the ego vehicle. However, the corresponding L-ZCBF is formed and incorporated in quadratic programming when it reaches the lane that the ego vehicle is moving in. This setup can be adjusted based on the application. For example, one might decide to design a more conservative controller and incorporate the L-ZCBF at the time of observation.

L-ZCBFs are defined using (4.14) and (4.39) as

$$\frac{\partial l_2}{\partial y_e}\dot{y_e} + \frac{\partial l_2}{\partial v_e}\dot{v_e} + \frac{\partial l_2}{\partial \hat{y}_2}\dot{\hat{y}}_2 + \alpha_2(\min(l_2(y_e, y_2), l_2(y_e, \hat{y}_2))) \geq 0$$
$$\frac{\partial l_3}{\partial y_e}\dot{y_e} + \frac{\partial l_3}{\partial v_e}\dot{v_e} + \frac{\partial l_3}{\partial \hat{y}_3}\dot{\hat{y}}_3 + \alpha_3(\min(l_3(y_e, y_3), l_3(y_e, \hat{y}_3))) \geq 0$$

For the performance purposes, LQR problem is solved as mentioned in Remark 4.11. Then, the overall controller is performed using Algorithm 3. The values of parameters used in the simulation can be found in Table 4.1.

Table 4.1: Simulation Parameters

Parameter	Value
Q	$I_{2\times 2}$
R	1
α_2, α_3	15
H, F, ρ	$I_{2\times 2}, 0, 1$
a	0.7
$\hat{W}_{2_0}$	-0.2
$\hat{W}_{3_0}$	0.1

4.6 Simulation Results

Simulation is performed for the aforementioned control scenario in three sub-scenarios. First, an accurate network model with zero reconstruction error is employed which can

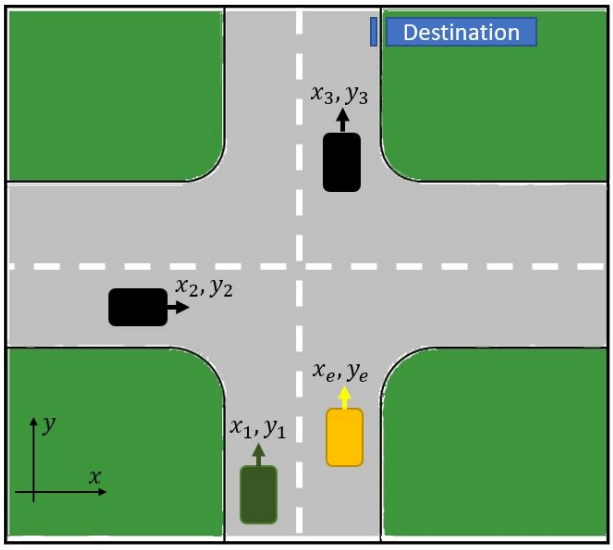

Figure 4.3: Control scenario

converge to the exact vehicle model. In the second sub-scenario, an inaccurate model is used, which cannot converge to the exact model. The third sub-scenario adds more complexity by considering an agent with a more complicated behavior in front of the ego vehicle, which the employed NN cannot accurately model. The purpose of using an inaccurate model is to demonstrate the strength of the proposed approach in guaranteeing safety in case of modeling error.

4.6.1 Zero Modeling Error Scenario

The network defined in (4.40) is assumed to be accurate without any reconstruction error, so after learning its weights, it converges to the exact model. Fig. 4.4 shows the results, where y as coordination of the ego vehicle and two risky vehicles 2, 3 are demonstrated. Without loss of generality, the destination of the ego vehicle is assumed to be located at the origin. The ego vehicle starts from its initial position, but a crossing vehicle is reaching, so the ego vehicle slows down and proceeds in a smooth maneuver when the crossing vehicle passes. After passing the crossroad, the ego vehicle faces another slow-moving vehicle in front of it; as a result, it slows down to adapt to the flow of traffic. As can be seen in Fig. 4.4, because

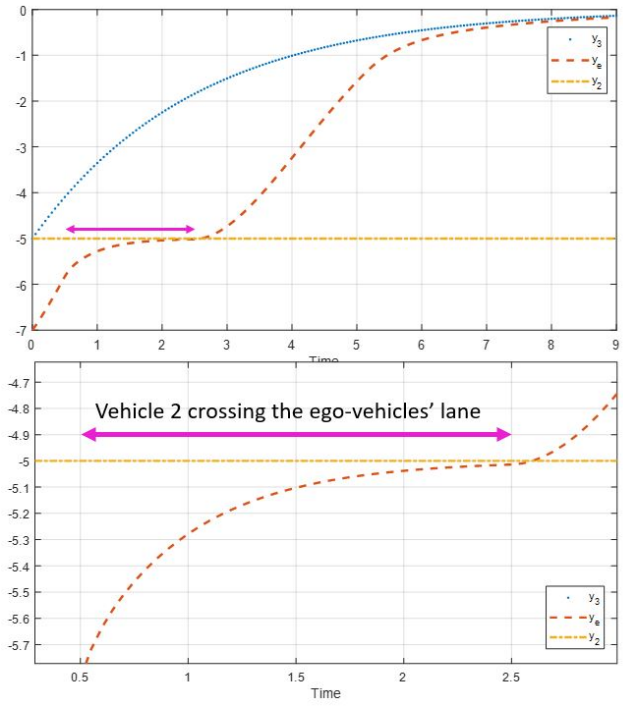

Figure 4.4: Position of vehicles in 'y' coordinate (Scenario1)

of the presence of vehicle 3, the ego vehicle could not reach the destination; but, it reached as close as possible while safety is still ensured. Fig. 4.5a shows the convergence of the weights of networks. The LQR performance of the system in lack of safety considerations is demonstrated in Fig. 4.5b. As can be seen in this figure, without safety consideration, the ego vehicle would crash with either vehicle 2 or vehicle 3. To further clarify the advantage of employing the proposed learning method, a simulation is conducted to compare the weight convergence with and without using the past stored data in the update law as depicted in Fig. 4.7. As seen in this figure, the network weight has a fast and exponential convergence under the proposed method.

4.6.2 First Non-zero Modeling Error Scenario

In this sub-scenario, the same network as (4.40) is employed while vehicle 3 demonstrates a different behavior as $\dot{y}_3 = a$. In other words, the allocated network is not a proper one

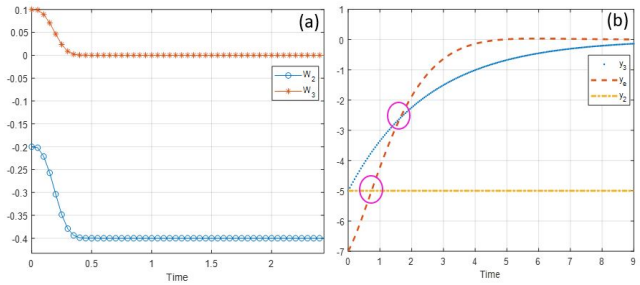

Figure 4.5: (a) NN weights (Scenario 1), (b) Optimal solution without CBF

for modeling the dynamics of vehicle 3. Fig. 4.6 shows the convergence of both networks' weights. As can be seen, W_3 could not converge. Fig. 4.8 shows the locations of the ego vehicle and vehicles 2 and 3 in this scenario. Despite inaccurate modeling, the safety of the system is ensured. The ego vehicle slows down to avoid crash with vehicle 2, and after that, it accelerates to reach the destination. However, it has faced vehicle 3 and has adjusted its velocity accordingly until it gets to the destination safely.

4.6.3 Second Non-zero Modeling Error Scenario

One of the big challenges of safe urban driving is unpredicted and hard-to-model dynamics such as the jump of an animal to the road or human behavior. The proposed method is functional in handling these unpredicted behaviors. To further analyze the result of employing this method, an agent with a more complicated dynamics is considered to be the only risky agent which is moving in front of the ego vehicle with dynamics of

$$\dot{y} = 0.4t + 1.2 - 0.7\sin(2\pi y) + 0.1y \tag{4.41}$$

The network (4.40) is employed for the identification of this dynamics, which has non-zero reconstruction error. The network weight update is shown in Fig. 4.9 which could not converge. The y-coordinate of both agents is shown in Fig. 4.10. As can be seen in this figure, despite the complexity in the behavior of the external agent and the existence of

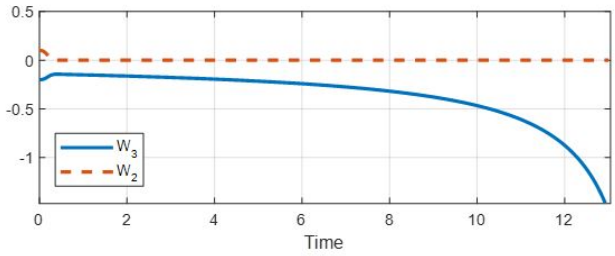

Figure 4.6: NN Weights (Scenario 2)

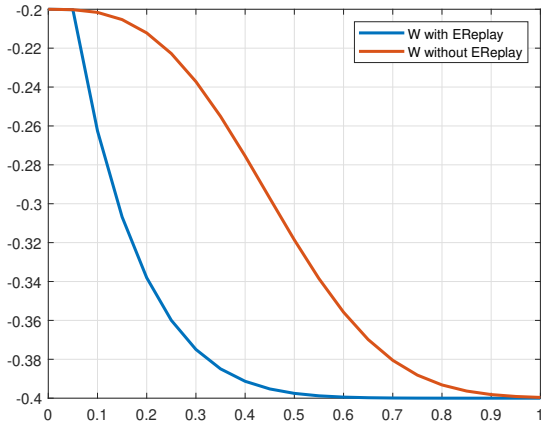

Figure 4.7: NN Weights with and without Experience replay

reconstruction error, the ego vehicle has a safe maneuver.

Remark 4.13. The purpose of this simulation is to demonstrate the capability of the method for guaranteeing safety in case of facing an agent whose dynamics cannot be modeled using pre-defined networks.

4.6.4 Discussion

The efficacy of the proposed method is examined in three different scenarios: 1) the assigned NN properly captures the dynamics of the external agent, but safety and performance are in conflict. It is shown that the agent has a safe maneuver during and after learning and gets close to its destination as far as it is safe. 2) there exists a reconstruction error, and the as-

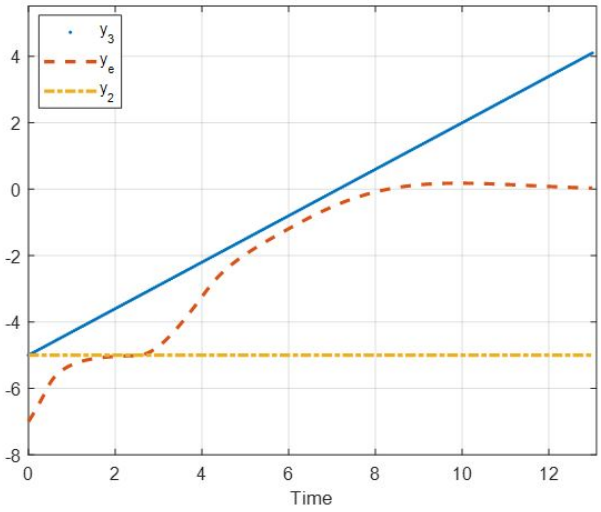

Figure 4.8: Position of vehicles in 'y' coordinate (Scenario2)

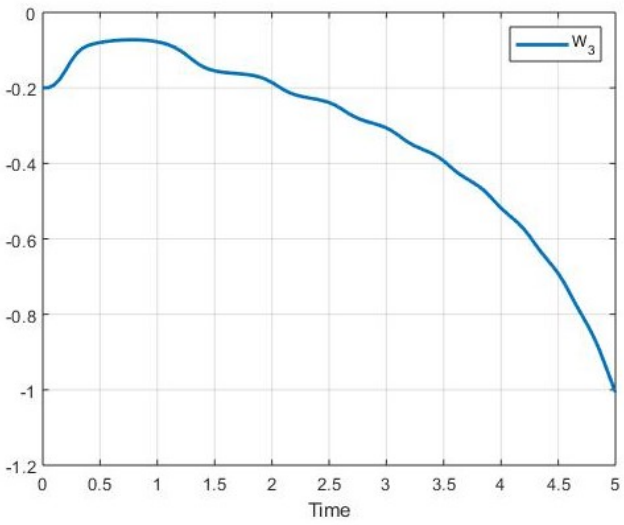

Figure 4.9: NN Weight (Scenario 3)

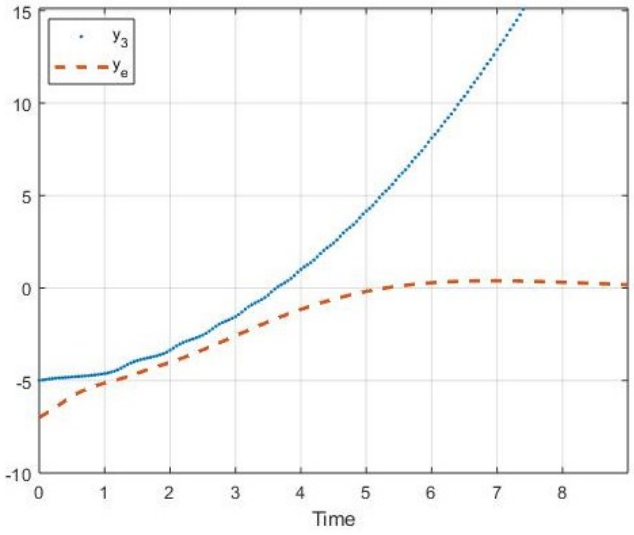

Figure 4.10: Position of agents in 'y' coordinate (Scenario 3)

signed NN cannot fully capture the dynamics of the external agents. It is shown that despite this error, the ego-vehicle still maintains a safe maneuver. This is of significant applicability since, in many real applications having an NN that fully captures the unknown dynamics is not always possible or tractable therefore, usage of a simplified model is facilitated. 3) A more complicated dynamical behavior in the presence of reconstruction error is considered in this scenario. It is shown that despite complex dynamics, still the safety of the ego-vehicle is ensured.

4.7 Conclusion and Future Work

In this chapter, a learning-enabled ZCBF controller for safety-critical systems under uncertainty has been proposed. It has been proved that the proposed method is capable of ensuring safety in complicated and uncertain environments in the presence of external agents with unknown dynamics. It has been also demonstrated that safety during learning and even with inaccurate modeling of external agents is guaranteed. As a result, this approach has provided a practical method in control scenarios that accurate modeling needs a great number

of data and computationally expensive learning schemes while still un-predicted objects are expected such as autonomous driving in an urban area. Meanwhile, having a better model has enabled the controller to take a less conservative action and has resulted in a better performance. To achieve this goal, a modified experience replay method has been proposed that identifies the external agents dynamic to minimize the difference between the safe set and its approximation. This method provides fast convergence and ensures a bounded error to the exact model even with inaccurate modeling which are both crucial in safety-critical control systems. Future work includes consideration of disturbance in the ego system's dynamics and extension to a robust framework. Furthermore, the reciprocal behavior of agents in the environment can be considered.

Chapter5

Robust Satisficing Cooperative Control Barrier Functions for Multi-Robots Systems using Information-Gap Theory

Contents of this chapter first appeared as [52] and have been reformatted to fit the requirements of this dissertation.

5.1 Introduction

Successful deployment of multi-robot and swarm systems demands safety guarantee of agents. While a centralized control approach can be used to design safe controllers for all agents in a swarm, the communication and computation complexity are expensive and do not scale up with the size of the swarm. Therefore, it is desired to prevent collision in a distributed manner using only local information exchange among agents, either explicitly through communications or implicitly through internal sensors. However, this local information is not certain and accurate due to imperfect communication, measurement errors, aging of sensors, weather conditions, and even failure of the sensing system. As a result, the failure of one agent in avoiding collision due to these uncertainties can lead to catastrophic failure of the whole system. This necessitates taking the uncertainty on the local information into account in any safe control design. In addition, it is desired to leverage the cooperative capability of the swarm system in ensuring safety via sharing responsibility in avoiding collision and

compensating for uncertainties as much as possible.

Common approaches for solving collision avoidance problems include conflict resolution [72], model predictive control (MPC) [73], potential field function[74], geometric guidance [75], and barrier function-based methods [33, 76, 77, 78]. Conflict resolution approaches, such as reachability-based methods rely on the availability of trajectories of other agents to find an obstacle-free route. MPC solves an optimization problem at every sample and accounts for state and input constraints. Collision avoidance criteria are also represented as the state constraint, and thus, MPC is employed to address this problem. In the potential field approach, each agent follows the gradient of potentials from which the target is attracting to and obstacles are repelling from. Geometric guidance methods such as collision cone [79] and velocity obstacle [75] result in less computational cost than the MPC and the conflict resolution approach. However, coping with sensors' errors which directly affects the safety criterion and thus the safety of the system and extension to decentralized safe frameworks which only rely on local information rather than global functions, remain as challenges [77].

Control barrier function (CBF)-based methods prevent collision between agents by ensuring the forward invariance of a safe set. This provides safety without neither the need for computing a safe reachable set in reachability-based methods nor solving an online nonlinear optimization for every instance of time in MPC-based methods, and, thus, is more computationally tractable. In addition, it can be applied to the control loop in a minimally invasive manner, in the sense that a nominal (e.g., optimal) controller can be modified as little as possible to ensure safety. Therefore, it is a versatile method that can be integrated into a variety of control approaches [27, 50, 80]. This is superior to the collision avoidance methods, which employ a secondary controller in the face of collision risk. This is because switching between primary and secondary controllers can delay or even prevent safe task completion, especially in a dense environment. In [34], the robustness of CBFs under model perturbation is investigated and asymptotic stability of the safe set is established. In [81], a distributed CBF-based approach is presented for multi-agent systems. This work is then extended in

[33] and [76] to heterogeneous swarm systems in which the maximum acceleration of agents is not equal, and the CBF is shared between agents based on their maximum acceleration. In [82], each agent is modeled as an ellipsoid and a distributed CBF is used to ensure the safety of a swarm system while disturbance and parametric uncertainty in gravity term of Lagrangian dynamics is considered and estimated. However, in a distributed CBF approach, which relies on the local information, measurement uncertainty (i.e., inaccuracy in the relative distance to other agents measured by the ego-agent) might jeopardize the safety of the overall system, which is not considered in the existing distributed CBF methods. Therefore, it is desired to investigate the effect of measurement uncertainty on the safe performance of swarm systems. Considering a worst-case uncertainty limits the action of agents and results in an overly conservative controller. Especially in a situation that error bound might be substantial, unknown, and time-varying, the worst-case method might not be feasible at all.

In addition, when the measurement confidence of agents is not uniform, higher uncertainty in one agent's measurement might lead to catastrophic failure of the overall system; therefore, agents need to cooperatively decide on their roles in ensuring pairwise safety. In addition, sources of measurement uncertainty are not always known a priori. For example, different weather and lighting conditions can alter the accuracy of the reading or make it unreliable. Therefore, to tackle this issue, rather than considering the probabilistic model or bound of the measurement error, we propose to use the cooperative capability of agents to maximize the horizon of uncertainty under which the safety of the overall system is ensured. When agents have different measurement confidences, the agents with higher certainty decide to take more responsibility for ensuring the safety of the overall system. Even in an extreme case that one agent fails in sensing, other agents can compensate by taking responsibility for the safety guarantee. This is achieved by sharing CBFs using information-gap (IG) theory to maximize the safe horizon of uncertainty between every two agents. IG theory is a decision-making tool for prioritizing alternatives when neither the probabilistic distribution of uncertainty nor its worst case is available. The uncertainty in IG theory rather is

modeled with an ambiguity set of possible outcomes with an unknown bound or uncertainty horizon. The robust satisficing IG method takes action that results in the highest horizon of uncertainty up to which a critical requirement is satisfied. IG theory is employed in various engineering problems [83, 84, 85, 86]. However, its application is less investigated in conjunction with control theory.

In this chapter, a safe and robust satisficing control protocol is proposed for the multi-agent collision avoidance problem in the presence of measurement uncertainty. In the proposed approach, it is assumed that agents are unaware of their neighboring agents' trajectories, and only uncertain local measurement information about neighbors is available (e.g., through embedded sensors) in the sense that neither a probabilistic model of the uncertainty nor its worst case is known. To maximize the robustness of the system to measurement uncertainty and satisfy the safety of the overall system, the IG theory along with the CBF approach is employed to determine the contribution of each agent in constructing shared CBFs between agents. It is shown that based on the IG theory, agents with more certain measurements must take more responsibility to ensure pairwise safety, and consequently, the overall safety of the multi-agent system, as they are allowed to have more agile behaviors and have more influence on the overall agility of the system. It is also shown that using the proposed approach, agents with more accurate measurements can cooperatively compensate for agents with less accurate measurements without sacrificing performance. In a nutshell, the contributions of the chapter are as follows.

1. Accounting for the measurement uncertainty in a distributed multi-agent barrier certified control framework.

2. Employing the cooperative capability of agents for safety guarantee and compensating for each other's measurement inaccuracy.

3. Presenting a robust satisficing approach to maximize acceptable horizon of uncertainty using information gap theory which enables a non-conservative robust design.

5.1.1 Organization of the Chapter

Section II is allocated for problem overview and background information on CBFs and IG theory. The problem statement and the proposed framework of robust-satisficing distributed safe control are presented in Section III. Section IV represents the simulation results and Section V concludes the chapter.

5.2 Problem Overview and Background

The problem overview and motivation are first presented in this section. Then, background on CBFs and IG theory is provided.

5.2.1 Problem Overview

The controller for safety-critical multi-agent systems that share an environment must be carefully designed to not only achieve their tasks but also satisfy coupled safety constraints. Satisfying these coupled safety constraints under uncertainty (e.g., measurement uncertainty) is challenging and without the agent's collaboration (e.g., shared collision avoidance strategy), might result in conservative control design and even infeasibility. To achieve this cooperation in ensuring safety, CBFs and IG theory are integrated.

CBFs are employed to ensure forward invariance of the safe set. However, since the trajectory information of involving states is needed in each pairwise CBF (i.e., dynamics of each two agents, which collision should be avoided between them). Therefore, the pairwise CBF is broken into two distributed CBFs in which each of them only relies on each agent's trajectory and local information. Satisfaction of distributed CBFs for each two agents in the vicinity of each other results in collision avoidance between them. However, measurement uncertainty leads to inaccuracy of distributed CBFs and, therefore collision. Considering the worst-case uncertainty severely limits the action of agents and leads to the conservatism

of the controller. In addition, the worst-case uncertainty is not always known. Therefore, as the main contribution of this chapter, another approach to tackle this problem is employed. Share of each agent in ensuring safety is determined in a cooperative manner using IG theory. In other words, CBF is shared between agents such that the system tolerates the highest horizon of uncertainty, while usage of uncertainty information is avoided in the formation of CBFs, resulting in agile and non-conservative controllers. It is shown that agents with more accurate measurements are able to compensate inaccuracy of other agents by taking more responsibility in ensuring safety. The formulation of this method unfolds in subsequent sections.

5.2.2 Background

In this section, the background on CBFs and IG theory is briefly provided. An affine nonlinear system is considered as

$$\dot{x} = f(x) + g(x)u \tag{5.1}$$

where $x \in \mathbb{R}^n$ and $u \in \mathscr{U} \subset \mathbb{R}^m$ are the state of the system and the control input, respectively. It is assumed that f and g are locally Lipschitz and the equilibrium point of the system is stabilizable.

5.2.2.1 Control Barrier Functions

CBFs are used to guarantee the forward invariance of a predefined set in a control system. Zeroing CBF (ZCBF), as one major form of CBFs, is positive within a set of interest and reaches zero at the set's boundary. Imposing a proper condition on its derivative prevents the system trajectories from passing the boundary of the set of interest, which guarantees its forward invariance.

The safety criterion is represented as $h(x) \geq 0$, where $h(x) : \mathbb{R}^n \to \mathbb{R}$ is a smooth

function. The safe set is defined accordingly as

$$\mathscr{C} = \{x \in \mathbb{R}^n | h(x) \geq 0\} \tag{5.2}$$

Definition 5.1. [55, 34] A continuous function $\beta : (-b, a) \to (-\infty, \infty)$ with $a, b > 0$ is an extended class $\mathscr{K}$ function, if it is strictly increasing and $\beta(0) = 0$. ■

Definition 5.2. [28] Considering the dynamical system (5.1) and the set $\mathscr{C} \subset \mathbb{R}^n$ (5.2) defined using a $h(x) \in C^1$ function, if there exists a locally Lipschitz extended class $\mathscr{K}$ function β such that

$$\sup_{u \in \mathscr{U}} [L_f h(x) + L_g h(x) u + \beta(h(x))] \geq 0, \quad \forall x \in \mathscr{D} \tag{5.3}$$

then, the function $h(x)$ is a ZCBF on domain of interest $\mathscr{D}$ with $\mathscr{C} \subseteq \mathscr{D} \subset \mathbb{R}^n$. ■

The set of feasible control inputs for $h(x)$ is formed accordingly as

$$\mathscr{U}_m(x) = \{u \in \mathscr{U} | L_f h(x) + L_g h(x) u + \beta(h(x)) \geq 0\}$$

Ensuring the forward invarinace of a set using ZCBFs is the result of the following theorem.

Theorem 5.1. *[28]. Given dynamical system* (5.1) *and the set* $\mathscr{C} \subseteq \mathscr{D}$ (5.2) *defined for* $h(x) \in C^1$*, if* h *is a ZCBF on* $\mathscr{D}$*, any Lipschitz continuous controller* $\{u : \mathscr{D} \to \mathbb{R} | u \in \mathscr{U}_m(x)\}$ *renders the set* $\mathscr{C}$ *forward invariant.*

Remark 5.1. Note that based on the definition of the safe set (5.2) and the ZCBF criterion defined in (5.3), if the initial state of the system (5.1) is inside the safe set, i.e., the initial condition $x(0)$ satisfies $h(x(0)) > 0$, then even if $h(x)$ decreases and the system's trajectory gets close to the safety boundary, since, the derivative of $h(x(t))$ is positive in the boundary, i.e., $L_f h(x) + L_g h(x) u \geq 0$, the value of $h(x)$ starts increasing. This pushes back the systems'

trajectories inside the safe set for which $h(x(t)) > 0$. Furthermore, $\beta(h(x))$ determines how fast the states of the system can reach the safety boundary.

5.2.2.2 Information-Gap Theory

Uncertainty is typically modeled by either probability distribution or its worst case. However, in a scenario in which the system changes over time and future variations in the condition is poorly known in advance, a poor or overly-conservative controller might be induced. In such scenarios, IG theory can be employed to drastically improve robustness and performance of the system. Robust satisficing IG theory is a non-probabilistic decision making method that prioritizes the choices to maximize robustness against uncertainty. First, an ambiguity set is leveraged to model the uncertainty. Based on the application and the available knowledge about the uncertain entity, different IG models can be employed such as hybrid IG model, slope bound model, and Fourier bound model [83, 84]. Note that the horizon of uncertainty in the ambiguity set is unknown.

To clarify this, assume that the parameter K is the entity of interest while limited knowledge about it is available and let $\hat{K}$ be a rough estimation of K, while the exact value of deviation from the true value is unknown. The following fractional IG model can be used as an ambiguity set

$$U(s,t) = \{K \big| |\frac{K - \hat{K}}{\phi(t)}| < s\} \tag{5.4}$$

where the parameter s is the horizon of uncertainty which is unknown. The true value of K can deviate from the available estimation by at most $\pm s|\phi(t)|$. Note that $\frac{1}{\phi(t)}$ is considered as the measurement confidence which is a rough measure on validity of the sensed or measured data. The sensed data with higher measurement confidence is more reliable. Depending on the application, there are different methods in the literature to quantify measurement confidence such as sensor fusion-based methods [87, 88]. In the simplest form, with no external

processing method to extract the reliability of measurement, measurement confidence is the measurement accuracy of the instrument in which the manufacturer has provided. Measurement confidence depends on the situation, the operating environment, and the accuracy of the sensing equipment of agents. For example, a change in the weather condition and the discrepancy between different sensors' measurements of an agent result in less confident measurements, and, therefore, the agent demands recommunication. An illustrative example is provided in the subsequent section to clarify the basis for calculation and change of the measurement confidence.

Note that IG approach is entirely different from the standard practice in robust control for which the uncertainty bounds are known and the goal is to design a controller that satisfies the system's requirement within known uncertainty boundaries. Instead, the goal of IG is to maximize the uncertainty horizon under which the system achieves its requirement. All IG models including fractional-error model (5.4) have two properties of contraction and nesting. Contraction property states that $U(s)$ is a singleton when $s = 0$, while nesting property means that increment in s makes $U(s)$ more inclusive and

$$s_2 < s_1 \implies U(s_2) \subset U(s_1)$$

This reveals the importance of our desire to maximize the horizon of uncertainty. When horizon of uncertainty gets bigger, the ambiguity set (5.4) expands and thus more uncertainty will be tolerated.

5.3 Robust-Satisficing Control Barrier Function

This section presents the problem of collision-free and safe control of multi-agent systems using CBF and IG theory in the presence of uncertainty in the agent's local measurement information. To cope with uncertainty, a robust satisficing approach is proposed to determine the share of each agent in the CBF between each two agents to tolerate the highest horizon

of uncertainty under which the pairwise safety is still ensured.

Considering the system model (5.1) and the uncertainty model (5.4) where the uncertain entity $h(x)$ is the relative distance between each pair of agents,

$$U(x,s,t) = \{h(x)\big|\big|\frac{h(x) - \hat{h}(x)}{\phi(t)}| < s\}$$

by determining the minimum performance requirement that must be satisfied, the IG safety robustness is defined as

$$s^*(x_c) = \max_s\{s|(\min_{h(x)\in U(s,t)} h(x)) \geq x_c\} \tag{5.5}$$

where $h(x)$ is used to represent safety requirement and x_c is a critical value (which is 0 here) from which $h(x)$ should not exceed. Eq. (5.5) implies that our goal is to maximize the horizon of uncertainty s^* which in the worst case, the system requirement $h \geq 0$ is satisfied.

Illustrative Example:

As mentioned earlier, the quantification of measurement confidence depends on the system's application and sensing capability. Thus, a variety of methods, such as sensor fusion approaches, can be employed to examine the reliability of the sensed data. Here we provide a simple example to clarify the concept further.

Assume the system is equipped with two different distance sensors, a laser scanner with the measured value of $d_{li}, i = 1, ..., N$, where N is the number of the nearby objects; and a radar with the measured distance to the nearby objects of $d_{ri}, i = 1, ..., N$. To determine the reliability of the sensed data d_{li}, one might examine the norm of the discrepancy between these two readings as

$$\phi_i(t) = (||d_{li} - d_{ri}||) + e_{li}$$

where e_{li} is the nominal measurement error of the laser scanner. As the discrepancy between two measurements increases, the measurement confidence $\frac{1}{\phi_i}$ decreases. On the other hand, if this discrepancy is negligible, then the measurement confidence simplifies to $\frac{1}{e_{li}}$. To have a discrete index, one might define this function as

$$\phi_i(t) = f(||d_{li} - d_{ri}||) + e_{li}$$

where

$$f(x) = \begin{cases} a_1, \ if \ 0 < x < b_1 \\ \vdots \\ a_i, \ if \ b_{i-1} < x < b_i \end{cases}$$

It will be shown later that agents need to recommunicate in case of a change in measurement confidence. Therefore, by having a discrete index they only need to re-communicate if a critical value of discrepancy is passed resulting in a reduced communication cost.

5.3.1 Problem Formulation

Consider a swarm system with N agents and index set of $\mathscr{M} = \{1, ..., N\}$. Each agent is modeled as a single integrator

$$\dot{\mathbf{p}}_i(t) = \mathbf{u}_i(t), \ \forall i \in \mathscr{M} \tag{5.6}$$

For simplicity, in the rest of paper, $\mathbf{p}_i(t)$ and $\mathbf{u}_i(t)$ are written as $\mathbf{p}_i$ and $\mathbf{u}_i$, respectively. $\mathbf{p}_i \in \mathbb{R}^2$ is the position vector in the Cartesian space and $\mathbf{u}_i \in \mathbb{R}^2$ is the velocity of the agent which is considered as the control input. It is desired to satisfy the following pairwise safety

criterion for collision avoidance between each two agents i and j

$$\|\Delta \mathbf{p}_{ij}\| \geq D_s, \ \ \forall i, j \in \mathscr{M}, \, i \neq j$$

where $\Delta \mathbf{p}_{ij} = \mathbf{p}_i - \mathbf{p}_j$ is the relative position between agents i and j and D_s is the minimum safe distance. The pairwise safety criterion h_{ij} between each two agents i and j is then defined as

$$h_{ij} = \|\Delta \mathbf{p}_{ij}\| - D_s \geq 0, \ \ \forall i, j \in \mathscr{M}, \, i \neq j \tag{5.7}$$

which specifies that the pairwise distances between agents should be kept above a critical value D_s, resulting in collision avoidance. According to (5.7), the following pairwise safety sets are formed

$$\mathscr{C}_{ij} = \{(\mathbf{p}_i, \mathbf{p}_j) | h_{ij} \geq 0\}, \ \ \forall i, j \in \mathscr{M}, \, i \neq j \tag{5.8}$$

The pairwise ZCBF constraint which ensures forward invariance of (5.8) and consequently, their pairwise safety based on Theorem 5.1 and using (5.3) with taking $\beta(h_{ij}) = \alpha_{ij} h_{ij}$ is

$$\frac{\Delta \mathbf{p}_{ij}^T}{||\Delta \mathbf{p}_{ij}||} \Delta \mathbf{u}_{ij} + \alpha_{ij}(||\Delta \mathbf{p}_{ij}|| - D_s) \geq 0, \\ \forall i, j \in \mathscr{M}, \, i \neq j \tag{5.9}$$

where $\alpha_{ij} > 0$ is a design parameter which determines how fast the trajectories of the system can approach the safety boundary and $\Delta \mathbf{u}_{ij} = \mathbf{u}_i - \mathbf{u}_j$. The overall safety set is formed using (5.8) as follows [33]

$$\mathscr{C} = \prod_{i \in \mathscr{M}} \{ \bigcap_{j \in \mathscr{M}, j \neq i} \mathscr{C}_{ij} \} \tag{5.10}$$

This implies that in order to have a collision-free maneuver for the overall system, the collision should be avoided between each two agents. This result is presented in the following theorem.

Lemma 5.1. *[33]. The multi-agent system represented by $\mathscr{M}$ is safe and $\mathscr{C}$ in* (5.10) *is forward invariant, if the control input $\boldsymbol{u} = [{\boldsymbol{u}_1}^T, ..., {\boldsymbol{u}_N}^T]^T$ satisfies all pairwise ZCBF constraints* (5.9).

As can be seen in (5.9), the information about trajectories of both agents i and j is needed in ZCBF inequality constraint. However, in reality, the exact trajectories of agents are not available due to measurement inaccuracy or communication noise and sensor or communication failure. As a result, alternative ZCBFs should be employed for which each agent takes responsibility on ensuring safety based on its own trajectory information and local uncertain measurements about the position of other agents.

5.3.2 Robust-satisficing Distributed CBF

In this subsection, the effect of measurement uncertainty in certifying the safety of the system is investigated. Afterwards, the idea of distributing ZCBF constraints in order to achieve highest robustness considering the measurement uncertainty is presented.

5.3.2.1 Distributed ZCBF

It is desired to guarantee safety of the overall system in a distributed network using only local information. By considering the fact that CBF provides safe and admissible set of inputs which can be divided into safe and admissible subsets, [33] and [76] propose to distribute the pairwise ZCBF constraint (5.9) between agents i and j as follows

$$ZCBF_i: \quad \frac{\Delta \mathbf{p}_{ij}^T}{||\Delta \mathbf{p}_{ij}||}\mathbf{u}_i + \alpha_i \cdot (||\Delta \mathbf{p}_{ij}|| - D_s) \geq 0 \tag{5.11}$$

$$ZCBF_j: \quad \frac{-\Delta \mathbf{p}_{ij}^T}{||\Delta \mathbf{p}_{ij}||}\mathbf{u}_j + \alpha_j \cdot (||\Delta \mathbf{p}_{ij}|| - D_s) \geq 0 \tag{5.12}$$

where

$$\alpha_i + \alpha_j = \alpha_{ij} \tag{5.13}$$

with α_{ij} as a design parameter set in advance to achieve a specific performance while safety is ensured. $ZCBF_i$ and $ZCBF_j$ are ZCBF constraints that agent i and agent j follow based on their local information. Thus, if each agent's controller $\mathbf{u}_i$ and $\mathbf{u}_j$ are designed such that (5.11) and (5.12) hold, then their summation, which is the pairwise ZCBF constraint (5.9), is satisfied as well. In this formulation, each agent only needs its own trajectory and local measurement information about the relative distance to surrounding agents to satisfy the corresponding ZCBF constraint; therefore, a distributed implementation is feasible. Parameters α_i and α_j specify how the pairwise ZCBF constraint (5.9) is shared between agents. The greater α_i and therefore the faster $\alpha_i \cdot (||\Delta\mathbf{p}_{ij}|| - D_s)$ gets close to zero, the faster the derivative terms become positive, which pushes harder and faster the trajectory of the system back into the safe set. In other words, the agent with greater allocated α_i is allowed to have a more agile maneuver. Since measurement uncertainty is inevitable and must be considered when designing safe controllers, it is desired to determine a method for allocation of these parameters to achieve the best possible robustness against measurement uncertainty. This will be covered in the following subsection.

5.3.2.2 Robust-satisficing distributed ZCBF

In a distributed safe control framework, each agent relies on its own local measurement information. However, measurement uncertainty and accuracy reduction due to aging of sensors or uncertainty due to imperfect communication can affect the safety of the overall system. Therefore, it is important to model and incorporate uncertainty in control design. In this chapter, IG theory is employed to address the question of how to design distributed ZCBFs which are capable of tolerating the highest horizon of uncertainty while avoiding

collision.

It is assumed that agent i is capable of measuring instantaneous relative position of surrounding agents. However, the local information of agent i about this relative position is uncertain

$$\Delta\hat{\mathbf{p}}_i = \Delta\mathbf{p}_{ij} + \mathbf{e}_i(t) \tag{5.14}$$

where $\Delta\hat{\mathbf{p}}_i$ is a rough estimation of agent i from $\Delta\mathbf{p}_{ij}$ and the measurement error is denoted by $\mathbf{e}_i(t)$ which is unknown to the agent i. Therefore, an ambiguity set is employed instead, to model measurement uncertainty

$$U(s_i, t) = \{\Delta\mathbf{p}_{ij} | ||\frac{\Delta\hat{\mathbf{p}}_i - \Delta\mathbf{p}_{ij}}{\phi_i(t)}|| \leq s_i\} \tag{5.15}$$

where s_i is the horizon of measurement uncertainty of agent i, and $\frac{1}{\phi_i(t)}$ indicates the confidence of agent i from its measurement. Note that $\Delta\mathbf{p}_{ij}, \Delta\hat{\mathbf{p}}_i$ are vectors, and therefore, radial and angular uncertainties are reflected in (5.15).

Assumption 5.1. The measurement error $\mathbf{e}_i(t)$ is bounded as

$$||\frac{\mathbf{e}_i(t)}{\phi_i(t)}|| \leq s_i$$

Note that in deeply uncertain scenarios, the exact value of error $\mathbf{e}_i(t)$, is unknown.

In contrast with robust framework in which a known worst-case horizon of uncertainty is respected, in here the goal is to make decisions that maximize the unknown horizon of uncertainty and then the highest possible worst case is derived based on made robust satisficing decisions.

Due to uncertainty, agents don't have access to the exact value of $||\Delta\mathbf{p}_{ij}||$, instead they have an uncertain measurement of it denoted by $||\Delta\hat{\mathbf{p}}_i||, ||\Delta\hat{\mathbf{p}}_j||$ and thus each agent perceives

safety criterion (5.7) differently as

$$h_i = ||\Delta\hat{\mathbf{p}}_i|| - D_s$$
$$h_j = ||\Delta\hat{\mathbf{p}}_j|| - D_s \tag{5.16}$$

where h_i and h_j are the perception of agents i and j from safety criterion h_{ij}, respectively. Note that in case of no measurement uncertainty $h_j = h_i = h_{ij} = ||\Delta\mathbf{p}_{ij}|| - D_s$; however, in the presence of measurement uncertainty, exact value of h_{ij} is not available to agents, and this uncertainty is reflected in the perception of agents from safety criterion as (5.16).

The distributed ZCBF constraints with uncertain measurements become

$$\overline{ZCBF}_i: \quad \frac{\Delta\hat{\mathbf{p}}_i^T}{||\Delta\hat{\mathbf{p}}_i||}\mathbf{u}_i + \alpha_i \cdot (||\Delta\hat{\mathbf{p}}_i|| - D_s) \geq 0 \tag{5.17}$$

$$\overline{ZCBF}_j: \quad \frac{-\Delta\hat{\mathbf{p}}_j^T}{||\Delta\hat{\mathbf{p}}_j||}\mathbf{u}_j + \alpha_j \cdot (||\Delta\hat{\mathbf{p}}_j|| - D_s) \geq 0 \tag{5.18}$$

where $\overline{ZCBF}_i$ and $\overline{ZCBF}_j$ are uncertain interpretation of $ZCBF_i$ and $ZCBF_j$, respectively. The robust design of (5.17) and (5.18) to guarantee the safety criterion (5.7) is the result of the following problem.

Problem 5.1. Consider the multi-agent system (5.6) and define the measurements ambiguity sets $U(s_i, t)$ and $U(s_j, t)$ for agents i and j by (5.15). Consider the pairwise ZCBF constraints (5.17) and (5.18). The goal is to distribute ZCBF constraint between agents i and j by assigning α_i and α_j to design a robust safe control mechanism that maximizes the uncertainty horizons in $U(s_i, t)$ and $U(s_j, t)$ under which the system still remains safe.

The goal in Problem 5.1 can be achieved by solving the following max-min problem

$$(\max_{S} \min_{\Delta\hat{\mathbf{p}}_i \in U(s_i,t), \Delta\hat{\mathbf{p}}_j \in U(s_j,t)} [\overline{ZCBF}_i + \overline{ZCBF}_j]) \geq 0 \tag{5.19}$$

where $S_{ij} = \sqrt{s_i \cdot s_j}$ is the pairwise horizon of uncertainty based on each agent's horizon of

uncertainty. Having a uniform horizon of uncertainty in which $s_i = s_j$ is desired, because lack of robustness in one agent affects the safety of the overall system. The inner minimization in (5.19) gives the worst case of ZCBF constraint considering the ambiguity set (5.15). In this scenario, positiveness of ZCBF constraint ensures safety of the system. It translates the worst case to the smallest value of ZCBF constraint. The outer maximization term gives the maximum horizon of uncertainty under which ZCBF constraint still remains positive.

Theorem 5.2. *The highest horizon of uncertainty that guarantees uniform robustness for all agents is obtained when the agility parameter α_{ij} is distributed between agents based on their measurement confidence as*

$$\alpha_i(t) = \alpha_{ij}(1 - \frac{|\phi_i(t)|}{|\phi_i(t)| + |\phi_j(t)|})$$
$$\alpha_j(t) = \alpha_{ij}(1 - \frac{|\phi_j(t)|}{|\phi_i(t)| + |\phi_j(t)|})$$

Proof. IG robustness is the highest horizon of uncertainty under which the safety of the system is ensured. $\overline{ZCBF}_i + \overline{ZCBF}_j$ in (5.19) based on the available uncertain measured values is

$$\overline{ZCBF}_i + \overline{ZCBF}_j = \\ \frac{\Delta\hat{\mathbf{p}}_i^T}{||\Delta\hat{\mathbf{p}}_i||}\mathbf{u}_i - \frac{\Delta\hat{\mathbf{p}}_j^T}{||\Delta\hat{\mathbf{p}}_j||}\mathbf{u}_j + \alpha_i(t)||\Delta\hat{\mathbf{p}}_i|| + \alpha_j(t)||\Delta\hat{\mathbf{p}}_j|| - \alpha_{ij}D_s \tag{5.20}$$

Define the inner minimization problem in (5.19) as $m(s_i, s_j)$. That is,

$$m(s_i, s_j) = \min_{\Delta\hat{\mathbf{p}}_i \in U(s_i,t), \Delta\hat{\mathbf{p}}_j \in U(s_j,t)} [\overline{ZCBF}_i + \overline{ZCBF}_j]$$

Note that $m(s_i, s_j)$ is the minimum value of (5.20) obtained when smallest values of $||\Delta\hat{\mathbf{p}}_i||$

and $||\Delta\hat{\mathbf{p}}_j||$ within the ambiguity set (5.15) occurred, which using triangular inequality are

$$\begin{aligned} ||\Delta\hat{\mathbf{p}}_i|| &= ||\Delta\mathbf{p}_{ij}|| - |\phi_i(t)||s_i| \\ ||\Delta\hat{\mathbf{p}}_j|| &= ||\Delta\mathbf{p}_{ij}|| - |\phi_j(t)||s_j| \end{aligned} \tag{5.21}$$

Therefore, by substituting (5.21) into (5.20) and some manipulations, one has

$$\begin{aligned} m(s_i, s_j) &= \frac{\Delta\hat{\mathbf{p}}_i^T}{||\Delta\hat{\mathbf{p}}_i||}\mathbf{u}_i - \frac{\Delta\hat{\mathbf{p}}_j^T}{||\Delta\hat{\mathbf{p}}_j||}\mathbf{u}_j \\ &+ \alpha_{ij} \cdot (||\Delta\mathbf{p}_{ij}|| - D_s) - \alpha_i(t)|\phi_i(t)||s_j| - \alpha_j(t)|\phi_j(t)||s_i| \end{aligned} \tag{5.22}$$

Note that $\frac{\Delta\hat{\mathbf{p}}_i^T}{||\Delta\hat{\mathbf{p}}_i||}\mathbf{u}_i = \frac{\Delta\mathbf{p}_{ij}^T}{||\Delta\mathbf{p}_{ij}||}\mathbf{u}_i \cos\theta_i$, where θ_i is deviation on direction of $\Delta\hat{\mathbf{p}}_i$ from $\Delta\mathbf{p}_{ij}$. Since $\theta_i << 1$ then $\cos\theta_i \approx 1$ and therefore $\frac{\Delta\hat{\mathbf{p}}_i^T}{||\Delta\hat{\mathbf{p}}_i||}\mathbf{u}_i = \frac{\Delta\mathbf{p}_{ij}^T}{||\Delta\mathbf{p}_{ij}||}\mathbf{u}_i$. Considering (5.22), problem (5.19) is simplified to

$$\max_S m(s_i, s_j) \geq 0 \tag{5.23}$$

Since the coefficients corresponding to s_i and s_j in (5.22) are negative, their maximum occurs when $m(s_i, s_j) = 0$. Therefore, by denoting maximum of s_i and s_j as ${s_i}^*$ and ${s_j}^*$, respectively, one has

$$\begin{aligned} &(\alpha_i(t)|\phi_i(t)|){s_i}^* + (\alpha_j(t)|\phi_j(t)|){s_j}^* = \\ &\frac{\Delta\hat{\mathbf{p}}_i^T}{||\Delta\hat{\mathbf{p}}_i||}\mathbf{u}_i - \frac{\Delta\hat{\mathbf{p}}_j^T}{||\Delta\hat{\mathbf{p}}_j||}\mathbf{u}_j + \alpha_{ij} \cdot (||\Delta\mathbf{p}_{ij}|| - D_s) \end{aligned} \tag{5.24}$$

To maximize the pairwise horizon of the uncertainty previously defined as $S_{ij} = \sqrt{s_i \cdot s_j}$, one

must solve the following optimization problem

$$\max_{\alpha_i,\alpha_j} {s_i}^* {s_j}^*$$
$$s.t. \quad (5.24)$$

which is a maximization problem over multiplication of two parameters while a linear relation exists between them. Therefore, by denoting the right-hand side of (5.24) by v, one has

$${s_i}^* = \frac{v}{2\alpha_i(t)|\phi_i(t)|}, \ {s_j}^* = \frac{v}{2\alpha_j(t)|\phi_j(t)|} \tag{5.25}$$

In addition, it is desired to have a uniform robustness for all agents, i.e., ${s_i}^* = {s_j}^*$. That is,

$$\alpha_i(t)|\phi_i(t)| = \alpha_j(t)|\phi_j(t)| \tag{5.26}$$

By considering (5.13) and (5.26), one has

$$\begin{aligned} \alpha_i(t) &= \alpha_{ij}(1 - \frac{|\phi_i(t)|}{|\phi_i(t)| + |\phi_j(t)|}) \\ \alpha_j(t) &= \alpha_{ij}(1 - \frac{|\phi_j(t)|}{|\phi_i(t)| + |\phi_j(t)|}) \end{aligned} \tag{5.27}$$

This completes the proof. □

Equation (5.27) provides a rule to share the ZCBF constraint between two agents to achieve the highest robustness against measurement uncertainty. Based on (5.27), the agent with higher measurement confidence takes a higher responsibility in ensuring pairwise safety, and behaves in an agile manner while the agent with lower confidence behaves conservatively and this leads to higher overall robustness.

Remark 5.2. The proposed method employs the cooperative capability of agents to deal with measurement uncertainty. It is also applicable to extreme cases. Assume that the

sensing system of agent i fails, and the agent realizes this through the discrepancy between measurements of two different sensors. This results in having very small measurement confidence $\frac{1}{\phi_i(t)}$. Therefore, according to (5.27), $\alpha_i = 0$ and $\alpha_j = \alpha_{ij}$. This means that agent j takes the whole responsibility in ensuring pairwise safety and compensates for the failure of agent i. This example clarifies the advantage of the proposed method to handle rare failure cases with infinity bound of uncertainty which cannot be obtained using worst-case analysis.

Remark 5.3. Note that even if agents share their positions through a communication network, the uncertainty in the information needs to be considered because knowledge of agents about their own positions might be drifted and also reliability and accuracy of communication network should be considered.

Remark 5.4. Note that the exact measurement error might be greater than $\phi(t)$ and measurement confidence $\frac{1}{\phi(t)}$ is the best knowledge of agents on the reliability of their measurements which does not affect the strict safety of the system as long as the horizon of uncertainty is not exceeded. However, more accurate measurement confidence leads to a more robust distribution of ZCBF constraint between them.

5.3.2.3 Discussion

The proposed approach can be extended to a more general dynamics. Assume that dynamics of agents i, j with states of x_i, x_j are, respectively, modeled as $\dot{x}_i = f_i(x_i) + g_i(x_i)u_i$ and $\dot{x}_j = f_j(x_j) + g_j(x_j)u_j$ with the safety criterion (5.7). Then, the distributed ZCBF conditions similar to (5.11) and (5.12) are

$$\begin{aligned} ZCBF_i: \quad & \frac{\partial h_{ij}}{\partial x_i}(f_i(x_i) + g_i(x_i)u_i) + \alpha_i \cdot (h_{ij}) \geq 0 \\ ZCBF_j: \quad & \frac{\partial h_{ij}}{\partial x_j}(f_j(x_j) + g_j(x_j)u_j) + \alpha_j \cdot (h_{ij}) \geq 0 \end{aligned}$$

Now, if $\frac{\partial h_{ij}}{\partial x_i}$ is only a function of x_i and $\frac{\partial h_{ij}}{\partial x_j}$ is only a function of x_j, then the distributed ZCBFs are functions of each agent's states and its local measurement. Therefore, by having

an ambiguity set for $\hat{h}_i$ and $\hat{h}_j$, one can form the uncertain interpretations of ZCBFs similarly as (5.17) and (5.18); which is used in solving optimization Problem 1 and similar results apply. However, if the partial derivative of h_{ij} with respect to x_i and x_j is a function of both states, such as higher-order derivatives of norm functions, then it demands further calculations to derive distributed formulation based on the agent state itself and the local measurement. This generalized formulation is one of the future research directions.

5.3.3 Controller Design

Considering Figure 5.1, each agent is a graph node, and the measurement confidence of agents is the edges of the graph. Each agent communicates with its surrounding agents at the initiation time, and they exchange their measurement confidence through a bidirectional communication graph. After that, distributed ZCBF constraints are formed and agents no longer need to communicate and they rely on their own measurements until a change in measurement confidence (e.g., the discrepancy in measured data obtained from two different sensors) of an agent is observed or a new agent gets close to it. In this case, the agent would re-communicate and reset the pairwise safety responsibility based on (5.27). Therefore, only on an event that an agent's measurement confidence changes, communication and change of α_i and α_j is needed. Thus, α_i and α_j can be considered as piecewise constants that will remain constant between two events.

Designing a controller for solving a safe and robust collision avoidance problem using the proposed approach includes two loops: 1. an outer loop that determines the share of each agent in the pairwise ZCBF constraints. 2. an inner loop that solves an optimization problem to find a safe controller that satisfies the safety of the overall multi-agent system by imposing the ZCBF constraint obtained from the outer loop while minimizing the intervention with the optimal controller for each agent. Algorithm 4 shows the proposed approach.

Remark 5.5. Note that Algorithm 4 is simultaneously performed for each agent using its own local information. Therefore the obtained control input for the overall multi-agents

Algorithm 4 Safe and Robust Control Design for each Agent i

1: **Initialization:** Start with safe initial conditions for all agents.

2: **procedure**

3: **Outer Loop Control Design.** Use Theorem 5.2 to find the distributed pairwise ZCBF constraints, i.e., the responsibility of each agent in ensuring pairwise safety for each two agents in the vicinity of each other. The overall distributed ZCBF constraint for each agent i is then formed as

$$ZCBF_{it} = [\overline{ZCBF}_{i_1}; ...; \overline{ZCBF}_{i_{Ni}}]$$

where $ZCBF_{i_l}, l = \{1, ..., N_i\}$ is the pairwise ZCBF constraint between agent i and its neighboring agent l and N_i is the number of neighbors of agent i.

4: **Inner Loop Control Design.** Use the matrix of distributed ZCBF constraints for each agent obtained by the outer loop as a hard constraint on the control design and solve an optimization problem that finds a safe controller which is robust to measurement uncertainty and minimally invasive to the optimal controller found by the linear quadratic regulator (LQR). This formulation integrates the LQR controller and ZCBF for each agent using quadratic programming, inspired by [27]

$$\begin{aligned} &\mathbf{u}_i{}^* = \arg\min_{\mathbf{u}_i} \|\mathbf{u}_i - \hat{\mathbf{u}}_i\| \\ &s.t.\ \ ZCBF_{it} \geq 0 \end{aligned} \tag{5.28}$$

where $\hat{\mathbf{u}}_i$ is the nominal controller obtained from LQR and $\mathbf{u}_i{}^*$ is the safe controller which is the minimally altered version of $\hat{\mathbf{u}}_i$ for agent i such that ZCBF constraints and therefore safety is ensured. Note that the nominal controller $\hat{\mathbf{u}}_i$ can be designed based on any performance objective or any other control approach.

5: **end procedure**

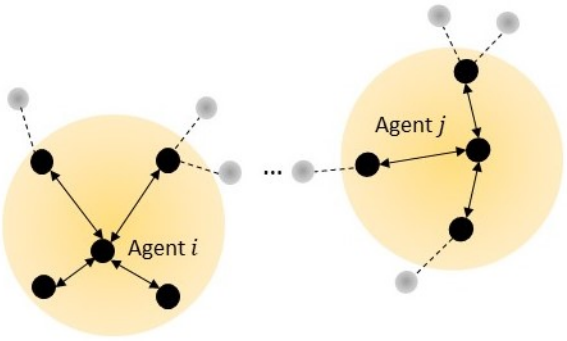

Figure 5.1: Graph Topology

system is

$$\mathbf{u}^* = [\mathbf{u}_1{}^*, \mathbf{u}_2{}^*, ..., \mathbf{u}_N{}^*]$$

With initial communication between agents in the vicinity of each other, ZCBFs sharing parameters are formed and the safe control input for each agent is obtained independently. After initial communication, they only need to re-communicate if the measurement confidence of one agent or the graph topology changes. In this condition, the agent informs its surrounding agents and they compromise again on the share of ZCBF constraint and this cycle continues. This provides a robust-satisficing distributed framework in which agents rely on their local measurements. This is an efficient method in which only occasional communication with surrounding agents is needed.

5.4 Simulation

A multi-robot system with five agents with integrator dynamics as (5.6) is considered. Agents are located around a circle with a radius of $r = 3$ and are supposed to get to the opposite point on the circle in a safe and collision-free manner in the sense that a pre-defined minimum safety distance $D_s = 0.3$ is respected between every two agents. Agents are aware of their destination and the LQR with $Q, R = 1$ is employed as the nominal controller for

this task objective. Simulation is conducted in three different scenarios; 1) Agents have accurate measurements and accurate distributed ZCBFs are employed and integrated into the controller as (5.28). 2) Measurement uncertainty is considered and its results compared to scenario (1) is investigated. 3) Measurement uncertainty is considered and the proposed method using IG is employed to share ZCBF constraints between agents to maximize robustness against uncertainty. Simulation results are given in two subplots, Figures(a), are trajectories of agents, in which their initial locations are depicted with triangle markers and their desired positions are depicted with star markers. Trajectories of agents are shown with dashed lines. Since this plot is given in $x - y$ plane, to have a sense of their maneuvers in time, the positions of agents in a time t_1 are also shown with filled circles. Figures.(b) demonstrate the pairwise distance $||\Delta \mathbf{p}_{ij}||$ between all agents from beginning to the end of the simulation. The minimum safety distance D_s is shown with a horizontal line. To have a safe maneuver, all pairwise distances should be higher than this minimum value.

Figure 5.2 depicts the result for the first scenario in which the measurement uncertainty is not considered, and accurate ZCBFs are available and equally shared between agents. Note that without the incorporation of ZCBFs, all agents would have crashed at the origin; however, in a barrier-certified fashion, agents get close to each other, turn around and move toward their desired positions. Figure 5.2 (b) demonstrates that pairwise distances have always been higher than minimum distance, and the safety of the system is preserved.

In the second scenario, the measurement uncertainty is incorporated as well. By clockwise numbering of agents in their initial positions, absolute measurement errors of agents 1, 2, 5 are 0.01. Measurement errors of agents 3 and 4 are considered 0.2 and 0.1, respectively. Note that these are high values of error considering that the minimum safety distance is 0.3. The exact values of errors are not known to agents, and their ZCBFs deviate from the actual value. The result of employing the same approach in the previous scenario and equally distributing ZCBF constraints between agents without using info-gap is depicted in Figure 5.3. As can be seen in this figure, agents approach the origin carelessly and

measurement error causes safety violation. The last scenario employs the proposed method to handle the measurement uncertainty. Measurement confidence of each agent is considered to be proportional to the inverse of its error, and pairwise ZCBF constraints are shared between agents using (5.27). The result is shown in Figure 5.4. As shown in Figure 5.4 (a), agents with higher confidence have more agile maneuvers and rapidly approach the origin, turn around and move toward their desired positions. Agent 4, with the next high measurement confidence, gives the right of way to agile agents while approaching the origin faster than agent 3 and agent 3, which has the lowest measurement confidence, moves slowly and conservatively until the path is clear. As it is shown in Figure 5.4 (b), the safety of the system is ensured despite inaccurate ZCBFs due to the measurement error. Note that measurement errors of agents 3 and 4 are significantly high and are 67 and 34 percent of the minimum safety distance, respectively. Note that the exact value of errors and their horizon are unknown to agents and the pairwise horizon of uncertainty based on each agent's horizon is $S_{ij} = \sqrt{s_i \cdot s_j}$. Thus, considering agents 3 and 4, which have the highest measurement uncertainty of 0.2 and 0.1, the pairwise uncertainty of at least 0.14 is safely tolerable. To show the effectiveness of the method for different values of safety distances, simulation is conducted for different values of $D_s = 0.1, 0.3, 0.6, 1$. The pairwise distances between every two agents are depicted in Figure 5.5. As can be seen in this figure, the pairwise distances are above the critical safety line, indicating that the task is accomplished safely. To better show the maneuvers of agents, a time-lapse is also shown in Figure 5.6.

Remark 5.6. Note that the proposed approach has a couple of advantages compared to the worst-case approach. First, in the worst-case approach, the information about the worst case of the measurement error is needed, and inaccuracy in this information results in violation of safety. Second, since the measurement error is high with respect to the minimum safety distance (67 percent for one of the agents), even if the worst-case is exactly known, still an overly conservative safety distance needs to be kept among agents which is not needed for certain agents. In addition, it makes the overall system slow and conservative and

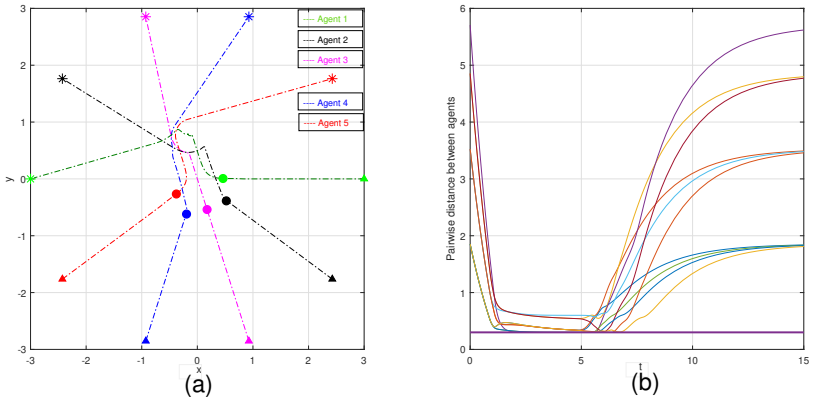

Figure 5.2: (a) Agents' trajectories, no measurement error (b) Corresponding $||\Delta \mathbf{p}_{ij}||$

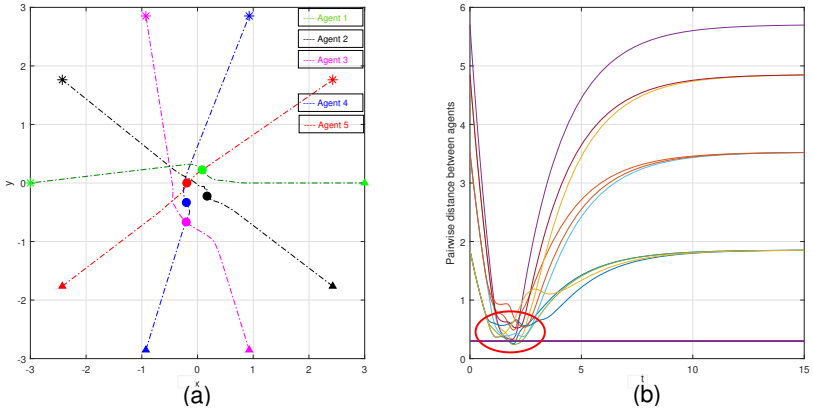

Figure 5.3: (a) Trajectories, measurement error without IG (b) Corresponding $||\Delta \mathbf{p}_{ij}||$

might result in the infeasibility of solution as well. Furthermore, the worst-case approach is not capable of rare cases of failure. Finally, the cooperative capability of agents in safety remains unused. However, in the proposed approach, non-conservative robustness against measurement uncertainty and rare cases of failure is achieved by proper distribution of ZCBF constraints between agents and giving them the benefit of cooperation for a safe maneuver.

5.5 Conclusion

In this chapter, designing safe controllers for collision-avoidance problem in multi-agent systems in the presence of measurement uncertainty is considered, and a robust-satisficing

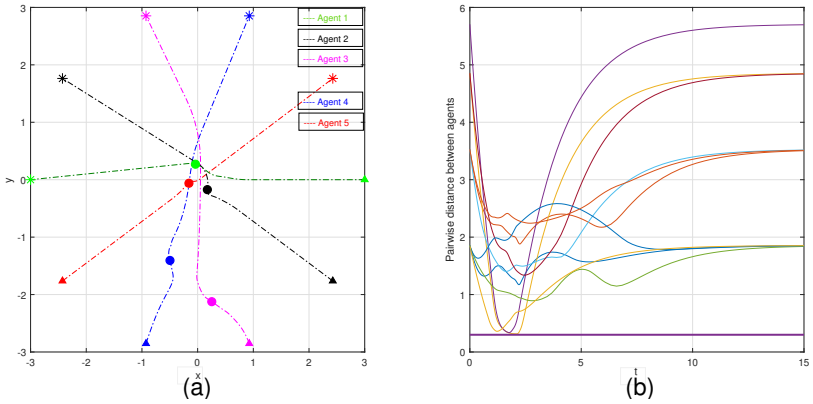

Figure 5.4: (a) Trajectories, measurement error with IG (b) Corresponding $||\Delta \mathbf{p}_{ij}||$

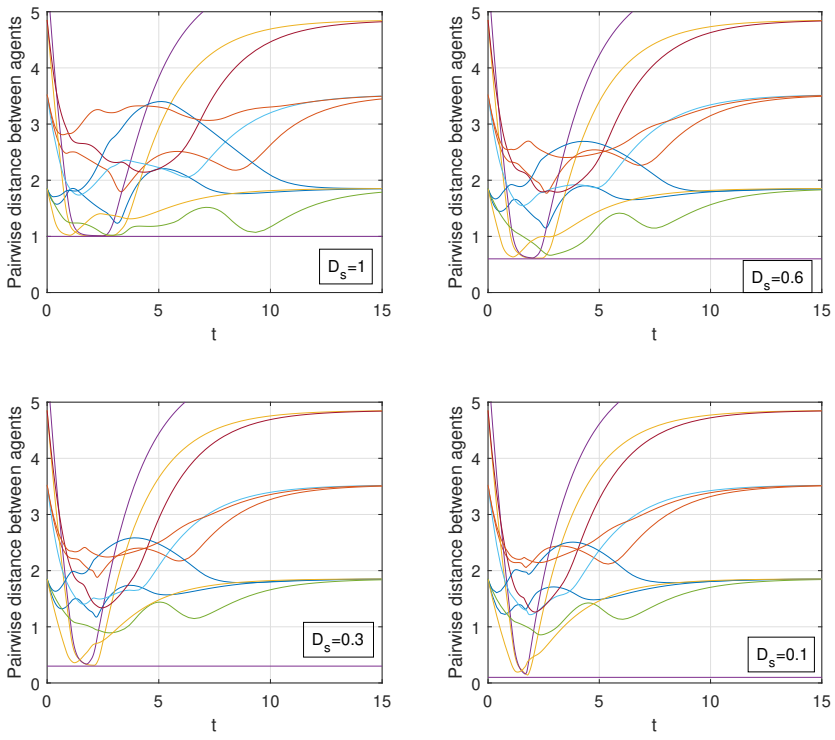

Figure 5.5: Pairwise distances between agents for different values of safety distance D_s

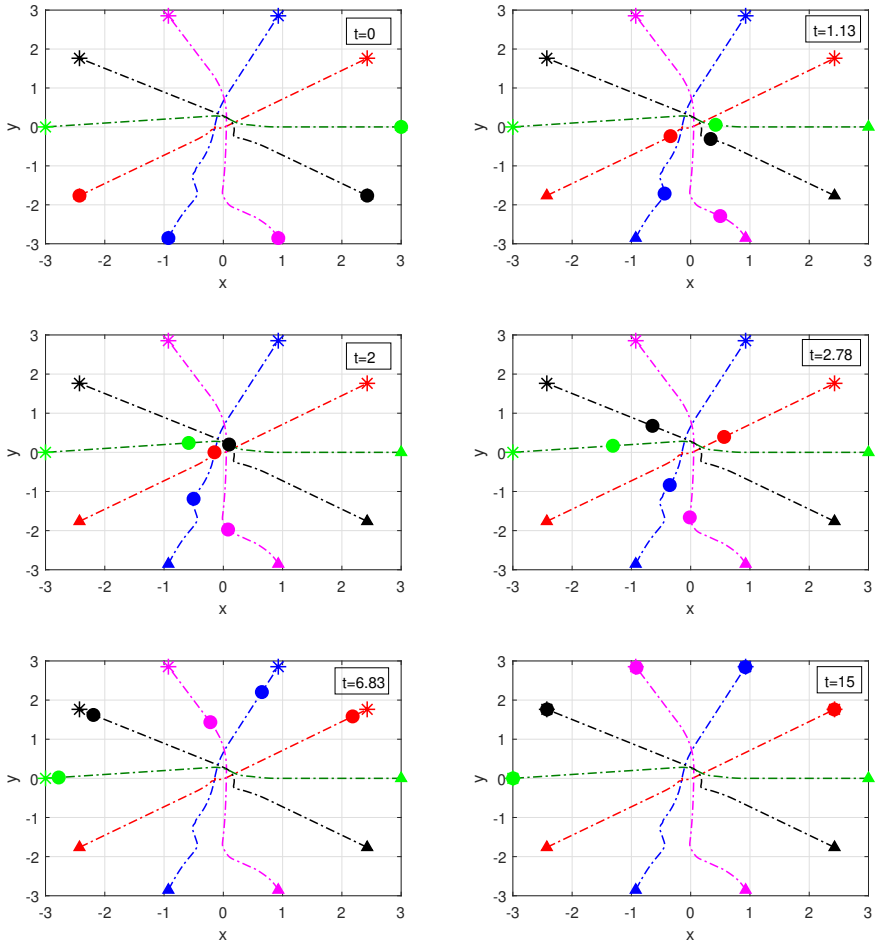

Figure 5.6: Time Lapse of agents' position with measurement error using IG method

CBF-based approach is proposed for safe cooperative maneuvers of agents. It is assumed that neither the probabilistic model of measurement nor its worst-case uncertainty is available. Then, IG theory is employed to achieve the best way of sharing safety in the sense of robustness toward uncertainty. It is shown that the certainty of one agent's measurement can compensate for the lack of accuracy of other agents. The simulation results with five agents in different scenarios are presented to show the performance of the proposed method. Future work includes providing a method for deriving and employing local confidence levels, which enables communication-free safety task assignment.

Chapter6

Conclusion

In this dissertation, the safety of the systems using CBFs in the face of two major challenges of uncertain system dynamics and uncertain environment is investigated. For the model uncertainty, a novel RL framework is proposed which augments the performance cost function with a barrier-type safety cost to form a safety-aware performance metric. It is shown that the presented performance also assures stability of the learned solution when there is no conflict between safety and stability. It is also shown that safety is guaranteed during successive approximation of control policies. A safe off-policy algorithm is employed to implement the proposed method. Afterward, the challenging problem of safe exploration is tackled with a barrier-certified safe RL framework which is obtained by means of efficient learning with prescribed performance along with a robutified safe and stabilizable controller throughout the algorithm including the data collection phase. Experience replay-based model approximation is employed, which ensures the exponential convergence of the learning error to zero after a mild rank condition is satisfied. This makes the learning error a vanishing perturbation to the approximated model, which facilitates designing stabilizing controller using the available rough knowledge of the system. The accurate bound of error is then employed in formation of a novel non-conservative AR-CBF which ensures safety during learning. AR-CBF and stabilizing controller are integrated through quadratic programming and is used for further data collection needed for off-policy iteration. The noisy input is modified accordingly to result in safe and stable action. After collecting safe rich data, the optimal

policy is approximated and then again is certified using AR-CBF for safe exploitation.

Afterward, the impact of uncertainty in the operating environment of the system is investigated. A learning-enabled ZCBF controller for safety-critical systems under uncertainty has been proposed. It has been proved that the proposed method is capable of ensuring safety in complicated and uncertain environments in the presence of external agents with unknown dynamics. It has been also demonstrated that safety during learning and even with inaccurate modeling of external agents is guaranteed. As a result, this approach has provided a practical method in control scenarios that accurate modeling needs a great number of data and computationally expensive learning schemes while still un-predicted objects are expected such as autonomous driving in an urban area. Meanwhile, having a better model has enabled the controller to take a less conservative action and has resulted in a better performance. To achieve this goal, a modified experience replay method has been proposed that identifies the external agents' dynamic to minimize the difference between the safe set and its approximation. This method provides fast convergence and ensures a bounded error to the exact model even with inaccurate modeling, which are both crucial in safety-critical control systems. Finally, the cooperative capability of multi-agent systems is employed for robust safety guarantee in the presence of measurement uncertainty. A robust-satisficing CBF-based approach is proposed for safe cooperative maneuvers of agents. It is assumed that neither the probabilistic model of measurement nor its worst-case uncertainty is available. Then, information-gap theory is employed to determine the share of agents in safety guarantee to achieve the highest robustness against uncertainty. It is shown that the certainty of one agent's measurement can compensate for the lack of accuracy of other agents, and even rare failure case of one agent can be compensated by others.

Future research direction includes the extension of safe exploratory RL framework to nonlinear systems and further employment of nonlinear theory in characterizing the learning behavior of learning-based controllers. It is also suggested to investigate the reciprocal interaction of agents in a cluttered environment and develop safe cooperative methods in

which conservatism is further reduced by efficient communication methodology.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] *Safety-critical Systems - Wikipedia @ONLINE.*

[2] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction (2nd Edition, in preparation)*. Vol. 1. MIT press Cambridge, 2017.

[3] F. L. Lewis and D. Vrabie. "Reinforcement learning and adaptive dynamic programming for feedback control". In: *IEEE Circuits and Systems Magazine* 9.3 (2009), pp. 32–50.

[4] B. Kiumarsi et al. "Optimal and Autonomous Control Using Reinforcement Learning: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.6 (2018), pp. 2042–2062.

[5] J García and F Fernández. "A comprehensive survey on safe reinforcement learning". In: *Journal of Machine Learning Research* 16 (Aug. 2015), pp. 1437–1480.

[6] A. Tamar et al. "Sequential Decision Making With Coherent Risk". In: *IEEE Transactions on Automatic Control* 62.7 (2017), pp. 3323–3338.

[7] O. Mihatsch and R. Neuneier. "Risk-sensitive reinforcement learning". In: *Machine Learning* 49.2 (2014), 267—290.

[8] T. Mannucci et al. "Safe Exploration Algorithms for Reinforcement Learning Controllers". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.4 (2018), pp. 1069–1081.

[9] Brenna D. Argall et al. "A survey of robot learning from demonstration". In: *Robotics and Autonomous Systems* 57.5 (2009), pp. 469 –483.

[10] Kurt Driessens and Sašo Džeroski. "Integrating Guidance into Relational Reinforcement Learning". In: *Machine Learning* 57.3 (2004), pp. 271–304.

[11] Brijen Thananjeyan et al. "Safety Augmented Value Estimation From Demonstrations (SAVED): Safe Deep Model-Based RL for Sparse Cost Robotic Tasks". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 3612–3619.

[12] Jingyu Niu et al. "Two-Stage Safe Reinforcement Learning for High-Speed Autonomous Racing". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2020, pp. 3934–3941.

[13] Canhuang Dai et al. “Reinforcement Learning with Safe Exploration for Network Security”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 3057–3061.

[14] Brijen Thananjeyan et al. “Recovery RL: Safe Reinforcement Learning With Learned Recovery Zones”. In: *IEEE Robotics and Automation Letters* 6.3 (2021), pp. 4915–4922.

[15] Shuo Li and Osbert Bastani. “Robust Model Predictive Shielding for Safe Reinforcement Learning with Stochastic Dynamics”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020, pp. 7166–7172.

[16] Zhaojian Li, Tianshu Chu, and Uroš Kalabić. “Dynamics-Enabled Safe Deep Reinforcement Learning: Case Study on Active Suspension Control”. In: *2019 IEEE Conference on Control Technology and Applications (CCTA)*. 2019, pp. 585–591.

[17] Shuojie Mo, Xiaofei Pei, and Chaoxian Wu. “Safe Reinforcement Learning for Autonomous Vehicle Using Monte Carlo Tree Search”. In: *IEEE Transactions on Intelligent Transportation Systems* (2021), pp. 1–8.

[18] Wei Wang et al. “Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems”. In: *IEEE Transactions on Smart Grid* 11.4 (2020), pp. 3008–3018.

[19] Yangyang Ge et al. “Safe Q-Learning Method Based on Constrained Markov Decision Processes”. In: *IEEE Access* 7 (2019), pp. 165007–165017.

[20] Zhehua Zhou et al. “A General Framework to Increase Safety of Learning Algorithms for Dynamical Systems Based on Region of Attraction Estimation”. In: *IEEE Transactions on Robotics* 36.5 (2020), pp. 1472–1490.

[21] I. M. Mitchell, A. M. Bayen, and C. J. Tomlin. “A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games”. In: *IEEE Transactions on Automatic Control* 50.7 (2005), pp. 947–957.

[22] A. K. Akametalu et al. “Reachability-based safe learning with Gaussian processes”. In: *53rd IEEE Conference on Decision and Control*. 2014, pp. 1424–1431.

[23] Jaime F. Fisac et al. “A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems”. In: *IEEE Transactions on Automatic Control* 64.7 (2019), pp. 2737–2752.

[24] M. Chen, S. Herbert, and C. J. Tomlin. “Exact and efficient Hamilton-Jacobi guaranteed safety analysis via system decomposition”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 87–92.

[25] Yifei Simon Shao et al. "Reachability-Based Trajectory Safeguard (RTS): A Safe and Fast Reinforcement Learning Safety Layer for Continuous Control". In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 3663–3670.

[26] "Barrier Lyapunov Functions for the control of output-constrained nonlinear systems". In: *Automatica* 45.4 (2009), pp. 918 –927. ISSN: 0005-1098.

[27] A. D. Ames, J. W. Grizzle, and P. Tabuada. "Control barrier function based quadratic programs with application to adaptive cruise control". In: *53rd IEEE Conference on Decision and Control*. 2014, pp. 6271–6278.

[28] A. D. Ames et al. "Control Barrier Function Based Quadratic Programs for Safety Critical Systems". In: *IEEE Transactions on Automatic Control* 62.8 (2017), pp. 3861–3876.

[29] M. Ohnishi et al. "Barrier-Certified Adaptive Reinforcement Learning With Applications to Brushbot Navigation". In: *IEEE Transactions on Robotics* 35.5 (2019), pp. 1186–1205.

[30] M. Srinivasan, S. Coogan, and M. Egerstedt. "Control of Multi-Agent Systems with Finite Time Control Barrier Certificates and Temporal Logic". In: *IEEE Conference on Decision and Control (CDC)*. 2018, pp. 1991–1996.

[31] Ayush Agrawal and Koushil Sreenath. "Discrete Control Barrier Functions for Safety-Critical Control of Discrete Systems with Application to Bipedal Robot Navigation". In: *Proceedings of Robotics: Science and Systems*. Cambridge, Massachusetts, 2017.

[32] L. Wang, E. A. Theodorou, and M. Egerstedt. "Safe Learning of Quadrotor Dynamics Using Barrier Certificates". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 2460–2465.

[33] L. Wang, A. Ames, and M. Egerstedt. "Safety barrier certificates for heterogeneous multi-robot systems". In: *2016 American Control Conference (ACC)*. 2016, pp. 5213–5218.

[34] Xiangru Xu et al. "Robustness of Control Barrier Functions for Safety Critical Control". In: *IFAC-PapersOnLine* 48.27 (2015), pp. 54 –61.

[35] A. J. Taylor and A. D. Ames. "Adaptive Safety with Control Barrier Functions". In: *2020 American Control Conference (ACC)*. 2020, pp. 1399–1405.

[36] B. T. Lopez, J. J. E. Slotine, and J. P. How. "Robust Adaptive Control Barrier Functions: An Adaptive and Data-Driven Approach to Safety". In: *IEEE Control Systems Letters* 5.3 (2021), pp. 1031–1036.

[37] Andrew Taylor et al. "Learning for Safety-Critical Control with Control Barrier Functions". In: *Proceedings of the 2nd Conference on Learning for Dynamics and Control*. Vol. 120. Proceedings of Machine Learning Research. PMLR, 2020, pp. 708–717.

[38] Richard Cheng et al. "End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 3387–3395.

[39] L. Wang, D. Han, and M. Egerstedt. "Permissive Barrier Certificates for Safe Stabilization Using Sum-of-squares". In: *American Control Conference (ACC)*. 2018, pp. 585–590.

[40] "Nearly optimal state feedback control of constrained nonlinear systems using a neural networks HJB approach". In: *Annual Reviews in Control* 28.2 (2004), pp. 239 –251. ISSN: 1367-5788.

[41] Y. Yang et al. "Safety-Aware Reinforcement Learning Framework with an Actor-Critic-Barrier Structure". In: *2019 American Control Conference (ACC)*. 2019, pp. 2352–2358.

[42] Adrian G. Wills and William P. Heath. "Barrier function based model predictive control". In: *Automatica* 40.8 (2004), pp. 1415–1422.

[43] C. Feller and C. Ebenbauer. "Weight recentered barrier functions and smooth polytopic terminal set formulations for linear model predictive control". In: *2015 American Control Conference (ACC)*. 2015, pp. 1647–1652.

[44] Z. Marvi and B. Kiumarsi. "Safety Planning Using Control Barrier Function: A Model Predictive Control Scheme". In: *2019 IEEE 2nd Connected and Automated Vehicles Symposium (CAVS)*. 2019, pp. 1–5.

[45] N. Wen et al. "UAV online path planning algorithm in a low altitude dangerous environment". In: *IEEE/CAA Journal of Automatica Sinica* 2.2 (2015), pp. 173–185.

[46] Dorsa Sadigh. "Safe and Interactive Autonomy: Control, Learning, and Verification". PhD thesis. EECS Department, University of California, Berkeley, 2017.

[47] Z. Marvi and B. Kiumarsi. "Safe Off-policy Reinforcement Learning Using Barrier Functions". In: *2020 American Control Conference (ACC)*. 2020, pp. 2176–2181.

[48] Zahra Marvi and Bahare Kiumarsi. "Barrier-certified Learning-based Control of Systems with Uncertain Safe Set". In: *2021 American Control Conference (ACC)*. 2021, pp. 3482–3487.

[49] Zahra Marvi and Bahare Kiumarsi. "Barrier-Certified Model-Learning and Control of Uncertain Linear Systems using Experience Replay Method". In: *2021 Conference on Decision and Control (CDC)*. 2021.

[50] Zahra Marvi and Bahare Kiumarsi. "Safe reinforcement learning: A control barrier function optimization approach". In: *International Journal of Robust and Nonlinear Control* (2020), pp. 1–18.

[51] Z. Marvi and B. Kiumarsi. "Barrier-certified learning-enabled safe control design for systems operating in uncertain environments". In: *IEEE/CAA J. Autom. Sinica* (2021), pp. 1–13.

[52] Zahra Marvi and Bahare Kiumarsi. "Robust Satisficing Cooperative Control Barrier Functions for Multi-Robots Systems using Information-Gap Theory". In: *International Journal of Robust and Nonlinear Control(Accepted)* (2021).

[53] Z. Marvi and B. Kiumarsi. "Reinforcement Learning based Control Design with Safety and Stability guarantees during Exploration". In: *(Under Review)* ().

[54] F.L. Lewis and V.L. Syrmos. *Optimal Control*. A Wiley-interscience publication. Wiley, 1995.

[55] H.K. Khalil. *Nonlinear Systems*. Pearson Education. Prentice Hall, 2002.

[56] G. N. Saridis and C. G. Lee. "An Approximation Theory of Optimal Control for Trainable Manipulators". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.3 (1979), pp. 152–159.

[57] Yu Jiang and Zhong-Ping Jiang. "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics". In: *Automatica* 48.10 (2012), pp. 2699–2704.

[58] Y. Jiang and Z. Jiang. "Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems". In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5 (2014), pp. 882–893.

[59] H. Modares, F. L. Lewis, and Z. Jiang. "H_∞Tracking Control of Completely Unknown Continuous-Time Systems via Off-Policy Reinforcement Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 26.10 (2015), pp. 2550–2562.

[60] Eric J. Rossetter and J. Christian Gerdes. "Lyapunov-based performance guarantees for the potential field lane-keeping assistance system". In: *Journal of Dynamic Systems, Measurement, and Control* 128.3 (Aug. 2005), pp. 510–522.

[61] C. Chen et al. “Reinforcement Learning-Based Adaptive Optimal Exponential Tracking Control of Linear Systems With Unknown Dynamics”. In: *IEEE Transactions on Automatic Control* 64.11 (2019), pp. 4423–4438.

[62] Mitio Nagumo. “Uber die Lage der Integralkurven gewo hnlicher Differentialgleichungen.” In: *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series* 24 (1942), pp. 551–559.

[63] F. Blanchini. “Set invariance in control”. In: *Automatica* 35.11 (1999), pp. 1747 –1767.

[64] Franco Blanchini and Stefano Miani. *Set-Theoretic Methods in Control*. Birkhäuser Basel, 2015.

[65] Georges Bouligand. *Introducion a la Geometrie Infinitesimale Directe*. Gauthiers–Villars, 1932.

[66] F. L. Lewis, A. Yesildirak, and Suresh Jagannathan. *Neural Network Control of Robot Manipulators and Nonlinear Systems*. Bristol, PA, USA: Taylor & Francis, Inc., 1998. ISBN: 0748405968.

[67] H. Modares, F. L. Lewis, and M. Naghibi-Sistani. “Adaptive Optimal Control of Unknown Constrained-Input Systems Using Policy Iteration and Neural Networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 24.10 (2013), pp. 1513–1525.

[68] Paul J. Werbos. “Approximate dynamic programming for real-time control and neural modeling”. In: *Handbook of Intelligent Control*. 1992.

[69] P. J. Werbos. “Neural networks for control and system identification”. In: *IEEE Conference on Decision and Control*. 1989, 260–265 vol.1.

[70] D. Zhao et al. “Experience Replay for Optimal Control of Nonzero-Sum Game Systems With Unknown Dynamics”. In: *IEEE Transactions on Cybernetics* 46.3 (2016), pp. 854–865.

[71] Katja Vogel. “A comparison of headway and time to collision as safety indicators”. In: *Accident Analysis and Prevention* 35.3 (2003), pp. 427–433.

[72] C. Tomlin, G. J. Pappas, and S. Sastry. “Conflict resolution for air traffic management: a study in multiagent hybrid systems”. In: *IEEE Transactions on Automatic Control* 43.4 (1998), pp. 509–521.

[73] Bassam Alrifaee, Kevin Kostyszyn, and Dirk Abel. “Model Predictive Control for Collision Avoidance of Networked Vehicles Using Lagrangian Relaxation”. In: *IFAC-*

PapersOnLine 49.3 (2016). 14th IFAC Symposium on Control in Transportation Systems CTS 2016, pp. 430 –435.

[74] O. Khatib. "Real-time obstacle avoidance for manipulators and mobile robots". In: *Proceedings. 1985 IEEE International Conference on Robotics and Automation*. Vol. 2. 1985, pp. 500–505.

[75] Paolo Fiorini and Zvi Shiller. "Motion Planning in Dynamic Environments Using Velocity Obstacles". In: *The International Journal of Robotics Research* 17.7 (1998), pp. 760–772.

[76] L. Wang, A. D. Ames, and M. Egerstedt. "Safety Barrier Certificates for Collisions-Free Multirobot Systems". In: *IEEE Transactions on Robotics* 33.3 (2017), pp. 661–674.

[77] Sunan Huang, Rodney Swee Huat Teo, and Kok Kiong Tan. "Collision avoidance of multi unmanned aerial vehicles: A review". In: *Annual Reviews in Control* 48 (2019), pp. 147 –164.

[78] S. Chung et al. "A Survey on Aerial Swarm Robotics". In: *IEEE Transactions on Robotics* 34.4 (2018), pp. 837–855.

[79] A. Chakravarthy and D. Ghose. "Obstacle avoidance in a dynamic environment: a collision cone approach". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 28.5 (1998), pp. 562–574.

[80] Fangyuan Xu, Li Tang, and Yan-Jun Liu. "Tangent barrier Lyapunov function-based constrained control of flexible manipulator system with actuator failure". In: *International Journal of Robust and Nonlinear Control* (2021), pp. 1–14.

[81] Urs Borrmann et al. "Control Barrier Certificates for Safe Swarm Behavior". In: *IFAC-PapersOnLine* 48.27 (2015). Analysis and Design of Hybrid Systems ADHS, pp. 68 – 73.

[82] C. K. Verginis and D. V. Dimarogonas. "Closed-Form Barrier Functions for Multi-Agent Ellipsoidal Systems With Uncertain Lagrangian Dynamics". In: *IEEE Control Systems Letters* 3.3 (2019), pp. 727–732.

[83] M. Majidi, B. Mohammadi-Ivatloo, and A. Soroudi. "Application of information gap decision theory in practical energy problems: A comprehensive review". In: *Applied Energy* 249 (2019), pp. 157 –165.

[84] V. Marchau et al. *Decision Making under Deep Uncertainty: From Theory to Practice*. Springer, Jan. 2019.

[85] Jun-Ming Hu, Hong-Zhong Huang, and Yan-Feng Li. "Reliability growth planning based on information gap decision theory". In: *Mechanical Systems and Signal Processing* 133 (2019), p. 106274.

[86] S. G. Pierce et al. "Evaluation of Neural Network Robust Reliability Using Information-Gap Theory". In: *IEEE Transactions on Neural Networks* 17.6 (2006), pp. 1349–1361.

[87] J. Frolik, M. Abdelrahman, and P. Kandasamy. "A confidence-based approach to the self-validation, fusion and reconstruction of quasi-redundant sensor data". In: *IEEE Transactions on Instrumentation and Measurement* 50.6 (2001), pp. 1761–1769.

[88] V. Zambianchi et al. "Distributed Nonasymptotic Confidence Region Computation Over Sensor Networks". In: *IEEE Transactions on Signal and Information Processing over Networks* 4.2 (2018), pp. 308–324.