Towards Safe, Strategic Multi-Agent Autonomy: A Game-Theoretic Perspective

Jingqi Li

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-157
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-157.html

August 14, 2025

Towards Safe, Strategic Multi-Agent Autonomy: A Game-Theoretic Perspective

by

Jingqi Li

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Claire J. Tomlin, Co-chair
Professor Somayeh Sojoudi, Co-chair
Professor Murat Arcak
Professor David Fridovich-Keil

Summer 2025

Towards Safe, Strategic Multi-Agent Autonomy: A Game-Theoretic Perspective

Abstract

Towards Safe, Strategic Multi-Agent Autonomy: A Game-Theoretic Perspective

by

Jingqi Li

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Claire J. Tomlin, Co-chair

Professor Somayeh Sojoudi, Co-chair

As autonomous systems increasingly operate in complex and uncertain environments, decentralized decision-making is essential for ensuring scalability, adaptability, and resilience. This dissertation integrates control theory, game theory, and reinforcement learning to advance safe, efficient, and strategic decision-making in multi-agent systems. The contributions are organized into three interconnected themes: safe multi-agent control, efficient computation of game-theoretic equilibria, and information asymmetry management.

The first theme focuses on safety-critical policy learning. It introduces a certifiable reachability learning framework based on a novel Lipschitz-continuous value function that guarantees safe operation. To address safety constraints more flexibly, an augmented Lagrangian reinforcement learning approach is proposed, enabling efficient policy optimization through adaptive penalty mechanisms. Building on these methods, a layered architecture integrates reachability-based filters with reinforcement learning to resolve conflicting constraints during multi-agent coordination.

The second theme addresses the computational challenges of game-theoretic decision-making. It introduces efficient algorithms for computing equilibria in dynamic games, including a primal-dual interior-point method for computing feedback Stackelberg equilibria and a parallelizable Alternating Direction Method of Multipliers (ADMM) algorithm for solving generalized Nash equilibria in stochastic settings. Leveraging these results, we apply stochastic game theory to energy systems, where we propose a nodal pricing mechanism using potential game structures to transform distributed coordination into tractable decision problems.

The third theme focuses on game-theoretic decision-making under incomplete information. It presents a method for inferring agents' objectives from partial observations in feedback settings, showing improved performance over traditional open-loop approaches. Additionally, it introduces an intent demonstration framework based on iterative linear-quadratic approximations, designed to

strategically influence agents' beliefs and enhance overall task performance.

Together, these contributions aim to provide a step toward designing safe, efficient, and strategically intelligent multi-agent systems. The proposed methods have potential applications in areas such as autonomous driving, aerial mobility, distributed energy systems, multi-robot manipulation, and human-robot collaboration.

To Ji, and to my parents, Arong Hu and Daming Li
— for their unwavering support and love.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my co-advisors, Professor Claire J. Tomlin and Professor Somayeh Sojoudi. Their mentorship, wisdom, and steadfast support have shaped every stage of my doctoral journey. From navigating complex research challenges to persevering through the uncertainties of the pandemic, they have guided me with both intellectual clarity and compassion.

Claire taught me how to identify research questions that are not only theoretically rich but also practically meaningful. Her vision of impactful research, which integrates rigorous theory with real-world application, has profoundly shaped the way I approach problem formulation and evaluation. One piece of advice she often shared with students, which has stayed with me, is the importance of including hardware demonstrations in job or dissertation talks to clearly convey real-world relevance. Claire also helped me grow as a communicator and collaborator. Every time I attend one of her talks, I gain a clearer sense of how to structure ideas, frame research problems thoughtfully, and present complex material in a way that is both accessible and compelling. Most importantly, she taught me to pay close attention to results that appear counterintuitive or contradict prevailing assumptions. She encouraged me to treat such findings not as anomalies to dismiss, but as opportunities to uncover deeper insights. Her intellectual rigor, receptiveness to the unexpected, and clear research vision continue to shape the way I think, reason, and conduct research.

Somayeh has played an equally important role in my development as a researcher. Through her detailed feedback and consistently high standards, she taught me how to reason precisely, articulate assumptions clearly, and communicate technical arguments with logical structure and clarity. Her mentorship was especially impactful during my early PhD years, when I was still learning how to transform loose ideas into focused, publishable research. She helped me build the discipline to write with rigor and the confidence to defend intricate results. As I progressed through the program, her guidance extended well beyond technical matters. Her advice on navigating the academic job market, especially during a period of uncertainty, was invaluable in helping me prepare strategically and move forward with confidence. Somayeh's emphasis on clarity, structure, and depth has left a lasting imprint on both my research practice and my professional growth.

I am equally thankful to my dissertation committee members, Professor Murat Arcak and Professor David Fridovich-Keil, for their thoughtful feedback and continued encouragement during my PhD years. David, in particular, has been a trusted collaborator and intellectual partner across several projects. I have learned a great deal from his ability to distill complex systems into tractable problems and to navigate research with clarity and creativity. I am excited to continue working with David as a postdoctoral researcher at UT Austin.

I would also like to acknowledge the professors and mentors whose advice had a lasting impact on my academic development. Professor Shankar Sastry helped me understand the essential traits of impactful research. Professor Jack Gallant taught me to distinguish between science and engineering and once shared a timeless insight: that every paper can be reduced by a third without losing its core message, and this resolved my always concern of how to fit a paper within the page limit. Jack also encouraged me to explore connections between reinforcement learning and neuroscience, particularly by aligning Q-functions with fMRI data in motor control. Although I could not fully pursue this direction during my PhD, I hope to revisit it in the future. I also thank Professor Negar

Mehr for her enthusiastic support of my many immature ideas in multi-agent games, Professor Hamsa Balakrishnan for helping me connect my work to air mobility research, and Professor Cathy Wu for offering practical wisdom on managing stress by setting priorities and focusing on what matters most.

Throughout my PhD, I have had the privilege to collaborate with and learn from many brilliant researchers whose ideas and insights have shaped my thinking. I am especially thankful for the collaborations and discussions with the following researchers (listed in alphabetical order): Anand Siththaranjan, Andrea Bajcsy, Andrew Lee, Brendon Anderson, Chinmay Maheshwari, Chris Strong, Dimitris Papadimitriou, Donggun Lee, Ebonye Smith, Eli Brock, Elizabeth Glista, Fernando Palafox, Filippos Fotiadis, Frank Chih-Yuan Chiu, Gabriel Chenevert, Gechen Qu, Haimin Hu, Jason Choi, Jasmine Jerry Aloor, Jesse Milzman, Joe Li, Karthik Rajgopal, Kaylene Stocking, Lasse Peters, Maria Gabriela Mendoza (Gaby), Marsalis Gibson, Maulik Bhatt, Michael Lim, Milad Shafaie, Mustafa Karabag, Sampada Deglurkar, Sangjae Bae, Sara Pohland, Sylvia Herbert, Somil Bansal, Tanmay Gautam, Tianjiao Zhang, Xinjie Liu, Yaodong Yu, Yatong Bai, Yixiao Huang, Zihan Liao, and Ziye Ma.

I am especially grateful to Somil Bansal, who met with me weekly during my first year to discuss research ideas and offer guidance. I thank Donggun Lee for the countless days and nights we spent working side by side in Cory 301, for teaching me advanced reachability analysis techniques, and for his persistence and collaboration in pushing projects forward. I am also indebted to Andrea Bajcsy for staying up late into East Coast hours to revise papers with me, and for teaching me how to visualize experimental results with clarity and elegance. I appreciate Chris Strong for working together in the SDH 7-th floor bug-free room and for helping me learn how to organize codebases and manage version control effectively. Finally, I thank Jason Choi not only for co-leading our reading group on safe reinforcement learning and sustaining insightful discussions across many sessions, but also for co-advising two students with me and helping guide their work into publishable research.

I am grateful for the support and intellectual exchange with my labmates in Somayeh's group: Yatong Bai, Sam Pfrommer, Hyunin Lee, Yixiao Huang, George Ma, and Eli Brock. I also thank my labmates in Claire's group: Alonso Marco, Chams Eddine Mballo, Ebonye Smith, Ellis Ratner, Gabriel E. Colon Reyes, Ian Chuang, Jason Choi, Katie Kang, Kaylene Stocking, Marius Wiggert, Marsalis Gibson, Michael Lim, Sampada Deglurkar, Sara Pohland, and Yarden Goraly. Their support and the stimulating lab environment have shaped both my research and personal growth. I will carry these experiences forward with appreciation and lasting gratitude.

I also made use of ChatGPT-4 [240] to assist with grammar checking during the writing process of this thesis; all content and ideas are entirely my own.

Finally, I want to express my heartfelt thanks to Ji Sun for your unwavering support during the most difficult moments of my PhD. Your care and encouragement meant more than words can express. To my parents, Arong Hu and Daming Li, thank you for your endless patience, unconditional love, and belief in me. Although I could not return home for several years due to visa constraints, your presence and support remained constant. I would not have reached this milestone without you.

Chapter 1

Introduction

Autonomous robotic systems have demonstrated remarkable success in structured and predictable environments, such as automated manufacturing plants and robotic warehouses. In these settings, the predictability of dynamics, consistency of interactions, and reliability of communication greatly simplify control and coordination. Robots in such environments can effectively synchronize their actions, enhancing productivity, efficiency, and operational safety.

However, extending autonomous systems to unstructured, dynamic, and human-centric environments remains a substantial challenge. These domains are inherently uncertain, involve frequent interaction with unpredictable agents (including humans), and often lack reliable communication. For instance, despite rapid technological progress, autonomous vehicles still face significant obstacles to widespread deployment due to safety concerns when navigating in complex and dynamically evolving traffic conditions. Similarly, robotic assistants in domestic or healthcare settings raise pressing safety challenges as they must operate safely in close proximity to humans. These scenarios highlight the need for adaptable, robust, and provably safe decision-making strategies.

A central difficulty in these domains lies in the decentralized nature of multi-agent decision-making. In many practical systems, multiple agents must coordinate despite having distinct or even conflicting objectives, varying levels of information, and limited ability to communicate. The optimal action for one agent may inadvertently compromise the safety or performance of others, potentially leading to unsafe interactions or system-level failures. These challenges are further complicated by information asymmetry, where agents lack full knowledge of each other's states, intentions, capabilities, or goals. Classical methods in control theory and multi-agent systems often assume full observability, shared knowledge, or ideal communication, which are assumptions that usually do not hold in practice. Realistic scenarios frequently involve noisy observations, strategic behavior, communication limitations, or privacy constraints. If such asymmetries are not explicitly accounted for, the resulting decisions may be suboptimal, uncoordinated, or even unsafe.

This dissertation systematically addresses these challenges through three interconnected research themes: safety assurance in multi-agent systems, efficient computation of dynamic game-theoretic equilibria, and strategic reasoning under information asymmetry. By integrating insights from control theory, game theory, and reinforcement learning, this work aims to develop principled and practical solutions for safe and intelligent multi-agent coordination under uncertainty. A key

unifying theme is the synthesis of model-based optimization with model-free learning, combining formal guarantees with data-driven adaptability.

1.1 Overview of the Dissertation

The first research theme of this dissertation focuses on safety assurance in complex multi-agent systems. Classical methods such as reachability analysis and model predictive control struggle with scalability when applied to high-dimensional systems with nonlinear dynamics and nonconvex constraints. Meanwhile, learning-based approaches such as deep reinforcement learning lack explicit safety guarantees due to their black-box structure and reliance on exploratory policies. To bridge this gap, this dissertation proposes a reachability learning framework that combines reinforcement learning with formal verification techniques. The resulting method provides certifiable safety guarantees under bounded environmental disturbances and is experimentally validated in real-world drone racing scenarios. Additionally, the dissertation adapts classical optimization techniques, such as the augmented Lagrangian method, to reinforcement learning settings, enabling effective handling of discrete action spaces and non-differentiable objectives. These techniques are further extended to multi-agent settings relevant to applications like air mobility, where safety is critical.

The second theme addresses the computational challenges of equilibrium analysis in dynamic multi-agent systems. Solving for Nash or Stackelberg equilibria is computationally intractable in general, especially when feedback and dynamics are involved. Most existing methods focus on static formulations and do not adequately capture the recursive structure of strategic interactions. This dissertation introduces efficient computational frameworks for feedback Stackelberg games using a nested Karush-Kuhn-Tucker (KKT) formulation. These methods ensure consistency and feasibility across multiple decision-making stages. To improve scalability in large systems such as power grids, air traffic networks, or transportation infrastructure, the work exploits structural properties like potential game formulations and network sparsity. This enables the design of scalable policy gradient algorithms capable of handling large agent populations and complex coupling constraints.

The third theme explores decision-making under information asymmetry in decentralized environments. Classical inverse game-theoretic methods often assume static or fully observable scenarios, which fail to capture the richness of dynamic feedback-based interactions. In realistic settings, agent behavior may reflect both immediate goals and long-term strategic intent. Moreover, agents may choose to maintain or exploit information asymmetry for strategic advantage. This dissertation develops new methods for inferring agent objectives from partial and dynamic behavioral observations, aligning inferred models more closely with true underlying intentions. It also introduces an intent demonstration framework, enabling agents not only to be understood but also to shape the beliefs and subsequent actions of others, thereby enhancing coordination and system-level performance.

Together, these contributions support the development of decentralized, safety-aware, and strategically intelligent multi-agent systems. The proposed methods offer practical tools for a range of real-world applications, including autonomous transportation, human-robot collaboration, aerial mobility, and robotic logistics.

Each chapter of this dissertation centers on one of the core research themes, presenting the associated technical and empirical results. The dissertation is based on a series of publications, including [182, 179, 60, 181, 184, 38, 180], as well as a conditionally accepted manuscript [183], from many co-authors. These will be cited as appropriate at the start of each chapter.

Part I

Safe Learning-Based Control

Chapter 2

Certifiable Reachability Learning for Nonlinear Dynamical Systems

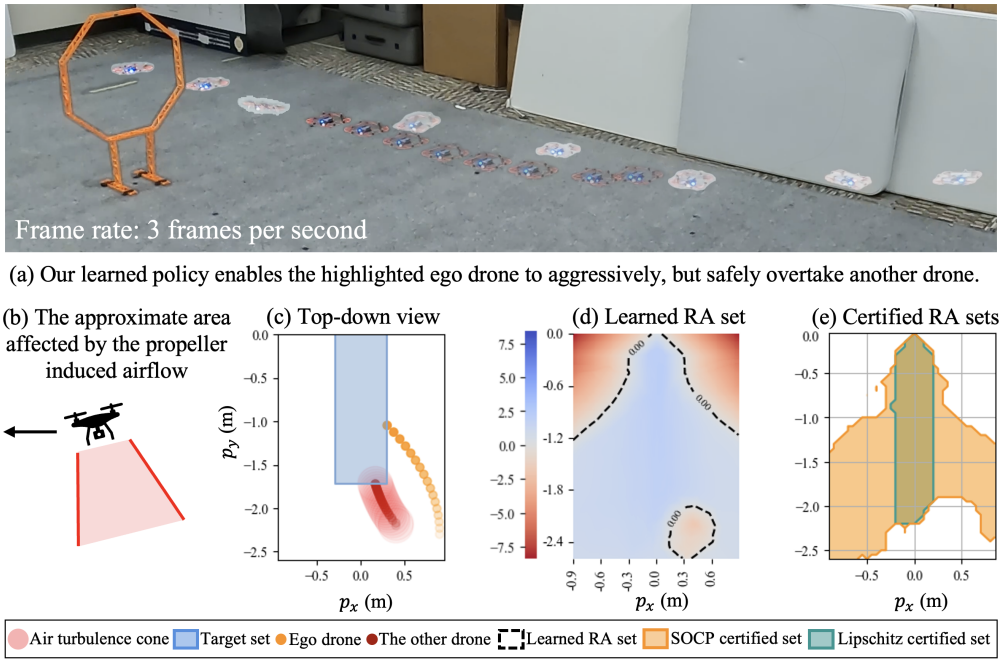
In this chapter, we focus on providing robust safety assurance for multi-agent systems from a zero-sum game perspective, where the autonomous agent plays against the worst-case behaviors of other agents and the environment. Our goal is to learn both safe control policies and a reachability set that provides global insight into the system's safety guarantees. This chapter is based on the published work [182], co-authored with Donggun Lee, Jaewon Lee, Kris Shengjun Dong, Somayeh Sojoudi, and Claire J. Tomlin.

2.1 Background

Ensuring the safe and reliable operation of robotic systems in uncertain environments is a critical challenge as autonomy is introduced into everyday systems. For instance, we would like humanoid robots to safely work close to humans. As a second example, new concepts for air taxis will need real-time synthesis of safe trajectories in crowded airspace. These safety-critical applications are typically characterized by sequences of tasks, and knowing the set of states from which the task can be safely completed despite unpredictable disturbances is important. Reachability analysis addresses this challenge by determining the *reach-avoid set*—a set of states that can safely reach a target set under all possible disturbances within a specified bound, as well as the corresponding control.

Traditional Hamilton-Jacobi reachability analysis methods [295, 206, 99] leverage dynamic programming to synthesize the optimal control and a reachability value function, whose sign indicates whether or not a state can safely reach the target set. Though theoretically sound, they suffer from the *curse of dimensionality* [28]: as the system's dimension increases, the computational complexity grows exponentially, making these methods impractical for real-world applications without approximation or further logic to manage the problem size.

There has been interest in using machine learning techniques to estimate reachability value functions for high-dimensional systems [19, 98, 134, 135, 132, 133, 190]. However, a major



(a) Our learned policy enables the highlighted ego drone to aggressively, but safely overtake another drone.

Figure 2.1: Applying our reachability analysis framework to drone racing. In (a), hardware experiments demonstrate that our learned control policy enables an ego drone to safely overtake another drone, despite unpredictable disturbances in the other drone's acceleration. In (b), we illustrate the concept of the propeller induced airflow [102], which can affect other drones' flight. In (c), we apply our learned control policy in a simulation with randomly sampled disturbances. In (d), we project the learned reach-avoid value function onto the (x, y) position of the ego drone. The super-zero level set, outlined by dashed curves, indicates our learned reach-avoid (RA) set. In (e), we plot the certified RA sets using Lipschitz and second-order cone programming certification.

drawback of existing reachability learning methods is the lack of deterministic safety guarantees. Recent work [193, 194] provides probabilistic safety guarantees for the learned reach-avoid sets. Additionally, safety filter approaches [133, 190, 234] have been proposed, which offer point-wise guarantees by ensuring safety for individual states.

In this chapter, we propose a novel method for learning reach-avoid sets for high-dimensional nonlinear systems with deterministic assurances. Our method involves learning a new reach-avoid value function (Sections 2.4 and 2.5), and then conducting set-based certification to ensure that all states in the certified set safely reach the target set despite disturbances (Section 2.6). Specifically:

1) We propose a new reach-avoid value function that is provably Lipschitz continuous. Though it is not based on a Lagrange-type objective function (cumulative rewards over time) as in classical reinforcement learning (RL) [31], we prove its Bellman equation is still a contraction mapping, removing the need for contractive Bellman equation approximation commonly used in prior reachability works [133, 190]. Moreover, the control policy derived from our value function tends to reach the target set rapidly. We apply deep RL to learn this new value function.

2) We develop two reach-avoid set certification methods. The first uses the Lipschitz constant of the dynamics to certify the safety of a subset of the learned reach-avoid set, ensuring all its

elements can safely reach the target set under disturbances. The second employs second-order cone programming to do the same. Both methods offer deterministic assurances. They can be applied online to verify if a neighboring set around the current state can safely reach the target set, or offline for verifying the same for a larger user-defined set.

3) We show the computational benefits of our new value function and the assurance of our (real-time) certification methods through simulations and hardware experiments. We empirically justify that the Lipschitz continuity of our new value function can accelerate value function learning.

2.2 Related works

Hamilton-Jacobi reachability learning. DeepReach [19] is a pioneering work on learning finite-horizon reachability value functions. In other studies, such as [98, 134, 132, 135, 133, 234, 190], the assumption of a known horizon is relaxed and infinite horizon reachability learning problems are considered. In this work, we also consider the infinite horizon case. We introduce a new value function which is provably Lipschitz continuous and whose Bellman operator is a contraction mapping, offering computational efficiency when compared with prior work.

Verification of learning-based control. Recent work [193, 194, 282] has provided probabilistic safety guarantees for DeepReach. In this chapter, we introduce methods that provide deterministic reach-avoid guarantees and we show how they could be used locally in real time. Other studies [133, 234, 190] provide point-wise safety filters. However, we propose set-based reach-avoid certification methods to verify if all states in a set can safely reach the target set under potential disturbances. Our certification methods also differ from existing set-based approaches for verifying neural network-controlled systems, including those that verify regions of attraction [235, 158, 273], forward reachability sets [67, 178, 87, 139, 164], and safe sets using barrier certificates [256, 210, 311]. To the best of the authors' knowledge, our work is the first to certify if a set of initial states is within the ground truth reach-avoid set.

Constrained optimal control. Control barrier functions (CBFs) offer safety guarantees [9, 292, 257], but they tend to be more conservative than model predictive control (MPC) [264, 198], which optimally balances task performance and safety. However, constructing or learning CBFs and solving MPC can be difficult for nonlinear systems with complex constraints. By leveraging deep neural networks, constrained reinforcement learning (CRL) [1, 62, 321] learns control policies that maximize task rewards while adhering to complex constraints. CBFs, along with the typical value functions defined in MPC and CRL, do not provide the information about whether a state can safely reach the target set. In contrast, our new value function not only provides the optimal control for the worst-case disturbance, but also indicates, based on its sign, whether or not a state can safely reach the target set.

2.3 Problem Formulation

We consider uncertain nonlinear dynamics described by

$$x_{t+1} = f(x_t, u_t, d_t), \tag{2.1}$$

where $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathcal{U} \subseteq \mathbb{R}^{m_u}$ is the control, and $d_t \in \mathcal{D} \subseteq \mathbb{R}^{m_d}$ represents the disturbance, such as model mismatch or uncertain actions of other agents. We assume that both $\mathcal{U}$ and $\mathcal{D}$ are compact and connected sets. The disturbance bound could be estimated from prior data or a physical model. We define a state trajectory originating from an initial condition x_0 under a control policy $\pi : \mathbb{R}^n \to \mathcal{U}$ and a disturbance policy $\phi : \mathbb{R}^n \times \mathcal{U} \to \mathcal{D}$ as $\xi_{x_0}^{\pi,\phi} := \{x_t\}_{t=0}^{\infty}$, where $x_{t+1} = f(x_t, \pi(x_t), \phi(x_t, \pi(x_t)))$, $\forall t \in \{0, 1, 2, \cdots\}$. Let $\mathcal{T} \subseteq \mathbb{R}^n$ be an open set, representing a *target set*. We assume that there exists a Lipschitz continuous, bounded reward function $r : \mathbb{R}^n \to \mathbb{R}$ indicating if a state x is in the target set,

$$r(x) > 0 \iff x \in \mathcal{T}. \tag{2.2}$$

We consider a finite number of Lipschitz continuous, bounded constraint functions $c_i(x) > 0, \forall i \in \mathcal{I}$, where $\mathcal{I}$ is the set of indexes of constraints. Throughout this chapter, we define Lipschitz continuity with respect to the ℓ_2 norm. We can simplify the representation of constraints by considering their pointwise minimum $c(x) := \min_{i \in \mathcal{I}} c_i(x)$. We define the constraint set as $\mathcal{C} := \{x \in \mathbb{R}^n : c(x) > 0\}$, and we have

$$c(x) > 0 \iff x \in \mathcal{C}. \tag{2.3}$$

We look for states that can be controlled to the target set safely under *the worst-case disturbance*, with dynamics given by (2.1). We refer to this set as the ***reach-avoid (RA) set*** [294, 206]:

$$\mathcal{R} := \left\{ \begin{array}{l} x_0 : \exists \pi \text{ such that } \forall \phi,\ \exists T < \infty, \\ \quad \big(r(x_T) > 0 \ \wedge\ \forall t \in [0, T], c(x_t) > 0\big) \end{array} \right\}, \tag{2.4}$$

which includes all the states that can reach the target set safely in finite time despite disturbances within the set $\mathcal{D}$.

Running example, safe take-over in drone racing: We model the drone take-over example in Figure 2.1 as an RA problem, where two crazyflie drones [109] compete to fly through an orange gate. The first drone (ego agent) starts behind and aims to overtake the other drone. The second drone flies directly to the gate using an LQR controller, but its acceleration is uncertain to the first drone. We model the uncertain part of the other drone's acceleration by a disturbance $\|d_t\|_2 \leq \varepsilon_d := 0.1$ m/s^2. We compute the RA set to ensure the ego drone can safely overtake the other despite this disturbance.

We consider a 12-dimensional dynamics [253], where the i-th drone's state is

$$x_t^i = [p_{x,t}^i, v_{x,t}^i, p_{y,t}^i, v_{y,t}^i, p_{z,t}^i, v_{z,t}^i]. \tag{2.5}$$

In each of the (x, y, z) axes, the i-th drone is modeled by double integrator dynamics, and the control is its acceleration $u_t^i = [a_{x,t}^i, a_{y,t}^i, a_{z,t}^i]$, with $\|u_t^i\|_\infty \leq \varepsilon_u := 1$m/s^2. We model the center of

the gate as the origin. The radius of the orange gate is 0.3 meters, and the radius of the crazyflie drone is 0.05 meters. We consider a target set for the ego drone:

$$\mathcal{T} = \begin{Bmatrix} x : p_y^1 - p_y^2 > 0, & v_y^1 - v_y^2 > 0, \\ |p_x^1| < 0.3, & |p_z^1| < 0.3 \end{Bmatrix}. \tag{2.6}$$

To ensure the ego drone flies through the gate, we constrain:

$$\pm p_{x,t}^1 - p_{y,t}^1 > -0.05, \quad \pm p_{z,t}^1 - p_{y,t}^1 > -0.05. \tag{2.7}$$

To ensure safe flight, the ego drone should avoid the area affected by the airflow from the other drone, as depicted in Figure 2.1, using the constraint:

$$\left\| \begin{bmatrix} p_{x,t}^1 - p_{x,t}^2 \\ p_{y,t}^1 - p_{y,t}^2 \end{bmatrix} \right\|_2^2 > \Big(1 + \max(p_{z,t}^2 - p_{z,t}^1, 0)\Big) \times 0.2, \tag{2.8}$$

where the required separation distance between the ego drone and the other drone increases as their height difference grows. Numerically computing the RA set directly for this problem is computationally infeasible [20]. We introduce our new reachability learning method in the following sections.

2.4 A New Reach-Avoid Value Function

In this section, we propose a new RA value function for evaluating if a state belongs to the RA set. Unlike prior works [98, 134, 132, 135, 133, 234, 190], our value function incorporates a time-discount factor. This results in a Lipschitz-continuous value function, which appears to accelerate the learning process, and establishes a contractive Bellman equation, eliminating the need for the contractive Bellman equation approximation commonly used in prior works [98, 134, 132, 135, 133, 234, 190]. Furthermore, we show that the control policy derived from this new value function tends to reach the target set rapidly.

We begin the construction of our new value function by first introducing the concept of *RA measure*, which assesses whether a trajectory can reach the target set safely. Let $\xi_{x_0}^{\pi,\phi}$ be a trajectory that enters the target set safely at a stage t. We have $r(x_t) > 0$ and $c(x_\tau) > 0$ for all $\tau \in \{0, 1, \ldots, t\}$. In other words, the *RA measure* $g(\xi_{x_0}^{\pi,\phi}, t)$, defined as

$$g(\xi_{x_0}^{\pi,\phi}, t) := \min \Big\{ r\big(x_t\big), \min_{\tau=0,\ldots,t} c\big(x_\tau\big) \Big\}, \tag{2.9}$$

is positive, $g(\xi_{x_0}^{\pi,\phi}, t) > 0$, if and only if there exists a trajectory from x_0 reaching the target set safely.

An *RA value function* $\bar{V}(x)$ has been proposed in prior works [98, 134, 132, 135, 133, 234, 190], and it evaluates the maximum RA measure under the worst-case disturbance:

$$\bar{V}(x) := \max_{\pi} \min_{\phi} \sup_{t=0,\ldots} g(\xi_{x_0}^{\pi,\phi}, t), \tag{2.10}$$

where $\bar{V}(x) > 0$ if and only if $x \in \mathcal{R}$. We compute $\bar{V}(x)$ by solving its Bellman equation. However, $\bar{V}(x)$ has a non-contractive Bellman equation, whose solution may not recover $\mathcal{R}$, as shown in Figure 2.2. To address this, prior works [98, 134, 132, 135, 133, 234, 190] include a time-discount factor γ to create a contractive Bellman equation approximation. For each $\gamma \in (0, 1)$, there is a unique solution to the approximated Bellman equation, which converges to $\bar{V}(x)$ as γ is gradually annealed to 1.

Inspired by previous studies, we enhance computational efficiency by designing a new *time-discounted* RA value function. Ours incorporates a time-discount factor into the value function formulation, resulting in a contractive Bellman equation without the need for any approximation. This improvement eliminates the requirement of the γ-annealing process commonly used in prior works [98, 134, 132, 135, 133, 234, 190], where N approximated Bellman equations are solved sequentially with γ values converging to 1. Theoretically, computing our new value function requires only $\frac{1}{N}$ of the time needed in prior works.

The central part of our new value function is the *time-discounted RA measure* $g_\gamma(\xi_{x_0}^{\pi,\phi}, t)$, for a $\gamma \in (0, 1)$,

$$g_\gamma(\xi_{x_0}^{\pi,\phi}, t) := \min\Big\{\gamma^t r\big(x_t\big), \min_{\tau=0,\dots,t} \gamma^\tau c\big(x_\tau\big)\Big\}. \tag{2.11}$$

This yields a new time-discounted RA value function

$$V_\gamma(x) := \max_\pi \min_\phi \sup_{t=0,\dots} g_\gamma(\xi_{x_0}^{\pi,\phi}, t). \tag{2.12}$$

For all $\gamma \in (0, 1)$ and any finite stage t, we have

$$g_\gamma(\xi_{x_0}^{\pi,\phi}, t) > 0 \iff g(\xi_{x_0}^{\pi,\phi}, t) > 0. \tag{2.13}$$

Therefore, for all $\gamma \in (0, 1)$, the super-zero level set of $V_\gamma(x)$, defined as $\mathcal{V}_\gamma := \{x : V_\gamma(x) > 0\}$, is equal to the RA set $\mathcal{R}$ in (2.4), and it includes all possible states that can reach the target set safely in finite time under the worst-case disturbance.

In what follows, we present the advantages of our new value function. First, we show that the Bellman equation for $V_\gamma(x)$ is a contraction mapping, with $V_\gamma(x)$ as its unique solution.

Theorem 1 (Contraction mapping)**.** *Let $\gamma \in (0, 1)$ and $V : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary bounded function. Consider the Bellman operator $B_\gamma[V]$ defined as,*

$$B_\gamma[V](x) := \max_u \min_d \min\big\{c(x), \max\{r(x), \gamma V(f(x, u, d))\}\big\}.$$

Then, we have $\|B_\gamma[V_\gamma^1] - B_\gamma[V_\gamma^2]\|_\infty \le \gamma\|V_\gamma^1 - V_\gamma^2\|_\infty$, for all bounded functions V_γ^1 and V_γ^2, and $V_\gamma(x)$ in (2.12) *is the unique solution to the Bellman equation $V(x) = B_\gamma[V](x)$.*

Proof. Let π^* and ϕ^* be the optimal control and the worst-case disturbance policies. Observe

$$\begin{aligned} V_\gamma(x_0) &= \max_\pi \min_\phi \min\big\{c(x_0), \max\{r(x_0), \gamma \sup_{\tau=0,\dots} g(\xi_{x_1}^{\pi^*,\phi^*}, \tau)\}\big\} \\ &= \max_\pi \min_\phi \min\big\{c(x_0), \max\{r(x_0), \gamma V_\gamma(x_1)\}\big\}, \end{aligned} \tag{2.14}$$

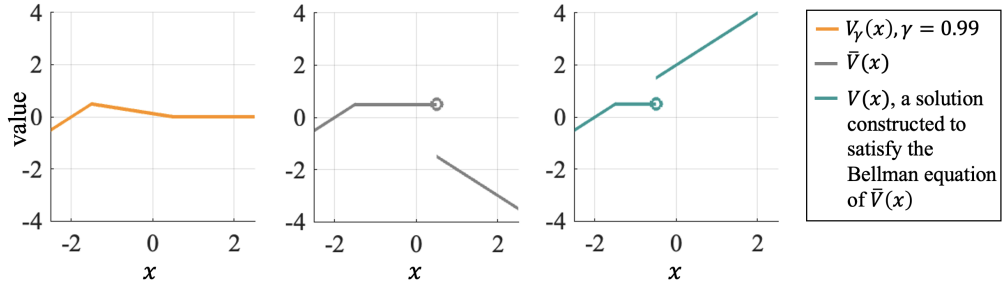


Figure 2.2: Comparing $V_\gamma(x)$ with $\bar{V}(x)$ from (2.10) and $V(x)$, a constructed solution to the Bellman equation of $\bar{V}(x)$ in prior works [98, 134, 132, 135, 133, 234, 190]. Consider a 1-dimensional dynamics: $x_{t+1} = 1.01x_t + 0.01(u_t + d_t)$, with $|u_t| \leq 1$ and $|d_t| \leq 0.5$. We associate $\mathcal{T} = \{x : x < -1\}$ and $\mathcal{C} = \{x : x > -2\}$ with bounded, Lipschitz continuous functions $r(x) = \max(\min(-(x+1), 10), -10)$ and $c(x) = \max(\min(x+2, 10), -10)$, respectively. For all $\gamma \in (0, 1)$, our super-zero level set $\{x : V_\gamma(x) > 0\}$ equals the RA set $\mathcal{R} = \{x : -2 < x < 0.5\}$. By Theorem 2, $V_\gamma(x)$ is Lipschitz continuous if $\gamma \in (0, 0.99009)$. The super-zero level set of $\bar{V}(x)$ also recovers $\mathcal{R}$, but $\bar{V}(x)$ is discontinuous at $x = 0.5$ because the control fails to drive the state to $\mathcal{T}$ under the worst-case disturbance when $x_t \geq 0.5$. Finally, in the third subfigure, we show that the Bellman equation in prior works [98, 134, 132, 135, 133, 234, 190] has non-unique solutions, whose super-zero level set may not equal $\mathcal{R}$.

where $x_1 = f(x_0, \pi(x_0), \phi(x_0, \pi(x_0)))$ is the only variable affected by π and ϕ. Following [31, p.234], it can be rewritten as $V_\gamma(x_0) = \max_{u_0} \min_{d_0} \min\left\{c(x_0), \max\{r(x_0), \gamma V_\gamma(x_1)\}\right\}$.

Thus, $V_\gamma(x)$ is a valid solution to the Bellman equation $V = B_\gamma[V]$. We show it is a unique solution by proving that $B_\gamma[V]$ is a contraction mapping when $\gamma \in (0, 1)$, i.e., $\|B_\gamma[V_1] - B_\gamma[V_2]\|_\infty \leq \gamma\|V_1 - V_2\|_\infty$, where V_1 and V_2 are two arbitrary bounded functions. Let x be an arbitrary state. We have $\|B_\gamma[V_1](x) - B_\gamma[V_2](x)\|_\infty \leq \|\gamma \max_u \min_d V_1(f(x, u, d)) - \gamma \max_u \min_d V_2(f(x, u, d))\|_\infty$. Since the max-min operator is non-expansive, we have, for all x, $\|B_\gamma[V_1](x) - B_\gamma[V_2](x)\|_\infty \leq \gamma\|V_1(x) - V_2(x)\|_\infty$. □

Theorem 1 suggests that annealing γ to 1 is unnecessary in our method because, for all $\gamma \in (0, 1)$, our Bellman equation admits $V_\gamma(x)$ as the unique solution, and the super-zero level set of $V_\gamma(x)$ equals the ground truth RA set.

Furthermore, we show in the following result that our new value function can be constructed to be Lipschitz continuous, which facilitates efficient learning when approximating high-dimensional value functions using neural networks [113, 314].

Theorem 2 (Lipschitz continuity)**.** *Suppose that $r(\cdot)$ and $c(\cdot)$ are L_r- and L_c-Lipschitz continuous functions, respectively. Assume also that the dynamics $f(x, u, d)$ is L_f-Lipschitz continuous in x, for all $u \in \mathcal{U}$ and $d \in \mathcal{D}$. Let $L := \max(L_r, L_c)$. Then, $V_\gamma(x)$ is L-Lipschitz continuous if $\gamma L_f < 1$.*

Proof. Consider two arbitrary initial states $x_0, x_0' \in \mathbb{R}^n$. Let π^* and ϕ^* be the optimal control and the worst-case disturbance policies. For each $t \in \{0, 1, 2, \dots\}$, define $x_{t+1} := f(x_t, u_t, d_t')$, $u_t := \pi^*(x_t)$, $x_{t+1}' := f(x_t', u_t, d_t')$, and $d_t' := \phi^*(x_t', u_t)$. Let $\mathbf{u} := \{u_t\}_{t=0}^\infty$, $\mathbf{d}' := \{d_t'\}_{t=0}^\infty$, $\mathbf{x} := \{x_t\}_{t=0}^\infty$, and $\mathbf{x}' := \{x_t'\}_{t=0}^\infty$. Given an arbitrarily small $\varepsilon > 0$, there exists a $\bar{t} < \infty$

such that $V_\gamma(x_0) \leq g_\gamma(\mathbf{x}, \bar{t}) + \varepsilon$. By definition, we have $V_\gamma(x_0') \geq g_\gamma(\mathbf{x}', \bar{t})$. Combining two inequalities, we have $V_\gamma(x_0') - V_\gamma(x_0) + \varepsilon \geq \min\{\gamma^{\bar{t}}(r(x_{\bar{t}}') - r(x_{\bar{t}})), \min_{\tau=0,\dots,\bar{t}} \gamma^\tau (c(x_{\bar{t}}') - c(x_{\bar{t}}))\} \geq -\max\{L_r \gamma^{\bar{t}} L_f^{\bar{t}}, \max_{\tau=0,\dots,\bar{t}} L_c \gamma^\tau L_f^\tau\} \|x_0 - x_0'\|_2$. The condition $\gamma L_f < 1$ implies $(\gamma L_f)^{\bar{t}} < 1, \forall \bar{t}$. As a result, $V_\gamma(x_0') - V_\gamma(x_0) + \varepsilon \geq -L\|x_0 - x_0'\|_2$. Similarly, we can show that $V_\gamma(x_0) - V_\gamma(x_0') + \varepsilon \geq -L\|x_0 - x_0'\|_2$. Combining these two inequalities, we prove Theorem 2. □

The main idea of the proof is that a small perturbation in the state x leads to a bounded change of the time-discounted RA measure value. Theorem 2 suggests that we can ensure the Lipschitz continuity of $V_\gamma(x)$ by selecting $\gamma < \frac{1}{L_f}$. In contrast, the classical RA value function $\bar{V}(x)$ can be discontinuous, as shown in Figure 2.2.

Moreover, the control policy derived from $V_\gamma(x)$ reaches the target set quickly, as a trajectory that reaches the target set rapidly incurs a high time-discounted RA measure value.

Theorem 3 (Fast reaching)**.** *Let x be in the RA set and ϕ be an arbitrary disturbance policy. Let $\gamma \in (0, 1)$. Suppose (π_1, t_1) and (π_2, t_2) are two control policies and corresponding times to maximize the discounted RA measure g_γ, $(\pi_1, t_1), (\pi_2, t_2) \in \arg\max_{\pi,t} g_\gamma(\xi_x^{\pi,\phi}, t)$. Moreover, suppose $t_1 < t_2$, then for all $\check{\gamma} \in (0, \min\{\gamma, \frac{V_\gamma(x)}{\max_x r(x)}\})$, we have $g_{\check{\gamma}}(\xi_x^{\pi_2,\phi}, t_2) < g_{\check{\gamma}}(\xi_x^{\pi_1,\phi}, t_1)$.*

Proof. From the definitions of $\check{\gamma}$, g_γ, and $V_\gamma(x)$, and the boundedness of $r(x)$, we have $\check{\gamma} < 1$, and $g_{\check{\gamma}}(\xi_x^{\pi_2,\phi}, t_2) \leq \check{\gamma}^{t_1} \check{\gamma} \max_x r(x) \leq \check{\gamma}^{t_1} V_\gamma(x) < \left(\frac{\check{\gamma}}{\gamma}\right)^{t_1} V_\gamma(x) \leq \left(\frac{\check{\gamma}}{\gamma}\right)^{t_1} g_\gamma(\xi_x^{\pi_1,\phi}, t_1) \leq g_{\check{\gamma}}(\xi_x^{\pi_1,\phi}, t_1)$. □

Theorem 3 also suggests that a control policy reaching the target set slowly may become suboptimal when γ is decreased.

While a small time-discount factor in $V_\gamma(x)$ offers numerous benefits, it is not conclusive that γ should always be near zero. In theory, for all $\gamma \in (0, 1)$, the super-zero level set of $V_\gamma(x)$ recovers the exact RA set. However, in practice, a near-zero γ can lead to a conservatively estimated RA set, where a trajectory reaching the target set at a late stage may have a near-zero or even negative time-discounted RA measure due to numerical errors. We will explore the trade-offs of selecting various γ values in section 2.8.

2.5 Learning the New RA Value Function

Since the optimal RA control policy is deterministic [294], we adapt max-min Deep Deterministic Policy Gradient (DDPG) [186], a deep RL method for learning deterministic policies and their value functions, to learn π, ϕ and V_γ.

Let γ be an arbitrary time discount factor in $(0, 1)$. Similar to prior works [133, 234, 190], we approximate the optimal control policy $\pi^*(x)$ and the worst-case disturbance policy $\phi^*(x, \pi^*(x))$ by neural network (NN) policies $\pi_{\theta_u}(x)$ and $\phi_{\theta_d}(x)$, respectively, with θ_u and θ_d being their parameters. We define an NN Q function as $Q_{\theta_q} : \mathbb{R}^n \times \mathcal{U} \times \mathcal{D} \to \mathbb{R}$, where θ_q represents the NN's parameter vector. Substituting π_{θ_u} and ϕ_{θ_d} into Q_{θ_q}, we can derive an NN value function $V_\theta(x) := Q_{\theta_q}(x, \pi_{\theta_u}(x), \phi_{\theta_d}(x))$, where θ is the concatenation of parameters θ_q, θ_u and θ_d. Let $\mathbb{P}$

be a sampling distribution with a sufficiently large support in $\mathbb{R}^n$ that covers at least a part of the target set. In max-min DDPG, we learn π_{θ_u}, ϕ_{θ_d} and Q_{θ_q} by alternatively optimizing the following problems:

We learn π_{θ_u} by maximizing the Q value over θ_u:

$$\max_{\theta_u} \mathbb{E}_{x\sim\mathbb{P}}\, Q_{\theta_q}(x, \pi_{\theta_u}(x), \phi_{\theta_d}(x)). \tag{2.15}$$

We learn ϕ_{θ_d} by minimizing the Q value over θ_d:

$$\min_{\theta_d} \mathbb{E}_{x\sim\mathbb{P}}\, Q_{\theta_q}(x, \pi_{\theta_u}(x), \phi_{\theta_d}(x)). \tag{2.16}$$

We learn Q_{θ_q} by minimizing the *critic loss*, also known as the Bellman equation error, over θ_q:

$$\min_{\theta_q} \mathbb{E}_{x\sim\mathbb{P}} \|V_\theta(x) - B_\gamma[V_\theta(x)]\|_2^2. \tag{2.17}$$

We define the *learned RA set* $\hat{\mathcal{R}}$ as the super zero-level set of $V_\theta(x)$. When DDPG converges to an optimal solution, $V_\theta(x)$ converges to $V_\gamma(x)$ due to Theorem 1. However, in practice, like other deep RL methods, DDPG often converges to a suboptimal solution with a near-zero critic loss. When $V_\theta(x)$ is a suboptimal solution, $\hat{\mathcal{R}}$ cannot be reliably considered as the ground truth RA set $\mathcal{R}$. This motivates us to use a suboptimal learning result to certify a trustworthy RA set, as detailed in the following section.

2.6 Certifying Learned RA Sets with Guarantees

In this section, we propose two methods to certify if a set of states belongs to the ground truth RA set. Both methods use a learned control policy, which is not necessarily optimal.

Certification using Lipschitz constants

We leverage a learned control policy π_{θ_u} and the Lipschitz constants of dynamics, reward and constraint to construct a theoretical lower bound of the ground truth value function $V_\gamma(x)$. *If such a lower bound of $V_\gamma(x)$ is greater than zero for all states in the neighboring set of x_0, $\mathcal{E}_{x_0} := \{x : \|x - x_0\|_2 \le \varepsilon_x\}$, then $V_\gamma(x) > 0$, $\forall x \in \mathcal{E}_{x_0}$. We claim that the set $\mathcal{E}_{x_0}$ is within the ground truth RA set $\mathcal{R}$.*

We begin constructing a lower bound of $V_\gamma(x)$ by considering a T-stage, disturbance-free, nominal trajectory $\{\bar{x}_t\}_{t=0}^T$,

$$\bar{x}_{t+1} = f(\bar{x}_t, \bar{u}_t, 0), \bar{u}_t = \pi_{\theta_u}(\bar{x}_t), \;\; \forall t = 0, \dots, T-1 \tag{2.18}$$

and a disturbed state trajectory $\{\tilde{x}_t\}_{t=0}^T$ under $\{\bar{u}_t\}_{t=0}^{T-1}$ using an arbitrary $d_t \in \mathcal{D}$, $\forall t = 0, \dots, T-1$:

$$\tilde{x}_{t+1} = f(\tilde{x}_t, \bar{u}_t, d_t), \;\; \forall t = 0, \dots, T-1. \tag{2.19}$$

Note that if we can verify that a trajectory starting at state x reaches the target set safely despite disturbances within T stages, then it suffices to claim $x \in \mathcal{R}$, where the certification horizon T can be set arbitrarily. Ideally, we would set $T = \infty$, but it is impractical to evaluate an infinitely

long trajectory. Therefore, during certification, we consider a finite, user-defined T. This T should preferably be long enough to allow initial states to reach the target set. A short T results in a conservative certification since it overlooks the possibility that the trajectory might safely reach the target set at a later time.

We assume that the dynamics f is Lipschitz continuous and there exists an upper bound on the disturbance in $\mathcal{D}$ at each stage t, i.e., $\|d_t\|_2 \leq \varepsilon_d$. Let L_{f_x} and L_{f_d} be the Lipschitz constants of the dynamics f with respect to the state x and disturbance d, respectively. At time $t = 1$, we observe

$$\begin{aligned}\|\bar{x}_1 - \tilde{x}_1\|_2 &= \|f(\bar{x}_0, \bar{u}_0, 0) - f(\tilde{x}_0, \bar{u}_0, d_0)\|_2 \\ &\leq L_{f_x}\|\bar{x}_0 - \tilde{x}_0\|_2 + L_{f_d}\|0 - d_0\|_2 \leq L_{f_x}\varepsilon_x + L_{f_d}\varepsilon_d.\end{aligned}$$

At time $t = 2$,

$$\begin{aligned}\|\bar{x}_2 - \tilde{x}_2\|_2 &= \|f(\bar{x}_1, \bar{u}_1, 0) - f(\tilde{x}_1, \bar{u}_1, d_1)\|_2 \\ &\leq L_{f_x}\|\bar{x}_1 - \tilde{x}_1\|_2 + L_{f_d}\varepsilon_d.\end{aligned}$$

By induction, we have

$$\|\bar{x}_t - \tilde{x}_t\|_2 \leq L_{f_x}^t\varepsilon_x + \Sigma_{\tau=0}^{t-1}\, L_{f_x}^\tau L_{f_d}\varepsilon_d \quad =: \Delta x_t. \tag{2.20}$$

We define a convex outer approximation of the set of dynamically feasible states as $\mathcal{X}_{t,\bar{x}_0}^L := \{x_t : \|x_t - \bar{x}_t\|_2 \leq \Delta x_t\}$, and we check if for all $x_t \in \mathcal{X}_{t,\bar{x}_0}^L$, $r(x_t) > 0$ and $c(x_t) > 0$. By Lipschitz continuity of the reward function, we have,

$$\forall x_t \in \mathcal{X}_{t,\bar{x}_0}^L, \quad \|r(\bar{x}_t) - r(x_t)\|_2 \leq L_r\Delta x_t$$

which yields a lower bound of $r(x_t)$, for all $x_t \in \mathcal{X}_{t,\bar{x}_0}^L$:

$$\check{r}_t^L := r(\bar{x}_t) - L_r\Delta x_t \leq r(x_t). \tag{2.21}$$

Similarly, we have a lower bound of $c(x_t)$, for all $x_t \in \mathcal{X}_{t,\bar{x}_0}^L$:

$$\check{c}_t^L := c(\bar{x}_t) - L_c\Delta x_t \leq c(x_t). \tag{2.22}$$

Using $\check{r}_t^L$ and $\check{c}_t^L$, we can construct a lower bound $\check{V}_\gamma^L(\bar{x}_0, T)$ for $V_\gamma(x_0)$, for all $x_0 \in \mathcal{E}_{\bar{x}_0}$:

$$\check{V}_\gamma^L(\bar{x}_0, T) := \max_{t=0,\ldots,T} \min\{\gamma^t\check{r}_t^L, \min_{\tau=0,\ldots,t} \gamma^\tau\check{c}_\tau^L\} \leq V_\gamma(x_0). \tag{2.23}$$

This implies

$$\check{V}_\gamma^L(\bar{x}_0, T) > 0 \Longrightarrow V_\gamma(x_0) > 0, \forall x_0 \in \mathcal{E}_{\bar{x}_0}. \tag{2.24}$$

Thus, when $\check{V}_\gamma^L(\bar{x}_0, T) > 0$, the set $\mathcal{E}_{\bar{x}_0}$ is *certified* to be within the ground truth RA set $\mathcal{R}$. Moreover, $\{\bar{u}_t\}_{t=0}^{T-1}$ are the *certified control inputs* that can drive all $x \in \mathcal{E}_{\bar{x}_0}$ to the target set $\mathcal{T}$ safely despite disturbances in $\mathcal{D}$.

Certification using second-order cone programming

Lipschitz certification is fast to compute. However, the lower bound $\check{V}_\gamma^L$ can be conservative. In this subsection, we propose another certification method using second-order cone programming (SOCP), aiming to provide a less conservative RA certification. Our key idea is to construct SOCPs that search over a tight, convex outer approximation set of the dynamically feasible trajectories and verify whether the state trajectory reaches the target set safely under all disturbances, within a user-defined finite certification horizon T.

The construction of these SOCPs involves two steps.

First, we formulate a *surrogate RA problem*: A subset of the original target set $\mathcal{T}$, represented by the interior of a polytope $\check{\mathcal{T}} := \{x : P_i x - k_i > 0, i \in \mathcal{I}_{\check{\mathcal{T}}}\}$, is defined as a *surrogate target set* $\check{\mathcal{T}}$, where $\mathcal{I}_{\check{\mathcal{T}}}$ is a finite index set of the polytope's edges. Additionally, we define a subset of the original constraint set $\mathcal{C}$, represented by the intersection of a finite number of positive semidefinite quadratic functions' super-zero level sets $\check{\mathcal{C}} := \{x : \frac{1}{2}x^\top Q_i x + q_i^\top x + b_i > 0, i \in \mathcal{I}_{\check{\mathcal{C}}}\}$, as a *surrogate (nonconvex) constraint set* $\check{\mathcal{C}}$. The set $\check{\mathcal{C}}$ could be nonconvex when approximating collision avoidance constraints.

Subsequently, we leverage SOCPs to verify if a state trajectory from x_0 can reach $\check{\mathcal{T}}$ while staying within $\check{\mathcal{C}}$ despite disturbances. We achieve this by sequentially minimizing each function defined in $\check{\mathcal{T}}$ and $\check{\mathcal{C}}$, iterating from the stage $\tau = 0$ to $\tau = T$. *If their minimum is positive at a stage t and there is no intermediate stage $\tau < t$ such that x_τ is outside $\check{\mathcal{C}}$, we claim that the initial state x_0 can safely reach the original target set $\mathcal{T}$, under all possible disturbances.*

To be more specific, we can check if there exists a stage $t \in \{0, 1, \dots, T\}$ such that, for all disturbances, all dynamically feasible states x_t, originating from the initial states set $\mathcal{E}_{\bar{x}_0} := \{x : \|x - \bar{x}_0\|_2 \leq \varepsilon_x\}$, are within $\check{\mathcal{T}}$ and for all stages $\tau \leq t$, x_τ are within $\check{\mathcal{C}}$. If this condition is met, we claim that $\mathcal{E}_{\bar{x}_0}$ is within the ground truth RA set $\mathcal{R}$, i.e., $\mathcal{E}_{\bar{x}_0} \subseteq \mathcal{R}$. Otherwise, some of its elements could be outside $\mathcal{R}$, and therefore we do not claim $\mathcal{E}_{\bar{x}_0} \subseteq \mathcal{R}$. To make the analysis tractable, we define the nominal state trajectory $\{\bar{x}_t\}_{t=0}^T$ and nominal control trajectory $\{\bar{u}_t\}_{t=0}^{T-1}$ as in (2.18). The nominal state and control trajectories allow us to formulate a convex set $\mathcal{X}_{t,\bar{x}_0}^S$ for outer approximating the set of dynamically feasible states:

$$
\begin{aligned}
&\mathcal{X}_{t,\bar{x}_0}^S := \{x_t : \exists \{d_\tau\}_{\tau=0}^{t-1} \text{ and } x_0 \text{ such that } \forall \tau \leq t-1, \\
&x_{\tau+1} \leq \hat{A}_\tau x_\tau + \hat{B}_\tau \bar{u}_\tau + \hat{D}_\tau d_\tau + \hat{c}_\tau, && \text{(Upper bound on } f\text{)} \\
&x_{\tau+1} \geq \check{A}_\tau x_\tau + \check{B}_\tau \bar{u}_\tau + \check{D}_\tau d_\tau + \check{c}_\tau, && \text{(Lower bound on } f\text{)} \\
&\|d_\tau\|_2 \leq \varepsilon_d, && \text{(Disturbance bound)} \\
&\|x_0 - \bar{x}_0\|_2 \leq \varepsilon_x\} && \text{(Initial state bound)}
\end{aligned}
$$

where the first two inequalities are element-wise and the bounds on the dynamics f can be derived using its Lipschitz constant or a Taylor series[1]. At a stage t, we can verify if $x_t \in \check{\mathcal{T}}$ by solving a sequence of SOCPs iterating over all $i \in \mathcal{I}_{\check{\mathcal{T}}}$, and checking if their minimum is positive:

[1]For simplicity, we can also use $\mathcal{X}_{t,\bar{x}_0}^L$ to define the convex outer approximation set of the dynamically feasible states in SOCP certifications, i.e., $\mathcal{X}_{t,\bar{x}_0}^S := \mathcal{X}_{t,\bar{x}_0}^L$.

$$\check{r}_t^S := \min_{i\in\mathcal{I}_{\check{\mathcal{T}}}} \big\{ \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} P_i x_t - k_i \big\}. \tag{2.25}$$

Similarly, for checking whether $x_t \in \check{\mathcal{C}}$, we can evaluate if the following term is positive:

$$\check{c}_t^S := \min_{i\in\mathcal{I}_{\check{\mathcal{C}}}} \big\{ \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} \frac{1}{2} x_t^\top Q_i x_t + q_i^\top x_t + b_i \big\}. \tag{2.26}$$

Combining the above two terms, we can construct a conservative certificate $\check{V}_\gamma^S$ for verifying if a state $\bar{x}_0$ and its neighboring set $\mathcal{E}_{\bar{x}_0}$ are within the ground truth RA set $\mathcal{R}$,

$$\check{V}_\gamma^S(\bar{x}_0, T) := \max_{t=0,\dots,T} \min\{\gamma^t \check{r}_t^S, \min_{\tau=0,\dots,t} \gamma^\tau \check{c}_\tau^S\}. \tag{2.27}$$

This suggests

$$\check{V}_\gamma^S(\bar{x}_0, T) > 0 \Longrightarrow V_\gamma(x_0) > 0, \forall x_0 \in \mathcal{E}_{\bar{x}_0} \tag{2.28}$$

and $\{\bar{u}_t\}_{t=0}^{T-1}$ are the *certified control inputs*, capable of driving all $x \in \mathcal{E}_{\bar{x}_0}$ to the target set $\mathcal{T}$ safely despite disturbances.

Running example (continued). At stage t, we evaluate $\check{r}_t^S = \min_i\{\check{r}_{t,i}^S\}_{i=1}^6$ by solving the following SOCPs, where each $\check{r}_{t,i}^S$ corresponds a function in the definition of $\mathcal{T}$ in (2.6):

$$\begin{aligned}
\check{r}_{t,1}^S &= \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} p_{y,t}^1 - p_{y,t}^2, \quad \check{r}_{t,2}^S = \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} v_{y,t}^1 - v_{y,t}^2 \\
\check{r}_{t,3}^S &= \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} 0.3 - p_{x,t}^1, \quad \check{r}_{t,4}^S = \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} 0.3 + p_{x,t}^1 \\
\check{r}_{t,5}^S &= \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} 0.3 - p_{z,t}^1, \quad \check{r}_{t,6}^S = \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} 0.3 + p_{z,t}^1
\end{aligned}$$

Similarly, we can evaluate $\check{c}_t^S = \min_i\{\check{c}_{t,i}^S\}_{i=1}^5$ by considering the following SOCPs, where each $\check{c}_{t,i}^S$ corresponds to a constraint function in (2.7) and (2.8):

$$\begin{aligned}
\check{c}_{t,i}^S &= \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} (-1)^i \times p_{x,t}^1 - p_{y,t}^1 + 0.05, \;\; i \in \{1,2\}, \\
\check{c}_{t,i}^S &= \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} (-1)^i \times p_{z,t}^1 - p_{y,t}^1 + 0.05, \;\; i \in \{3,4\}.
\end{aligned}$$

We overapproximate the maximum height difference between two drones via

$$\Delta_{z,t}^{21} := \max_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} p_{z,t}^2 - p_{z,t}^1, \tag{2.29}$$

and consider

$$\check{c}_{t,5}^S = \min_{x_t\in\mathcal{X}_{t,\bar{x}_0}^S} \left\| \begin{bmatrix} p_{x,t}^1 - p_{x,t}^2 \\ p_{y,t}^1 - p_{y,t}^2 \end{bmatrix} \right\|_2^2 - (1 + \max(\Delta_{z,t}^{21}, 0)) \times 0.2$$

Remark 1. *1) SOCP certification can be less conservative than Lipschitz certification when $\check{\mathcal{T}}$ and $\check{\mathcal{C}}$ can represent $\mathcal{T}$ and $\mathcal{C}$ exactly, and $\mathcal{X}^S_{t,\bar{x}_0}$ is a subset of the Lipschitz dynamically feasible set $\mathcal{X}^L_{t,\bar{x}_0}$, as defined in Section 2.6, for all $t \leq T$. This is because Lipschitz certification adds extra conservatism when estimating lower bounds of $r(x)$ and $c(x)$ using the Lipschitz constant, as shown in* (2.21) *and* (2.22)*; 2) However, Lipschitz certification is faster to compute than SOCP because calculating $\check{r}^L_t$ and $\check{c}^L_t$ is easier than evaluating $\check{r}^S_t$ and $\check{c}^S_t$; 3) SOCP certification employs convex over-approximations of the dynamically feasible state sets, similar to tube MPC [198]. However, it differs from tube MPC in that we compute the worst-case (nonconvex) constraint violation and target set deviation rather than the control inputs that optimize an objective.*

Remark 2. *The computational complexity of evaluating $\check{V}^L_\gamma(x,T)$ and $\check{V}^S_\gamma(x,T)$ scales **polynomially** with both the dimension of the dynamical system and the length of T.*

2.7 Combining reachability learning and certification

We integrate the reachability learning and certification into a new framework of computing trustworthy RA sets, as described in Algorithm 1. The super zero-level set of V_θ provides an estimation of the ground truth RA set. We use π_{θ_u} to certify a set of states, ensuring deterministic RA guarantees there. In particular, we can apply certification either online or offline:

Online certification: Let x be an arbitrary state. We can use the RA certificates $\check{V}^L_\gamma(x,T)$ in (2.23) or $\check{V}^S_\gamma(x,T)$ in (2.27) as online RA certification methods, verifying if all elements in $\mathcal{E}_x = \{x' : \|x' - x\|_2 \leq \varepsilon_x\}$ can reach the target set safely despite disturbances. We can compute $\check{V}^L_\gamma(x,T)$ and $\check{V}^S_\gamma(x,T)$ in real-time (10 Hz in our examples), as shown in Figure 2.5. It can also be integrated into the safety filter proposed in [133], offering more robust RA certification than [133] by verifying that all states in $\mathcal{E}_x$ can reach the target set safely, thereby enabling RA capability verification without perfect state estimation.

Offline certification: We consider a finite set of states $\mathcal{L} := \{x^{(i)}\}_{i=1}^N$ such that the union of their neighboring sets $\mathcal{E}_{x^{(i)}} = \{x : \|x - x^{(i)}\|_2 \leq \varepsilon_x\}$ covers the set of states that we aim at certifying. For example, this includes the area near the orange gate in drone racing, as show in Figure 2.1. We enumerate each element $x \in \mathcal{L}$ and certify whether $\mathcal{E}_x \subseteq \mathcal{R}$ by checking if $\check{V}^L_\gamma$ in (2.23) or $\check{V}^S_\gamma$ in (2.27) is positive. The union $\mathcal{S}$ of those certified sets constitutes a subset of the ground truth RA set $\mathcal{R}$. For all elements in $\mathcal{S}$, we guarantee that they can reach the target set safely under potential disturbances.

2.8 Experiments

We test our reachability learning and certification methods[2] in a 12-dimensional drone racing hardware experiment (Figure 2.3), and a triple-vehicle highway take-over simulation. In the

[2]Experiment code and hardware drone racing video are available at https://github.com/jamesjingqili/Lipschitz_Continuous_Reachability_Learning.git.

Algorithm 1: Certifiable Reachability Learning:

Require: an arbitrary $\gamma \in (0, 1)$, a finite list of states $\mathcal{L}$, and a certification horizon $T < \infty$;
Initialization: certified RA set $\mathcal{S} \leftarrow \{\}$;
Learn π_{θ_u}, ϕ_{θ_d} and V_θ via max-min DDPG [186];
`// Certifications:`
for $x_0 \in \mathcal{L}$ **do**
  if $\check{V}^L_\gamma(x_0, T) > 0$ or $\check{V}^S_\gamma(x_0, T) > 0$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x : \|x_0 - x\|_2 \leq \varepsilon_x\}$
return certified RA set $\mathcal{S}$

Initial states uniform sampling	Drone racing			Highway			
	DDPG-L	SAC-L	Ours	DDPG-L	SAC-L	CPO	Ours
In a large bounded state set	0.5716	0.7291	**0.7655**	0.7512	0.6343	0.5552	**0.8782**
In the learned reach-avoid set	0.6276	0.8006	**0.8889**	0.9430	0.9149	0.9566	**0.9924**
In the SOCP certified set	0.9673	0.9948	**1.0000**	0.9111	0.8850	0.9451	**1.0000**

Table 2.1: Success rates table. Our method achieves a 1.0 success rate when the initial states are sampled from the SOCP certified set. CPO fails to converge for the drone racing experiment due to the complex and nonconvex constraints.

highway simulation, we control one ego vehicle, modeled with nonlinear unicycle dynamics, to safely overtake another vehicle while avoiding a third vehicle driving in the opposite direction, as shown in Figure 2.4. Figures 2.1 and 2.4 demonstrate the high quality of the learned and certified RA sets.

Hypothesis 1: Our learned policy has a higher success rate than state-of-the-art constrained RL methods

We compare our learned policy π_{θ_u} with Deep Deterministic Policy Gradient-Lagrangian (DDPG-L)[62], Soft Actor Critic-Lagrangian (SAC-L) [321], and Constrained Policy Optimization (CPO) [1]. We summarize the results in Table I. The success rate is estimated by computing the ratio of sampled initial states that can reach the target set safely under randomly generated disturbances. Our method achieves a 1.0 success rate when the initial states are sampled from the SOCP certified set, validating the deterministic guarantee that all elements in the certified sets can safely reach the target set despite disturbances.

Hypothesis 2: Our online RA set certification methods can be computed in real-time

Figure 2.5 shows that $\check{V}^L_\gamma(x, T)$ and $\check{V}^S_\gamma(x, T)$ can be computed in real-time to certify if all elements in $\mathcal{E}_x = \{x' : \|x' - x\|_2 \leq 0.1\}$ can safely reach the target set, under all potential disturbances. This

(a) Ego drone's initial (*x, y, z*) position: (-0.5, -2.5, 0.0) (b) Ego drone's initial (*x, y, z*) position: (0.9, -2.3, 0.0)

Figure 2.3: We sampled 50 initial states from the SOCP certified set shown in Figure 2.1. A few crashes occurred due to insufficient battery charge or Vicon sensor failures caused by natural light. These instances were excluded as outliers. With a fully charged battery and no Vicon system failures, the ego drone successfully overtook the other drone from each of the 50 initial states, despite the latter's uncertain acceleration. We visualize two hardware experiments in the above subfigures. The remaining 9-dimensional initial state includes $[v^1_{x,t}, v^1_{y,t}, v^1_{z,t}, p^2_{x,t}, p^2_{y,t}, p^2_{z,t}, v^2_{x,t}, v^2_{y,t}, v^2_{z,t}] = [0, 0.7, 0, 0.4, -2.2, 0, 0, 0.3, 0]$.

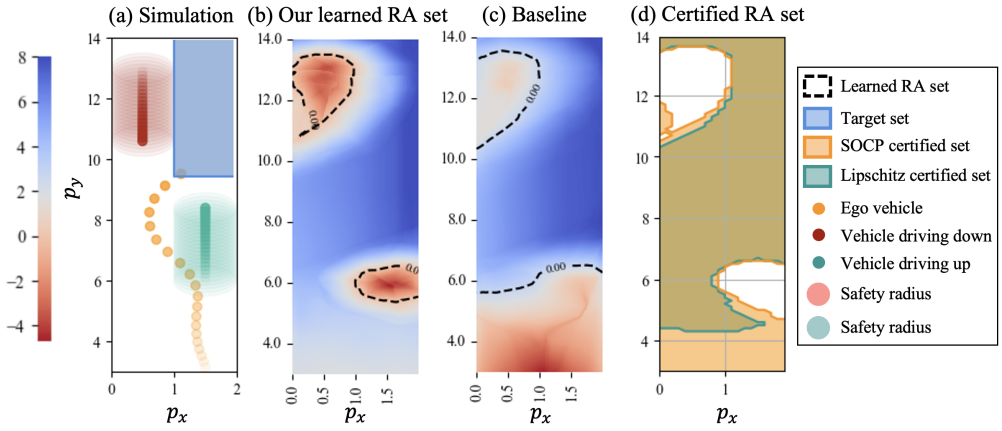


Figure 2.4: Highway reachability analysis: In (a), we simulate the nonlinear dynamics with the learned policy π_{θ_u} and randomly sampled disturbances on other vehicles' acceleration. The 10-dimensional state space includes $[p^1_{x,t}, p^1_{y,t}, v^1_t, \theta^1_t, p^2_{x,t}, p^2_{y,t}, v^2_{y,t}, p^3_{x,t}, p^3_{y,t}, v^3_{y,t}]$. The p_y-axis movement of the red and green agents is modeled using double integrator dynamics, while their initial p_x positions are sampled randomly and remain stationary during simulation. In (b), we project our learned value function, with $\gamma = 0.95$, onto the (x, y) position of the ego vehicle. In (c), we plot the RA set learned using the state-of-the-art method [190, 234] with $\gamma = 0.95$. As suggested in [134], annealing $\gamma \to 1$ is necessary for prior works; otherwise, the learned RA sets in prior works are conservative. In (d), we plot our certified RA sets.

enables real-time online certification.

Hypothesis 3: The Lipschitz continuity of our new value function appears to accelerate learning

In Figure 2.6, we compare the critic loss (defined in (2.17) to measure the Bellman equation error) under different Bellman equations and time-discount factor γ values. Our critic loss converges rapidly when γ is chosen to ensure the Lipschitz continuity of our new value function.

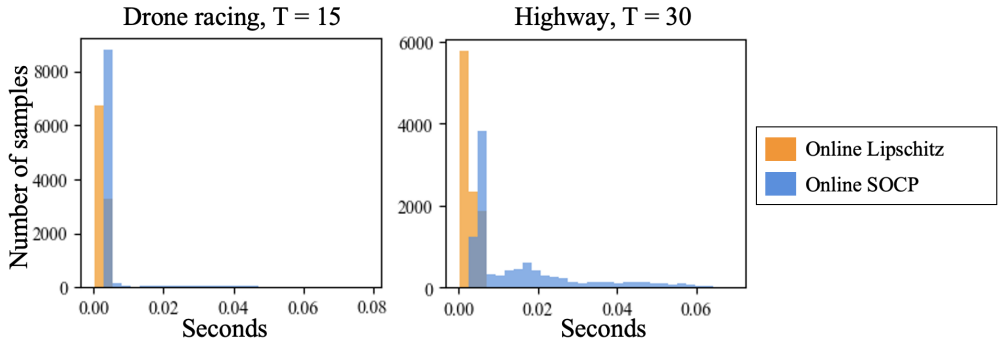


Figure 2.5: Histogram of the time required for computing $\check{V}_\gamma^L(x,T)$ and $\check{V}_\gamma^S(x,T)$ for each of the 10,000 randomly sampled states x. The certification horizons for drone racing and highway are $T = 15$ and $T = 30$, respectively.

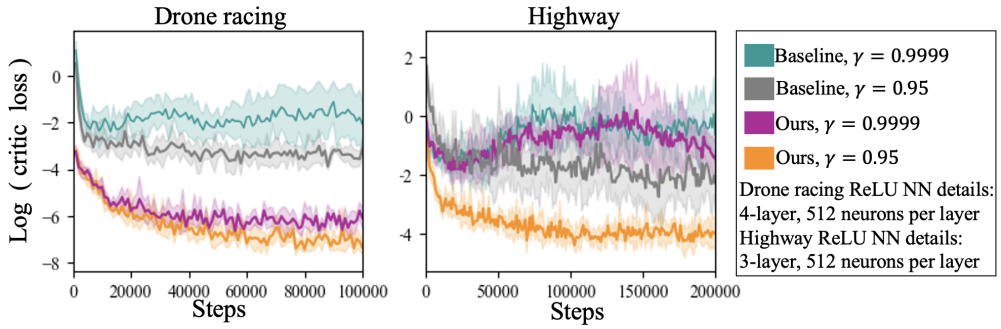


Figure 2.6: We compare the convergence of the critic loss under our new Bellman equation with the baseline from previous works [234, 190], using different γ values but identical training parameters. Our critic loss with $\gamma = 0.95$ converges faster than with $\gamma = 0.9999$, likely due to the Lipschitz continuity of $V_\gamma(x)$ at $\gamma = 0.95$. The training speed is around 1700 steps per second.

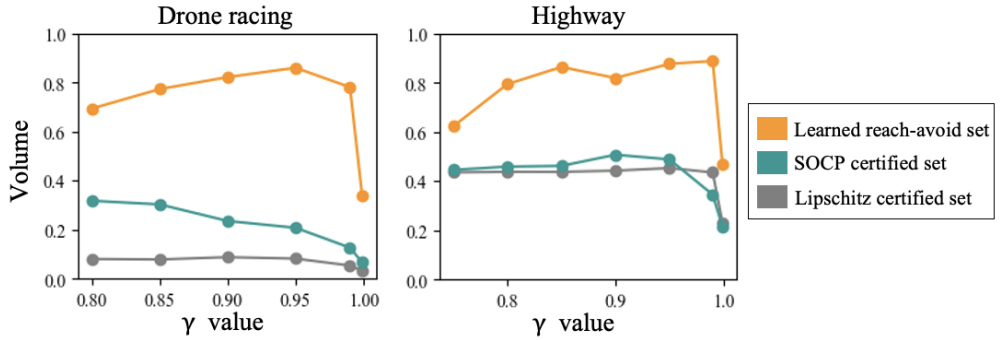


Figure 2.7: The volumes of the learned RA set, SOCP certified set, and the Lipschitz certified set change as γ varies. We estimate the set volumes using the Monte Carlo method with 10,000 random samples in the state space.

The trade-off of selecting a time-discount factor γ

We summarize our result in Figures 2.7 and 2.8. With a small γ, the learned RA sets can be conservative due to numerical errors, as an initial state whose optimal trajectory reaches the target

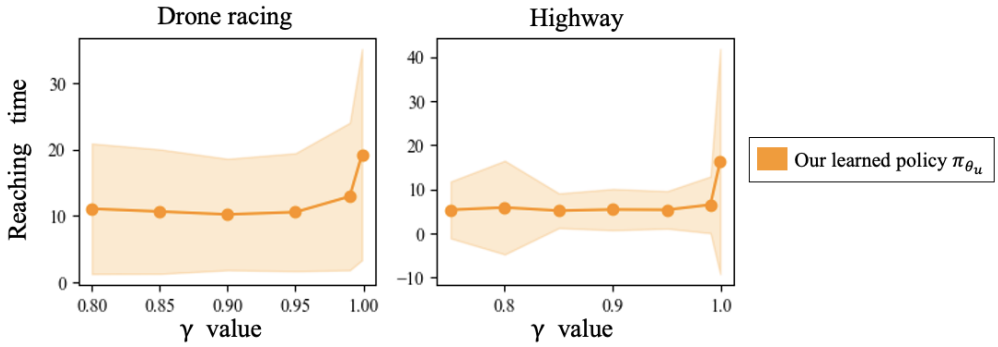


Figure 2.8: The average time taken for reaching the target set grows as γ increasing.

set at a later stage may have near-zero time-discounted RA measures. However, the optimal policy tends to drive the state to the target set rapidly, as depicted in Figure 2.8. Conversely, a large γ can induce discontinuities in the value function, destabilizing learning and yielding suboptimal solutions. In the drone racing and highway experiments, we find that $\gamma = 0.95$ ensures the Lipschitz continuity of $V_\gamma(x)$, thereby enhancing learning efficiency, and also mitigates unnecessary conservatism.

2.9 Conclusion and Future Work

We propose a new framework for learning trustworthy reach-avoid (RA) sets. Our method features a newly designed RA value function that offers improved computational efficiency. We employ max-min DDPG to learn our value functions and propose two efficient methods to certify whether a set of states can safely reach the target set with deterministic guarantees. We validate our methods through drone racing hardware experiments and highway take-over simulations. Our certification methods can be performed in real time, but they rely on offline value function learning beforehand. Future research may explore online reachability learning, as well as more efficient RA set certification methods.

Chapter 3

Augmented Lagrangian Safe Reinforcement Learning

In this chapter, we extend our previous work on reach-avoid reinforcement learning, which focuses on safely reaching a target set, to address more general constrained optimal control problems. In this broader setting, the objective can encode complex task requirements, including trade-offs between competing goals, while ensuring that safety constraints are satisfied. To achieve this, we adapt the augmented Lagrangian method from classical convex optimization to the context of safe reinforcement learning. This chapter is based on the published work [179], co-authored with David Fridovich-Keil, Somayeh Sojoudi, and Claire J. Tomlin.

3.1 Background

Deep reinforcement learning algorithms have achieved state-of-the-art performance in many domains [216, 177, 118]. In standard reinforcement learning (RL), the ultimate goal is to optimize the expected sum of rewards or costs, and the agent can freely explore in order to improve the current policy. RL methods have been widely used to learn optimal policies for agents with complicated or even unknown dynamics. RL has successfully solved a wide range of tasks, including the game of Go [280], robotic control [77], and traffic control [310].

There is a well-known trade-off between exploration and exploitation in RL. To optimize the overall reward, the agent must balance whether to take a sequence of actions similar to what it has already tried (i.e., exploitation) or to try a new combination of actions (i.e., exploration). Since most RL problems are non-convex, pure exploitation leads to a suboptimal policy leading to a poor local maximum of the reward function. To encourage the agent to find a better policy, various methods have been proposed for promoting exploration, such as using an Upper Confidence Bound [14], Inverse Entropy [124], or designing a Variational Auto-encoder [10]. Nevertheless, in many applications such as autonomous driving [277] and surgical robotics [85], exploration can be dangerous because violating certain constraints even by a small amount may have significant consequences. Thus, ensuring safety is of great importance in real-world applications.

A natural way of encoding safety in RL is through constraints. Here, there are two types of constraints: cumulative constraints (e.g., average vehicle speed) and instantaneous constraints (e.g., collision avoidance at each time). A cumulative constraint requires that an infinite-horizon or a finite-horizon discounted sum of a constraint cost function lie within a certain bound. By contrast, an instantaneous constraint must hold at all time instants. For both problems, the horizon could be either infinite or finite.

One common formulation of RL with cumulative constraints is the Constrained Markov Decision Process (CMDP) framework [7], where the agent optimizes an objective while satisfying constraints on the expectation of an infinite-horizon discounted sum of auxiliary costs. A classical approach to solving CMDPs is the Lagrangian dual method [7]. The Lagrangian approach allows us to transform a constrained control problem to an equivalent minmax unconstrained control problem. Recently, it has been shown that under certain regularity conditions there is no duality gap for infinite-horizon RL problems with cumulative constraints, despite their non-convex nature [245]. This result theoretically justifies the effectiveness of popular Lagrangian-relaxation-based CMDP algorithms, such as Constrained Policy Optimization (CPO) [1], Primal-Dual Policy Optimization (PDO) [63] and Lyapunov-based safe learning [61].

As will be shown in Section 3.2, the satisfaction of cumulative constraints may not lead to the satisfaction of instantaneous constraints. Therefore, it is crucial to develop methods for solving instantaneously constrained RL problems. The authors of [98] propose to solve instantaneously constrained RL problems by optimizing a smoothed version of the worst constraint violation rather than an explicitly constrained objective. One line of work devoted to safe RL with instantaneous constraints is projection-based Safe RL [119, 252, 116], where at each step the agent selects one action from a pre-computed safe action set. However, one potential drawback of this approach is that the pre-computed safe action set could be conservative, leading to a suboptimal policy [151, 25].

In this chapter, we consider an infinite-horizon optimal control problem with instantaneous safety constraints. We adapt the classical Augmented Lagrangian method [238] to obtain a safe policy satisfying instantaneous safety constraints. Our work is closely related to [197], where an interior-point method is adapted to solve the safe RL problem. One major difference is that we relax the assumption of an initial safe policy, which is required in [197].

We first extend the strong duality results in [245] to instantaneously constrained RL, and propose a sufficient condition for the strong duality of the instantaneously constrained RL problem. Inspired by the Augmented Lagrangian method, we then design a surrogate objective function, and we show that under certain conditions on the feasible policy set, the policy returned by optimizing the surrogate function converges to an optimal policy for the original problem. We propose a primal-dual algorithm for optimizing this surrogate function and our empirical results show that the proposed method is more data-efficient than the existing Lagrangian dual method. Our empirical results also suggest that this method reduces the total constraint violation, highlighting the potential of our method for promoting safety throughout learning.

The rest of the chapter is organized as follows. In Section 3.2, we formulate the Instantaneously Constrained RL problem. We present our main theoretical results in Sections 3.3 and 3.4, with proofs provided in the Appendix. In Section 3.5, we present three illustrative examples: a tabular

learning example, and an OpenAI constrained pendulum and half-cheetah example. Finally, we conclude and discuss future directions in Section 3.6.

3.2 From Cumulative to Instantaneous Constraints

In this section, we first review Constrained Markov Decision Processes, and then motivate and introduce our instantaneously constrained RL problem formulation.

A Markov Decision Process (MDP) is a tuple $(\mathcal{Z}, \mathcal{A}, \gamma, r, p_z, p_0)$, where $\mathcal{Z}$ and $\mathcal{A}$ are compact state and action spaces, $\gamma \in [0, 1)$ is a discounting factor, $r(z, a) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ is the immediate cost function, $p_z(\cdot|z, a)$ is the transition probability distribution density, and p_0 is the initial state distribution density. In addition, let $g(z, a) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ be the constraint function. A function $f : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ is bounded if there exists a constant $c \in \mathbb{R}$ such that $f(z, a) \leq c$, for $\forall (z, a) \in \mathcal{Z} \times \mathcal{A}$. The agent chooses actions sequentially based on a policy $\pi \in \mathcal{P}(\mathcal{Z})$, where $\mathcal{P}(\mathcal{Z})$ is the space of probability measures on $(\mathcal{A}, \mathcal{B}(\mathcal{A}))$ parametrized by elements of $\mathcal{Z}$, where $\mathcal{B}(\mathcal{A})$ are the Borel sets of $\mathcal{A}$.

A Constrained Markov Decision Process was introduced in [7] by incorporating an additional inequality constraint:

$$\begin{aligned} \max_{\pi} \ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi\right] \\ \text{s.t.} \ & z_{t+1} \sim p_z(\cdot|z_t, a_t), a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0, \\ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t) \middle| \pi\right] \leq 0. \end{aligned} \tag{3.1}$$

where $\mathbb{E}[\cdot]$ is the expectation operator. The z_t and a_t are state and action at time $t \in \{0, 1, \dots\}$, respectively.

In what follows, we will illustrate with a simple 2D example that a cumulative constraint does not generally provide any guarantees for the associated instantaneous constraints, i.e., solving CMDPs may not be sufficient to ensure the satisfaction of instantaneous constraints.

Example 1. *Consider a linear dynamical system $z_{t+1} = Az_t + Ba_t$, where $A \in \mathbb{R}^{2\times 2}$ and $B \in \mathbb{R}^{2\times 1}$ are specified as*

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \ B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{3.2}$$

Let $\mathcal{K} \subseteq \mathbb{R}^2$ be the feasible policy class. Given an initial point $\check{z}_0 \in \mathbb{R}^2$, we consider the infinite-

horizon constrained optimal control problem

$$\begin{aligned} K^* := \arg\max_{K\in\mathcal{K}} \Bigg[& -\sum_{t=0}^{\infty} \big(z_t^\top Q z_t + a_t^\top R a_t \big) \Bigg] \\ & \textit{s.t. } z_{t+1} = A z_t + B a_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad a_t = K z_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad z_0 = \check{z}_0, \\ & \quad \sum_{t=0}^{\infty} z_t \leq 0 \end{aligned} \tag{3.3}$$

and the instantaneously constrained RL problem

$$\begin{aligned} \tilde{K}^* := \arg\max_{K\in\mathcal{K}} \Bigg[& -\sum_{t=0}^{\infty} \big(z_t^\top Q z_t + a_t^\top R a_t \big) \Bigg] \\ & \textit{s.t. } z_{t+1} = A z_t + B a_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad a_t = K z_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad z_0 = \check{z}_0, \\ & \quad z_t \leq 0, \quad \forall t \in \{0,1,\dots\}. \end{aligned} \tag{3.4}$$

with the parameters $Q = I_2$ and $R = 1$. If we assume the policy class to be $\mathcal{K} = \mathbb{R}^2$, then the optimal feedback matrix K^* may be found by solving the well-known LQR Riccati equation and recognizing that the constraint $\sum_{t=0}^{\infty} z_t \leq 0$ in (3.3) *is inactive for K^*. Pick $\tilde{K} \in \mathbb{R}^2$ such that $(A + B\tilde{K})$ has real positive eigenvalues with magnitude strictly smaller than 1 and $(A + B\tilde{K})$ has two eigenvectors $v_1 \leq 0$ and $v_2 \leq 0$ whose convex hull contains $\check{z}_0$. By Proposition 4 in the Appendix, we can show that $\tilde{K}$ is a feasible solution for* (3.4)*. We plot the state trajectories under the two feedback controllers $a_t = K^* z_t$ and $\tilde{a}_t = \tilde{K}\tilde{z}_t$. In Figure 3.1, the state trajectory under the controller $a_t = K^* z_t$ violates the constraint $z_t \leq 0$ at time $t = 2$ while the trajectory under $\tilde{a}_t = \tilde{K}\tilde{z}_t$ does not.* ■

As illustrated in Example 1, enforcing a constraint cumulatively does not imply that it holds at each time. We emphasize that constraints may be arbitrary functions of state. In this way, an instantaneous constraint may be understood to encode desired safety configurations, such as in collision avoidance [153], human-robot interaction [195], and aerospace control [2]. Motivated by the above discussion, we formulate the Instantaneously Constrained RL problem as follows.

Problem 1 (Instantaneously Constrained RL Problem)**.** *Consider an MDP with transition dynamics $z_{t+1} \sim p_z(\cdot|z_t, a_t)$ and initial state distribution p_0, along with a bounded reward function $r(z, a)$ and a bounded constraint function $g(z, a)$. The objective is to find a policy π^* that solves the*

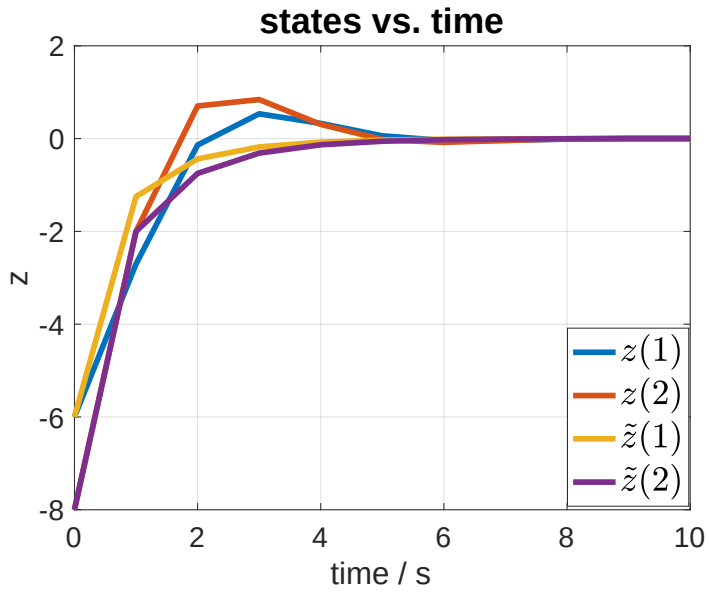


Figure 3.1: State trajectories comparison between the two controllers $a_t = K^* z_t$ and $\tilde{a}_t = \tilde{K}\tilde{z}_t$.

following constrained optimization problem over the infinite horizon:

$$\begin{aligned} \max_{\pi} \ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi\right] \\ \textit{s.t.} \ & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0, \\ & \mathbb{E}[g(z_t, a_t)|\pi] \leq 0, \quad \forall t \in \{0, 1, \dots\}. \end{aligned} \tag{3.5}$$

We remark here that an optimal policy feasible for (3.5) could be a conservative but feasible solution for (3.1). Therefore, a policy learned from (3.5) is also safe with respect to the constraint in (3.1). In addition, although we only consider one set of instantaneous constraints in (3.5), the results of this chapter could be extended to the general case with multiple sets of instantaneous constraints, by associating each constraint with a Lagrange multiplier and carrying out an analysis similar to the single constraint case (3.5).

3.3 Augmented Lagrangian Surrogate Function

In this section, we introduce our Augmented Lagrangian Surrogate Function. We first propose a sufficient condition under which strong duality holds for (3.5), and then design a new surrogate function which could promote safety during the learning phase.

Since most of the existing results on RL deal with unconstrained problems, it is beneficial to work with the unconstrained Lagrangian dual of the primal problem (3.5) given below,

$$\begin{aligned} \min_{\substack{\{\lambda_t\}_{t=0}^{\infty} \\ \lambda_t \leq 0}} \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda_t g(z_t, a_t)) \middle| \pi\right] \\ \text{s.t. } z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ z_0 \sim p_0. \end{aligned} \tag{3.6}$$

where λ_t is the Lagrange multiplier associated with the scalar constraint $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$.

It is known that strong duality holds in the case of cumulative constraints [245]. For completeness, we first introduce Assumption 2, and then build on the result of [245].

Assumption 1. *Suppose that the feasible policy set for* (3.5) *has a non-empty relative interior. Furthermore, suppose that for any π satisfying $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$, π also satisfies $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$.*

Remark 3. *We note that, under Assumption 1, Problem 1 could be equivalently considered as a special subclass of CMDPs in which the cumulative constraints could approximate instantaneous constraints. In what follows, we will show that Assumption 1 permits us to characterize the strong duality of Problem 1, and thereby design a new surrogate objective for* (3.5) *which yields superior empirical performance than existing primal-dual approach.*

Proposition 1. *Under Assumption 1, strong duality holds for* (3.5)*.*

A natural question that arises is whether Assumption 1 is stringent. To ensure the condition $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$ implies $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, for all $t \in \{0, 1, \dots\}$, we propose two approaches. First, we propose a "clipping" method whereby constraint values at safe states are set to zero. For example, suppose that we have an instantaneous constraint $\mathbb{E}[h(z_t, a_t)|\pi] \leq 0$ with a bounded function $h(z_t, a_t) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$, computing only the positive part, i.e., $\mathbb{E}[\text{Relu}(h(z_t, a_t))|\pi] \leq 0$, where $\text{Relu}(\cdot) : \mathbb{R} \to \mathbb{R}$ is defined as $\text{Relu}(x) = x$ if $x \geq 0$ and 0 if $x < 0$. The choice of the Relu function is not strictly necessary, i.e., it could be replaced by other non-negative activation functions such as Softplus or Sigmoid [258]. The other approach is to restrict the feasible policy class, as highlighted in the following 2D example:

Example 1 (Continued)**.** *Suppose that the policy class $\mathcal{K} \subseteq \mathbb{R}^2$ is such that for any $K \in \mathcal{K}$, the closed loop dynamics $(A + BK) \in \mathbb{R}^{2\times 2}$ has two real positive eigenvalues, and it has two eigenvectors $v_1 \leq 0$ and $v_2 \leq 0$ whose convex hull contains the point z_0. We will show in Proposition 4 in the Appendix that under any policy $K \in \mathcal{K}$, the constraint $\sum_{t=0}^{\infty} z_t \leq 0$ and the condition $z_t \leq 0$, $\forall t \in \{0, 1, \dots\}$, are always satisfied. In this simple instance where $g(z_t, a_t) \equiv z_t$, the constraint $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$ implies the constraint $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, for all $t \in \{0, 1, \dots\}$.*

Remark 4. *However, as indicated in [245], strong duality is only proved for CMDPs with arbitrary stochastic policies. Characterizing strong duality in parametric, restricted policy classes such as that of Problem 1 is an important direction for future research.*

Building upon the Lagrangian dual (3.6), by the linearity of the expectation operator, at each time $t \in \{0, 1, \dots\}$, (3.6) suggests an instantaneous reward function $r_t(z_t, a_t) = r(z_t, a_t) + \lambda_t g(z_t, a_t)$. This function depends upon the Lagrange multiplier λ_t and hence is time-varying. However, infinite-horizon RL algorithms typically assume time-invariant reward functions. We next show that, under Assumption 1, a set of optimal Lagrange multipliers $\{\lambda_t^*\}$ could share the same value. That is, we may presume that all $\{\lambda_t\}_{t=0}^{\infty}$ are equal to some constant λ, and therefore obtain a time-invariant reward function. This time-invariance permits us to apply existing RL algorithms to find the best policy maximizing the time-invariant instantaneous reward function.

Proposition 2. *Let* $(\{\lambda_t\}_{t=0}^{\infty}, \pi^*)$ *be an optimal solution of* (3.6)*. Let* $(\lambda^*, \tilde{\pi}^*)$ *be a pair of optimal solutions to*

$$\begin{aligned} \min_{\lambda \leq 0} \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(r(z_t, a_t) + \lambda^\top \left(\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t) \right) \right) \middle| \pi \right] \\ \textit{s.t. } z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ z_0 \sim p_0. \end{aligned} \tag{3.7}$$

Let $\tilde{\lambda}_t = \lambda^*$*, for all* $t \in \{0, 1, \dots\}$*. Under Assumption 1, we have that* $(\{\tilde{\lambda}_t\}_{t=0}^{\infty}, \tilde{\pi}^*)$ *is also a pair of optimal solution to* (3.6)*.*

Remark 5. *Proposition 2 does not preclude the existence of optimal Lagrange multipliers* $\{\lambda_t^*\}_{t=0}^{\infty}$ *of* (3.6) *which are time-varying.*

Building upon the above results and by assuming $\lambda_t = \lambda$, for all $t \geq 0$, we design the time-invariant instantaneous reward inspired by Augmented Lagrangian Method [238],

$$\tilde{r}(z_t, a_t) := r(z_t, a_t) + \lambda \cdot \text{Relu}(g(z_t, a_t)) - \frac{\rho}{2} \cdot \text{Relu}(g(z_t, a_t))^2, \tag{3.8}$$

and subsequently we obtain the infinite-horizon objective function

$$\begin{aligned} R(\pi, \lambda, \rho) := \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda \cdot \text{Relu}(g(z_t, a_t)) \\ - \frac{\rho}{2} \cdot \text{Relu}(g(z_t, a_t))^2 \Big) \Bigg| \pi \Bigg]. \end{aligned} \tag{3.9}$$

Remark 6. *Under Assumption 1,* (3.9) *can be interpreted as a new surrogate function for a special subclass of CMDPs in which the cumulative constraints could approximate instantaneous constraints. We will show in Section 3.4 that by optimizing* (3.9)*, we can find a high-quality policy within fewer iterations and smaller constraint violation throughout learning than a current primal-dual method. That is, for this special subclass of CMDPs,* (3.9) *serves as an alternative surrogate function with a superior empirical performance than the existing primal-dual method.*

Algorithm 2: Augmented Lagrangian RL

Pick $c_\rho \in [1, \infty)$, dual ascent stepsize $\ell \in \mathbb{R}_+$, and convergence tolerance $\varepsilon > 0$;
Initialize $\rho^{(0)} \in \mathbb{R}_+$, $\lambda^{(0)} = 0$;
Randomly initialize the policy π_0;
for $k = 0, 1, 2, \ldots$ **do**
  $\pi_k \leftarrow \arg\max_\pi R(\lambda^{(k)}, \rho^{(k)}, \pi)$;
  $\lambda^{(k+1)} \leftarrow \Big\lfloor \lambda^{(k)} - \ell\Big(\mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t \text{Relu}(g(z_t, a_t))|\pi_k\big]\Big)\Big\rfloor_-$
  $\rho^{(k+1)} \leftarrow c_\rho \rho^{(k)}$;
  return π_k if $||\pi_k - \pi_{k-1}||_\infty \leq \varepsilon$.

Notice that $R(\pi, \lambda, 0)$ is not equivalent to the objective function in (3.7), $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda g(z_t, a_t))|\pi]$, because the constraint $\mathbb{E}[\text{Relu}(g(z_t, a_t))|\pi] \leq 0$, is a sufficient but not necessary condition for the constraint $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$.

We will show in Section 3.4 that under certain conditions, as $\rho \to \infty$, any infeasible policy would become sub-optimal when we maximize the function $R(\pi, \lambda, \rho)$, with λ fixed. Thus, an optimal policy returned by optimizing (3.9) for both π and λ would eventually become safe and optimal as we increase ρ. In addition, we remark here that the introduction of the Relu function or other non-negative activation functions in (3.9) is necessary because otherwise, it is not generally true that an optimal policy for (3.9) is also optimal for (3.5), due to the fact that under an optimal policy π^* of problem (3.5), $\mathbb{E}[g(z_t, a_t)^2|\pi^*]$ may be nonzero and therefore $R(\pi, \lambda, \rho) \to -\infty$, as $\rho \to \infty$.

Following the same spirit of the primal-dual algorithm in constrained optimization [238, 245, 63], we propose Algorithm 2.

In Algorithm 2, we initialize $\lambda^{(0)} = 0$ and $\rho^{(0)} \in \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of non-negative real numbers. At the k-th iteration, we first find a policy $\pi_k \in \arg\max_\pi R(\lambda^{(k)}, \rho^{(k)}, \pi)$, which could be done by any unconstrained RL algorithm in the literature (e.g., SAC [121], DDPG [192], TRPO [272]). Then, we update the Lagrange multiplier by dual ascent $\lambda^{(k+1)} = \lfloor \lambda^{(k)} - \ell(\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \text{Relu}(g(z_t, a_t))|\pi_k]) \rfloor_-$, where the function $\lfloor \cdot \rfloor_- : \mathbb{R} \to \mathbb{R}_-$ is defined as follows:

$$\lfloor x \rfloor_- = \begin{cases} 0 & \text{if } x > 0, \\ x & \text{otherwise.} \end{cases} \tag{3.10}$$

We also update $\rho^{(k+1)} = c_\rho \rho^{(k)}$, where $c_\rho \in [1, \infty)$ is the increasing rate of the quadratic penalty coefficient $\rho^{(k)}$ as the iteration index k grows.

3.4 Convergence Analysis

In this section, we show that under certain conditions on the feasible policy set, by optimizing the surrogate function (3.9) we recover an optimal policy for (3.5).

Proposition 3. *Under Assumption 1, consider the primal maximization of* (3.7)*, denoted by* $d_\rho(\lambda) : \mathbb{R} \to \mathbb{R}$ *and defined as*

$$\begin{aligned} d_\rho(\lambda) := \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda \cdot \textit{Relu}(g(z_t, a_t)) - \frac{\rho}{2} \cdot \textit{Relu}(g(z_t, a_t))^2) \middle| \pi \right] \\ \textit{s.t. } z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ z_0 \sim p_0. \end{aligned} \tag{3.11}$$

Let $\lambda^*_\rho := \arg\min_{\lambda \leq 0} d_\rho(\lambda)$*. Suppose that under an optimal policy* π^* *of problem* (3.5)*,* $g(z_t, a_t) \leq 0$*,* $\forall t \in \{0, 1, \dots\}$*. We define a policy* $\pi^*_\rho(\lambda)$ *as*

$$\begin{aligned} \pi^*_\rho(\lambda) := \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r(z_t, a_t) + \lambda \cdot \textit{Relu}(g(z_t, a_t)) - \frac{\rho}{2} \cdot \textit{Relu}(g(z_t, a_t))^2 \Big) \middle| \pi \right] \\ \textit{s.t. } z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ z_0 \sim p_0. \end{aligned} \tag{3.12}$$

Then, as $\rho \to \infty$*, we have,*

$$\left\| \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi^* \right] - \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi^*_\rho(\lambda^*_\rho) \right] \right\|_2 \to 0. \tag{3.13}$$

The condition that under an optimal policy π^* in the original problem (3.5), $g(z_t, a_t) \leq 0$, $\forall t \in \{0, 1, \dots\}$, is equivalent to the condition $\mathbb{E}[g(z_t, a_t)|\pi^*] \leq 0$, $\forall t \in \{0, 1, \dots\}$, if we have a deterministic dynamical system. In addition, this condition could be easy to meet for safety-critical systems, due to the fact in many control applications we have a safe but sub-optimal base controller, e.g., Autopilot [270], safe robot-human interaction [230], autonomous driving [115].

We remark here that the analysis in Proposition 3 is conservative. However, in Section 3.5 we consider instantaneous constraints g which we do not know a priori are deterministically satisfiable for each t. That is, we consider g for which there may not be a policy π for which $g(z_t, a_t) \leq 0, \forall t$. Still, our empirical results suggest that when the parameter ρ is sufficiently large, Algorithm 2 returns a high-quality safe policy.

3.5 Experiments

We validate Algorithm 2 in experiments with different initial values of ρ in the settings of a tabular MDP [286], inverted pendulum [39], and half-cheetah [39]. In all experiments, we assume that the constraints are of the form $\mathbb{E}[\text{Relu}(h(z_t, a_t))|\pi] \leq 0$ for some function $h : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$, and

therefore when $\rho_0 = 0$, $R(\pi, \lambda, 0)$ recovers the classical Lagrangian dual method adopted in [7, 1, 63, 61, 245].

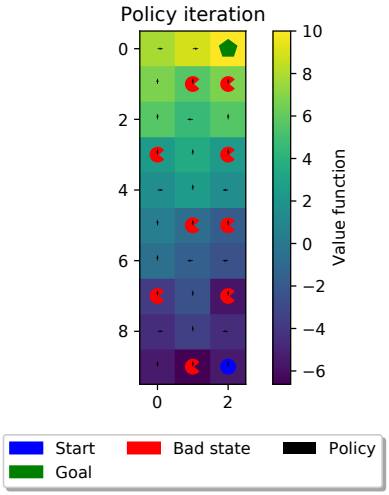
We first consider a constrained tabular MDP in Figure 3.2a, where we have 10×3 states, each corresponding to a grid cell of a table. The agent starts from an initial state and tries to reach the goal state. At each grid cell, the agent can stay at the same cell or move up, down, left, or right. For those grid cells on the boundary, no action moving out of the table is permitted. The constraint function $g(z_t, a_t)$ takes the value 1 if z_t is considered unsafe and 0 otherwise. The agent receives a reward $r(s, a) = 10$ for reaching the goal state (which is terminal) and a reward $r(s, a) = -1$ otherwise. In this experiment, we keep the quadratic penalty coefficient fixed at each iteration in Algorithm 2, and therefore we pick the parameter $c_\rho = 1$. At the k-th iteration of Algorithm 2, we apply the classical tabular Policy Iteration [286] to find the policy π_k.

In Figure 3.2b, we show that the duality gap eventually goes to zero as we update the Lagrange multiplier at each iteration, which empirically validates Proposition 1. In Figure 3.2c, we observe that as ρ_0 grows, the speed at which the policy returned by Algorithm 2 converges to the optimal policy increases. In Figure 3.2d, we measure the accumulated constraint $\sum_{t=0}^{\infty} \mathbb{E}[g(z_t, a_t)|\pi]$, and we observe that it decreases as we increase ρ_0. This implies that the surrogate function (3.9) could promote safety during learning, compared with the case $\rho_0 = 0$, i.e., the Lagrangian dual approach in [7, 1, 63, 61, 245].

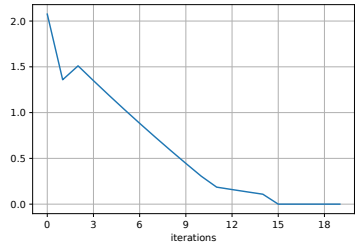
Subsequently, we consider a constrained pendulum example, where we add an additional constraint corresponding to avoiding collision with an obstacle near the pendulum, i.e., $\theta_t \notin [\frac{\pi}{2}, \pi]$, to the OpenAI Gym "Pendulum-v0" environment [39]. To satisfy Assumption 1, we reformulate this constraint as $\mathbb{E}[g(\theta_t)|\pi] \leq 0$, where $g(\theta) = 1$ if $\theta \in [\frac{\pi}{2}, \pi]$ and $g(\theta) = 0$ otherwise. Unlike the previous tabular MDP example where we can find a globally optimal policy, we may only obtain a locally optimal policy due to the non-convexity of RL problems. In line 2 of Algorithm 2, we find a locally optimal policy by running a fixed number of steps of Deterministic Deep Policy Gradient (DDPG) [192]. By picking $c_\rho = 1.15$, we slowly increase the quadratic penalty coefficient ρ as the iteration number grows. We update the parameters λ and ρ almost after 6×10^4 steps of DDPG, as indicated by the vertical dashed lines in Figure 3.3.

We run experiments with different random seeds and show the average performance and the standard deviation in Figure 3.3. We see that as ρ_0 increases, the rate at which the policy converges increases and the constraint violations decreases in Figure 3.3. In particular, we observe that the standard deviation of the constraint violations dramatically decreases as ρ_0 grows in Figure 3.3b, which suggests that optimizing the surrogate function (3.9) promotes both safety and stability during learning.

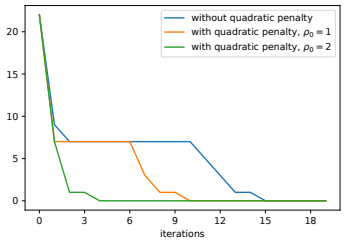
Finally, we consider the constrained half-cheetah example [197], which is adapted from the OpenAI gym "Half-cheetah-v0" environment [39] by adding an additional constraint $|v_x(t)| \leq 1$ on the horizontal velocity of the cheetah. We reformulate the constraint as $\mathbb{E}[\text{Relu}(|v_x(t)| - 1)|\pi] \leq 0$. In line 2 of Algorithm 2, we find a locally optimal policy by running a fixed number of steps of Soft Actor Critic (SAC) [121]. By picking $c_\rho = 1.5$, we rapidly increase the quadratic penalty coefficient ρ as the iteration number grows. Similar to the constrained Pendulum experiments, we update λ and ρ every 1.6×10^5 steps of SAC, as indicated by the vertical dashed lines in Figure 3.4. As we increase ρ_0, we see in Figure 3.4 that the speed at which the policy converges increases and the



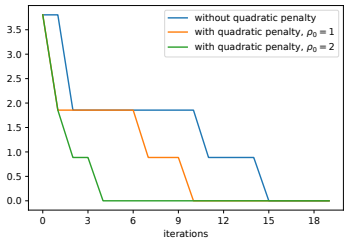
(a) tabular MDP Experiment



(b) duality gap



(c) distance between the learned policy and the optimal policy vs. iterations



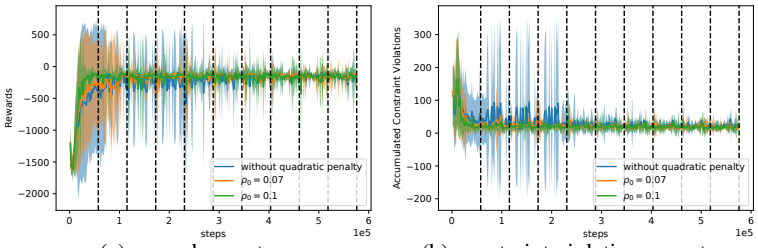
(d) constraint violations vs. iterations

Figure 3.2: Tabular MDP Results

constraint violations decreases. However, when ρ_0 is too large, e.g., $\rho_0 = 2.0$ in Figure 3.3a, we observe that the exploration is inhibited, which suggests that by tuning ρ_0 we could control the trade-off between exploration and safety. We also plot the state trajectory under polices learned from Algorithm 2 with different values of ρ_0 in Figure 3.5. We observe that as long as ρ_0 is sufficiently large, the learned policies perform safely with a similar performance quality. When ρ_0 is too large, the learned policy becomes conservative possibly due to poor exploration.

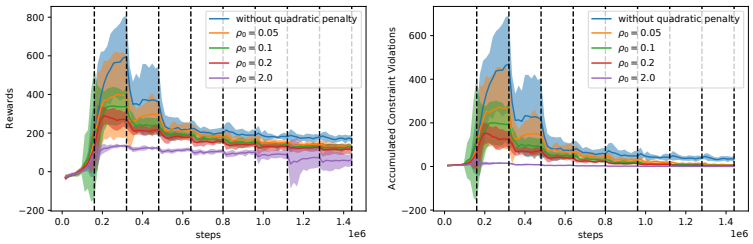
3.6 Conclusion

In this chapter, we considered Instantaneously Constrained RL problems. We first extended a recent result on Cumulatively Constrained RL problems to characterize the strong duality of Instantaneously Constrained RL problems. Inspired by the Augmented Lagrangian method, we proposed a new surrogate function that can promote safety for instantaneous constraints, i.e., reducing the constraint violations during learning. Our surrogate function can be optimized using



(a) rewards vs. steps (b) constraint violations vs. steps

Figure 3.3: Constrained pendulum results



(a) rewards vs. steps (b) constraint violations vs. steps

Figure 3.4: Constrained half-cheetah results

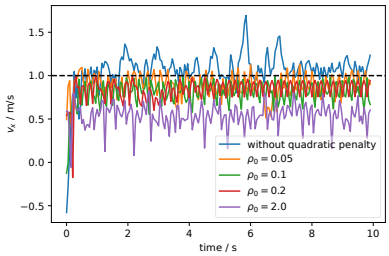


Figure 3.5: The state trajectory $v_x(t)$ of constrained half-cheetah under learned policies corresponding to different values of ρ_0. The horizontal black dashed line indicates the constraint $|v_x| \leq 1$.

common unconstrained RL algorithms. We provided theoretical results to justify the use of unconstrained algorithms, which requires stationary Lagrange multipliers to yield time-invariant rewards in the (augmented) Lagrangian. Theoretical results also show that under certain conditions we can recover an optimal policy. Finally, our empirical results suggested that our surrogate function could promote safety during learning. Additionally, we observed that our surrogate function reliably yielded a faster convergence relative to a standard Lagrangian dual approach.

Appendix

Proposition 4. *Consider a linear system $z_{t+1} = Az_t$, where $A \in \mathbb{R}^{2\times 2}$. Consider a fixed initial condition $z_0 \leq 0$. Suppose that the eigenvalues of A are real and positive, and there exist two eigenvectors $v_1 \leq 0$ and $v_2 \leq 0$ whose convex hull contains z_0. Then, we have $z_t \leq 0$, $\forall t \in \{0, 1, \dots\}$ and $\sum_{t=0}^{\infty} z_t \leq 0$.*

Proof. Denote by λ_i the i-th eigenvalue of A. Let $v_i = [v_{i1}, v_{i2}]$ be an eigenvector associated with λ_i. Consider initial condition $z_0 = [z_{01}, z_{02}] \leq 0$. Construct

$$c_1 = \frac{v_{22}z_{01} - v_{12}z_{02}}{v_{22}v_{11} - v_{12}v_{21}}, c_2 \quad = \frac{v_{11}z_{02} - v_{21}z_{01}}{v_{22}v_{11} - v_{12}v_{21}}. \tag{3.14}$$

We can verify that $z_0 = c_1v_1 + c_2v_2$. Suppose that there exists a $t' \in \{0, 1, \dots\}$ such that $z_{t'} \leq 0$ is not true. Then, it follows that $z_{t'} = c_1\lambda_1^{t'}v_1 + c_2\lambda_2^{t'}v_2 \leq 0$ is not true, i.e., either $c_1 < 0$ or $c_2 < 0$. However, since z_0 is in the convex hull of v_1 and v_2, we have $\frac{v_{21}}{v_{22}} \leq \frac{z_{01}}{z_{02}} \leq \frac{v_{11}}{v_{12}}$, which yields $v_{22}z_{01} - v_{12}z_{02} \geq 0$ and $v_{11}z_{02} - v_{21}z_{01} \geq 0$. Recall that v_1 and v_2 are in the third quadrant. We have $v_{22}v_{11} - v_{12}v_{21} \geq 0$. This suggests that $c_1 \geq 0$ and $c_2 \geq 0$, which presents a contradiction. Thus, we have $z_t \leq 0$, for all $t \in \{0, 1, \dots\}$, and it also yields $\sum_{t=0}^{\infty} z_t \leq 0$. □

Proof of Proposition 1. By definition, the condition $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$, implies the condition $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$. By combining the condition in Proposition 1, we have that $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$, if and only if $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$.

Let π^* be an optimal policy for (3.5). Define $p^* := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t)|\pi^*]$. From Theorem 1 in [245], we have

$$\begin{aligned} p^* = \min_{\lambda} \max_{\pi} & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda g(z_t, a_t))\middle|\pi\right] \\ \text{s.t. } & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0. \end{aligned} \tag{3.15}$$

By construction, for any non-positive Lagrange Multipliers $\{\lambda_t\}_{t=0}^{\infty}$, we have

$$\begin{aligned} & \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda_t g(z_t, a_t))\middle|\pi\right] \\ \geq & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda_t g(z_t, a_t))\middle|\pi^*\right] \\ \geq & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t)\middle|\pi^*, z_0\right] + \sum_{t=0}^{\infty} \gamma^t \lambda_t \mathbb{E}[g(z_t, a_t)|\pi^*] \\ \geq & p^* \end{aligned} \tag{3.16}$$

which also implies

$$\min_{\{\lambda_t\}_{t=0}^{\infty}} \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) + \lambda_t g(z_t, a_t) \middle| \pi\right] \geq p^*. \tag{3.17}$$

Let λ^* be an optimal solution of (3.15). Subsequently, suppose $\tilde{\lambda}_t := \lambda^*, \forall t \in \{0, 1, \dots\}$. Then, we have

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \tilde{\lambda}_t g(z_t, a_t)) \middle| \pi\right] = p^*, \tag{3.18}$$

which implies that

$$\min_{\{\lambda_t\}_{t=0}^{\infty}} \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda_t g(z_t, a_t)) \middle| \pi\right] \leq p^* \tag{3.19}$$

It follows from (3.17) and (3.19) that strong duality holds for (3.5). □

Proof of Proposition 2. Observe that the problem (3.7) is the Lagrangian relaxation of

$$\begin{aligned} \max_{\pi}\ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi\right] \\ \text{s.t. } & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0, \\ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t) \middle| \pi\right] \leq 0. \end{aligned} \tag{3.20}$$

Recall that the condition $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$, yields $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$. Suppose that under any closed-loop dynamics $z_{t+1} \sim p_z(\cdot|z_t, a_t)$ with $a_t \sim \pi(z_t)$, the constraint $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$ implies that $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$. Then, the problem (3.20) shares the same feasible domain with problem (3.6). From Theorem 1 in [245], strong duality holds for (3.7). By Proposition 1, strong duality also holds for (3.6). Therefore, a pair of optimal solutions to problem (3.7) implies that $(\{\tilde{\lambda}_t\}_{t=0}^{\infty}, \tilde{\pi}^*)$ is a pair of optimal solutions to (3.6). □

Before we present the proof of Proposition 3, we first introduce the following Lemma, which builds the foundation for the proof of Proposition 3.

Lemma 1. *Under Assumption 1, consider the function* $d(\lambda) : \mathbb{R}_- \to \mathbb{R}$ *defined as*

$$\begin{aligned} d(\lambda) := \max_{\pi}\ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (r(z_t, a_t) + \lambda(g(z_t, a_t)) \middle| \pi\right] \\ \textit{s.t. } & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0, \end{aligned} \tag{3.21}$$

and the function $d_\rho(\lambda) : \mathbb{R}_- \to \mathbb{R}$ *defined as*

$$\begin{aligned} d_\rho(\lambda) := \max_\pi \mathbb{E}\Bigg[& \sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) + \lambda \cdot \textit{Relu}(g(z_t, a_t)) \\ & -\frac{\rho}{2} \cdot \textit{Relu}(g(z_t, a_t))^2 \Big| \pi \Bigg] \\ \textit{s.t. } & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0. \end{aligned} \tag{3.22}$$

Let $\lambda^* := \arg\min_{\lambda \le 0} d(\lambda)$ *and* $\lambda^*_\rho := \arg\min_{\lambda \le 0} d_\rho(\lambda)$. *Suppose that under an optimal policy* π^* *of problem* (3.5)*,* $g(z_t, a_t) \le 0$ *for all* $t \ge 0$ *under an optimal policy* π^*. *Then, we have* $d(\lambda^*) \le d_\rho(\lambda^*_\rho) \le d_\rho(\lambda^*) \le d(\lambda^*)$.

Proof. On one hand, we observe that for any $\rho \ge 0$, $d_\rho(\lambda) \le d_{\rho=0}(\lambda) \le d(\lambda)$, and therefore, $d_\rho(\lambda^*) \le d(\lambda^*)$. By definition, $d_\rho(\lambda^*_\rho) \le d_\rho(\lambda^*)$. Moreover, observe that π^* is a feasible solution to problem (3.11), and under the policy π^*,

$$\begin{aligned} d_\rho(\lambda^*_\rho) \ge & \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \big(r(z_t, a_t) + \lambda^* \cdot \text{Relu}(g(z_t, a_t)) \\ & - \frac{\rho}{2} \cdot \text{Relu}(g(z_t, a_t))^2 \big) \Big| \pi^* \Big] \\ = & \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \Big| \pi^* \Big] = d(\lambda^*). \end{aligned} \tag{3.23}$$

Thus, $d(\lambda^*) \le d_\rho(\lambda^*_\rho) \le d_\rho(\lambda^*) \le d(\lambda^*)$. □

Proof of Proposition 3. We aim to show that as $\rho \to \infty$, any infeasible policy π' will become suboptimal for problem (3.12). Given $\rho \ge 0$, suppose that $\pi^*_\rho(\lambda^*_\rho)$ is infeasible. Then,

$$\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t) \Big| \pi^*_\rho(\lambda^*_\rho) \Big] > 0, \tag{3.24}$$

because otherwise $\pi^*_\rho(\lambda^*_\rho)$ would be feasible.

Define the function

$$J(\pi) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \Big| \pi \Big]. \tag{3.25}$$

There are only two cases: either $\lambda^*_\rho < \lambda^*$ or $\lambda^*_\rho \ge \lambda^*$.

For the first case that $\lambda^*_\rho < \lambda^*$, since $\pi^*_\rho(\lambda^*_\rho)$ is an optimal policy in problem (3.12), we have

$$\begin{aligned} d(\lambda^*) \geq & \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda^* g(z_t,a_t))\Big|\pi_\rho^*(\lambda_\rho^*)\Big] \\ = & J(\pi_\rho^*(\lambda_\rho^*)) + \lambda^*\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\Big|\pi_\rho^*(\lambda_\rho^*)\Big] \\ > & J(\pi_\rho^*(\lambda_\rho^*)) + \lambda_\rho^*\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\Big|\pi_\rho^*(\lambda_\rho^*)\Big] \\ > & d_\rho(\lambda_\rho^*) \end{aligned} \tag{3.26}$$

which contradicts that $d(\lambda^*) = d_\rho(\lambda_\rho^*)$, as shown in Lemma 1.

For the second case that $\lambda_\rho^* \geq \lambda^*$, we can pick $\rho' \geq 0$ such that

$$J(\pi_\rho^*(\lambda_\rho^*)) - \frac{\rho'}{2}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t \text{Relu}(g(z_t,a_t))^2\Big|\pi_\rho^*(\lambda_\rho^*)\Big] < J(\pi^*). \tag{3.27}$$

Subsequently, we have

$$\begin{aligned} d_{\rho'}(\lambda_{\rho'}^*) = & J(\pi^*) \\ > & J(\pi_\rho^*(\lambda_\rho^*)) - \frac{\rho'}{2}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t \text{Relu}(g(z_t,a_t))^2\Big|\pi_\rho^*(\lambda_\rho^*)\Big] \\ \geq & J(\pi_\rho^*(\lambda_\rho^*)) - \lambda_{\rho'}^*\mathbb{E}\Big[\sum_{t=0}^{\infty}\text{Relu}(\gamma^t g(z_t,a_t))\Big|\pi_\rho^*(\lambda_\rho^*)\Big] \\ & - \frac{\rho'}{2}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t \text{Relu}(g(z_t,a_t))^2\Big|\pi_\rho^*(\lambda_\rho^*)\Big] \end{aligned} \tag{3.28}$$

which implies that $\pi_\rho^*(\lambda_\rho^*)$ becomes a sub-optimal solution in problem (3.12), as ρ increases to ρ'.

For any infeasible policy π', there exists a sufficiently large but finite ρ' such that π' is sub-optimal for problem (3.12), $\forall \rho \geq \rho'$. Recall that π^* is an optimal policy. Thus, as $\rho \to \infty$, it holds that

$$\Big\|\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\Big|\pi^*\Big] - \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\Big|\pi_\rho^*(\lambda_\rho^*)\Big]\Big\|_2 \to 0. \tag{3.29}$$

□

Chapter 4

Layered Safety Approaches to Multi-Agent Reinforcement Learning for Air Mobility

In this chapter, we integrate reachability analysis with safe reinforcement learning to develop an efficient layered architecture that enhances the safety of multi-agent coordination tasks, such as air mobility in dense airspace. This approach enables scalable and safe decision-making in complex, safety-critical environments. The chapter is based on the published work [60], co-authored with Jason J. Choi, Jasmine Jerry Aloor, Jingqi Li, Maria G. Mendoza, Hamsa Balakrishnan, and Claire J. Tomlin.

4.1 Background

Motivation

Collision-free operation is a fundamental requirement for multi-robot coordination tasks, such as formation control [254], multi-robot payload transport [196], and autonomous navigation [64]. When only two agents interact, there is a single collision-avoidance constraint, which can be easily managed using a safety filter. However, with multiple nearby agents, the resolution of a constraint between two agents can conflict with a constraint involving a third agent. These conflicts may result in suboptimal task performance, such as creating a severe gridlock that prevents agents from taking actions to achieve their tasks. More crucially, the inability to simultaneously satisfy all constraints can result in an agent taking an action that makes collision inevitable. In particular, this issue poses a significant safety risk in high-density scenarios like air taxi operations for Advanced Air Mobility (AAM) [3].

Prior works have addressed safe multi-robot coordination problems by using model-based control methods like control barrier functions (CBFs) [306] and reachability analysis [307]. Although CBFs and reachability provide a framework for safety assurance, they generally offer rigorous guarantees only when a single safety constraint is considered. The fundamental challenge in extending these methods to the multi-agent case is that the intersection of the safe sets corresponding to individual

constraints (each derived from a pair of agents) does not necessarily represent the true safe set when all constraints are considered together (see Figure 4.2 for an example). In the Hamilton-Jacobi (HJ) reachability literature, the gap between these two regions is referred to as the “leaky corner" [215]. Agents that enter a leaky corner can no longer satisfy all safety constraints simultaneously and are inevitably forced to violate at least one. Unfortunately, identifying leaky corners without recomputing the reachability analysis from scratch while incorporating all constraints remains an open problem [172, 147]. Performing reachability analysis or designing CBFs for all possible combinatorial interaction scenarios is computationally intractable. In summary, the fundamental challenge in achieving scalability with such control-theoretic methods in multi-agent settings lies in handling conflicting constraints.

In this chapter, we combine the control barrier-value function (CBVF) [59], which is a CBF design method based on Hamilton-Jacobi reachability, with multi-agent reinforcement learning (MARL) into a layered safety architecture. This integration is driven by the essential role MARL can play in learning to strategically optimize task performance in multi-agent scenarios while proactively navigating potential conflicting constraints, which helps achieve safer and more effective behaviors. As a result, our approach enhances both safety and performance to a level that neither safe control methods nor MARL alone could achieve.

Contributions

1. *Architecture:* We propose a layered architecture that combines safety-informed MARL-based policy and CBVF-based safety filtering mechanism (Figure 4.1), which can significantly mitigate the issues arising from conflicting constraints, such as inefficiency due to gridlock and the leaky corner problem.

2. *Training method:* We propose a method to incorporate a CBVF-based safety filter into the training of MARL, considering two key aspects. First, a main challenge in this safety-constrained training is that the conservativeness introduced by safety filtering can hinder the exploration necessary for MARL to learn an effective policy. To address this, we introduce curriculum learning into the application of the safety filter, carefully balancing safety and exploration. Second, based on reachability analysis, we derive a conservative estimate of the safe region that is free from the issue of conflicting constraints (represented by the range r_{conflict} in Figure 4.1). Based on this estimate, the MARL policy is informed to minimize entry into this region, thereby avoiding potential conflicting constraints. Crucially, unlike many existing methods [27, 330], our proposed training approach does not impose safety through penalty terms directly penalizing the safety violation. Instead, MARL learns to enhance safety by making strategic decisions that mitigate conflicting constraints. This indirect approach significantly reduces unnecessary conservativeness, a common side effect of safe reinforcement learning-based methods.

3. *Experimental validation*: We conduct hardware experiments using Crazyflie drones and perform simulations of high-density AAM scenarios to validate our hypothesis.

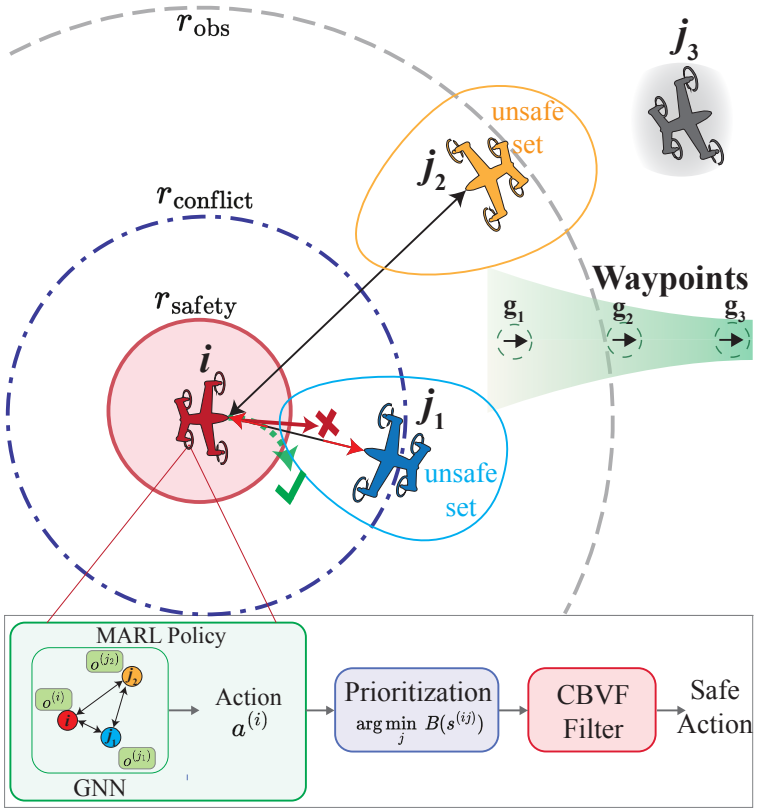


Figure 4.1: The figure shows our approach using an example scenario of four agents. Agent i must reach the waypoints shown on the right. Our Layered Safe MARL framework consists of three key components, and we describe it as applied through agent i: 1) The MARL policy generates an action based on the observation within the range r_{obs} while aiming to reduce the likelihood of entering other agents' potential conflict range r_{conflict}. 2) The prioritization module identifies the most critical neighboring agent in a potential collision scenario by evaluating the CBVF. In this example, agent j_1 is within the potential conflict region and forms a *potential collision pair*. 3) The CBVF safety filter adjusts the action to ensure safe navigation.

The remainder of this chapter is organized as follows. Section 4.2 provides background on safety for multi-agent coordination. Section 4.3 describes the system, environment, and problem statement. Section 4.4 presents the safety analysis of multi-agent problems under collision avoidance constraints. Sections 4.5 and 4.6 present our proposed Layered Safe MARL approach, the experiments performed, and the results obtained. We discuss some limitations of our approach in Section 4.7. Finally, we conclude and propose future work in Section 4.8.

4.2 Related works

Core related works—safety for multi-agent problems

Classic Control barrier function (CBF)-based approaches. CBFs [9] are used to design safe controllers via the principle of set invariance, and their application to safe multi-agent coordination has been explored in [306, 146, 111]. A primary challenge in employing CBFs lies in constructing a valid CBF, which often requires system-specific, handcrafted designs [306]. In [111, 146], a more generic design principle based on exponential CBFs [236] is employed; however, this approach does not address control input bounds. Another common limitation of existing methods is the treatment of multiple, potentially conflicting CBF constraints, which can lead to infeasibility. To address this, we adopt the CBVF-based framework to construct valid CBFs and handle multiple safety constraints in multi-agent coordination via a layered safety architecture.

Neural CBF-based approaches. While learning-based methods [257, 330, 331, 332] are proposed to design approximate CBFs, they lack deterministic safety guarantees due to the nonconvexity of the learning problems. Graphical CBF (GCBF) in [330, 331] offers a CBF based on local observations under multi-agent interaction, but how it learns to handle multiple constraints is not explicitly examined. Discrete Graphical Proximal Policy Optimization (DG-PPO) [332] proposes a model-free approach to learning decentralized CBFs and a safe control policy optimizing the task objective. Unlike DG-PPO, our approach leverages model-based information to compute Control Barrier Value Functions (CBVFs) [59] for pairwise collision avoidance, thereby ensuring deterministic safety guarantees. Finally, the aforementioned methods focus on learning safety certificates and policies for uncertain dynamical systems, often with high-dimensional system states. In contrast, our work specifically addresses the challenge of conflicting constraints in multi-agent interactions—a critical issue that persists even when each agent's dynamics can be effectively represented by simple, low-dimensional models.

Reachability for multi-agent interaction. Classical HJ reachability analysis computes the set of states that are guaranteed to be safe by computing the optimal control value function with dynamic programming. Prior works have investigated the reachability analysis for the special case of three-agent interaction [50] and using the value function to guide the planner, and its optimal control law is used in the tracking controller for safe multi-agent interaction [307]. Due to the curse of dimensionality in dynamic programming [28], the applicability of HJ reachability to high-dimensional systems is inherently limited. Recent work leveraged deep learning techniques to learn high-dimensional reachability [19, 134, 133, 336, 182], demonstrating their use in multi-agent collision avoidance scenarios. However, the learned solution does not generalize to new scenarios involving different agents. Additionally, the question of how to certify the safety of the learned safe set is still an open research topic [323, 194, 182]. Finally, alternative methods for solving reachable sets with over-approximation have been used in the context of multi-agent problems and air traffic management [291, 30].

Safe multi-agent reinforcement learning (MARL). A common approach to safe MARL is through constrained Markov decision processes (CMDPs) [7]. In theory, CMDPs have no duality gap under certain assumptions [245], but in practice, training with PPO-Lagrangian [260], and its multi-agent

variant [117] often suffers from instability due to suboptimal policies and inaccurate Lagrange multipliers. Another approach is shielded MARL, which uses safety filters [132, 27] to enforce safety during training and deployment. Originally introduced for single-agent RL [5], this method has been extended to multi-agent settings [80]. However, designing a shielding policy remains challenging due to the curse of dimensionality.

Other related works

Multi-agent reinforcement learning. Multi-agent extensions of single-agent RL algorithms, such as PPO [271] and DDPG [192], include MA-PPO [325] and MA-DDPG [201], both of which assume full observability. However, in many real-world applications, such as autonomous driving, agents have to make decisions based on their local information and coordinate effectively with other agents. A key challenge in MARL is the decentralized decision-making under partial observation. InforMARL [231] leverages graphical neural networks for information sharing to develop an efficient, coordinated learning framework for acquiring a high-performance MARL policy. Our approach builds on InforMARL to allow agents to make decentralized decisions based on their local observations.

Control and game-theoretic methods. In collaborative multi-agent settings, Model Predictive Control (MPC) has been used to ensure safe control [283, 95, 82, 334, 111] and its integration with MARL is explored in [71]. However, the complexity of the constrained optimization involved in MPC often limits its real-time execution in complex systems. When agents pursue distinct objectives, the problem becomes a non-cooperative game, with various solution approaches proposed in prior work [24, 86, 227, 104, 33, 138]. However, treating the safety constraints in the game-theoretic solutions remains an open research challenge [167, 181].

Collision avoidance & conflict resolution for air traffic control. With the growing interest in AAM applications such as drone deliveries and air taxi services, developing a scalable low-altitude air traffic management system that is automated or semi-automated has become an urgent need. Compared to current aviation, AAM operations are expected to be large-scale, ad hoc, on-demand, and dynamic. These characteristics motivate the development of a new air traffic management (ATM) framework that can achieve scalable, efficient, and collision-free operations [3, 157].

Existing work on collision avoidance and conflict resolution for ATM is categorized into *strategic deconfliction*, which focuses on preemptive deconfliction, and *tactical deconfliction*, which focuses on imminent proactive collision avoidance. A substantial body of work leverages control theory to design methods for strategic deconfliction. An early work proposed a flight mode switching framework derived from a hybrid automaton and reachability-based analysis [295]. As this method suffers from the computational complexity of HJ reachability, [50] uses a mixed integer program to assign avoidance responsibilities and resolve conflicts cooperatively. The work in [51] alternatively organizes vehicles into platoons on structured air highways, treating each platoon as a coordinated entity. While these methods provide strong safety guarantees, they rely on a predefined set of coordination rules for those guarantees to hold. Additionally, the approach in [68] uses preemptive strategic speed adjustments to prevent perceived conflicts without requiring controller intervention. Finally, a negotiation-based framework is introduced in [318] for collision-free strategic planning.

In parallel, the aviation community employs tactical collision avoidance modules as the final layer for safety. The Traffic Alert and Collision Avoidance System (TCAS) is an onboard system developed in the 1980s for conventional airliners, designed to detect and prevent collisions through vertical separation [163]. A method for tactical collision avoidance through horizontal resolution is also proposed in [83]. The successor of TCAS, the Airborne Collision Avoidance System (ACAS) X, integrates predictive modeling with real-time sensor inputs [130] using a partially observable Markov decision process framework. These existing methods crucially rely on the assumption that no more than two vehicles are involved in a single conflict resolution. This assumption is typically upheld by the upstream strategic deconfliction decisions.

Various methods in both strategic and tactical deconfliction are integrated further into layered, hierarchical decision-making architectures, enhancing the safety of ATM [296, 289, 84]. Our work is inspired by these layered approaches in the aviation community; however, the separations between layers underlying the existing approaches do not directly apply to high-volume AAM scenarios. As such, we have to consider how to achieve safe collision avoidance in instances of simultaneous multi-vehicle engagement.

Finally, MARL-based methods have also been explored in air traffic control to ensure tactical deconfliction through preconditioned strategic planning [53], demonstrating improved safety and efficiency over rule-based methods. However, the available actions of each agent in this work are limited to the adjustment of speed or position.

4.3 Problem Formulation

In this section, we define the system, environment, each agent's dynamics, and their safety requirement, and the problem statement.

Preliminaries

We formulate our multi-agent system as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) defined by the tuple $\langle N, S, O, \mathcal{A}, P, R, \gamma \rangle$, where:

- N is the number of agents
- $s^{(i)} \in \mathbb{R}^D$ is the state of each agent with D as the state dimension, including their position variables,
- $s \in S = \mathbb{R}^{N \times D}$ is the environment state, which is the concatenation of each agent's states and the state space of the environment, respectively,
- $o^{(i)} = O(s^{(i)}) \in \mathbb{R}^d$ is the observation of agent i,
- $a^{(i)} \in \mathcal{A}$ is the action space for agent i. $\boldsymbol{a}^{(i)}$ denotes the sequence of actions for timesteps $k = 0, 1, \cdots$,

- $P(s'|s,a)$ is the transition probability from s to s' given the joint action a, the concatenation of each agent's actions,
- $R(o^{(i)}, a^{(i)})$ is the common reward function of all agents,
- $\gamma \in [0,1)$ is the discount factor.

The objective is to find a policy $\Pi = \left(\pi^{(1)}, \cdots, \pi^{(N)}\right)$, where $\pi^{(i)}\left(a^{(i)}|o^{(i)}\right)$ is agent i's policy that selects an action based on its observation.

Agent's dynamics & safety constraint

We consider each agent's dynamics as a sampled data system, meaning that their underlying physical dynamics evolve continuously in time, but their actions are updated at discrete timesteps. Their continuous dynamics are given as

$$\dot{s}^{(i)}(t) = f^{(i)}(s^{(i)}(t), a^{(i)}(t)), \quad s^{(i)}(0) = s_0^{(i)}, \tag{4.1}$$

and their action is updated at every sampling time Δt—i.e. the action sequence $\boldsymbol{a}^{(i)}$ maps to the signal in time given as $a^{(i)}(t) \equiv a_k^{(i)}$ for $t \in [k\Delta t, (k+1)\Delta t)$. Their discrete-time state is given as $s_k^{(i)} = s^{(i)}(k\Delta t)$.

The primary safety constraint we consider in this work is the collision avoidance between agents. For all time $t \geq 0$, agents must satisfy

$$\mathrm{dist}(s^{(i)}(t), s^{(j)}(t)) \geq r_{\text{safety}}, \quad \text{for } \forall i \neq j, \tag{4.2}$$

where r_{safety} is the safety distance.

In the subsequent safety analysis, we consider the relative dynamics between a pair of agents, (i,j). We define the relative state between the agents, which can be given as $s^{(ij)} := \mathrm{rel}(s^{(i)}, s^{(j)})$, where rel is a mapping from two agents' states to the relative state. We assume that relative position variables are part of $s^{(ij)}$; thus, dist can be defined based on $s^{(ij)}$. The dynamics of the relative state are described by

$$\dot{s}^{(ij)}(t) = f^{(ij)}\left(s^{(ij)}(t), a^{(i)}(t), a^{(j)}(t)\right), \tag{4.3}$$

which is derived from (4.1).

Observations

For each agent to learn an effective policy for performance and safety, the observations $o^{(i)}$ need to contain adequate information. We make the following assumptions that are generic for many multi-agent robot tasks.

- Each agent i's observation $o^{(i)}$ consists of its local observations of other agents and entities relevant to their task goals (e.g., goal location) within their observation range defined as r_{obs} and any additional information needed for its task. Thus, the reward given to agent i at each timestep, $R(o^{(i)}, a^{(i)})$, is defined based on its local observation and action.

- We define $I(i) : \mathbb{N} \to 2^{\mathbb{N}}$ as the index set of the agents within the observation range of agent i. We assume that $o^{(i)}$ contains information that can be used to reconstruct $s^{(ij)}$ from $o^{(i)}$ for all $j \in I(i)$. Thus, for agent i, with its observation, it is feasible to execute a feedback policy on $s^{(ij)}$ if agent j is within its observation range. This assumption will be used in the design of our safety framework.

Problem statement

To sum up, the decentralized multi-agent coordination problem, subjected to the collision avoidance constraint we consider in this work, can be described as

$$\begin{aligned} \max_{\pi^{(i)}} \ & \mathbb{E}\left[\sum_{k=0}^{\infty} R(o_k^{(i)}, a_k^{(i)})\right] \\ \text{s.t. } & s_{k+1} \sim P(s \mid s_k, a_k) \\ & a_k^{(i)} \sim \pi^{(i)}(a^{(i)} \mid o_k^{(i)}) \\ & \operatorname{dist}(s^{(i)}(t), s^{(j)}(t)) \geq r_{\text{safety}}, \\ & \text{for } \ \forall i \neq j, \forall t \geq 0, \end{aligned} \tag{4.4}$$

where each agent's action $a_k^{(i)}$ is determined by their policy $\pi^{(i)}$, based on their local observations. The agent learns to maximize its objective subject to its collision avoidance constraint.

4.4 Safety Analysis

In this section, we present the safety analysis of the multi-agent problem under collision avoidance constraints. Specifically, we first derive the safety analysis for a pair of agents and then investigate how it applies to the multi-agent scenario.

Collision avoidance for a pair of agents

To ensure $\operatorname{dist}(s^{(ij)}(t)) \geq r_{\text{safety}}$ for all $t \geq 0$, we consider the following cost function, which captures the closest relative distance along the trajectory:

$$J(s_0^{(ij)}, \boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) := \min_{t \in [0,\infty)} \operatorname{dist}\big(s^{(ij)}(t)\big). \tag{4.5}$$

If $J(s_0^{(ij)}, \boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) \geq r_{\text{safety}}$, the agents i and j are rendered safe (collision-free) by their actions.

Reachability analysis for computing the maximal safe set

The agent pair prioritizing safety would want to maximize (4.5) to move away from each other. From this intuition, we can consider the following optimal control problem

$$V(s_0^{(ij)}) := \max_{\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}} J(s_0^{(ij)}, \boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}). \tag{4.6}$$

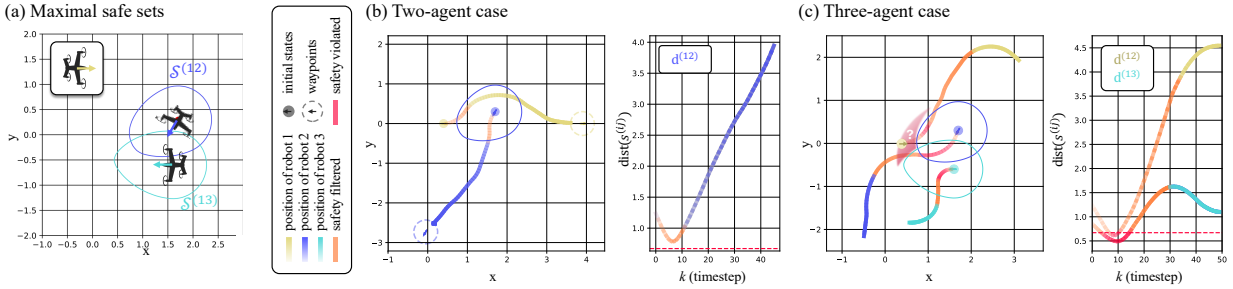


Figure 4.2: Running example illustrating the CBVF-based safe sets, safety filtering, and the leaky corner issue. (a) Visualization of the ego agent ($s^{(1)} = [0.4\text{km}, 0\text{km}, 0^o, 110\text{ kt}]$)'s maximal safe sets (exterior of the level sets) against two agents, $s^{(2)} = [1.7\text{km}, 0.3\text{km}, -120^o, 110\text{kt}]$ and $s^{(3)} = [1.7\text{km}, -0.6\text{km}, -180^o, 60\text{kt}]$. (b) In the two-agent case, each agent executing their CBVF safety filters (4.12) successfully prevents collision. (c) In the three-agent case, although agent 1 started inside the intersection of $\mathcal{S}^{(12)}$ and $\mathcal{S}^{(13)}$, it is not able to prevent safety violation. This is because the initial state of robot 1 is in the leaky corner.

Solving V is a specific type of reachability problem called the minimal Backward Reachable Tube (BRT) problem [302]. To see this, consider $\mathcal{L}^{(ij)} = \{s^{(ij)} \mid \text{dist}(s^{(ij)}) < r_{\text{safety}}\}$, the near-collision region, as the target set. The minimal BRT of $\mathcal{L}^{(ij)}$ is defined as

$$\mathcal{BRT}(\mathcal{L}^{(ij)}) := \{s_0^{(ij)} \mid \forall \boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}, \exists t \geq 0 \text{ s.t. } s^{(ij)}(t) \in \mathcal{L}^{(ij)}\}, \tag{4.7}$$

which encapsulates a region from which no action sequence can prevent the relative state from entering the near-collision region $\mathcal{L}^{(ij)}$. Using the definition in (4.6), we can express $\mathcal{BRT}(\mathcal{L}^{(ij)})$ as $\{s_0^{(ij)} \mid V(s_0^{(ij)}) < r_{\text{safety}}\}$.

The complement of $\mathcal{BRT}(\mathcal{L}^{(ij)})$ becomes the *maximal* safe set from which the agent pair can avoid collisions since it encompasses all the states from which there exist action sequences $\boldsymbol{a}^{(i)}$ and $\boldsymbol{a}^{(j)}$ that can avoid collision. This maximal safe set is denoted as

$$\mathcal{S}^{(ij)} := \{s_0^{(ij)} \mid \exists \boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}, \text{ s.t. } \forall t \geq 0, s^{(ij)}(t) \notin \mathcal{L}^{(ij)}\}, \tag{4.8}$$

and satisfies

$$\mathcal{S}^{(ij)} = \left(\mathcal{BRT}(\mathcal{L}^{(ij)})\right)^c = \{s_0^{(ij)} \mid V(s_0^{(ij)}) \geq r_{\text{safety}}\}. \tag{4.9}$$

We use the open-source library in [165] to compute (4.6) and $\mathcal{S}^{(ij)}$, which computes the Hamilton-Jacobi (HJ) partial differential equation (PDE) associated with the BRT problem [20].

Running example We consider the reduced-order dynamics of an autonomous air taxi given as

$$\dot{x} = v\cos\theta, \quad \dot{y} = v\sin\theta, \quad \dot{\theta} = \omega, \quad \dot{v} = \text{a}, \tag{4.10}$$

where the robot state consists of $s^{(i)} = [x; y; \theta; v]$, representing the positions, heading, and speed. The allowable actions are $a^{(i)} = [\omega, \text{a}]$, corresponding to the angular rate and the longitudinal acceleration, respectively. The speed is limited to the range of $[v_{\min}, v_{\max}]$. The action space is defined as $\mathcal{A} = [-\omega_{\max}, \omega_{\max}] \times [\text{a}_{\min}, \text{a}_{\max}]$. The parameters we use are defined in Table 4.1 and are explained in more detail in Section 4.6.

The relative state $s^{(ij)} = [x^{(ij)}; y^{(ij)}; \theta^{(ij)}; v^{(i)}; v^{(j)}]$ includes the relative position and heading of agent j from agent i's perspective, where x-axis is in the direction of the agent i's heading. The relationship between $s^{(ij)}$ and ($s^{(i)}$, $s^{(j)}$) and the relative state dynamics are given in Appendix 4.8.

The computation of V was completed within an hour using an Nvidia RTX A4500 GPU. [1] The computed maximal safe set $\mathcal{S}^{(ij)}$, defined in the relative state space of $s^{(ij)}$, can be projected to the position space of the ego agent (agent i), which incorporates all safe positions of the agent that can ensure collision avoidance, given its heading, speed, and the opponent agent (agent j)'s state. Examples of $\mathcal{S}^{(ij)}$ are visualized in Fig. 4.2 (a) with respect to two different opponent states.

Control barrier-value function-based safety filtering

Next, we investigate how the computed value function V can be used to constrain the relative state $s^{(ij)}$ to stay within the safe set $\mathcal{S}^{(ij)}$. Since each agent makes their primary decision based on their MARL policy in our framework, we consider how to filter the MARL action if it is potentially unsafe.

To achieve this safety filtering mechanism, we consider the barrier constraint-based mechanism of the CBFs. For a generic state variable s and its dynamics $\dot{s} = f(s, a)$, if a function $B(s)$ satisfies the barrier constraint given by

$$\nabla B(s) \cdot f(s, a) + \gamma B(s) \geq 0, \tag{4.11}$$

for every state inside the zero-superlevel set of B, i.e. $\forall s \in \{s \mid B(s) \geq 0\}$, and for some constant $\gamma > 0$, we can guarantee that $B(s(t)) \geq 0$, for all $t \geq 0$ [9]. Thus, the state can be maintained to stay within $\{s \mid B(s) \geq 0\}$.

If the computed reachability value function V in (4.6) is almost-everywhere differentiable, we can construct a CBF by taking $B(s^{(ij)}) = V(s^{(ij)}) - r_{\text{safety}}$. This choice of B satisfies the barrier constraint (4.11) almost everywhere, and results in the maximal safe set to be represented as the CBF zero-superlevel set, $\mathcal{S}^{(ij)} = \{s^{(ij)} \mid B(s^{(ij)}) \geq 0\}$. Such usage of the reachability value function as the CBF is referred to as the Control Barrier-Value Function (CBVF) [59].

Remark 7. *In [332], V is denoted as a constraint-value function and is used to learn Graphical CBF (GCBF) for uncertain dynamics. In this work, we consider its exact computation for a hard safety guarantee. However, our approach can be combined with the learning methods proposed in [332] or other learning-enabled approaches [98, 56] to be extended to agents subjected to uncertain dynamics.*

Remark 8. *For certain types of dynamics, the value function can be discontinuous without introducing a discount factor in time to the cost function* (4.5) *[58].*

Finally, the CBVF-based safety filtering can be implemented in a decentralized manner, with each agent executing its own safety filter. Here, we assume that the agents are *cooperative for*

[1] The computation time is not a critical concern in our setting, as the value function is computed offline rather than during real-time deployment.

safety, meaning that although their unfiltered actions can be selfish, their final filtered actions are coordinated to avoid collision with each other. To achieve this coordination, agent i and agent j can individually solve the identical optimization program defined as

---

CBVF Safety Filter (cooperative case):

$$\begin{aligned}(a^{(i)}_{\text{safe}},\ a^{(j)}_{\text{safe}}) &= \arg\min_{(a^{(i)},a^{(j)})\in\mathcal{A}} ||a^{(i)}-a^{(i)}_{\text{marl}}||^2 + ||a^{(j)}-a^{(j)}_{\text{marl}}||^2 \\ \text{s.t. } &\nabla B(s^{(ij)})\cdot f^{(ij)}\big(s^{(ij)},a^{(i)},a^{(j)}\big)+\gamma B(s^{(ij)}) \geq 0, \end{aligned} \tag{4.12}$$

---

and then execute their own action. If the dynamics $f^{(ij)}$ are affine in actions, the optimization becomes a quadratic program [9, 59].

If the opponent agent is non-cooperative for safety, agent i can solve for its own safe action considering the worst-case possible action of the opponent:

CBVF Safety Filter (non-cooperative case):

$$\begin{aligned}a^{(i)}_{\text{safe}} &= \arg\min_{a^{(i)}\in\mathcal{A}} ||a^{(i)}-a^{(i)}_{\text{marl}}||^2 \\ \text{s.t. } &\min_{a^{(j)}\in\mathcal{A}} \nabla B(s^{(ij)})\cdot f^{(ij)}\big(s^{(ij)},a^{(i)},a^{(j)}\big)+\gamma B(s^{(ij)}) \geq 0, \end{aligned}$$

---

where now B has to be constructed based on a value function for a differential game, which considers the opponent's worst-case actions [214], given as

$$V_{\text{worst}}(s_0^{(ij)}) := [\min_{\{a_k^{(j)},a_k^{(i)}\}_{k\geq 0}} \max]\ J(s_0^{(ij)},\boldsymbol{a}^{(i)},\boldsymbol{a}^{(j)}). \tag{4.13}$$

Here, $[\min\max]$ denotes the alternating operation $\min$ (over $a_k^{(j)}$) and $\max$ (over $a_k^{(i)}$). The computation of this worst-case value function can be done similarly to the computation of V by solving the min-max HJ PDE [20].

Running example In the two-agent case, in Figure 4.2 (b), the initial relative state between agents 1 and 2 is set near the boundary of the maximal safe set $\mathcal{S}^{(12)}$. By each agent applying the CBVF safety filter, both agents reach their goals safely under the safety-filtered MARL actions.

Analysis of the multi-agent case

We begin the analysis of this section by continuing with the running example of the multi-agent case:

Running example In Figure 4.2 (c), we now consider the case where a third agent is introduced. The relative states still remain within the pairwise maximal safe sets $\mathcal{S}^{(12)}$, $\mathcal{S}^{(13)}$, and $\mathcal{S}^{(23)}$. Despite all agents actively attempting to avoid collisions, their relative distances fall below r_{safety}. As mentioned in the introduction, this demonstrates the issue of conflicting constraints, implying that although agent 1's initial state did not cross the boundaries of the individual safe sets, it may

already be outside the true safe set when considering all interactions simultaneously. Computing this true safe set requires defining the relative dynamics of the three agents, which increases the system's dimensionality. While approximations of this set have been computed, such as in [19], the computation of this multiple-agent safe set is challenging.

As can be seen in the above running example, it is crucial to prevent the agents from falling into the region in which one safety constraint can potentially conflict with the other, i.e., the leaky corners. Although their exact computation is hard, we define the region that can tightly overapproximate this potential conflict region.

Proposition 5. *Define*

$$\begin{aligned}\hat{\mathcal{S}}^{(i)} := \Big\{ \{s^{(j)}\}_{j\in I(i)} \mid \forall j \in I(i), V(s^{(ij)}) \geq r_{safety}, \,\&\, \\ \nexists j_1, j_2 \in I(i) \text{ s.t. } V_{worst}(s^{(ij_1)}) < r_{safety} \,\&\, V_{worst}(s^{(ij_2)}) < r_{safety} \Big\}\end{aligned} \tag{4.14}$$

where V is defined in (4.6)*, and V_{worst} is defined in* (4.13)*. Note the difference between V and V_{worst}. Then for any opponent agent states $\{s^{(j)}\}_{j\in I(i)} \in \hat{\mathcal{S}}^{(i)}$, there exists $\boldsymbol{a}^{(i)}$ and $\boldsymbol{a}^{(j)}$ for all $j \in I(i)$, such that $\forall t \geq 0$, $s^{(ij)}(t) \notin \mathcal{L}^{(ij)}$ for all $j \in I(i)$. In other words, set $\hat{\mathcal{S}}^{(i)}$ can be maintained forward invariant.*

Proof. The second condition in (4.14) allows at most one opponent agent to enter the area in which $V_{\text{worst}}(s^{(ij)}) < r_{\text{safety}}$. We denote this agent as j_{near}. For (i, j_{near}), since $V(s^{(ij_{\text{near}})}) \geq r_{\text{safety}}$ based on the first condition in (4.14), agent i and agent j_{near} are within their CBVF safe set $\mathcal{S}^{(ij_{\text{near}})}$ and can select their action sequences $\boldsymbol{a}^{(i)}$ and $\boldsymbol{a}^{(j_{\text{near}})}$, such that $s^{(ij_{\text{near}})}(t) \notin \mathcal{L}^{(ij_{\text{near}})}$ for all $t \geq 0$.

Next, we consider all other agents $j \in I(i) \setminus \{j_{\text{near}}\}$. Based on [100, Proposition 4], for any Lipschitz continuous V_{worst}, its level set is a robust control invariant set. Thus, for all $j \in I(i)\setminus\{j_{\text{near}}\}$, there exists $\boldsymbol{a}^{(j)}$ that results in $V_{\text{worst}}(s^{(ij)}(t)) \geq r_{\text{safety}}$ for all $t \geq 0$, regardless of $\boldsymbol{a}^{(i)}$, ensuring $s^{(j)}(t) \in \hat{\mathcal{S}}^{(i)}$. □

Intuitively, the set $\hat{\mathcal{S}}^{(i)}$ prevents conflict of constraints by allowing only one agent to coordinate for collision avoidance with the ego agent and by prohibiting the other agents from entering the worst-case safe set. These other agents are able to stay away from the pair (i, j_{near}) due to the robust invariance property of the level set of V_{worst}.

Practical implementation: In practice, enforcing each agent to stay within $\hat{\mathcal{S}}^{(i)}$ can be computationally expensive since we have to evaluate V and V_{worst} for all pairs of interaction. In the next section, we use this analysis to inform MARL to implicitly learn not to enter this region. For this, we define the *potential conflict range* as below:

$$\begin{aligned} r_{\text{conflict}} := \min_{r} \; & r \\ \text{s.t.} \quad & V_{\text{worst}}(s^{(ij)}) \geq r_{\text{safety}} \;\; \forall s^{(ij)} \text{ s.t. } \text{dist}(s^{(ij)}) \geq r. \end{aligned} \tag{4.15}$$

Then, the set of opponent agent states is defined as

$$\tilde{\mathcal{S}}^{(i)} := \Big\{ \{s^{(j)}\}_{j\in I(i)} \mid \forall j \in I(i), V(s^{(ij)}) \geq r_{\text{safety}}, \& \\ \nexists j_1, j_2 \in I(i) \text{ s.t. } \text{dist}(s^{(ij_1)}) < r_{\text{conflict}} \ \& \ \text{dist}(s^{(ij_2)}) < r_{\text{conflict}} \Big\} \tag{4.16}$$

This is an underapproximation of the true conflict-free set $\hat{\mathcal{S}}^{(i)}$ by definition (4.15). To ensure safety, we want to restrict the number of opponent agents entering this region to be at most one.

Our analysis requires that the observation range be larger than the potential conflict range, $r_{\text{obs}} > r_{\text{conflict}}$. Rather than a restriction, this serves as a design guideline for the observation stack of the robot for safe multi-robot coordination. As shown in Figure 4.1, the concept of the potential conflict range divides a robot's proximity into three layers: (1) the range $\text{dist}(s^{(ij)}) < r_{\text{safety}}$, where collision is imminent; (2) the range $r_{\text{safety}} < \text{dist}(s^{(ij)}) < r_{\text{conflict}}$, where engaging with multiple vehicles may introduce safety risks; and (3) the region $r_{\text{conflict}} < \text{dist}(s^{(ij)})$, where the maneuvers of other agents pose minimal safety concerns. A similar three-layer structure was proposed and manually designed in [108]. However, our approach provides a theoretical foundation for defining these boundaries based on reachability analysis.

Remark 9. *(Limitation) The set $\hat{\mathcal{S}}^{(i)}$ is a conflict-free safe set only from agent i's perspective. In other words, it does not guarantee that the collision-avoidance maneuvers of other agents $j \in I(i)$ will not interfere with one another. Addressing this issue requires analyzing the combinatorial number of possible interactions, which remains an open problem. In our work, we address this challenge by training the MARL policy to learn strategies that mitigate these conflicts.*

4.5 Multi-Agent Reinforcement Learning with Layered Safety

Extending InforMARL for improved decentralized decisions

Our work builds upon the InforMARL architecture [231], a MARL algorithm that solves the multi-agent navigation problem by using a graph representation of the environment, enabling local information-sharing across the edges of the graph. InforMARL uses graph neural networks (GNNs) to process neighborhood entity observations, allowing the framework to operate with any number of agents and provide scalability without changing the model architecture. Each agent has a set of neighboring agents within its observation range, r_{obs}, and shares its relative position, speed, and goal information with these neighbors. Agents are tasked to navigate to their respective goal positions. When agents reach their respective goals, they get a goal reward $\mathcal{R}_{\text{goal}}(o_k^{(i)}, a_k^{(i)})$.

The extensions we make to the baseline InforMARL to incorporate the layered safety framework and to make it more practical for multi-robot navigation tasks are as follows:

1) Sequential goal point tracking: In the updated framework, the agents navigate to a sequence of waypoints, each specified by its position and the desired direction and speed, leading to the final goal (as shown by the green circles in Figure 4.1). At each time step, an agent gets the following additional rewards, $\mathcal{R}_{\text{tracking}}(o_k^{(i)}, a_k^{(i)})$ which are computed based on the heading and speed of the

agent relative to the current target waypoint. The details of these terms are presented in Appendix 4.8.

2) Model architecture enhancements: To improve the algorithm and generalize it over diverse scenarios, we update the observations to incorporate rotation-invariant relative distances of the ego agent to goals and neighbors. Once an agent crosses a waypoint, we no longer consider the waypoint in its observation, and the agent moves to the subsequent waypoint. We introduce dynamics-aware action spaces that are updated based on the dynamics model, angular rate, and longitudinal acceleration, as in the running example in Section 4.4, ensuring agents respect motion constraints specific to their dynamics.

3) Curriculum training: The training framework also incorporates curriculum learning where we progressively make the training environment harder [228] for improving agents' performance, refining safety rewards, and updating the safety distance r_{safety} used in the safety filter. This is detailed in the subsequent sections.

Safety filter design for multiple agents

For multiple agents, the CBVF $B(s^{(ij)})$ is evaluated for each agent i and any neighboring agent j within the observation range r_{obs}. A smaller value of B indicates that the near-collision is more imminent and safety is at greater risk. The neighbor agent with the minimum pairwise $B(s^{(ij)})$ is selected as the agent whose actions will be curtailed. We term the module that selects this prioritized constraint as the *prioritization module*. If an agent pair (i, j) has each other as the minimum pairwise $B(s^{(ij)})$, then we call them a *potential collision pair*.

Safety-informed training

Curriculum update

We start the training routine without any safety filter or penalty applied for the first half of the training steps. This is done to optimize the task performance of MARL unconstrained by any safety parameters. Once training reaches half the number of total training steps, we activate the safety filter. Additionally, we introduce the following safety parameters, which are updated using the curriculum learning framework during training. First, the safety distance r_{safety} is initialized to zero during the start of model training, allowing agents to approach each other at close ranges. As the training progresses, we increase r_{safety} to the desired value. Similarly, we scale the conflict radius r_{conflict} computed using Eq. (4.15) based on the value of the r_{safety}. This setup allows agents to explore the environment early on in the training and prevents them from converging to overly conservative behavior.

Safety-informed reward

In addition to the heading, speed, and goal rewards, we introduce some additional penalties. When more than two agents are within the conflict radius r_{conflict}, we apply a potential conflict penalty

$$\mathcal{C}_{\text{conflict}} := \sum_{j \in \{j | \text{dist}(s^{(ij)}) < r_{\text{conflict}}\}} \max\{0, r_{\text{conflict}} - \text{dist}(s^{(ij)})\} \times \max\Big\{0, \underbrace{-\begin{bmatrix} x^{(ij)} & y^{(ij)} \end{bmatrix} \begin{bmatrix} v_x^{(ij)} \\ v_y^{(ij)} \end{bmatrix}}_{\text{relative distance change}}\Big\}, \tag{4.17}$$

which evaluates whether the agent j is within the potential conflict range and is approaching towards agent i. Based on Proposition 5, we do not apply the penalty when there is just one agent within r_{conflict}.

The penalty $\mathcal{C}_{\text{conflict}}$ is carefully designed to mitigate the risks associated with potential conflicting constraints when multiple agents enter the range, while simultaneously minimizing the conservatism it may introduce. This penalty is an **indirect** safety penalty, as it is not incurred based on explicit safety violations but rather indirectly through the proximity of multiple agents.

The final reward structure is

$$\mathcal{R}_{\text{total}}(o_k^{(i)}, a_k^{(i)}) = \mathcal{R}_{\text{tracking}}(o_k^{(i)}, a_k^{(i)}) + \rho_{\text{goal}}\mathcal{R}_{\text{goal}}(o_k^{(i)}, a_k^{(i)}) - \rho_{\text{conflict}}\mathcal{C}_{\text{conflict}} \tag{4.18}$$

where ρ_{goal} is a binary indicator when the agent is at the goal, and ρ_{conflict} is a binary indicator when the number of other agents within the potential conflict region is more than one.

4.6 Results

The main robotic application we focus on is the safe autonomous navigation of aerial vehicles. We apply our framework to Crazyflie drones navigating through waypoints in both simulation and hardware experiments, as well as to the simulation of air taxi operations in realistic settings.

Experiment Setup

Considered dynamics. We consider two types of dynamics, one for the quadrotors and the other for the air taxi vehicle in a wing-borne flight. The parameters for both dynamics are summarized in Table 4.1 and are set to match the industry standard [149, 12, 309]. For instance, we use an angular rate bound of 0.1 rad/s for the air taxi dynamics, as it results in the lateral acceleration $0.5g$ under the nominal speed, which amounts to the maximum tolerable lateral acceleration for passenger comfort in NASA market studies [279].

The quadrotor dynamics in the horizontal plane are represented as simple double integrators with

$$\dot{x} = v_x, \quad \dot{y} = v_y, \quad \dot{v}_x = a_x, \quad \dot{v}_y = a_y, \tag{4.19}$$

where $s^{(i)} = [x, y, v_x, v_y]$ and $a^{(i)} = [a_x, a_y]$. The quadrotor runs the low-level onboard flight controller to track the commanded actions.

Table 4.1: Parameter Summary for Different Vehicle Dynamics

Parameter	**Air taxi (Sim)**	**Crazyflie**
Groundspeed		
$v_{\min}$	60 knot (30 m/s)	-1.0 m/s
$v_{\max}$	175 knot (90 m/s)	1.0 m/s
v_{nominal}	110 knot (57 m/s)	0.5 m/s
Acceleration		
$a_{\min}$	-3.3 ft/s^2 (-1.0 m/s^2)	-0.5 m/s^2
$a_{\max}$	6.6 ft/s^2 (2.0 m/s^2)	0.5 m/s^2
Angular Rate ($\omega_{\max}$) (rad/s)	0.1	-
Sampling Rate (s)	1.0	0.1
Waypoint Thresholds ($\pm$)		
Distance to Goal	0.186 miles (0.3 km)	0.2 m
Heading	45^o	45^o
Speed	38.9 knot (20 m/s)	0.1 m/s
Observation Range ($r_{\textbf{obs}}$)	3.1 mi. (5.0 km)	4.0 m
Safety Distance ($r_{\textbf{safety}}$)	500 - 2200 ft (0.152 - 0.671 km)	0.5 m
Potential Conflict Range ($r_{\textbf{conflict}}$)	4600 ft (for r_{safety}=2200 ft)	1.0 m

The air taxi dynamics in the horizontal plane are represented using the kinematic vehicle model in (4.10) of the running example in Section 4.4. Three features of the air taxi dynamics considered in this work make its safety assurance more challenging and interesting. First, the vehicle cannot stop as it has to maintain the wing-borne flight ($v_{\min} > 0$). Next, the dynamics are nonholonomic, meaning that its control towards the lateral direction can be achieved only by changing its direction. Finally, due to small acceleration or deceleration authority, the vehicle often has to employ turning maneuvers for deconfliction. This is common for fixed-wing and hybrid mode vehicles like vertical

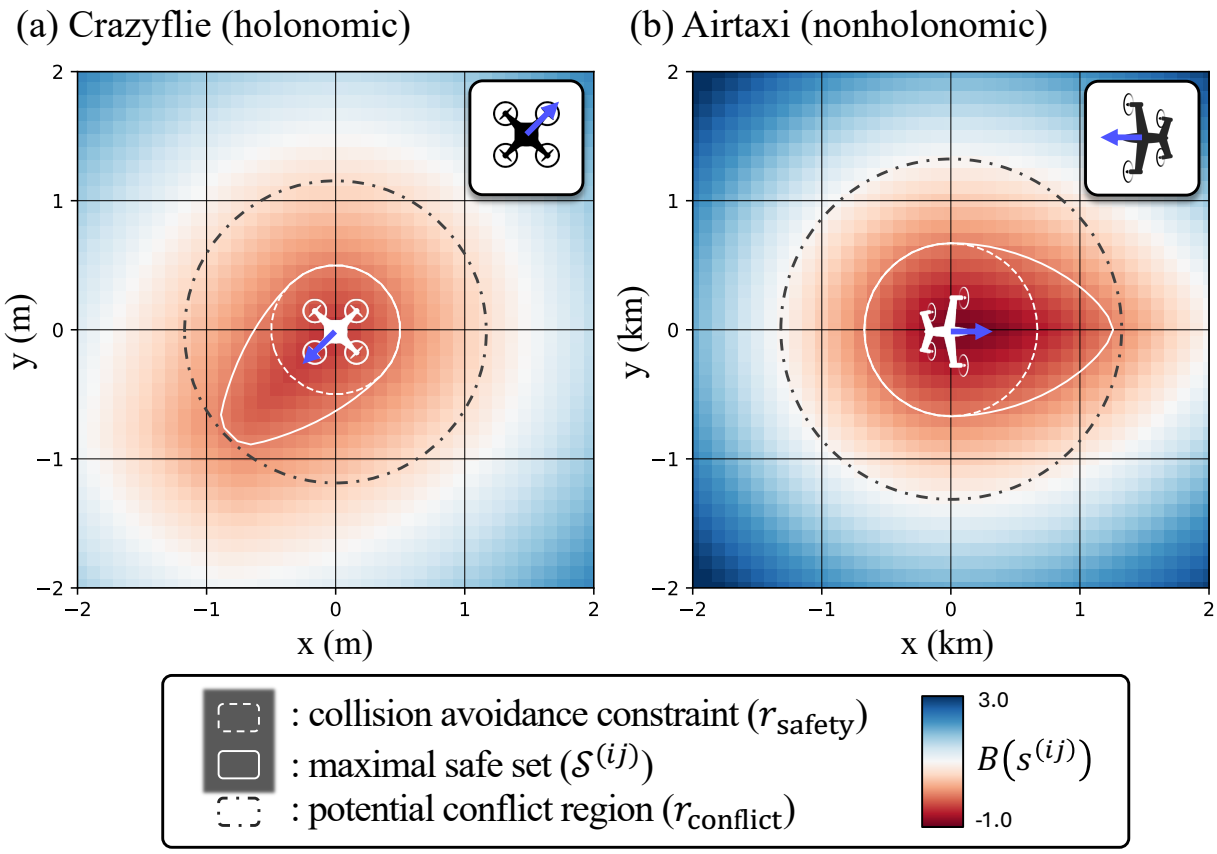


Figure 4.3: Maximum safe sets (exterior of the white level sets), potential conflict region, and CBVF (colormap) for each vehicle dynamics, displayed in the relative position space when (a) relative velocity is $(v_x, v_y) = (1, 1)$ [m/s], (b) relative speed and heading is 220 knots and 180^o, respectively.

takeoff-and-landing vehicles (VTOLs), those envisioned for AAM operations [8, 11, 93].

Due to these challenges, the advantages of our method for enhancing safety are particularly evident for the air taxi dynamics (Section 4.6). In contrast, for the quadrotors (Section 4.6), our safety filter design consistently ensures safety across all evaluated methods; thus, we focus more on how our approach achieves performance enhancement. The safe sets, CBVFs, and the potential conflict range computed using HJ reachability are visualized in Figure 4.3.

Task & Training environment. We modify Multi Particle Environments (MPE) [201] to incorporate agents to follow the dynamics as specified before and the safety filter. In our navigation task setup, the drone must pass through a waypoint with its state satisfying the threshold conditions specified in Table 4.1 to proceed to the next waypoint. The main values that define the training environments are the number of agents N, the number of waypoints per agent M, and the size of the environment L. At every episode, the initial positions of the agents, the waypoints' locations, and the headings are set randomly. The episode is terminated if all agents reach their goal, the last waypoint. After training the MARL policies, we test them in various evaluation scenarios with

values of N, M, L different from those of the training environment. For the quadrotor, we use N=4, M=2, L=4, and for the air taxi dynamics, we use N=4, M=2, L=6 for the training.

Comparison Studies

We first conduct two sets of simulation experiments to compare our method against: 1) approaches that do not employ a safety filter or curriculum during training, and instead rely on alternative reward designs for safety, and 2) methods from prior works based on model-based CBF design and model-free safe MARL for multi-agent coordination. Both studies are conducted in the quadrotor simulation environment.

Ablation study for safety-informed training

First, we designed our experiments to evaluate the value of (1) introducing the safety filter during training, (2) using the curriculum, and (3) the effectiveness of a potential conflict penalty term for safety, as described in Section 4.5. To evaluate the effect of employing the safety filter during training, we compare the results of those trained with and without the filter. To evaluate the effect of the curriculum, we compare our method against a policy trained without the curriculum update in Section 4.5. Finally, to evaluate the effectiveness of the potential conflict penalty term, we compare it against three alternative penalty terms for safety suggested in the literature:

- Hinge loss for constraint violation:

  $$\mathcal{C}_{\text{plain}} := \max\{0, r_{\text{safety}} - \text{dist}(s^{(ij)})\},$$

  This is the most typical penalty term, introduced in the safe RL literature [1].

- CBVF-based hinge loss:

  $$\mathcal{C}_{\text{cbvf}} := \max\{0, -B(s^{(ij)})\},$$

  This penalizes the agent for entering the zero-sublevel set of the CBVF, the unsafe set. The use of reachability value functions as a safety penalty in RL has been explored in previous works such as [13].

- Penalty occurring when safety filter intervenes:

  $$\mathcal{C}_{\text{norm.diff}} := ||a^{(i)}_{\text{safe}} - a^{(i)}_{\text{marl}}||,$$

  based on (4.12). This is the main penalty term used in the method of [27] to inform MARL with safety.

When we introduce each penalty term, its weight is carefully tuned to maximize task and safety metrics. In total, we test nine variants of the training methods based on the activation of the safety filter, curriculum, and choice of the reward term, which are detailed in Appendix 4.8.

We evaluate each method in three scenarios. In addition to the random scenario same as the training environment, in the second scenario, we test how the MARL policy interacts with a larger number of agents and a more challenging waypoint configuration by setting N=6, M=3, and L=6 while also placing the first two waypoints at the same positions, representing the air corridor. The third scenario reconstructs our hardware experiment environment, which will be detailed in Section 4.6, in simulation.

Below is the summary of the key aspects of the result, while more details and the table of the comparison for the performance metrics are presented in Appendix 4.8:

- *Effect of using the safety filter in training:* Methods that incorporate the safety filter during training consistently outperform their counterparts trained without the filter across all metrics.
- *Effect of curriculum learning:* The curriculum learning can significantly enhance performance by reducing the conservativeness of the trained policy.
- *Effect of potential conflict penalty $C_{conflict}$ compared to other penalty candidates:* Our method achieves the best performance in most cases. Importantly, our method outperforms other methods, especially when there is a larger number of agents (the second scenario).

Comparison to other methods

Next, we compare our method to (1) DG-PPO [332] and (2) a safety filter designed based on the exponential CBF (ECBF) [236], used for multi-agent collision avoidance in [111]. We use N=4, M=1, L=3 for the training of all three methods, which enables fair comparison, especially with results we can obtain from the DG-PPO source code. We run the training of both DG-PPO and our MARL policy with the same number of environment steps (1e7) and gradient steps (epoch_ppo=1), where the initial and goal positions are randomly generated.

We evaluate the trained policies in two environments for 25 episodes each: (1) same random environment with world size increased to L=6. (2) environment with an increased number of agents N=8 and world size L=6, where initial and goal positions are arranged in two parallel lines in random order. The results are reported in Tables 4.2 and 4.3. While all methods perform well when the number of agents remains the same as in the training environment, our method is the only method that guarantees 100% safety when N increases, whereas the percentage of near-collision events increases significantly for DG-PPO and ECBF. It must be noted that DG-PPO is a model-free method that learns a neural CBF during its training, whereas our method uses the CBVF computed based on the system dynamics model. As observed in [331], such model-free methods are vulnerable to generalization in scenarios with a large number of agents.

Hardware experiments with quadrotors

Next, as illustrated in Figure 4.4, we conduct hardware experiments with three Crazyflie 2.0 drones using a Vicon system for localization. Each drone is controlled by a hierarchical framework: its high-level system is modeled with double integrator dynamics, and the resulting high-level state is passed

Table 4.2: Simulation results for Crazyflie dynamics with N**=4**, with time horizon 51.2s. We evaluate goal reach rate (%) for performance and the percentage of near-collision events ($\text{dist}(s^{(ij)}) < r_{\text{safety}}$) in the timestamped trajectory data (Near collision %) for safety.

Methods	Goal reach(%)	Near collision(%)
DG-PPO	96 ± 11.8	0.04 ± 0.16
Exponential CBF	100 ± 0	0.0 ± 0.0
Our Method	100 ± 0	0.0 ± 0.0

Table 4.3: Simulation results with N**=8**, and initial & goal positions arranged in lines under random order. Videos are available in the supplementary material.

Methods	Goal reach(%)	Near collision(%)
DG-PPO	100 ± 0	9.1 ± 2.7
Exponential CBF	93 ± 8.9	8.8 ± 10.7
Our Method	$\mathbf{100 \pm 0}$	$\mathbf{0.0 \pm 0.0}$

to onboard PID tracking controllers. We define high-level feedback control (acceleration in the x and y axes) based on real-time (100 Hz) Vicon system data. To approximate decentralized control, we run distinct decentralized policies for each drone on a single ThinkPad laptop, transmitting high-level control commands every 0.1 seconds.

In the experimental scenario, each drone is required to pass through two shared waypoints—representing an air corridor—before reaching its designated landing location (N=3, M=3, L=3m). We compare the policy learned under our method to the baseline, which is trained without a safety filter in hardware experiments, with the recorded trajectories shown in Figure 4.5. Under our approach, three drones smoothly avoid conflicts and safely navigate their individual waypoints, completing the task in 12.95 seconds. In contrast, the baseline policy requires one drone to perform a second pass, after missing its first waypoint due to the safety filter preventing it from approaching other agents passing the waypoint, thus extending the total completion time to 25.29 seconds. These results demonstrate how our Layered Safe MARL framework enhances task performance through efficient deconfliction.

Simulation of decentralized air taxi operations

Finally, we evaluate the application of our framework to decentralized air taxi operations. Although there is no single consensus on how the air traffic management (ATM) system will function for advanced air mobility (AAM) operations, each vehicle will likely be required to have fallback autonomy systems in place, for instance, to ensure safety in case the centralized system fails.

To conduct the study with a realistic traffic volume, we use the results of the Urban Air Mobility (UAM) demand analysis from [40, 174], which estimates how much ground traffic could be replaced by AAM considering various factors including different demographics of riders, socioeconomic factors, and historical commuting patterns. By combining these insights, our study derives a reasonable estimate of how traffic density will evolve once UAM operations reach full implementation. From their results, we consider a peak-density scenario in which each vertiport

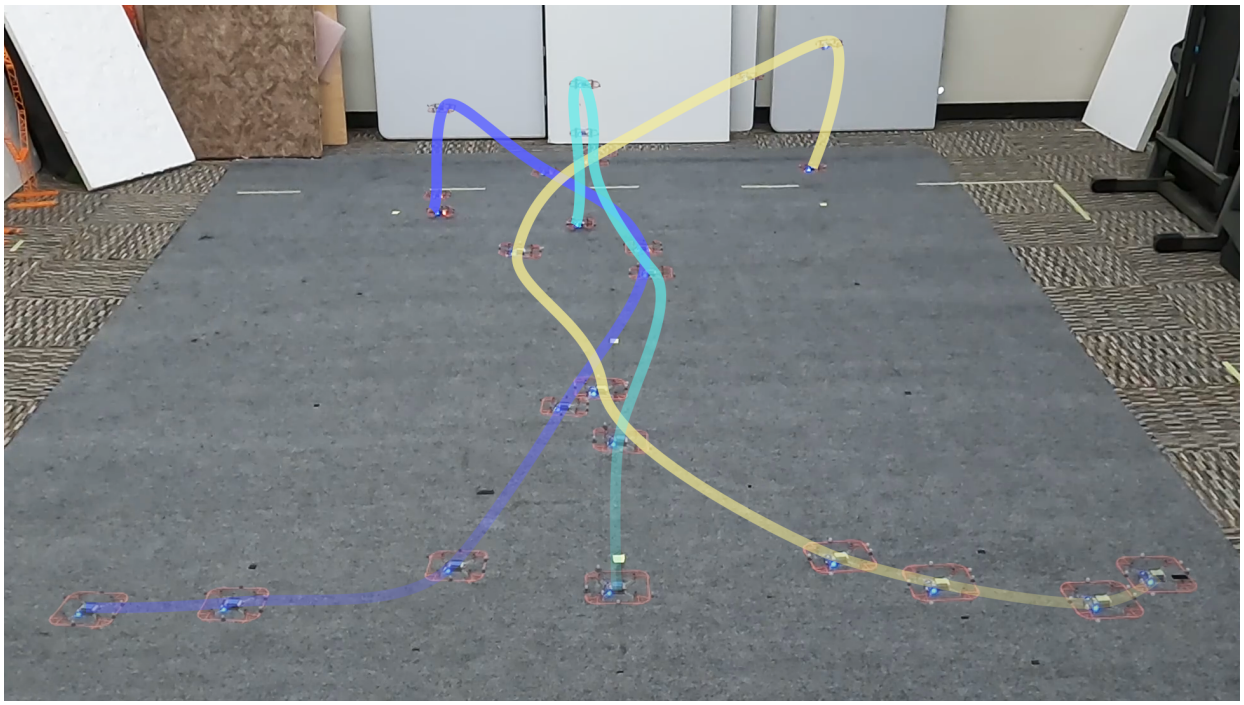


Figure 4.4: Crazyflie hardware experiment with the MARL policy learned by our method. The three drones have to pass through two common waypoints to get to their landing location. The trajectories corresponding to the video footage are visualized in Fig. 4.5 (b).

serves 500 passengers per hour during peak hours, equating to two operations (takeoffs and landings) per minute, with each operation accommodating a 4-passenger aircraft. This corresponds to about 125 trips per hour. We chose vertiports from multiple locations in the Bay Area with high travel demand and designed the air corridors with a lateral separation of 1500 ft based on a preliminary analysis of the separation standards for UAM [174]. Waypoints are created to connect the trails of these corridors spaced 3-4 km apart. For simplicity and clearer visualization, our simulations focus on aerial vehicles traveling westward (from the East Bay to San Francisco and to the South Bay); we assume outbound trips use a separate fixed altitude, thus our study addresses only horizontal deconfliction.

An important modeling assumption is the required separation distance between vehicles. We base these restrictions on industry standards that define the minimum safe distance between the aircraft and potential hazards to maintain an acceptable collision risk [279]. Although there is no single global standard for the separation distance for AAM vehicles yet, we adopt a set of parameters from a variety of literature: the separation distance from dynamic obstacles, maximum accelerations, and bank angle to ensure safety and passenger comfort, [92, 279, 174], which guides the minimum horizontal distances between UAVs to range from 500 to 2200 feet. In our study, a horizontal separation of 1500 ft was used, as it aligns with NASA's UAM corridor design and provides a good balance between collision risk and operational efficiency.

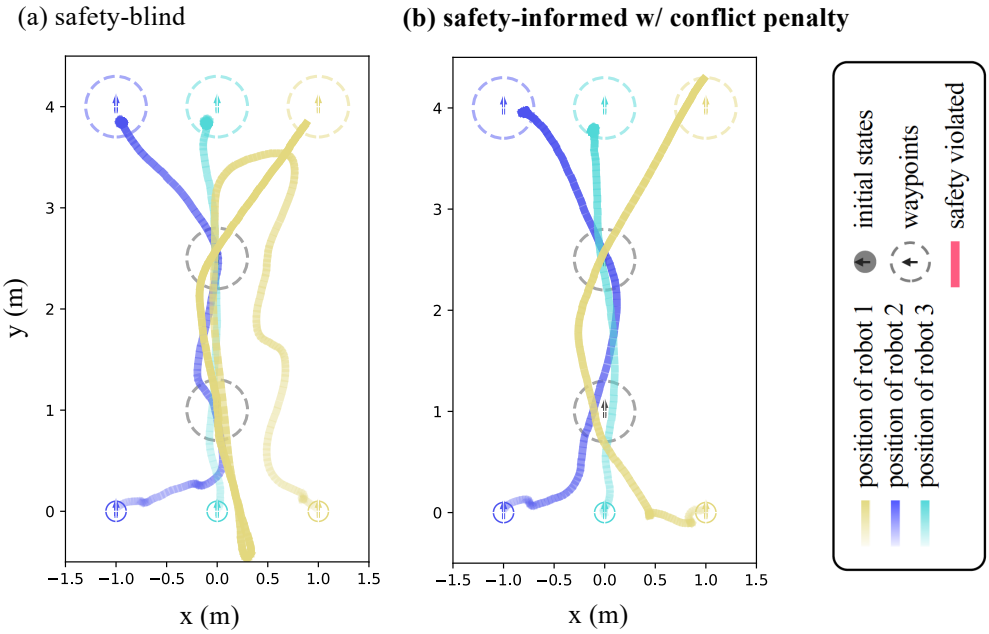


Figure 4.5: We compare the recorded Crazyflie hardware experiment trajectories under our method and the baseline policy trained without the safety filter. With our approach, the drones smoothly deconflict and efficiently complete the task. In contrast, under the baseline policy, the yellow Crazyflie misses a waypoint and must make a second pass. These results demonstrate that incorporating layered safety information during training improves the performance of the MARL policy.

Table 4.4: Simulation results of air taxi operations emulating potential peak traffic around the Bay Area—a scenario in which all vehicles merge into the city-inbound corridor. For performance, we evaluate the mean travel time (s). For safety, we evaluate the percentage of near-collision events ($\text{dist}(s^{(ij)}) < r_{\text{safety}}$) in the timestamped trajectory data (Near collision %), and the percentage of instances having multiple agents encountered within the potential conflict range ($\text{dist}(s^{(ij)}) < r_{\text{conflict}}$) (Conflict %).

Methods	Merging Scenario (N=8, M=5)		
	Travel t(s)(↓)	Near collision(%)(↓)	Conflict(%)(↓)
Safety-blind	675.6	0.055	2.4
No penalty	617.9	0.042	5.5
Proposed	**450.5**	**0.021**	**3.2**

We evaluate three methods: MARL trained without the safety filter and no safety penalty (**Safety blind**), MARL trained with the safety filter under the proposed curriculum but with no safety penalty (**No penalty**), and the proposed safety-informed method employing the safety filter, curriculum, and the potential conflict penalty (**Proposed**) in two high-density air traffic scenarios. The two scenarios, illustrated in Figure 4.6, represent different commuting patterns in the Bay Area. In the left scenario (**Merge Scenario**), multiple air routes (eight in total) from the northern Bay merge into a single

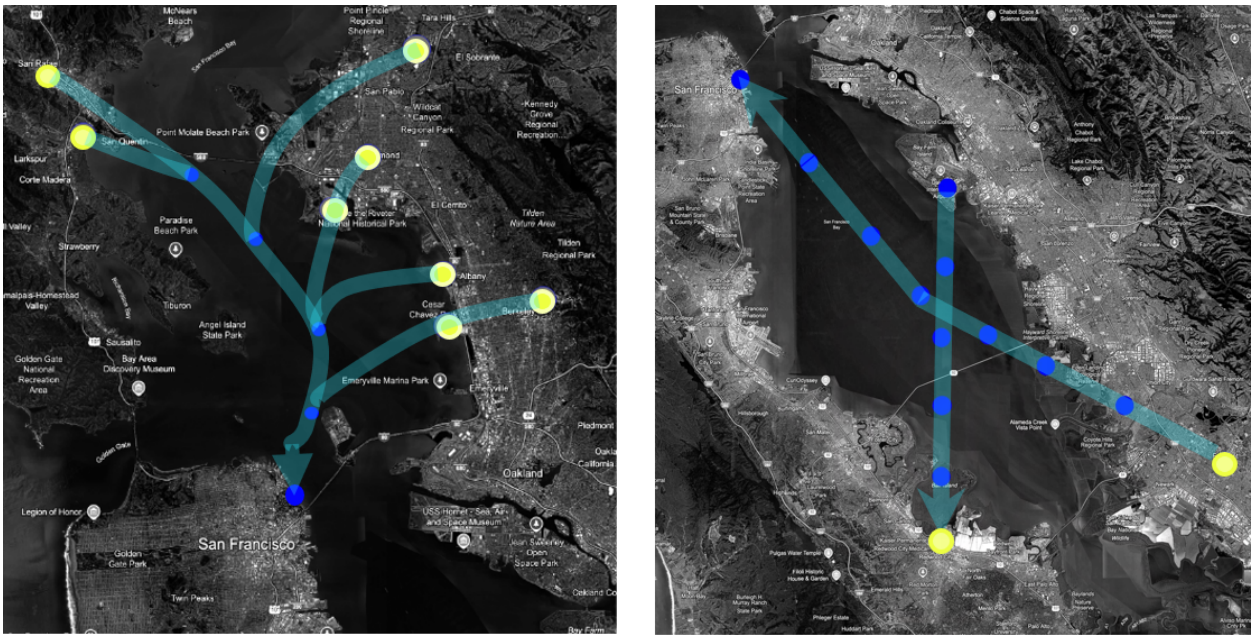


Figure 4.6: Bay Area case scenarios. The left panel illustrates routes where multiple air taxi vehicles would travel from the North and East Bay toward San Francisco, merging into a single air corridor. The right panel shows intersecting air corridors: one where the vehicles would travel from Fremont (southeast) to San Francisco, and another from Oakland (northeast) to Redwood City. The blue dots represent the waypoints that UAVs follow, while the yellow dots indicate the departure or an incoming waypoint of the corridor.

Table 4.5: Simulation results of air taxi operations—a scenario in which two air corridors intersect with each other.

Methods	Intersection Scenario (N=16, M=6)		
	Travel t(s)(↓)	Near collision(%)(↓)	Conflict(%)(↓)
Safety blind	987.4	0.058	2.1
No penalty	780.5	0.129	3.8
Proposed	**660.8**	**0.056**	**1.6**

corridor leading to San Francisco. The departure time of each vehicle varies randomly within a 60-second range. In the scenario shown to the right (**Intersection Scenario**), two westbound air corridors intersect. Here, we set the UAVs to leave the origin every 90 ± 15 seconds to intentionally induce congestion at the intersection.

We evaluate 25 random episodes for each method and report the results in Table 4.4 and Table 4.5 for each scenario, respectively. The results show that the proposed method achieves both the highest performance, measured by the shortest mean travel time, and the lowest percentage of near-collision events.

Examples of vehicle trajectories for each scenario and method are visualized in Figure 4.7. In the merging scenario, our method demonstrates the most efficient deconfliction of trajectories when multiple vehicles merge into the air corridor. In the intersection scenario, near the intersection, the

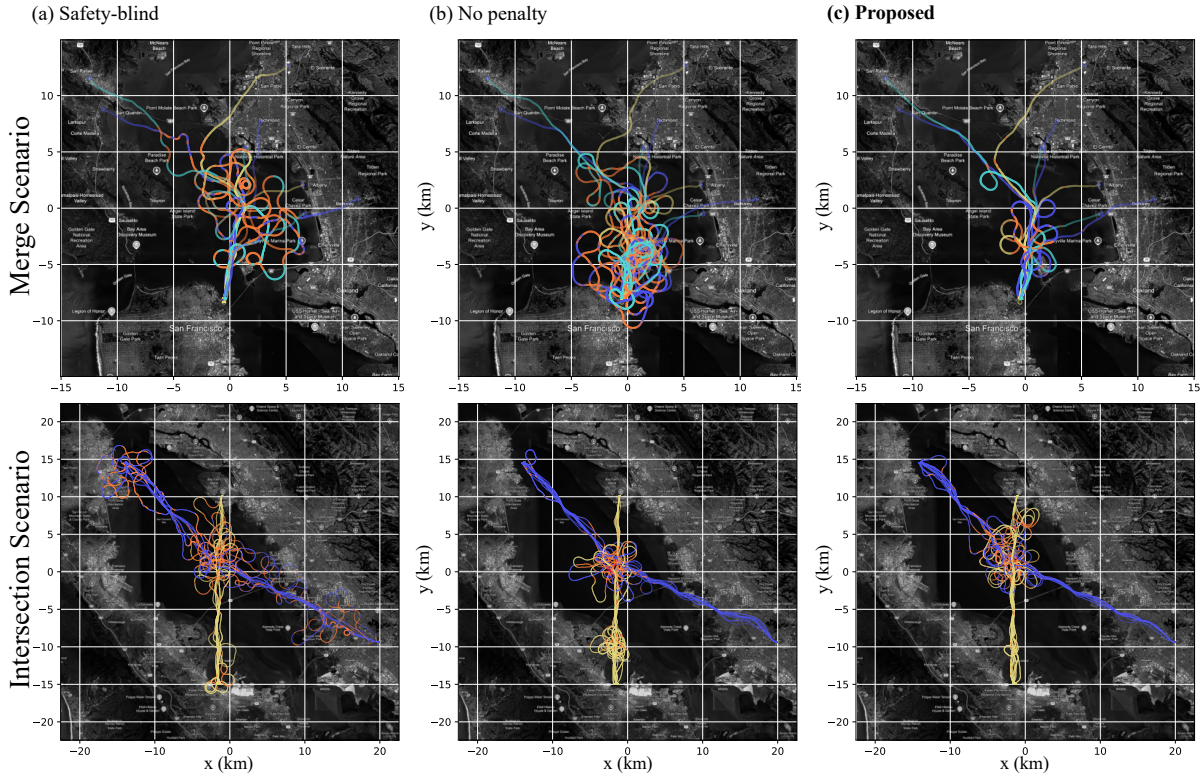


Figure 4.7: Comparison of air taxi trajectories in merging and crossing scenarios: The top row illustrates the single-lane merging scenario, where UAVs converge into a shared inbound air corridor, while the bottom row depicts intersecting air corridors. In the merging scenario, our method achieves the most efficient deconfliction of trajectories, minimizing congestion near the corridor. In the crossing scenario, our method demonstrates a wider safety buffer around intersections, as UAVs actively maintain greater separation to mitigate conflicts. Videos are available in the supplementary material.

region occupied by vehicles as they maneuver to avoid collisions is noticeably larger in our method compared to the second method, trained with no penalty. This indicates that vehicles using our approach proactively maintain greater separation to mitigate the conflicting constraints.

While we expect that a fully operational ATM system for AAM will be significantly more efficient, seamless, and safer than our simulation study suggests, we present this study as an initial guideline for resolving hypothetical emergency scenarios. For instance, the situations we simulated can emerge when the airspace congestion coincides with the loss of centralized traffic control, requiring each agent to make independent, safe decisions.

4.7 Limitations

While our framework shows significant improvements in achieving safety and performance, there are limitations, which we list below.

Table 4.6: Results of policies trained under various methods for Crazyflie dynamics: We evaluate mean travel time (s) and number of reached waypoints (Waypoint #) for performance, and the percentage of the events involving multiple agents encountered within the potential conflict range in the trajectory data (Conflict %) for safety risk. Note that in these simulations, the agent never violated safety for all methods due to our safety filter, except in the training scenario when the agent is initialized at the safety-violating states. (N=number of agents, M=number of waypoints, L=world size)

Methods	Scenario 1 (Training) (N=4, M=2, $L=4$)			Scenario 2 (N=6, M=3, $L=6$)			Scenario 3 (N=3, M=3, $L=3$)		
	Travel time(s)(↓)	Waypoint#(↑)	Conflict(%)(↓)	Travel t	Waypoint #	Conflict	Travel t	Waypoint #	Conflict
1 (safety-blind)	18.33	1.62 ± 0.25	8.7	29.18	2.11 ± 0.24	21.9	19.08	2.14 ± 0.53	16.2
2	18.21	1.67 ± 0.19	7.1	29.77	2.06 ± 0.24	19.7	18.92	2.40 ± 0.47	16.0
3	17.73	1.76 ± 0.16	6.6	28.59	2.26 ± 0.21	19.7	18.44	2.41 ± 0.46	14.9
4	18.73	1.58 ± 0.23	7.2	29.11	2.20 ± 0.21	17.1	18.79	2.17 ± 0.47	13.4
5 (no penalty)	17.56	1.75 ± 0.17	5.3	28.46	2.33 ± 0.20	16.9	16.09	2.78 ± 0.28	11.8
6	18.31	1.67 ± 0.17	5.9	28.92	2.27 ± 0.21	17.0	17.69	2.42 ± 0.42	15.0
7	17.90	1.71 ± 0.18	5.3	28.98	2.32 ± 0.18	15.8	17.59	2.54 ± 0.34	11.2
8	20.52	1.26 ± 0.21	2.7	31.25	1.74 ± 0.27	6.9	18.36	2.20 ± 0.36	8.6
9 (proposed)	**17.81**	$\mathbf{1.78 \pm 0.18}$	**5.4**	**28.59**	$\mathbf{2.42 \pm 0.20}$	**15.1**	**16.91**	$\mathbf{2.71 \pm 0.24}$	**10.8**

1. *Scalability to higher dimensions*: Our current framework is designed for 2D scenarios, and needs extending it to 3D environments.

2. *Guarantees for multiple engagements*: Our safety guarantees are currently limited to pairwise interactions. While our method is designed to significantly mitigate collision risks in multi-agent interactions based on theoretical analysis, it does not provide formal guarantees for scenarios involving multi-agent engagements. For further details, see Remark 9.

3. *Hardware experimentation constraints*: In our hardware experiments, local observations were emulated using a motion capture (mocap) system rather than being obtained through onboard sensing. This simplification may not fully reflect real-world operational constraints and should be addressed in future implementations. Additionally, fixed-wing vehicles were not included in our experiments.

4. *Communication range constraints*: The impact of communication range limitations in our hardware experiments was not analyzed.

4.8 Conclusions

In this chapter, we presented a layered architecture combining a CBVF-based safety filtering mechanism with a MARL policy, demonstrating its effectiveness in ensuring both safety and efficiency. Our approach enables MARL to navigate conflicts proactively while benefiting from safety-informed reward signals. Along with the safety filter introduced during training using a curriculum learning approach, the Layered Safe MARL framework achieved shorter travel times and reached more waypoints with fewer conflicts. The key components of our approach, curriculum

learning, and cost terms that inform potential conflict zones are agnostic to the choice of the MARL algorithm. We validated our method by applying it to two distinct dynamics—quadrotor and fixed-wing AAM flight dynamics—and evaluated it in progressively complex scenarios. We also conducted hardware experiments on three Crazyflie drones, highlighting the applicability of our method in real-world systems.

Our method integrates model-based safety tools from control theory (CBVFs) with learning-based methods (MARL), together forming a framework that addresses two major challenges in multi-agent problems—safety and efficient coordination. While deep reinforcement learning has faced skepticism in safety-critical applications such as air traffic management, recent advances—including our work—demonstrate the viability of hybrid approaches that combine learning and control, and illustrate how RL can be responsibly applied in safety-critical settings.

Future research could investigate decomposition techniques and learning-based reachability analysis (e.g., DeepReach [19]) to extend safety verification to higher-dimensional settings. Adapting to other methods, including other MARL algorithms (e.g., MAPPO or even further refining DG-PPO), MPC, or game-theoretic solutions, by treating the potential conflict zone as a soft constraint, is an exciting future work direction. Further investigation is needed to assess how communication constraints affect coordination and safety in decentralized multi-agent systems. Finally, an important future direction is testing the proposed approach's applicability in various robotics domains, ranging from higher-order dynamics to complex environments and sensing constraints, such as ground robots, underwater autonomous vehicles, and space robots.

Acknowledgments For This Chapter

We thank Dr. Mir Abbas Jalali (Joby), George Gorospe (NASA), Dr. Anthony Evans (Airbus), Inkyu Jang (SNU) and Kanghyun Ryu (UC Berkeley) for the helpful discussions. Jasmine Aloor and Hamsa Balakrishnan were supported in part by NASA under Grant No. 80NSSC23M0220. Jason J. Choi, Jingqi Li and Claire J. Tomlin were supported in part by DARPA Assured Autonomy under Grant FA8750-18-C-0101, DARPA ANSR under Grant FA8750-23-C-0080, the NASA ULI Program in Safe Aviation Autonomy under Grant 62508787-176172, and ONR Basic Research Challenge in Multibody Control Systems under Grant N00014-18-1-2214. Jason J. Choi received the support of a fellowship from Kwanjeong Educational Foundation, Korea. Jasmine J. Aloor was supported in part by a Mathworks Fellowship. Maria G. Mendoza acknowledges support from NASA under the Clean Sheet Airspace Operating Design project MFRA2018-S-0471. The authors would like to thank the MIT SuperCloud [263] and the Lincoln Laboratory Supercomputing Center for providing high performance computing resources that have contributed to the research results reported within this chapter.

Appendix

Reward function design for goal reaching

We assume that, from the agent's local observation, it can evaluate its state, including the position and heading relative to the current target waypoint, denoted as $s^{(i)}_{\text{ref}}$. The additional reward $\mathcal{R}_{\text{tracking}}$ is designed to guide agents to efficiently navigate to the target waypoint. It uses a distance-like measure relative to the waypoint position, and can incorporate additional information like the errors from the desired heading angle and speed. The vehicle dynamics also inform the reward design. We design a specific reward for each quadrotor and air taxi dynamics. For the quadrotor, although the vehicle dynamics are holonomic, we want the vehicle to approach the waypoint from a specific target heading direction. To achieve this, we design a reference velocity field, $v_{\text{ref}}(s^{(i)}_{\text{ref}})$, around the waypoint, shaping it like the magnetic field around a solenoid. Then, the reward for the waypoint tracking is given as $\mathcal{R}_{\text{tracking}}(o^{(i)}_k, a^{(i)}_k) = ||v^{(i)} - v_{\text{ref}}||$. For the air taxi dynamics, shaping the reference velocity field is not straightforward, as the vehicle is nonholonomic and is mainly limited by its maximum turning radius, determined by its speed and yaw rate. Thus, we compute the time-to-reach function [320], which is the minimum time required to reach the target waypoint (satisfying the threshold conditions) subject to the vehicle dynamics. This time-to-reach reward is also used in [203] for RL-based navigation of mobile robots. The use of the time-to-reach reward guides the vehicle to learn how to perform a 360-degree turn when it misses its waypoint.

Air taxi dynamics: additional information

The relationship between $s^{(ij)}$ and ($s^{(i)}$, $s^{(j)}$) for the air taxi dynamics in (4.10) is given by

$$\begin{aligned} x^{(ij)} &= \cos\theta^{(ij)}(x^{(j)} - x^{(i)}) + \sin\theta^{(ij)}(y^{(j)} - y^{(i)}), \\ y^{(ij)} &= -\sin\theta^{(ij)}(x^{(j)} - x^{(i)}) + \cos\theta^{(ij)}(y^{(j)} - y^{(i)}), \\ \theta^{(ij)} &= \theta^{(i)} - \theta^{(j)}. \end{aligned} \tag{4.20}$$

The relative dynamics (4.3) can be derived from (4.10) and (4.20), and are express as

$$\begin{aligned} \dot{x}^{(ij)} &= -v^{(i)} + v^{(j)}\cos\theta^{(ij)} + y^{(ij)}\omega^{(i)} \\ \dot{y}^{(ij)} &= v^{(j)}\sin\theta^{(ij)} - x^{(ij)}\omega^{(i)} \\ \dot{\theta}^{(ij)} &= \omega^{(j)} - \omega^{(i)}. \end{aligned} \tag{4.21}$$

Ablation Study: Details

We conduct comparison studies among the following nine methods:

1. Policy trained without the safety filter and no safety penalty **(Safety blind)**
2. Policy trained without safety filter and with $\mathcal{C}_{\text{plain}}$

3. Policy trained without safety filter and with $\mathcal{C}_{\text{conflict}}$
4. Policy trained with the safety filter and without curriculum learning (with no penalty)
5. Policy trained with the safety filter and no safety penalty **(No penalty)**
6. Policy trained with the safety filter and with $\mathcal{C}_{\text{plain}}$
7. Policy trained with the safety filter and with $\mathcal{C}_{\text{cbvf}}$
8. Policy trained with the safety filter and with $\mathcal{C}_{\text{norm.diff}}$
9. **Policy trained with the safety filter and with $\mathcal{C}_{\text{conflict}}$ (Proposed)**

Note that methods 5-9 are trained with curriculum learning on r_{safety}, as described in Section 4.5. Every policy we compared has been trained for the same number of environment steps. Methods 1, 5, and 9 correspond to the methods we also evaluate in the air taxi operation simulation.

Table 4.6 summarizes the results of the simulation study, and Figure 4.8 visualizes example trajectories in Scenario 2 under the policies of methods 1, 5, and 9. Each method is evaluated using four random seeds, with 25 episodes per seed, totaling 100 random episodes. Note that in these simulations, the agent never violated safety for all methods due to our safety filter, except in the training scenario when the agent is initialized at the safety-violating states. Thus, the percentage of near-collision events (safety violation) is not reported, and only the rate of potential conflict (for instance, when more than two agents enter the potential conflict range r_{conflict}) is calculated.

The key aspects of the result in Table 4.6 are:

- *Effect of using the safety filter in training (1-3 vs 5-6, 9):* Methods that incorporate the safety filter during training consistently outperform their counterparts trained without the filter across all metrics.
- *Effect of curriculum learning (4 vs 5):* The curriculum learning can significantly enhance performance by reducing the conservativeness of the policy.
- *Effect of potential conflict penalty $\mathcal{C}_{conflict}$ compared to other penalty candidates (6, 7, 8 vs 9):* Although method 8 that uses $\mathcal{C}_{\text{norm.diff}}$ consistently shows the lowest rate of potential conflict, and its performance is significantly impaired by the penalty. Our method achieves the best performance in most cases. Importantly, our method outperforms other methods, especially when there is a larger number of agents (Scenario 2).

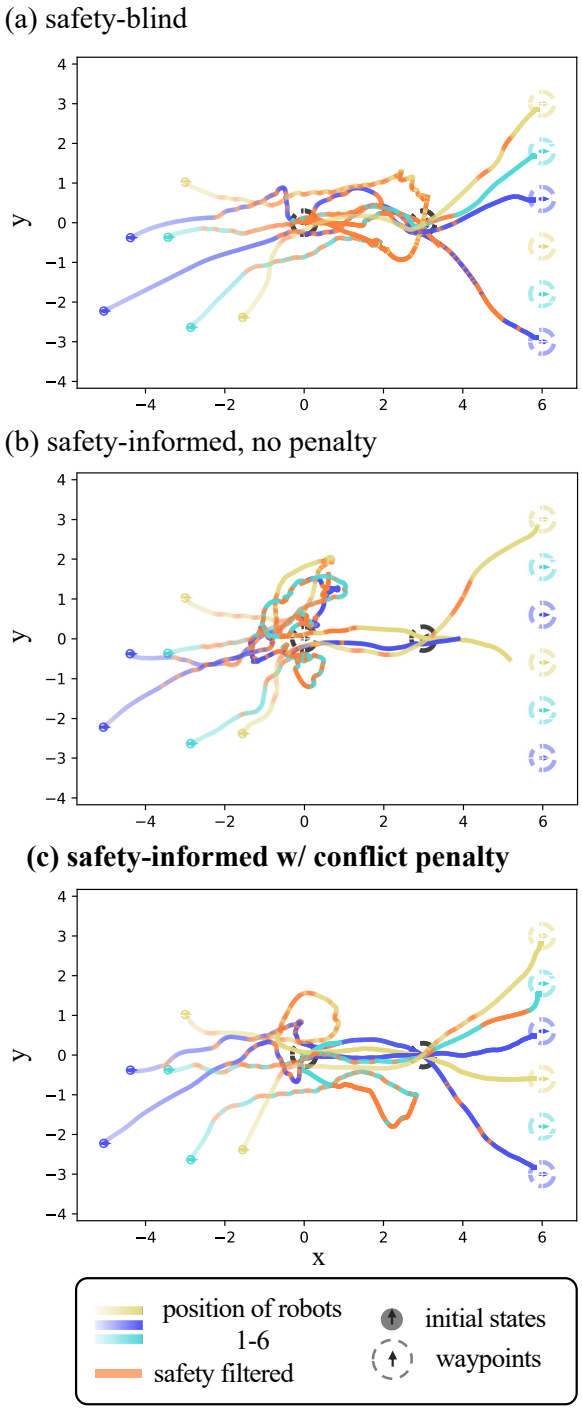


Figure 4.8: Simulation results of (a) safety-blind (method 1), (b) safety-informed with no penalty (method 5), and (c) safety-informed with potential conflict penalty (method 9) under Scenario 2 in Table 4.6, trained for double integrator dynamics. Agents are initialized at random positions and have to merge into a line formed by two waypoints before reaching their final waypoints. While our safety filter ensures safety for all cases, the MARL method trained with a potential conflict penalty shows the most efficient behavior for reaching waypoints. Videos are available in the supplementary material.

Part II

Game-Theoretic Decision-Making

Chapter 5

Primal Dual Interior Point Method for Nonlinear Feedback Stackelberg Games

As demonstrated in previous chapters, ensuring robust safety among agents often leads to overly conservative behavior when planning for worst-case scenarios in decentralized decision-making environments. However, in many practical applications, agents have well-defined individual objectives that can be understood and anticipated by other agents. This insight suggests that instead of conservatively planning for arbitrary worst-case scenarios, we can model agents as rational decision-makers with known objectives. Game theory provides a natural modeling framework for anticipating the actions of such rational agents when conflicting objectives or constraints exist among them. Nevertheless, computing game-theoretic equilibria remains computationally challenging. In this chapter, we focus on feedback Stackelberg equilibrium and demonstrate how its computation can be reformulated as solving a series of nested Karush-Kuhn-Tucker (KKT) condition equations. Leveraging this mathematical connection, we propose an efficient second-order algorithm using primal-dual interior point methods to compute game-theoretic equilibria under coupled constraints among agents. This chapter is based on published work [181], co-authored with Somayeh Sojoudi, Claire J. Tomlin, and David Fridovich-Keil.

5.1 Background

Dynamic game theory [24] provides tools for analyzing strategic interactions in multi-agent systems. It has broad applications in control [52], biology [168], and economics [126]. A well-known equilibrium concept in dynamic game theory is the *Nash equilibrium* [229], where players pursue strategies that are unilaterally optimal, and players make decisions simultaneously. However, this may not apply to a broad class of games where a decision hierarchy exists, such as lane-merging in highway driving [324], predator-prey competition in biology [18], and retail markets in economics [187]. These games could be more naturally formulated as *Stackelberg games* [301], where players act sequentially in a predefined order. For such games, the *Stackelberg equilibrium* is the appropriate equilibrium concept.

The formulation of Stackelberg equilibria depends on the information structure [24]. For instance, in scenarios where players lack access to the current game state, one can compute an *open-loop Stackelberg equilibrium* (OLSE). At such an equilibrium, players' decisions depend on the initial state of a game and followers' decisions are influenced by the leaders'. When players also have access to state information and their prior players' actions, it becomes appropriate to compute a *feedback Stackelberg equilibrium* (FSE), where each player's decision is contingent upon the current state and the actions of preceding players. One advantage of FSE over OLSE is its sub-game perfection, meaning that decision policies remain optimal for future stages, even if the state is perturbed at an intermediate stage. This feature is particularly beneficial in scenarios with feedback interactions among players, such as in lane merging during highway driving [288] and human-robot interactions [103]. In these situations, the sub-game perfection of FSE makes it a more suitable equilibrium concept than OLSE, as it allows players to adjust their decisions based on the current state information.

Though FSE is conceptually appealing, computing it poses significant challenges [129, 308, 312, 207]. Previous research has extensively explored the FSE problem in finite dynamic games, characterized by a finite number of states and actions [281, 24, 293, 160, 17, 305]. In contrast, infinite dynamic games–those with an infinite number of states and actions–have mostly been considered within the framework of linear quadratic (LQ) games, featuring linear dynamics and stage-wise quadratic costs [107, 24, 75, 297, 150]. The computation of FSE for more general nonlinear games is more challenging than for LQ games. A naive application of existing dynamic programming solutions in finite dynamic games necessitates gridding the continuous state and action spaces, often leading to computational intractability [28]. Recent works [226, 317] have proposed using approximate dynamic programming to compute an approximate FSE for input-affine systems. Additionally, several iterative linear-quadratic (LQ) approximation approaches have been proposed in [137, 154], but they lack convergence guarantees.

Moreover, existing approaches are ill-suited for handling coupled equality and inequality constraints on players' states and decisions, which frequently arise in safety-critical applications such as autonomous driving [284] and human-robot interaction [155]. For instance, existing iterative LQ game solvers [154, 137] cannot be directly integrated with the primal log barrier penalty method [250] to incorporate these constraints. The most relevant studies, such as [223, 89, 205], focus on computing OLSE in games under linear constraints. This chapter aims to bridge this gap in the literature.

Our contributions are threefold: (1) We first reformulate the N-player feedback Stackelberg equilibrium problem, characterized by N players making sequential decisions over time, into a sequence of nested optimization problems. This reformulation enables us to derive the Karush–Kuhn–Tucker (KKT) conditions and a second-order sufficient condition for the feedback Stackelberg equilibrium. (2) Using these results, we propose a Newton-style primal-dual interior point (PDIP) algorithm for computing a local FSE for LQ games. Under certain regularity conditions, we show the convergence of our algorithm to a local FSE. (3) Finally, we propose an efficient PDIP method for approximately computing a local FSE for more general nonlinear games under (nonconvex) coupled equality and inequality constraints. The computed feedback policy locally approximates the ground truth nonlinear policy. Theoretically, we characterize the approximation error of our method, and show

the exponential convergence under certain conditions. Empirically, we validate our algorithm in a highway lane merging scenario, demonstrating its ability to tolerate infeasible initializations and efficiently converge to a local FSE in constrained nonlinear games.

5.2 Related Works

Closely related to the feedback Stackelberg equilibrium (FSE), the feedback Nash equilibrium (FNE) has been extensively studied, for example, in [23, 24, 262, 167]. Our work builds upon [167], where the authors proposed KKT conditions for constrained FNE. However, the FNE KKT conditions in [167] fail to hold true for FSE due to the decision hierarchy in FSE. In our work, we introduce a set of new KKT conditions for FSE. Another key difference is that we adopt the primal-dual interior point method for solving LQ and nonlinear games, whereas [167] considers the active-set method. In general, the former has polynomial complexity, but the latter has exponential complexity [112]. Moreover, we are able to prove the exponential convergence of our algorithm under certain conditions. However, there is no such convergence proof in [167].

As highlighted in the literature, e.g., [24, 297, 226, 185, 317], the dominant approach to computing unconstrained FSE is using (approximate) dynamic programming. LQ games can be solved efficiently via exact dynamic programming; however, in more general nonlinear cases the value function could be hard to compute and, in general, has no analytical solution [226]. Compared with those works, our approach could be considered as computing an efficient local approximation of the value function along the state trajectory under a local FSE policy instead of approximating the value function everywhere as in [226].

Finally, to further motivate our work, we examine whether the Stackelberg equilibrium can be effectively approximated by the Nash equilibrium and whether the FSE can be accurately approximated by its open-loop counterpart. According to [160], in repeated matrix games, the Stackelberg equilibrium may coincide with the Nash equilibrium. However, in Appendix 5.9, we present a counterexample demonstrating that, in games with quadratic costs—reminiscent of oligopoly models in economics [298]—the Stackelberg equilibrium can deviate arbitrarily from the Nash equilibrium. Moreover, there is a recent trend of approximating feedback policies via receding horizon open-loop policies [170, 333], where an open-loop policy is re-solved at each time for future steps. However, we show in another counter-example in Appendix 5.9 that the trajectory under the feedback Stackelberg policy and the one under the receding horizon open-loop Stackelberg policy could be quite different, even if there is no state perturbation. Thus, it is essential to develop specific tools for computing the feedback Stackelberg equilibrium.

5.3 Constrained Feedback Stackelberg Games

In this section, we introduce the formulation of constrained feedback Stackelberg games. We formulate the problem by extending the N-player feedback Stackelberg games [107] to its constrained setting. We denote by $\mathbb{N}$ and $\mathbb{R}$ the sets of natural numbers and real numbers, respectively. Given

$j, k \in \mathbb{N}$, we denote by $\mathbf{I}_j^k = \{j, j+1, \ldots, k\}$ if $j \leq k$ and $\emptyset$ otherwise. Let $T \in \mathbb{N}$ be the time horizon over which the game is played. At each time t, we denote by x_t and $u_t^i \in \mathbb{R}^{m_i}$ the state of the entire game and the control input of player i, respectively. We define $u_t := [u_t^1, u_t^2, \ldots, u_t^N] \in \mathbb{R}^m$, with $m := \sum_{i=1}^N m_i$, to be the joint control input at time t. Moreover, at each time t, players make decisions in the order of their indices. We consider the time-varying dynamics

$$x_{t+1} = f_t(x_t, u_t), \tag{5.1}$$

where $f_t(x_t, u_t) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is assumed to be a twice differentiable function. Given a sequence of control inputs $\mathbf{u} := [u_0, u_1, \ldots, u_T] \in \mathbb{R}^{Tm}$, we denote by $\mathbf{x} := [x_0, x_1, \ldots, x_{T+1}] \in \mathbb{R}^{(T+1)n}$ a state trajectory under dynamics (5.1).

At each time $t \in \mathbf{I}_0^T$, we denote the stage-wise cost of player $i \in \mathbf{I}_1^N$ by $\ell_t^i(x_t, u_t) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, and associate with each player a terminal cost, $\ell_{T+1}^i(x_{T+1}) : \mathbb{R}^n \to \mathbb{R}$. Each player $i \in \mathbf{I}_1^N$ considers the following time-separable costs,

$$J^i(\mathbf{x}, \mathbf{u}) = \sum_{t=0}^{T} \ell_t^i(x_t, u_t) + \ell_{T+1}^i(x_{T+1}). \tag{5.2}$$

Moreover, let $n_{h,t}^i$ and $n_{g,t}^i$ be the number of equality and inequality constraints held by player $i \in \mathbf{I}_1^N$ at time t, respectively. We denote the equality and inequality constraint functions of player i by $h_t^i(x_t, u_t) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n_{h,t}^i}$ and $g_t^i(x_t, u_t) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n_{g,t}^i}$, respectively. We specify the stage-wise equality and inequality constraints of player $i \in \mathbf{I}_1^N$ as

$$0 = h_t^i(x_t, u_t),\ 0 \leq g_t^i(x_t, u_t). \tag{5.3}$$

At the terminal time $t = T+1$, we represent the equality and inequality constraint functions of player $i \in \mathbf{I}_1^N$ by $h_{T+1}^i(x_{T+1}) : \mathbb{R}^n \to \mathbb{R}^{n_{h,T+1}^i}$ and $g_{T+1}^i(x_{T+1}) : \mathbb{R}^n \to \mathbb{R}^{n_{g,T+1}^i}$, respectively. We consider the following equality and inequality constraints of player $i \in \mathbf{I}_1^N$ at the terminal time,

$$0 = h_{T+1}^i(x_{T+1}),\ 0 \leq g_{T+1}^i(x_{T+1}). \tag{5.4}$$

We remark that these definitions generate coupled dynamics and constraints among different players at each time $t \in \mathbf{I}_0^{T+1}$. We consider the following regularity assumption, following [167, 57].

Assumption 2. *The feasible set $\mathcal{F} := \{x \in \mathbb{R}^{(T+1)n}, u \in \mathbb{R}^{Tm} : h_t^i(x_t, u_t) = 0, g_t^i(x_t, u_t) \geq 0, h_{T+1}^i(x_{T+1}) = 0, g_{T+1}^i(x_{T+1}) \geq 0, x_{t+1} = f_t(x_t, u_t), \forall i \in \mathbf{I}_1^N, t \in \mathbf{I}_0^T\}$ is compact. The costs, dynamics, equality and inequality constraints are twice differentiable and bounded, but could be nonconvex in general.*

Local Feedback Stackelberg Equilibria

Before we formalize the decision process of feedback Stackelberg games, we introduce a few notations to compactly represent different players' control at different times. We define $u_{t:t'}^{i:i'} := \{u_\tau^j, \tau \in \mathbf{I}_t^{t'}, j \in \mathbf{I}_i^{i'}\}$. In particular, we define $u_t^{1:i-1} := \emptyset$ when $i = 1$ and $u_t^{i+1:N} := \emptyset$ when $i = N$. We also denote by $u_{t+1:T}^{1:i} := \emptyset$ when $t = T$.

The policy of each player can be defined as follows. At the t-th stage, since player 1 makes a decision first, its policy function $\pi_t^1(x_t) : \mathbb{R}^n \to \mathbb{R}^{m_1}$ depends only on the state x_t. For players $i \in \mathbf{I}_2^N$, the policies are modeled as $\pi_t^i(x_t, u_t^{1:i-1}) : \mathbb{R}^n \times \mathbb{R}^{\sum_{j=1}^{i-1} m_j} \to \mathbb{R}^{m_i}$. We will define the concept of *local feedback Stackelberg equilibria* in the remainder of this subsection.

At the terminal time $t = T + 1$, we define the state-value functions for a player $i \in \mathbf{I}_1^N$ as

$$V_{T+1}^i(x_{T+1}) := \begin{cases} \ell_{T+1}^i(x_{T+1}) & \text{if } \begin{smallmatrix} 0=h_{T+1}^i(x_{T+1}) \\ 0\le g_{T+1}^i(x_{T+1}) \end{smallmatrix} \\ \infty & \text{else.} \end{cases} \tag{5.5}$$

At time $t \le T$, we first construct the state-action-value function for player N:

$$Z_t^N(x_t, u_t^{1:N-1}, u_t^N) := \begin{cases} \ell_t^N(x_t, u_t) + V_{t+1}^N(x_{t+1}) & \text{if } \begin{smallmatrix} 0=x_{t+1}-f_t(x_t,u_t) \\ 0=h_t^N(x_t,u_t) \\ 0\le g_t^N(x_t,u_t) \end{smallmatrix} \\ \infty & \text{else.} \end{cases} \tag{5.6}$$

Given $(x_t, u_t^{1:N-1})$, there could be multiple u_t^N minimizing $Z_t^N(x_t, u_t^{1:N-1}, u_t^N)$. We define player N's local FSE policy π_t^N by picking an arbitrary local minimizer u_t^{N*},

$$\pi_t^N(x_t, u_t^{1:N-1}) := u_t^{N*} \in \arg\min_{\tilde{u}_t^N} Z_t^N(x_t, u_t^{1:N-1}, \tilde{u}_t^N). \tag{5.7}$$

We then construct the state-action-value function of player $i \in \mathbf{I}_2^{N-1}$,

$$Z_t^i(x_t, u_t^{1:i-1}, u_t^i) := \begin{cases} \ell_t^i(x_t, u_t) + V_{t+1}^i(x_{t+1}) & \text{if } \begin{smallmatrix} 0=x_{t+1}-f_t(x_t,u_t) \\ 0=h_t^i(x_t,u_t) \\ 0\le g_t^i(x_t,u_t) \\ u_t^j=\pi_t^j(x_t,u_t^{1:j-1}),\ j\in\mathbf{I}_{i+1}^N \end{smallmatrix} \\ \infty & \text{else,} \end{cases} \tag{5.8}$$

and its local FSE policy π_t^i by picking an arbitrary local minimizer $u_t^{i^*}$,

$$\pi_t^i(x_t, u_t^{1:i-1}) := u_t^{i*} \in \arg\min_{\tilde{u}_t^i} Z_t^i(x_t, u_t^{1:i-1}, \tilde{u}_t^i). \tag{5.9}$$

We finally construct the state-action-value function of the first player:

$$Z_t^1(x_t, u_t^1) := \begin{cases} \ell_t^1(x_t, u_t) + V_{t+1}^1(x_{t+1}) & \text{if } \begin{smallmatrix} 0=x_{t+1}-f_t(x_t,u_t) \\ 0=h_t^1(x_t,u_t) \\ 0\le g_t^1(x_t,u_t) \\ u_t^j=\pi_t^j(x_t,u_t^{1:j-1}),\ j\in\mathbf{I}_2^N \end{smallmatrix} \\ \infty & \text{else,} \end{cases} \tag{5.10}$$

and its local FSE policy

$$\pi_t^1(x_t) := u_t^{1*} \in \arg\min_{\tilde{u}_t^1} Z_t^1(x_t, \tilde{u}_t^1). \tag{5.11}$$

We define the state-value function of player $i \in \{1, 2, \ldots, N\}$ at time $t \le T$ as

$$V_t^i(x_t) = Z_t^i(x_t, u_t^{1*}, \ldots, u_t^{i*}), \tag{5.12}$$

where $u_t^{j*} = \pi_t^j(x_t, u_t^{1:(j-1)*}), \forall j \in \mathbf{I}_1^i$.

We formally define the local feedback Stackelberg equilibria as follows.

Definition 1 (Local Feedback Stackelberg Equilibria [24])**.** *Let* $\{\pi_t^i\}_{t=0,i=1}^{T,N}$ *be a set of policies defined in* (5.7)*,* (5.9) *and* (5.11)*, and define* $(\mathbf{x}^*, \mathbf{u}^*)$ *to be a state and control trajectory under the policies* $\{\pi_t^i\}_{t=0,i=1}^{T,N}$*, i.e.,*

$$x_{t+1}^* = f_t(x_t^*, u_t^*),\ u_t^{i*} = \pi_t^i(x_t^*, u_t^{1:(i-1)*}),\ \forall t \in \mathbf{I}_0^T,\ i \in \mathbf{I}_1^N. \tag{5.13}$$

We say that $(\mathbf{x}^*, \mathbf{u}^*)$ *is a* ***local feedback Stackelberg equilibrium trajectory*** *if there exists an* $\epsilon > 0$ *such that, for all* $t \in \mathbf{I}_0^T$*,*

$$\begin{aligned} Z_t^1(x_t^*, \tilde{u}_t^1) &\geq Z_t^1(x_t^*, u_t^{1*}), \\ &\vdots \\ Z_t^N(x_t^*, u_t^{1*}, \ldots, u_t^{(N-1)*}, \tilde{u}_t^N) &\geq Z_t^N(x_t^*, u_t^{1*}, \ldots, u_t^{(N-1)*}, u_t^{N*}) \end{aligned} \tag{5.14}$$

for all $\tilde{u}_t^1 \in \{u : \|u - u_t^{1*}\|_2 \leq \epsilon\}$*, . . . , and* $\tilde{u}_t^N \in \{u : \|u - u_t^{N*}\|_2 \leq \epsilon\}$*.*

The above definition encapsulates the traditional approach to computing feedback Stackelberg equilbiria. This involves optimizing over state-action-value functions, which are obtained by integrating other players' policies into each player's problem and then recording the overall costs.

Remark 10 (Existence of Local Feedback Stackelberg Equilibria)**.** *In general, it is difficult to establish a sufficient condition for the existence of a feedback Stackelberg equilibrium [29]. The main difficulty is that the decision problem of each player is nested within that of other players. It must be solved hierarchically. For example, the existence of feedback Stackelberg policies [202] of a player* $i \in \mathbf{I}_1^{N-1}$ *is related to the topological properties of the set of policies of players* $j \in \mathbf{I}_{i+1}^N$*. Even if all the players' costs are convex, the feedback Stackelberg policy of player* N *at the terminal time could be lower semi-continuous. Subsequently, the cost of player* $(N-1)$ *could become upper semi-continuous when substituting in the* N*-th player's policy into the* $(N-1)$*-th player's cost. Since there may not exist a solution when minimizing an upper semi-continuous function, there may not exist a feedback Stackelberg policy for player* $(N-1)$*. However, if we can show that the policy of each player is always continuous in the state and prior players' controls, and the continuous costs are defined on a compact domain, then there exist feedback Stackelberg equilibria [24].*

We will now proceed to characterize the feedback Stackelberg equilibria in greater detail in the subsequent subsection.

5.4 Necessary and Sufficient Conditions for Local Feedback Stackelberg Equilibria

We show in the following theorem that the dynamic programming problem, as described in Definition 1, can be reformulated as a sequence of nested constrained optimization problems. In this reformulation, the policies for other players are integrated as constraints within the problem of

each player i, instead of being directly substituted into the costs for computing state-action-value functions, as is typical in traditional optimal control literature. This approach enables us to establish KKT conditions for feedback Stackelberg games in the latter part of this subsection.

Theorem 4. *Under Assumption 2, for each $t \in \mathbf{I}_0^T$ and each $i \in \mathbf{I}_1^N$, a local feedback Stackelberg policy π_t^i can be equivalently represented as an optimization problem, given the knowledge of current state $\bar{x}_t$ and prior players' actions $\bar{u}_t^{1:i-1}$,*

$$\pi_t^i(\bar{x}_t, \bar{u}_t^{1:i-1}) = \tilde{u}_t^i \in \arg\min_{\substack{u_t^i \\ u_t^{i:N} \\ u_{t+1:T}^{1:N} \\ x_{t+1:T+1}}} \ell_t^i(\bar{x}_t, \bar{u}_t^{1:i-1}, u_t^{i:N}) + \sum_{\tau=t+1}^{T} \ell_\tau^i(x_\tau, u_\tau) + \ell_{T+1}^i(x_{T+1}) \tag{5.15a}$$

$$\text{s.t. } 0 = u_t^j - \pi_t^j(\bar{x}_t, \bar{u}_t^{1:i-1}, u_t^{i:j-1}), \qquad j \in \mathbf{I}_{i+1}^N \tag{5.15b}$$

$$0 = x_{t+1} - f_t(\bar{x}_t, \bar{u}_t^{1:i-1}, u_t^{i:N}), \tag{5.15c}$$

$$0 = h_t^i(\bar{x}_t, \bar{u}_t^{1:i-1}, u_t^{i:N}),\ 0 \le g_t^i(\bar{x}_t, \bar{u}_t^{1:i-1}, u_t^{i:N}) \tag{5.15d}$$

$$0 = u_\tau^j - \pi_\tau^j(x_\tau, u_\tau^{1:j-1}), \qquad \tau \in \mathbf{I}_{t+1}^T, j \in \mathbf{I}_1^N \setminus \{i\} \tag{5.15e}$$

$$0 = x_{\tau+1} - f_\tau(x_\tau, u_\tau), \qquad \tau \in \mathbf{I}_{t+1}^T \tag{5.15f}$$

$$0 = h_\tau^i(x_\tau, u_\tau),\ 0 \le g_\tau^i(x_\tau, u_\tau), \qquad \tau \in \mathbf{I}_{t+1}^T \tag{5.15g}$$

$$0 = h_{T+1}^i(x_{T+1}),\ 0 \le g_{T+1}^i(x_{T+1}) \tag{5.15h}$$

where we drop (5.15b) *when $i = N$, and we drop* (5.15e)*,* (5.15f) *and* (5.15g) *when $t = T$. The notation $\arg_u \min_{u,v}$ represents that we minimize over (u, v) but only return u as an output.*

Proof. The proof can be found in the Appendix. □

In what follows, we will characterize the KKT conditions of the constrained optimization problems in (5.15). Before doing that, we first introduce Lagrange multipliers, which facilitate the formulation of Lagrangian functions for all players.

Let $t \in \mathbf{I}_0^T$ and $i \in \mathbf{I}_1^N$. We denote by $\lambda_t^i \in \mathbb{R}^n$ the Lagrange multiplier for the dynamics constraint $0 = x_{t+1} - f_t(x_t, u_t)$. Let $\mathbb{R}_{\ge 0}$ be the set of non-negative real numbers. We define $\mu_t^i \in \mathbb{R}^{n_{h,t}^i}$ and $\gamma_t^i \in \mathbb{R}_{\ge 0}^{n_{g,t}^i}$ to be the Lagrange multipliers for the constraints $0 = h_t^i(x_t, u_t)$ and $0 \le g_t^i(x_t, u_t)$, respectively. When $t \le T$, the constrained problem (5.15) of player $i < N$ considers the feedback interaction constraint $0 = u_t^j - \pi_t^j(x_t, u_t^{1:j-1})$, $j \in \mathbf{I}_{i+1}^N$. Thus, we associate those constraints with multipliers $\psi_t^i := [\psi_t^{i,i+1}, \psi_t^{i,i+2}, \ldots, \psi_t^{i,N}]$, where $\psi_t^{i,j} \in \mathbb{R}^{m_i}$. Moreover, when $t < T$, the constrained problem (5.15) of a player $i \le N$ includes the feedback interaction constraints $0 = u_{\tau+1}^j - \pi_{\tau+1}^j(x_{\tau+1}, u_{\tau+1}^{1:j-1})$, for $\tau \ge t$ and $j \in \mathbf{I}_1^N \setminus \{i\}$. Thus, we associate those constraints with multipliers $\eta_t^i := [\eta_t^{i,1}, \ldots, \eta_t^{i,i-1}, \eta_t^{i,i+1}, \ldots, \eta_t^{i,N}]$, where $\eta_t^{i,j} \in \mathbb{R}^{m_j}$. Finally, we simplify the notation by defining $\lambda_t := [\lambda_t^1, \lambda_t^2, \ldots, \lambda_t^N]$, and define μ_t, γ_t, η_t, and ψ_t accordingly.

Subsequently, we define the Lagrangian functions of all the players. We first consider player $i \in \mathbf{I}_1^N$,

$$\begin{aligned} L_t^i(x_{t:t+1}, u_{t:t+1}, \lambda_t, \mu_t, \gamma_t, \eta_t, \psi_t) := & \ell_t^i(x_t, u_t) - \lambda_t^{i\top}(x_{t+1} - f_t(x_t, u_t)) \\ & - \mu_t^{i\top} h_t^i(x_t, u_t) - \gamma_t^{i\top} g_t^i(x_t, u_t) \\ & - \sum_{j \in \mathbf{I}_{i+1}^N} \psi_t^{i,j\top}(u_t^j - \pi_t^j(x_t, u_t^{1:j-1})) \\ & - \sum_{j \in \mathbf{I}_1^N \setminus \{i\}} \eta_t^{i,j\top}(u_{t+1}^j - \pi_{t+1}^j(x_{t+1}, u_{t+1}^{1:j-1})) \end{aligned} \tag{5.16}$$

where the right hand side terms represent player i's cost, dynamics constraint, equality and inequality constraints, and constraints encoding the feedback interaction among players at the current and future time steps.

Furthermore, at the terminal time $t = T$, for player $i \in \mathbf{I}_1^N$, we consider

$$\begin{aligned} L_T^i(x_{T:T+1}, u_T, \lambda_T, \mu_{T:T+1}, \gamma_{T:T+1}, \psi_T) := & \ell_T^i(x_T, u_T) + \ell_{T+1}^i(x_{T+1}) \\ & - \lambda_T^{i\top}(x_{T+1} - f_T(x_T, u_T)) \\ & - \mu_T^{i\top} h_T^i(x_T, u_T) - \mu_{T+1}^{i\top} h_{T+1}^i(x_{T+1}) \\ & - \gamma_T^{i\top} g_T^i(x_T, u_T) - \gamma_{T+1}^{i\top} g_{T+1}^i(x_{T+1}) \\ & - \sum_{j \in \mathbf{I}_{i+1}^N} \psi_T^{i,j\top}(u_T^j - \pi_T^j(x_T, u_T^{1:j-1})) \end{aligned} \tag{5.17}$$

where the right hand side terms represent player i's costs, dynamics constraint, equality and inequality constraints, and constraints encoding the feedback interaction among players at the terminal time T. Note that there is no more decision to be made at time $t = T + 1$, and therefore, there is no term representing the feedback interactions among players for future time steps in (5.17), which is different from (5.16).

For all time steps $t \in \mathbf{I}_0^T$ and players $i \in \mathbf{I}_1^N$, assuming the state x_t is given and each player $j < i$ has taken action u_t^j, we formulate the Lagrangian of the problem (5.15) of player i at the t-th stage as

$$\begin{aligned} \mathcal{L}_t^i(x_{t:T+1}, u_{t:T}, \lambda_{t:T}, \mu_{t:T+1}, \gamma_{t:T+1}, \eta_{t:T-1}, \psi_{t:T}) := & \sum_{\tau=t}^{T-1} L_\tau^i(x_{\tau:\tau+1}, u_{\tau:\tau+1}, \\ & \lambda_\tau, \mu_\tau, \gamma_\tau, \eta_\tau, \psi_\tau) + L_T^i(x_{T:T+1}, u_T, \lambda_T, \mu_{T:T+1}, \gamma_{T:T+1}, \psi_T) \end{aligned} \tag{5.18}$$

where for each $\tau \in \mathbf{I}_{t+1}^T$, the terms associated with ψ_τ in L_τ^i ensure constraints already addressed by the terms associated with $\eta_{\tau-1}$ in $L_{\tau-1}^i$ and can therefore be dropped when defining $\mathcal{L}_t^i$. We can concatenate the KKT conditions of each player at each stage, and summarize the overall KKT conditions for (5.15) in the following theorem.

Theorem 5 (Necessary Condition)**.** *Under Assumption 2, let* $(\mathbf{x}^*, \mathbf{u}^*)$ *be a local feedback Stackelberg equilibrium trajectory. Suppose that the Linear Independence Constraint Qualification (LICQ) [238] and strict complementarity condition [35] are satisfied at* $(\mathbf{x}^*, \mathbf{u}^*)$*. Furthermore, suppose* $\{\pi_t^i\}_{t=0,i=1}^{T,N}$ *is a set of local feedback Stackelberg policies and* π_t^i *is differentiable around* $(x_t^*, u_t^{1:(i-1)*})$*,* $\forall t \in \mathbf{I}_0^T$*,* $i \in \mathbf{I}_1^N$*. The KKT conditions of* (5.15) *can be formulated as, for all* $i \in \mathbf{I}_1^N$*,* $t \in \mathbf{I}_0^T$*,*

$$
\begin{aligned}
0 &= \nabla_{u_\tau^i} \mathcal{L}_t^i(x_{t:T+1}^*, u_{t:T}^*, \lambda_{t:T}, \mu_{t:T+1}, \gamma_{t:T+1}, \eta_{t:T-1}, \psi_{t:T}) && \forall \tau \in \mathbf{I}_t^T \\
0 &= \nabla_{x_\tau} \mathcal{L}_t^i(x_{t:T+1}^* u_{t:T}^*, \lambda_{t:T}, \mu_{t:T+1}, \gamma_{t:T+1}, \eta_{t:T-1}, \psi_{t:T}) && \forall \tau \in \mathbf{I}_{t+1}^{T+1} \\
0 &= \nabla_{u_t^j} \mathcal{L}_t^i(x_{t:T+1}^*, u_{t:T}^*, \lambda_{t:T}, \mu_{t:T+1}, \gamma_{t:T+1}, \eta_{t:T-1}, \psi_{t:T}) && \forall j \in \mathbf{I}_{i+1}^N \\
0 &= \nabla_{u_\tau^j} \mathcal{L}_t^i(x_{t:T+1}^*, u_{t:T}^*, \lambda_{t:T}, \mu_{t:T+1}, \gamma_{t:T+1}, \eta_{t:T-1}, \psi_{t:T}) && \forall j \in \mathbf{I}_1^N \setminus \{i\}, \forall \tau \in \mathbf{I}_{t+1}^T \\
0 &= x_{\tau+1}^* - f_\tau(x_\tau^*, u_\tau^*) && \forall \tau \in \mathbf{I}_t^T \\
0 &= h_\tau^i(x_\tau^*, u_\tau^*) && \forall \tau \in \mathbf{I}_t^T \\
0 &\leq \gamma_\tau^i \perp g_\tau^i(x_\tau^*, u_\tau^*) \geq 0 && \forall \tau \in \mathbf{I}_t^T \\
0 &= h_{T+1}^i(x_{T+1}^*) \\
0 &\leq \gamma_{T+1}^i \perp g_{T+1}^i(x_{T+1}^*) \geq 0
\end{aligned}
\tag{5.19}
$$

where $\perp$ *represents the complementary slackness condition [35]. Then, there exists Lagrange multipliers* $\boldsymbol{\lambda} := [\lambda_t]_{t=0}^T$*,* $\boldsymbol{\mu} := [\mu_t]_{t=0}^{T+1}$*,* $\boldsymbol{\gamma} := [\gamma_t]_{t=0}^{T+1}$*,* $\boldsymbol{\eta} := [\eta_t]_{t=0}^{T-1}$*, and* $\boldsymbol{\psi} := [\psi_t]_{t=0}^T$*, such that* (5.19) *holds true.*

Proof. The proof can be found in the Appendix. □

Constructing the KKT conditions in (5.19) requires the computation of policy gradients, $\{\nabla \pi_t^i\}_{t=0,i=1}^{T,N}$, which appear in the first four rows of (5.19). However, knowing the policy itself is not required, as any solution satisfying the KKT conditions obeys the corresponding feedback Stackelberg policy, as shown in the proof of Theorem 5. A key distinction between (5.19) and the FNE KKT conditions in [167] lies in the accommodation of a decision hierarchy among the N players at each stage. This is reflected in the terms $-\sum_{j \in \mathbf{I}_{i+1}^N} \psi_t^{i,j\top}(u_t^j - \pi_t^j(x_t, u_t^{1:j-1}))$ in the Lagrangian $\mathcal{L}_t^i$. Additionally, this decision hierarchy differentiates the construction of the FSE KKT conditions from those of FNE. We will outline a detailed procedure for constructing the FSE KKT conditions in Sections 5.5 and 5.6, with an example provided in Appendix 5.10.

Furthermore, we propose a sufficient condition for feedback Stackelberg equilibrium trajectories in the following theorem.

Theorem 6 (Sufficient Condition)**.** *Let* $(\mathbf{x}^*, \mathbf{u}^*)$ *be a trajectory and* $\{\pi_t^i\}_{t=0,i=1}^{T,N}$ *be the associated policies. Suppose there exist Lagrange multipliers* $\{\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\psi}\}$ *satisfying* (5.19) *and there exists*

an $\epsilon > 0$ such that, for all $i \in \mathbf{I}_1^N$, $t \in \mathbf{I}_0^T$, and nonzero $\{\Delta x_{T+1}\} \bigcup \{\Delta x_\tau, \Delta u_\tau\}_{\tau=t}^T$ satisfying

$$\begin{aligned}
0 &= \Delta u_t^j - \nabla \pi_t^j(x_t^*, u_t^{1:(j-1)*}) \begin{bmatrix} \Delta x_t \\ \Delta u_t^{1:j-1} \end{bmatrix}, \forall j \in \mathbf{I}_i^N \\
0 &= \Delta u_\tau^j - \nabla \pi_\tau^j(x_\tau^*, u_\tau^{1:(j-1)*}) \begin{bmatrix} \Delta x_\tau \\ \Delta u_\tau^{1:j-1} \end{bmatrix}, \forall j \in \mathbf{I}_1^N, \forall \tau \in \mathbf{I}_{t+1}^T \\
0 &= \Delta x_{\tau+1} - \nabla f_\tau(x_\tau^*, u_\tau^*) \begin{bmatrix} \Delta x_\tau \\ \Delta u_\tau \end{bmatrix}, \forall \tau \in \mathbf{I}_t^T \\
0 &= \nabla h_\tau^j(x_\tau^*, u_\tau^*) \begin{bmatrix} \Delta x_\tau \\ \Delta u_\tau \end{bmatrix}, 0 = \nabla h_{T+1}^j(x_{T+1}^*) \Delta x_{T+1}, \forall \tau \in \mathbf{I}_0^T, \forall j \in \mathbf{I}_1^N
\end{aligned} \tag{5.20}$$

we have

$$\sum_{\tau=t}^{T} \begin{bmatrix} \Delta x_\tau \\ \Delta u_\tau^i \end{bmatrix}^\top \nabla^2_{[x_\tau^*, u_\tau^{i*}]} L_\tau^i \begin{bmatrix} \Delta x_\tau \\ \Delta u_\tau^i \end{bmatrix} + \Delta x_{T+1}^\top \nabla^2_{[x_{T+1}^*]} L_T^i \Delta x_{T+1} > 0. \tag{5.21}$$

Then, $(\mathbf{x}^*, \mathbf{u}^*)$ constitutes a local feedback Stackelberg equilibrium trajectory.

Proof. The proof can be found in the Appendix. □

Remark 11. *The gap between the necessity condition in Theorem 5 and the sufficiency condition in Theorem 6 is due to the fact that a solution to* (5.19) *may not necessarily be a feedback Stackelberg equilibrium, and that there exist feedback Stackelberg equilibria where the cost functions possess zero second-order gradients.*

Theorems 5 and 6 establish conditions to certify whether a trajectory $(\mathbf{x}, \mathbf{u})$ constitutes a feedback Stackelberg equilibrium with a set of feedback Stackelberg policies $\{\pi_t^i\}_{t=0,i=1}^{T,N}$. However, computing feedback Stackelberg equilibria can be challenging. In the following sections, we will discuss how to approximately compute local feedback Stackelberg equilibria. We will first compute feedback Stackelberg equilibria for Linear Quadratic games and then extend the result to nonlinear games.

5.5 Constrained Linear Quadratic Games

We consider the linear dynamics

$$x_{t+1} = f_t(x_t, u_t) = A_t x_t + B_t^1 u_t^1 + \cdots + B_t^N u_t^N + c_t,\ t \in \mathbf{I}_0^T, \tag{5.22}$$

where $A_t \in \mathbb{R}^{n \times n}$, $B_t^i \in \mathbb{R}^{n \times m_i}$ and $c_t \in \mathbb{R}^n$. We denote by $B_t := [B_t^1, B_t^2, \ldots, B_t^N]$. The cost of the i-th player is defined as

$$\begin{aligned}
\ell_t^i(x_t, u_t) =& \frac{1}{2} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top \begin{bmatrix} Q_t^i & S_t^{i\top} \\ S_t^i & R_t^i \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix} + q_t^{i\top} x_t + r_t^{i\top} u_t,\ t \in \mathbf{I}_0^T, \\
\ell_{T+1}^i(x_{T+1}) =& \frac{1}{2} x_{T+1}^\top Q_{T+1}^i x_{T+1} + q_{T+1}^{i\top} x_{T+1},
\end{aligned} \tag{5.23}$$

where symmetric matrices $Q_t^i \in \mathbb{R}^{n\times n}$ and $R_t^i \in \mathbb{R}^{m\times m}$ are positive semidefinite and positive definite, respectively. The off-diagonal matrix is denoted as $S_t^i \in \mathbb{R}^{m\times n}$. In particular, we partition the structure of R_t^i, S_t^i and r_t^i as follows

$$R_t^i = \begin{bmatrix} R_t^{i,1,1} & R_t^{i,1,2} & \cdots & R_t^{i,1,N} \\ R_t^{i,2,1} & R_t^{i,2,2} & \cdots & R_t^{i,2,N} \\ \vdots & \vdots & \ddots & \vdots \\ R_t^{i,N,1} & R_t^{i,N,2} & \cdots & R_t^{i,N,N} \end{bmatrix}, S_t^i = \begin{bmatrix} S_t^{i,1} \\ S_t^{i,2} \\ \vdots \\ S_t^{i,N} \end{bmatrix}, r_t^i = \begin{bmatrix} r_t^{i,1} \\ r_t^{i,2} \\ \vdots \\ r_t^{i,N} \end{bmatrix}, \tag{5.24}$$

where $R_t^{i,j,k}$, $S_t^{i,j}$ and r_t^i represent the cost terms $u_t^{j\top} R_t^{i,j,k} u_t^k$, $u_t^{j\top} S_t^{i,j} x_t$ and $r_t^{i,j^\top} u_t^j$ in $\ell_t^i(x_t, u_t)$. The linear equality and inequality constraints are specified as,

$$\begin{aligned} 0 &= h_t^i(x_t, u_t) = H_{x_t}^i x_t + \sum_{j\in\mathbf{I}_1^N} H_{u_t^j}^i u_t^j + \bar{h}_t^i, \quad t \in \mathbf{I}_0^T \\ 0 &\le g_t^i(x_t, u_t) = G_{x_t}^i x_t + \sum_{j\in\mathbf{I}_1^N} G_{u_t^j}^i u_t^j + \bar{g}_t^i, \quad t \in \mathbf{I}_0^T \\ 0 &= h_{T+1}^i(x_{T+1}) = H_{x_{T+1}}^i x_{T+1} + \bar{h}_{T+1}^i, \\ 0 &\le g_{T+1}^i(x_{T+1}) = G_{x_{T+1}}^i x_{T+1} + \bar{g}_{T+1}^i. \end{aligned} \tag{5.25}$$

Computing Feedback Stackelberg Equilibria and Constructing the KKT Conditions for LQ Games

In this subsection, we introduce a process for deriving FSE and the KKT conditions for LQ games. When we have linear inequality constraints, the optimal policies of LQ games are generally piecewise linear functions of the state [32, 167]. However, this makes them non-differentiable at the facets. In our work, we propose to use the primal-dual interior point (PDIP) method [238] to solve constrained LQ games. The benefits of using PDIP are its polynomial complexity and tolerance of infeasible initializations. Critically, under certain conditions, PDIP yields a local differentiable policy approximation to the ground truth piecewise linear policy, as shown in the rest of this section and an example in Appendix 5.9.

To this end, we introduce a set of non-negative slack variables $\{s_t^i\}_{t=0,i=1}^{T+1,N}$ such that we can rewrite the inequality constraints as equality constraints for $t \in \mathbf{I}_0^{T+1}$ and $i \in \mathbf{I}_1^N$,

$$g_t^i(x_t, u_t) - s_t^i = 0,\ g_{T+1}^i(x_{T+1}) - s_{T+1}^i = 0. \tag{5.26}$$

In this chapter, we consider PDIP as a homotopy method as in [238]. Instead of solving the mixed complementarity problem (5.19) directly, we seek solutions to the homotopy approximation of the complementary slackness condition

$$\gamma_t^i \odot s_t^i = \rho\mathbf{1},\ s_t^i \ge 0,\ \gamma_t^i \ge 0 \tag{5.27}$$

where $\odot$ denotes the elementwise product and $\rho > 0$ is a hyper-parameter to be reduced to 0 gradually such that we recover the ground truth solution when $\rho \to 0$. In the following section, we will construct the KKT conditions where we replace the mixed complementarity condition with its approximation (5.27). For each $\rho > 0$, we denote its corresponding local feedback policy as $\{\pi^i_{t,\rho}\}^{T,N}_{t=0,i=1}$, if it exists.

As shown in Theorem 5, the construction of the KKT conditions for player i at stage t requires the policy gradients of subsequent players at the current stage and future stages. In what follows, we construct those KKT conditions in reverse player order and backward in time.

Player N at the T-th stage

Before constructing the KKT conditions, we first introduce the variables of player N at the terminal time T, $\mathbf{z}^N_T := [u^N_T, \lambda^N_T, \mu^N_{T:T+1}, \gamma^N_{T:T+1}, s^N_{T:T+1}, x_{T+1}]$. As shown in Theorem 5, the KKT conditions of player N at time T can be written as

$$0 = K^N_{T,\rho}(\mathbf{z}^N_T) := \begin{bmatrix} \nabla_{u^N_T} L^N_T \\ \nabla_{x_{T+1}} L^N_T \\ x_{T+1} - f_T(x_T, u_T) \\ h^N_T(x_T, u_T) \\ h^N_{T+1}(x_{T+1}) \\ g^N_T(x_T, u_T) - s^N_T \\ g^N_{T+1}(x_{T+1}) - s^N_{T+1} \\ \gamma^N_{T:T+1} \odot s^N_{T:T+1} - \rho \mathbf{1} \end{bmatrix}, \tag{5.28}$$

where the rows of $K^N_{T,\rho}(\mathbf{z}^N_T)$ represent the stationarity conditions with respect to u^N_T and x_{T+1}, dynamics constraint, equality constraints, inequality constraints, and relaxed complementarity conditions. To obtain a local policy and its policy gradient around a $\mathbf{z}^N_T$ satisfying (5.28), we build a first-order approximation to (5.28),

$$\nabla K^N_{T,\rho} \cdot \Delta \mathbf{z}^N_T + \nabla_{[x_T, u^{1:N-1}_T]} K^N_{T,\rho} \cdot \begin{bmatrix} \Delta x_T \\ \Delta u^{1:N-1}_T \end{bmatrix} + K^N_{T,\rho}(\mathbf{z}^N_T) = 0. \tag{5.29}$$

If there is no solution $\Delta \mathbf{z}^N_T$ to (5.29), then we claim there is no feedback Stackelberg policy. Suppose (5.29) has a solution $\Delta \mathbf{z}^N_T$, then we can define $\Delta \mathbf{z}^N_T$ as

$$\Delta \mathbf{z}^N_T = \underbrace{-\left(\nabla K^N_{T,\rho}\right)^+ \cdot \left(\nabla_{[x_T, u^{1:N-1}_T]} K^N_{T,\rho} \cdot \begin{bmatrix} \Delta x_T \\ \Delta u^{1:N-1}_T \end{bmatrix} + K^N_{T,\rho}(\mathbf{z}^N_T)\right)}_{F^N_T(\Delta x_T, \Delta u^{1:N-1}_T)}, \tag{5.30}$$

where $(\cdot)^+$ represents the pseudo-inverse and we denote $\Delta \mathbf{z}^N_T$ as a function F^N_T of $(\Delta x_T, \Delta u^{1:N-1}_T)$. Since Δu^N_T represents the first m_N entries of $\Delta \mathbf{z}^N_T$, we consider Δu^N_T as a function of $(\Delta x_T, \Delta u^{1:N-1}_T)$,

$$\Delta u^N_T = -\left[\left(\nabla K^N_{T,\rho}\right)^+\right]_{u^N_T} \cdot \left(\nabla_{[x_T, u^{1:N-1}_T]} K^N_{T,\rho} \cdot \begin{bmatrix} \Delta x_T \\ \Delta u^{1:N-1}_T \end{bmatrix} + K^N_{T,\rho}(\mathbf{z}^N_T)\right), \tag{5.31}$$

where $\left[\left(\nabla K^N_{T,\rho}\right)^+\right]_{u^N_T}$ represents the rows of the matrix $\left(\nabla K^N_{T,\rho}\right)^+$ corresponding to the variable u^N_T, i.e., the first m_N rows of the matrix $\left(\nabla K^N_{T,\rho}\right)^+$.

Furthermore, for some $x \in \mathbb{R}^n$ and $u^{1:N-1} \in \mathbb{R}^{\sum_{i=1}^{N-1} m_i}$, let $\Delta x_T = x - x_T$, $\Delta u_T^{1:N-1} = u^{1:N-1} - u_T^{1:N-1}$ and $\Delta u_T^N = u^N - u_T^N$. Substituting them into (5.31), we obtain a local policy $\tilde{\pi}^N_{T,\rho}$ for player N at time T,

$$\begin{aligned} u^N =& \tilde{\pi}^N_{T,\rho}(x, u^{1:N-1}) \\ :=& u_T^N - \left[\left(\nabla K^N_{T,\rho}\right)^+\right]_{u^N_T} \cdot \left(\nabla_{[x_T, u_T^{1:N-1}]} K^N_{T,\rho} \cdot \begin{bmatrix} x - x_T \\ u^{1:N-1} - u_T^{1:N-1} \end{bmatrix} + K^N_{T,\rho}(\mathbf{z}^N_T)\right). \end{aligned} \tag{5.32}$$

Suppose that $\nabla K^N_{T,\rho}(\mathbf{z}^N_T)$ has a constant row rank in an open set containing $\mathbf{z}^N_T$, then, by the constant rank theorem [145], the policy $\tilde{\pi}^N_{T,\rho}$ of player N at time T is locally differentiable with respect to $(x, u^{1:N-1})$, and its gradient over $(x, u^{1:N-1})$ is

$$\nabla \tilde{\pi}^N_{T,\rho} = -\left[\left(\nabla K^N_{T,\rho}\right)^+\right]_{u^N_T} \cdot \nabla_{[x_T, u_T^{1:N-1}]} K^N_{T,\rho}. \tag{5.33}$$

In the following subsection, we construct the KKT conditions of a player $i < N$ at stage T.

Players $i < N$ at the T-th stage

For player $i < N$, assuming that $\mathbf{z}_T^{i+1}$ has been defined and $\nabla \tilde{\pi}^{i+1}_{T,\rho}$ has been computed, we first introduce variables

$$\mathbf{y}^i_T := [u^i_T, \psi^i_T, \lambda^i_T, \mu^i_{T:T+1}, \gamma^i_{T:T+1}, s^i_{T:T+1}] \text{ and } \mathbf{z}^i_T := [\mathbf{y}^i_T, \mathbf{z}^{i+1}_T]. \tag{5.34}$$

The KKT conditions of player i at time T is

$$0 = K^i_{T,\rho}(\mathbf{z}^i_T) := \begin{bmatrix} \hat{K}^i_{T,\rho}(\mathbf{y}^i_T) \\ K^{i+1}_{T,\rho}(\mathbf{z}^{i+1}_T) \end{bmatrix}, \ \hat{K}^i_{T,\rho}(\mathbf{y}^i_T) := \begin{bmatrix} \nabla_{u^i_T} L^i_T \\ \nabla_{x_{T+1}} L^i_T \\ \nabla_{u^j_T} L^i_T, \forall j \in \mathbf{I}^N_{i+1} \\ h^i_T(x_T, u_T) \\ h^i_{T+1}(x_{T+1}) \\ g^i_T(x_T, u_T) - s^i_T \\ g^i_{T+1}(x_{T+1}) - s^i_{T+1} \\ \gamma^i_{T:T+1} \odot s^i_{T:T+1} - \rho \mathbf{1} \end{bmatrix}, \tag{5.35}$$

where the definition of L^i_T involves the policy $\pi^{i+1}_{T,\rho}$, as shown in (5.17). Building a first-order approximation to $0 = K^i_{T,\rho}(\mathbf{z}^i_T)$, we have

$$\nabla K^i_{T,\rho} \cdot \Delta \mathbf{z}^i_T + \nabla_{[x_T, u_T^{1:i-1}]} K^i_{T,\rho} \cdot \begin{bmatrix} \Delta x_T \\ \Delta u_T^{1:i-1} \end{bmatrix} + K^i_{T,\rho}(\mathbf{z}^i_T) = 0. \tag{5.36}$$

However, a drawback of PDIP is that the policy $\pi^{i+1}_{T,\rho}$ is nonlinear in state x_T and prior players' controls $u_T^{1:i}$, as shown in a simplified problem in Appendix 5.9. The computation of $\nabla K^i_{T,\rho}$

involves the evaluation of $\nabla(\nabla_{u_T^i} L_T^i)$, which requires the computation of $\nabla(\psi_T^{i,i+1}\nabla\pi_{T,\rho}^{i+1}) = \nabla\psi_T^{i,i+1}\cdot\nabla\pi_{T,\rho}^{i+1} + \psi_T^{i,i+1}\cdot\nabla^2\pi_{T,\rho}^{i+1}$. Furthermore, to evaluate $\nabla^2\pi_{T,\rho}^{i+1}$, we need the computation of $\nabla^3\pi_{T,\rho}^{i+2}$. In other words, the construction of $\nabla K_{T,\rho}^i$ needs the evaluation of $\nabla^2\pi_{T,\rho}^{i+1}$, $\nabla^3\pi_{T,\rho}^{i+2}$, ..., and $\nabla^{N-i+1}\pi_{T,\rho}^{N}$. The evaluation of high-order policy gradients is challenging in practice [167] because there is no closed-form solution to the KKT equation $0 = K_{T,\rho}^{i+1}(\mathbf{z}_T^{i+1})$.

We prove in Appendix 5.9 that the high-order policy gradients could decay to zero as $\rho \to 0$, when the ground truth policy is piecewise linear and differentiable around x_T. Motivated by this observation, we propose to approximate the nonlinear policy $\pi_{T,\rho}^{i+1}$ by its first-order approximation $\tilde{\pi}_{T,\rho}^{i+1}$ in (5.32). With this approximation, we have $\nabla(\psi_T^{i,i+1}\nabla\tilde{\pi}_{T,\rho}^{i+1}) = \nabla\psi_T^{i,i+1}\cdot\nabla\tilde{\pi}_{T,\rho}^{i+1}$. We refer to this policy $\tilde{\pi}_{T,\rho}^{i+1}$ as a ***quasi-policy***.

In the remainder of this section, we will always approximate the ground truth nonlinear policy by quasi-policy when we define the KKT conditions.

Solving equation (5.36), we can obtain $\Delta\mathbf{z}_T^i$ and $\nabla\tilde{\pi}_{T,\rho}^i$ as in (5.30) and (5.33), respectively. However, by construction, the dimension of $\Delta\mathbf{z}_T^i$ is higher than $\Delta\mathbf{z}_T^{i+1}$. Therefore, it is more expensive to compute $(\nabla K_{T,\rho}^i)^+$ than $(\nabla K_{T,\rho}^{i+1})^+$, and it is worthwhile to reduce the complexity of computing $\Delta\mathbf{z}_T^i$ by leveraging the computation that we have done for $\Delta\mathbf{z}_T^{i+1}$ and $\nabla\tilde{\pi}_{T,\rho}^{i+1}$. To this end, by exploiting the structure $\mathbf{z}_T^i = [\mathbf{y}_T^i, \mathbf{z}_T^{i+1}]$ in (5.34), we can rewrite (5.36) as,

$$\begin{cases} \nabla\hat{K}_{T,\rho}^i\cdot\Delta\mathbf{y}_T^i + \nabla_{[x_T,u_T^{1:i-1}]}\hat{K}_{T,\rho}^i\cdot\begin{bmatrix}\Delta x_T\\ \Delta u_T^{1:i-1}\end{bmatrix} + \hat{K}_{T,\rho}^i(\mathbf{y}_T^i) = 0 \\ \nabla K_{T,\rho}^{i+1}\cdot\Delta\mathbf{z}_T^{i+1} + \nabla_{[x_T,u_T^{1:i}]}K_{T,\rho}^{i+1}\cdot\begin{bmatrix}\Delta x_T\\ \Delta u_T^{1:i}\end{bmatrix} + K_{T,\rho}^{i+1}(\mathbf{z}_T^{i+1}) = 0 \end{cases} \tag{5.37}$$

Observe that we have solved the second equation of (5.37) in Section 5.5. What remains to be solved is the first equation in (5.37). We solve it as follows,

$$\begin{aligned} \Delta\mathbf{y}_T^i =& \underbrace{-\big(\nabla\hat{K}_{T,\rho}^i\big)^+\cdot\Big(\nabla_{[x_T,u_T^{1:i-1}]}\hat{K}_{T,\rho}^i\cdot\begin{bmatrix}\Delta x_T\\ \Delta u_T^{1:i-1}\end{bmatrix} + \hat{K}_{T,\rho}^i(\mathbf{y}_T^i)\Big)}_{\hat{F}_T^i(\Delta x_T,\Delta u_T^{1:i-1})}, \\ \Delta u_T^i =& -\big[\big(\nabla\hat{K}_{T,\rho}^i\big)^+\big]_{u_T^i}\cdot\Big(\nabla_{[x_T,u_T^{1:i-1}]}\hat{K}_{T,\rho}^i\cdot\begin{bmatrix}\Delta x_T\\ \Delta u_T^{1:i-1}\end{bmatrix} + \hat{K}_{T,\rho}^i(\mathbf{y}_T^i)\Big), \\ \nabla\tilde{\pi}_{T,\rho}^i =& -\big[\big(\nabla\hat{K}_{T,\rho}^i\big)^+\big]_{u_T^i}\cdot\nabla_{[x_T,u_T^{1:i-1}]}\hat{K}_{T,\rho}^i. \end{aligned} \tag{5.38}$$

Combining (5.38) and (5.30), we have

$$\Delta\mathbf{z}_T^i = \begin{bmatrix}\Delta\mathbf{y}_T^i\\ \Delta\mathbf{z}_T^{i+1}\end{bmatrix} = \begin{bmatrix}\hat{F}_T^i(\Delta x_T,\Delta u_T^{1:i-1})\\ F_T^{i+1}(\Delta x_T,\Delta u_T^{1:i})\end{bmatrix}. \tag{5.39}$$

Since Δu_T^i is also a function of $(\Delta x_T, \Delta u_T^{1:i-1})$, as shown in (5.38), we can represent (5.39) compactly as $\Delta\mathbf{z}_T^i = F_T^i(\Delta x_T, \Delta u_T^{1:i-1})$.

As such, given that the KKT conditions of player $(i+1)$ at time T have been constructed, we have finished the construction of the KKT conditions for player i at time T, and we introduced a computationally efficient way to compute $\nabla\tilde{\pi}^i_{T,\rho}$. We can derive the KKT conditions and quasi-policy gradient of player $i < N$ at time T, sequentially, from $i = N-1$ to $i = 1$.

Player N at a stage $t < T$

At a stage $t < T$, assuming that we have constructed the KKT conditions $0 = K^1_{t+1,\rho}(\mathbf{z}^1_{t+1})$, we are ready to derive the KKT conditions for player N at time t. We first introduce the variable $\mathbf{z}^N_t := [\mathbf{y}^N_t, \mathbf{z}^1_{t+1}]$, with $\mathbf{y}^N_t := [u^N_t, \eta^N_t, \lambda^N_t, \mu^N_t, \gamma^N_t, s^N_t, x_{t+1}]$. We construct the KKT conditions of player N at time t as follows,

$$0 = K^N_{t,\rho}(\mathbf{z}^N_t) := \begin{bmatrix} \nabla_{u^N_t} L^N_t \\ \nabla_{x_{t+1}} L^N_t \\ \nabla_{u^j_{t+1}} L^N_t, \forall j \in \mathbf{I}_1^{N-1} \\ x_{t+1} - f_t(x_t, u_t) \\ h^N_t(x_t, u_t) \\ g^N_t(x_t, u_t) - s^N_t \\ \gamma^N_t \odot s^N_t - \rho\mathbf{1} \\ K^1_{t+1,\rho}(\mathbf{z}^1_{t+1}) \end{bmatrix}. \tag{5.40}$$

Building a first-order approximation to the above equation, we can obtain quasi-policy gradient $\nabla\tilde{\pi}^N_{t,\rho}$ as in (5.38) when it exists.

Players $i < N$ at a stage $t < T$

Suppose that we have constructed the KKT conditions for the $(i+1)$-th player at the t-th stage, we are then ready to construct the KKT conditions for player i at the t-th stage. We introduce the variable $\mathbf{z}^i_t := [\mathbf{y}^i_t, \mathbf{z}^{i+1}_t]$ with $\mathbf{y}^i_t := [u^i_t, \psi^i_t, \eta^i_t, \lambda^i_t, \mu^i_t, \gamma^i_t, s^i_t]$. The KKT conditions of player i at time t is

$$0 = K^i_{t,\rho}(\mathbf{z}^i_t) := \begin{bmatrix} \nabla_{u^i_t} L^i_t \\ \nabla_{x_{t+1}} L^i_t \\ \nabla_{u^j_t} L^i_t, \forall j \in \mathbf{I}^N_{i+1} \\ \nabla_{u^j_{t+1}} L^i_t, \forall j \in \mathbf{I}^N_1 \setminus \{i\} \\ h^i_t(x_t, u_t) \\ g^i_t(x_t, u_t) - s^i_t \\ \gamma^i_t \odot s^i_t - \rho\mathbf{1} \\ K^{i+1}_{t,\rho}(\mathbf{z}^{i+1}_t) \end{bmatrix}. \tag{5.41}$$

Building a first approximation to the above equation, we can obtain the quasi-policy gradient $\nabla\tilde{\pi}^i_{t,\rho}$ as in (5.38), when it exists.

We observe that, by construction, the KKT conditions in (5.19) is equivalent to $0 = K^1_{0,\rho}(\mathbf{z}^1_0)$. To simplify notation, we define

$$\mathbf{z} := \mathbf{z}^1_0, \;\; K_\rho(\mathbf{z}) := K^1_{0,\rho}(\mathbf{z}). \tag{5.42}$$

Algorithm 3: Local Feedback Stackelberg Equilibrium via PDIP

Require: $\{f_t\}_{t=0}^{T}$, $\{\ell_t^i, h_t^i, g_t^i\}_{t=0,i=1}^{T+1,N}$, initial homotopy parameter ρ, contraction rate $\sigma \in (0,1)$, parameters $\beta \in (0,1)$ and $\kappa \in (0,1)$, tolerance ϵ, initial solution $\mathbf{z}_\rho^{(0)} := [\mathbf{x}_\rho^{(0)}, \mathbf{u}_\rho^{(0)}, \boldsymbol{\lambda}_\rho^{(0)}, \boldsymbol{\mu}_\rho^{(0)}, \boldsymbol{\gamma}_\rho^{(0)}, \boldsymbol{\eta}_\rho^{(0)}, \boldsymbol{\psi}_\rho^{(0)}, \mathbf{s}_\rho^{(0)}]$ with $\mathbf{s}_\rho^{(0)} > 0$ and $\boldsymbol{\gamma}_\rho^{(0)} > 0$

Ensure: policies $\{\tilde{\pi}_{t,\rho}^i\}_{t=0,i=1}^{T,N}$, converged solution $\mathbf{z}_\rho$

for $k^{\text{out}} = 1, 2, \ldots, k_{\max}^{\text{out}}$ **do**
  while the merit function $\|K_\rho(\mathbf{z}_\rho^{(k)})\|_2 > \epsilon$ **do**
    construct the first-order approximation of the KKT conditions $0 = \nabla K_\rho \cdot \Delta \mathbf{z}_\rho + K_\rho(\mathbf{z}_\rho^{(k)})$
    $\Delta \mathbf{z}_\rho \leftarrow -\big(\nabla K_\rho\big)^+ \cdot K_\rho(\mathbf{z}_\rho^{(k)})$
    initialize the step size for line search, $\alpha \leftarrow 1$
    while $\|K_\rho(\mathbf{z}_\rho^{(k)} + \alpha \Delta \mathbf{z}_\rho)\|_2 > \kappa \|K_\rho(\mathbf{z}_\rho^{(k)})\|_2$ or $\hat{\mathbf{z}}_\rho := (\mathbf{z}_\rho^{(k)} + \alpha \Delta \mathbf{z}_\rho)$ has a non-positive element in its sub-vector $[\hat{\mathbf{s}}_\rho, \hat{\boldsymbol{\gamma}}_\rho]$ **do**
      $\alpha \leftarrow \beta \cdot \alpha$
    end while
    if $\alpha == 0$ **then**
      claim **failure** to find a feedback Stackelberg equilibrium
    end if
    $\mathbf{z}_\rho^{(k+1)} \leftarrow \mathbf{z}_\rho^{(k)} + \alpha \Delta \mathbf{z}_\rho$
  end while
  $\rho \leftarrow \sigma \cdot \rho$
end for
construct $\{\tilde{\pi}_{t,\rho}^i\}_{t=0,i=1}^{T,N}$ as in (5.32) and record $\mathbf{z}_\rho \leftarrow \mathbf{z}_\rho^{(k)}$.
return $\{\tilde{\pi}_{t,\rho}^i\}_{t=0,i=1}^{T,N}$, $\mathbf{z}_\rho$

The KKT conditions (5.19) can be represented compactly as $0 = K_\rho(\mathbf{z})$. To more effectively illustrate the construction process of KKT conditions described above, we have included detailed examples of the KKT conditions for two-player LQ games in Appendix 5.10 as a reference.

Primal-Dual Interior Point Algorithm and Convergence Analysis in Constrained LQ Games

In this subsection, we propose the application of Newton's method to compute $\mathbf{z}^* = [\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, \boldsymbol{\gamma}^*, \boldsymbol{\eta}^*, \boldsymbol{\psi}^*, \mathbf{s}^*]$, ensuring $0 = K_\rho(\mathbf{z}^*)$. This approach guarantees that the associated quasi-policies form a set of local FSE policies, provided that we anneal the parameter ρ to zero and the sufficient condition in Theorem 6 is satisfied. We formalize our method in Algorithm 3.

In Algorithm 3, we gradually decay the homotopy parameter ρ to zero such that $\lim_{\rho \to 0} \mathbf{z}_\rho$ recovers an FSE solution. For each ρ, at the k-th iteration, we first construct the KKT conditions $0 = K_\rho(\mathbf{z})$ along the trajectory $\mathbf{z}_\rho^{(k)}$. We compute the Newton update direction $\Delta \mathbf{z} := -(\nabla K_\rho)^+ \cdot$

$K_\rho(\mathbf{z}_\rho^{(k)})$. Since we aim at finding a solution $\mathbf{z}^*$ to $0 = K_\rho(\mathbf{z}^*)$, a natural choice of merit function is $\|K_\rho(\mathbf{z})\|_2$. Given this choice of the merit function, we perform a line search to determine a step size α and update $\mathbf{z}_\rho^{(k+1)} = \mathbf{z}_\rho^{(k)} + \alpha\Delta\mathbf{z}$ until convergence. The converged solution is denoted as $\mathbf{z}_\rho^*$. Subsequently, we steadily decay ρ and repeat these Newton update steps. We characterize how the magnitude of the KKT residual value $\|K_\rho(\mathbf{z})\|_2$ influences the convergence rate of Algorithm 3 when solving LQ games in the following result.

Theorem 7. *Under Assumption 2, let $\mathcal{F}_z := \{\mathbf{z} = [\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\psi}, \mathbf{s}] : \boldsymbol{\gamma} \geq 0, \mathbf{s} \geq 0\}$ be the solution set. We denote by $\nabla K_\rho(\mathbf{z})$ and $\nabla^* K_\rho(\mathbf{z})$ the Jacobians of the KKT conditions with and without considering quasi-policy gradients, respectively. Suppose that $\nabla K_\rho(\mathbf{z})$ is invertible and there exist constants D and C such that*

$$\|(\nabla K_\rho(\mathbf{z}))^{-1}\|_2 \leq D, \qquad \forall i \in \mathbf{I}_1^N, \forall \mathbf{z} \in \mathcal{F}_z, \tag{5.43a}$$

$$\|\nabla^* K_\rho(\mathbf{z}) - \nabla^* K_\rho(\tilde{\mathbf{z}})\|_2 \leq C\|\mathbf{z} - \tilde{\mathbf{z}}\|_2, \qquad \forall i \in \mathbf{I}_1^N, \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{F}_z. \tag{5.43b}$$

Let $\hat{\alpha} \in [0, 1]$ be the maximum feasible stepsize for all $\mathbf{z} \in \mathcal{F}_\mathbf{z}$, i.e., $\hat{\alpha} := \max\{\alpha \in [0, 1] : \mathbf{z}, \mathbf{z} + \alpha\Delta\mathbf{z} \in \mathcal{F}_z\}$. Moreover, suppose $\|\nabla^* K_\rho(\mathbf{z}) - \nabla K_\rho(\mathbf{z})\|_2 \leq \delta$ for all $\mathbf{z} \in \mathcal{F}_\mathbf{z}$ and $D \cdot \delta < 1$. Then, for all $\mathbf{z} \in \mathcal{F}_z$, there exists $\alpha \in [0, \hat{\alpha}]$ such that

1. *if $\|K_\rho(\mathbf{z})\|_2 > \frac{1-D\delta}{D^2C\hat{\alpha}}$, then $\|K_\rho(\mathbf{z} + \alpha\Delta\mathbf{z})\|_2 \leq \|K_\rho(\mathbf{z})\|_2 - \frac{(1-D\delta)^2}{2D^2C}$;*
2. *if $\|K_\rho(\mathbf{z})\|_2 \leq \frac{1-D\delta}{D^2C\hat{\alpha}}$, then $\|K_\rho(\mathbf{z} + \alpha\Delta\mathbf{z})\|_2 \leq \big(1 - \frac{1}{2}\hat{\alpha}(1 - D\delta)\big) \cdot \|K_\rho(\mathbf{z})\|_2$, and we have exponential convergence.*

Proof. The proof can be found in the Appendix. □

Theorem 7 suggests that, under certain conditions, the merit function $\|K_\rho(\mathbf{z})\|_2$ decays to zero exponentially fast, and Algorithm 3 converges to a solution satisfying the KKT conditions considering the quasi-policy gradients. The above analysis can be considered as an extension of the classical PDIP convergence proof in [35] to constrained feedback Stackelberg games where we consider feedback interaction constraints $0 = u_t^i - \tilde{\pi}_{t,\rho}^i(x_t, u_t^{1:i-1})$ and the quasi-policy gradients. The condition (5.43a) equates to establishing a lower bound for the smallest nonzero singular value of $\nabla K_\rho(\mathbf{z})$. Practically, this can be achieved by adding a minor cost regularization term to the KKT conditions [57]. Moreover, the constant C in (5.43b) depends on the maximum singular values of the Hessians of costs, the Jacobian of constraints, and linear dynamics, which are all constant matricies in LQ games and can therefore be upper bounded.

Given a $\rho > 0$, a converged solution $\mathbf{z}_\rho^*$ renders $K_\rho(\mathbf{z}_\rho^*) = 0$. Note that the KKT conditions $0 = K_\rho(\mathbf{z}_\rho^*)$ reduce to the one in Theorem 5 when ρ decays to zero. As ρ approaches zero, the solution $\mathbf{z}_\rho^*$, when converged, recovers a solution to the KKT conditions in Theorem 5. When the sufficient conditions in Theorem 6 are also satisfied, the computed solution converges to a local feedback Stackelberg equilibrium.

5.6 From LQ Games to Nonlinear Games

In this section, we extend our solution for LQ games to feedback Stackelberg games with nonlinear dynamics. Without loss of generality, each player could have non-quadratic costs. Coupled nonlinear equality and inequality constraints could also exist among players.

Iteratively Approximating Nonlinear Games via LQ Games by Aligning Their KKT Conditions

In this subsection, we introduce a procedure which iteratively approximates the constrained nonlinear games using constrained LQ games, and computes approximate local feedback Stackelberg equilibria for the nonlinear games. These LQ game approximations are designed to ensure that the first-order approximations of their KKT conditions, expressed as $0 = \nabla K_\rho(\mathbf{z}) \cdot \Delta\mathbf{z} + K_\rho(\mathbf{z})$, align with those of the original nonlinear games, specifically considering the inclusion of quasi-policies. Our approach differs from the existing iterative LQ game approximation techniques [154, 137] for FSE policies, which linearize the dynamics and quadraticize only the costs. In contrast, our method linearizes the dynamics but quadraticizes the Lagrangian. This enables us to utilize the convergence results for LQ games, as discussed in the previous section, to analyze the convergence properties of our method in nonlinear games. Consequently, our work provides the first iterative LQ game approximation approach that has provable convergence guarantees for constrained nonlinear feedback Stackelberg games.

In what follows, we introduce local LQ game approximations of the original nonlinear game. Let $\mathbf{z}$ be a solution in the set $\mathcal{F}_z$. We first define the following linear approximation of the dynamics and constraints around $\mathbf{z}$, for all $t \in \mathbf{I}_0^T, i \in \mathbf{I}_1^N$,

$$
\begin{aligned}
&A_t := \nabla_{x_t} f_t(x_t, u_t), && B_t^i := \nabla_{u_t^i} f_t(x_t, u_t), && c_t := f_t(x_t, u_t) - x_{t+1},\\
&H_{x_t}^i := \nabla_{x_t} h_t^i,\ H_{u_t^j}^i := \nabla_{u_t^j} h_t^i, && G_{x_t}^i := \nabla_{x_t} g_t^i,\ G_{u_t^j}^i := \nabla_{u_t^j} g_t^i, && \forall j \in \mathbf{I}_1^N,\\
&\bar{h}_t^i := h_t^i(x_t, u_t), && \bar{g}_t^i := g_t^i(x_t, u_t), &&\\
&H_{x_{T+1}}^i := \nabla_{x_{T+1}} h_{T+1}^i, && G_{x_{T+1}}^i := \nabla_{x_{T+1}} g_{T+1}^i, &&\\
&\bar{h}_{T+1}^i := h_{T+1}^i(x_{T+1}), && \bar{g}_{T+1}^i := g_{T+1}^i(x_{T+1}). &&
\end{aligned}
\tag{5.44}
$$

For each $i \in \mathbf{I}_1^N$ and $t \in \mathbf{I}_0^T$, we represent the second order terms and cost-related terms in the Lagrangian $\mathcal{L}_t^i$ as quadratic costs (5.23), with parameters defined as follows,

$$
\begin{aligned}
Q_t^i :=& \nabla_{xx}^2 \ell_t^i + (\nabla_{xx}^2 f_t)^\top \lambda_t^i - (\nabla_{xx}^2 h_t^i)^\top \mu_t^i - (\nabla_{xx}^2 g_t^i)^\top \gamma_t^i,\\
S_t^i :=& \nabla_{ux}^2 \ell_t^i + (\nabla_{ux}^2 f_t)^\top \lambda_t^i - (\nabla_{ux}^2 h_t^i)^\top \mu_t^i - (\nabla_{ux}^2 g_t^i)^\top \gamma_t^i,\\
R_t^i :=& \nabla_{uu}^2 \ell_t^i + (\nabla_{uu}^2 f_t)^\top \lambda_t^i - (\nabla_{uu}^2 h_t^i)^\top \mu_t^i - (\nabla_{uu}^2 g_t^i)^\top \gamma_t^i,\\
Q_{T+1}^i :=& \nabla_{xx}^2 \ell_{T+1}^i - (\nabla_{xx}^2 h_{T+1}^i)^\top \mu_{T+1}^i - (\nabla_{xx}^2 g_{T+1}^i)^\top \gamma_{T+1}^i,\\
q_t^i :=& \nabla_x \ell_t^i, \qquad r_t^i := \nabla_u \ell_t^i, \qquad q_{T+1}^i := \nabla_x \ell_{T+1}^i.
\end{aligned}
\tag{5.45}
$$

We can modify Algorithm 3 to address nonlinear games by applying an LQ game approximation around the solution $\mathbf{z}_\rho^{(k)}$ in step 3 of Algorithm 3 and formulate the resulting approximate KKT conditions $0 = K_\rho(\mathbf{z}_\rho^{(k)})$ defined with terms in (5.44) and (5.45). Furthermore, this LQ game approximation is reiterated around $\mathbf{z}_\rho^{(k)} + \alpha\Delta\mathbf{z}_\rho$ in step 6, when we evaluate the merit function $\|K_\rho(\mathbf{z}_\rho^{(k)} + \alpha\Delta\mathbf{z}_\rho)\|_2$ during line search.

Quasi-Policies Approximation Error and Exponential Convergence Analysis in Nonlinear Games

In the above solution procedure, we approximate the ground truth nonlinear policies of nonlinear games by quasi-policies. However, different from LQ games, the ground truth feedback Stackelberg policies for nonlinear games could have nonzero high-order policy gradients. Thus, it is worthwhile to characterize the error caused by the quasi-policy gradients. Essentially, there are two error sources. The first type of error is due to the fact that we have neglected high-order policy gradients when evaluating the KKT Jacobian $\nabla K_{t,\rho}^i(\mathbf{z})$, and the second form of error is how these changes propagate into the expression of KKT conditions $0 = K_{t,\rho}^i(\mathbf{z})$ for earlier players and stages. Suppose those two error sources could be upper bounded; then, we can characterize their impact on the policy gradients error in the following proposition.

Proposition 6. *Under Assumption 2, let* $\mathbf{z}$ *and* $\tilde{\mathbf{z}}$ *be two elements in the solution set* $\mathcal{F}_z$*. We denote by* $\{\pi_{t,\rho}^i\}_{t=0,i=1}^{T,N}$ *a set of policies around* $\mathbf{z}$ *and* $\{\tilde{\pi}_{t,\rho}^i\}_{t=0,i=1}^{T,N}$ *a set of quasi-policies around* $\tilde{\mathbf{z}}$*, respectively. We denote by* $\{K_{t,\rho}^i(\mathbf{z})\}_{t=0,i=1}^{T,N}$ *and* $\{K_{t,\rho}^{i*}(\mathbf{z})\}_{t=0,i=1}^{T,N}$ *the KKT conditions with and without quasi-policies, respectively. Let* $i \leq N$ *and* $t \leq T$*. Suppose that the Jacobian matrices* $\nabla K_{t,\rho}^i(\tilde{\mathbf{z}})$*,* $\nabla K_{t,\rho}^{i*}(\tilde{\mathbf{z}})$ *and* $\nabla K_{t,\rho}^{i*}(\mathbf{z})$ *are invertible. Let* $\epsilon_{\mathbf{z},\tilde{\mathbf{z}}} > 0$ *be an upper error bound such that*

$$\begin{aligned}\max\Big\{&\|\nabla K_{t,\rho}^i(\tilde{\mathbf{z}}) - \nabla K_{t,\rho}^{i*}(\tilde{\mathbf{z}})\|_2, \quad \|\nabla K_{t,\rho}^{i*}(\tilde{\mathbf{z}}) - \nabla K_{t,\rho}^{i*}(\mathbf{z})\|_2, \\ &\|K_{t,\rho}^i(\tilde{\mathbf{z}}) - K_{t,\rho}^i(\mathbf{z})\|_2, \qquad \|K_{t,\rho}^i(\mathbf{z}) - K_{t,\rho}^{i*}(\mathbf{z})\|_2\Big\} \leq \epsilon_{\mathbf{z},\tilde{\mathbf{z}}}.\end{aligned} \tag{5.46}$$

Then, the error between the quasi-policy gradient and the policy gradient can be bounded as follows,

$$\begin{aligned}\|\nabla\tilde{\pi}_{t,\rho}^i(\tilde{\mathbf{z}}) - \nabla\pi_{t,\rho}^i(\mathbf{z})\|_2 \leq \epsilon_{\mathbf{z},\tilde{\mathbf{z}}} \cdot \Big(&2\|\nabla K_{t,\rho}^{i*}(\mathbf{z})^{-1}\|_2 + \\ &\big(\|\nabla K_{t,\rho}^i(\tilde{\mathbf{z}})^{-1}\|_2 + \|\nabla K_{t,\rho}^{i*}(\mathbf{z})^{-1}\|_2\big) \cdot \|\nabla K_{t,\rho}^{i*}(\tilde{\mathbf{z}})^{-1}\|_2\|K_{t,\rho}^i(\tilde{\mathbf{z}})\|_2\Big).\end{aligned} \tag{5.47}$$

Proof. The proof can be found in Appendix. □

Proposition 6 suggests that the error introduced by the quasi-policy gradients is proportional to $\epsilon_{\mathbf{z},\tilde{\mathbf{z}}}$, as described in (5.47). However, it is challenging to obtain an analytical bound $\epsilon_{\mathbf{z},\tilde{\mathbf{z}}}$ because the evaluation of $K_{t,\rho}^{i*}(\mathbf{z})$ and $\nabla K_{t,\rho}^{i*}(\mathbf{z})$ requires computing the high-order policy gradients. The above analysis only provides a partial analysis for the policy gradient error introduced by the quasi-policy gradients. In principle, it is possible that the quasi-policy gradients could lead to a different feedback

Stackelberg policy from the ground truth feedback Stackelberg policy. However, it is intractable to compute high-order policy gradients when we have a long horizon game. In general, the quasi-policy is a local linear approximation to the ground truth nonlinear feedback Stackelberg policy, and when a state perturbation occurs at time t, such policies are only approximately optimal for the resulting sub-game. We believe that the local feedback Stackelberg quasi-policy is the closest computationally tractable approximation possible when we consider the first-order policy approximation techniques for long-horizon feedback Stackelberg games.

Furthermore, we can leverage the sufficient condition of the local FSE and the convergence analysis in Theorem 7 to show that we will converge to a local FSE of nonlinear games under certain conditions on the iterative LQ game approximations.

Theorem 8 (Exponential Convergence in Nonlinear Games)**.** *Suppose that there exist constants $(D, C, \delta, \hat{\alpha})$, as defined in Theorem 7, such that at each iteration k of Algorithm 3, the approximate LQ game defined in* (5.44) *and* (5.45) *satisfies the conditions of Theorem 7. Then, for each $\rho > 0$ and a sufficiently large k, $\mathbf{z}_\rho^{(k)}$ converges exponentially fast to a solution $\mathbf{z}_\rho^*$, which renders $\|K_\rho(\mathbf{z}_\rho^*)\|_2 = 0$. Moreover, if the limit $\mathbf{z}^* := \lim_{\rho \to 0} \mathbf{z}_\rho^*$ exists and Theorem 6, which provides a sufficient condition for local FSE trajectories, holds true at $\mathbf{z}_\rho^*$ for all $\rho > 0$, then the converged solution $\mathbf{z}^*$ recovers a local FSE trajectory.*

Proof. The proof can be found in the Appendix. □

5.7 Experiments

In this section, we consider a two-player feedback Stackelberg game modeling highway driving[1], where two highway lanes merge into one and the planning horizon $T = 20$. We associate with each player a 4-dimensional state vector $x_t^i = [p_{x,t}^i, p_{y,t}^i, v_t^i, \theta_t^i]$, where $(p_{x,t}^i, p_{y,t}^i)$ represents the (x, y) coordinate, v_t^i denotes the velocity, and θ_t^i encodes the heading angle of player i at time t. The joint state vector of the two players is denoted as $x_t = [x_t^1, x_t^2]$. Both players have nonlinear unicycle dynamics, $\forall t \in \mathbf{I}_0^T$, $\forall i \in \{1, 2\}$,

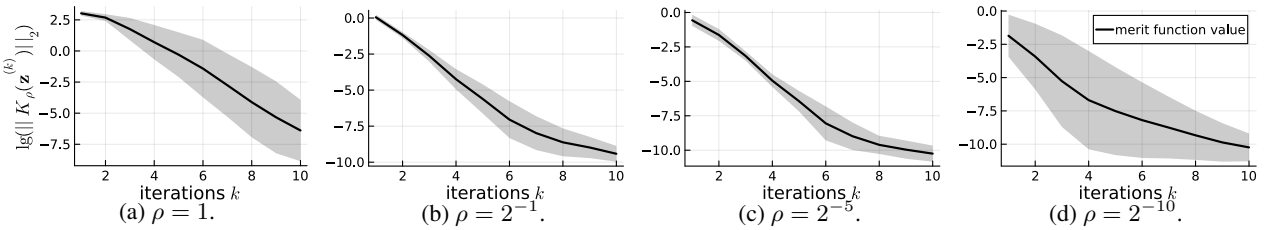
$$\begin{aligned} p_{x,t+1}^i &= p_{x,t}^i + \Delta t \cdot v_t^i \sin(\theta_t^i), \quad & p_{y,t+1}^i &= p_{y,t}^i + \Delta t \cdot v_t^i \cos(\theta_t^i), \\ v_{t+1}^i &= v_t^i + \Delta t \cdot a_t^i, & \theta_{t+1}^i &= \theta_t^i + \Delta t \cdot \omega_t^i. \end{aligned} \tag{5.48}$$

We consider the following cost functions, for all $t \in \mathbf{I}_0^T$,

$$\ell_t^1(x_t, u_t) = 10(p_{x,t}^1 - 0.4)^2 + 6(v_t^1 - v_t^2)^2 + 2\|u_t^1\|_2^2, \ell_t^2(x_t, u_t) = \|\theta_t^2\|_2^4 + 2\|u_t^2\|_2^2, \tag{5.49}$$

and the terminal costs $\ell_{T+1}^1(x_{T+1}) = 10(p_{x,T+1}^1 - 0.4)^2 + 6(v_t^1 - v_t^2)^2$ and $\ell_{T+1}^2(x_{T+1}) = \|\theta_{T+1}^2\|_2^4$. Note that we include a fourth-order cost term in player 2's cost at each stage to model its preference

[1]The code is available at `https://github.com/jamesjingqili/FeedbackStackelbergGames.jl.git`



(a) $\rho = 1$. (b) $\rho = 2^{-1}$. (c) $\rho = 2^{-5}$. (d) $\rho = 2^{-10}$.

Figure 5.1: Convergence of Algorithm 3 with iterative LQ game approximations under different values of the homotopy parameter ρ from 10 sampled initial states. The solid curve and the shaded area denote the mean and the standard deviation of the logarithm of the merit function values, respectively. By gradually annealing ρ to zero, the solution converges to a local FSE trajectory. Moreover, under each ρ, the plots above empirically support the linear convergence described in Theorem 8.

of small heading angle. We consider the following (nonconvex) constraints encoding collision avoidance, driving on the road, and control limits,

$$\begin{aligned} &\sqrt{\|p^1_{x,t} - p^2_{x,t}\|_2^2 + \|p^1_{y,t} - p^2_{y,t}\|_2^2} - d_{\text{safe}} \geq 0, && t \in \mathbf{I}_0^{T+1}, \\ &p^i_{x,t} - p_l \geq 0, \qquad p_r(p^i_{y,t}, p^i_{x,t}) \geq 0, \qquad \|u_t\|_\infty \leq u_{\max}, && t \in \mathbf{I}_0^{T+1}, i \in \{1,2\}, \end{aligned} \tag{5.50}$$

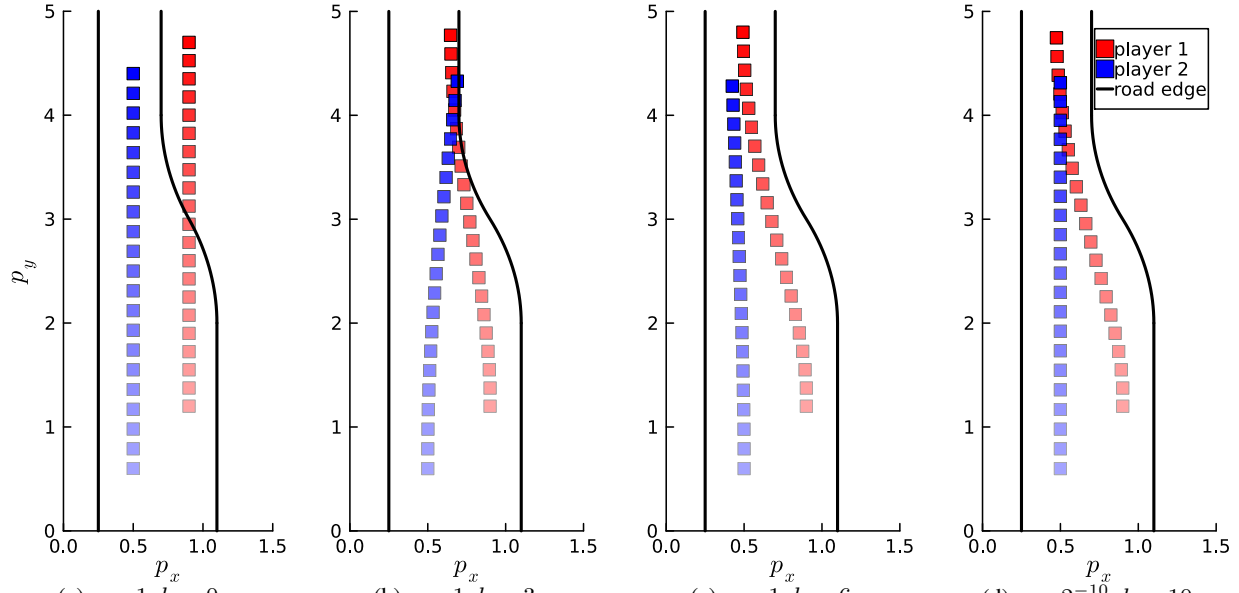
where we define $p_l \in \mathbb{R}$ to be the left road boundary and denote by $p_r(p^i_{x,t}, p^i_{y,t})$ the distance between player i and the right road boundary curve. We also consider the following equality constraints at the terminal time

$$v^1_{T+1} - v^2_{T+1} = 0,\ \theta^1_{T+1} = 0, \tag{5.51}$$

where the two players aim to reach a consensus on their speeds, with player 1 maintaining its heading angle pointing forwards.

The nominal initial states of two players are specified as $x_0^1 = [0.9, 1.2, 3.5, 0.0]$ and $x_0^2 = [0.5, 0.6, 3.8, 0.0]$, respectively. We randomly sample 10 initial states around $x_0 = [x_0^1, x_0^2]$ under a uniform distribution within the range of -0.1 to 0.1. From each sampled $\hat{x}_0$, we obtain an initial state trajectory $\mathbf{x}^{(0)}$ by simulating the nonlinear dynamics (5.48) with the initial controls $\mathbf{u}^{(0)} = \mathbf{0}$. Set the initial slack variables for the inequality constraints as $\mathbf{s}^{(0)} = \mathbf{1}$, along with the corresponding Lagrange multipliers $\boldsymbol{\gamma}^{(0)} = \mathbf{1}$. We set all other Lagrange multipliers $\{\boldsymbol{\lambda}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\eta}^{(0)}, \boldsymbol{\psi}^{(0)}\}$ to zeros. Consequently, we have constructed an initial solution $\mathbf{z}^{(0)} = [\mathbf{x}^{(0)}, \mathbf{u}^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\gamma}^{(0)}\boldsymbol{\mu}^{(0)}, \boldsymbol{\eta}^{(0)}, \boldsymbol{\psi}^{(0)}, \mathbf{s}^{(0)}]$. We repeat this initialization trajectory defining process for different sampled $\hat{x}_0$.

For each sampled initial state $\hat{x}_0$, we employ Algorithm 3 with iterative LQ game approximations to compute a local FSE trajectory. The convergence of our method under different sampled $\hat{x}_0$ is depicted in Figure 5.1. For each ρ, the merit function value decreases as the iterations continue. Furthermore, since the cost functions are strongly convex with respect to each player's controls, Theorem 6 ensures that our converged solution constitutes a local FSE trajectory. Moreover, we show our method can tolerate infeasible initialization in Figure 5.2, where the right road boundary constraint is initially violated by initialization $\mathbf{z}^{(0)}$, and as the algorithm progresses, subsequent iterates $\mathbf{z}^{(k)}$ become feasible.



(a) $\rho = 1, k = 0$. (b) $\rho = 1, k = 3$. (c) $\rho = 1, k = 6$. (d) $\rho = 2^{-10}, k = 10$.

Figure 5.2: Tolerance of an infeasible trajectory initialization and the converged trajectories of two players. In Figure 5.2a, we plot the initial state trajectories of two players, where player 1's trajectory is infeasible because it violates the road boundary constraint. When $\rho = 1$, we plot the state trajectories in the third and the sixth iterations in Figure 5.2b and Figure 5.2c, respectively. They become feasible at the sixth iteration. In Figure 5.2d, we plot the converged solution, with $\rho = 2^{-10}$.

5.8 Conclusions

In this chapter, we considered general-sum feedback Stackelberg dynamic games with coupled constraints among N players. We proposed a primal-dual interior point method to compute an approximate feedback Stackelberg equilibrium and the associated policies for all players. To the best of the authors' knowledge, this represents the first attempt to compute approximate local feedback Stackelberg equilibria in both LQ games and nonlinear games under general coupled equality and inequality constraints, within continuous state and action spaces. We theoretically characterized the approximation error and the exponential convergence of our algorithm. Numerical experiments suggest that the proposed algorithm can tolerate infeasible initializations and efficiently converge to a feasible equilibrium solution. Future research should investigate the potential benefits of higher-order policy gradient approximations. Additionally, extending our approach to solve other types of equilibria in dynamic games is also a promising direction for future research.

Acknowledgments For This Chapter

We would like to thank Professor Lillian Ratliff, Professor Forrest Laine, and Eli Brock for valuable discussions and comments.

5.9 Supplementary results

Proof of Theorem 4. At the terminal time $t = T$, for ease of notation, we define $x_T = \bar{x}_T$, and $u_T^{1:i-1} = \bar{u}_T^{1:i-1}$. We observe that, for each player $i \in \mathbf{I}_1^N$, the equation (5.15) can be rewritten as

$$\begin{aligned}\tilde{u}_T^i \in \arg_{u_T^i} \min_{u_T^i} \Big\{ \min_{\substack{u_T^{i+1:N} \\ x_{T+1}}} \ell_T^i(x_T, u_T) + V_{T+1}^i(x_{T+1}) \Big\} \\ \text{s.t. } 0 = u_T^j - \pi_T^j(x_T, u_T^{1:j-1}),\ 0 = x_{T+1} - f_T(x_T, u_T) \qquad j \in \mathbf{I}_{i+1}^N \\ 0 = h_T^i(x_T, u_T),\ 0 \leq g_T^i(x_T, u_T) \\ 0 = h_{T+1}^i(x_{T+1}),\ 0 \leq g_{T+1}^i(x_{T+1})\end{aligned}$$

which implies $\tilde{u}_T^i \in \arg_{u_T^i} \min_{u_T^i} Z_T^i(\bar{x}_T, \bar{u}_T^{1:i-1}, u_T^i)$. Moreover, for all $t \in \mathbf{I}_0^{T-1}$ and $i \in \mathbf{I}_1^N$, for the ease of notation, we assume $x_t = \bar{x}_t$, and $u_t^{1:i-1} = \bar{u}_t^{1:i-1}$. We observe

$$\begin{aligned}\tilde{u}_t^i \in \arg_{u_t^i} \min_{u_t^i} \Big\{ \min_{\substack{u_t^{i+1:N} \\ u_{t+1:T}^{1:N} \\ x_{t+1:T+1}}} \sum_{\tau=t}^{T} \ell_\tau^i(x_\tau, u_\tau) + \ell_{T+1}^i(x_{T+1}) \Big\} \\ \text{s.t. } 0 = u_t^j - \pi_t^j(x_t, u_t^{1:j-1}) \qquad j \in \mathbf{I}_{i+1}^N \\ 0 = x_{\tau+1} - f_\tau(x_\tau, u_\tau) \qquad \tau \in \mathbf{I}_t^T \\ 0 = u_\tau^j - \pi_\tau^j(x_\tau, u_\tau^{1:j-1}) \qquad \tau \in \mathbf{I}_{t+1}^T, j \in \mathbf{I}_1^N \setminus \{i\} \\ 0 = h_\tau^i(x_\tau, u_\tau),\ 0 \leq g_\tau^i(x_\tau, u_\tau) \qquad \tau \in \mathbf{I}_t^T \\ 0 = h_{T+1}^i(x_{T+1}),\ 0 \leq g_{T+1}^i(x_{T+1})\end{aligned}$$

The above can be further rewritten as

$$\begin{aligned}\tilde{u}_t^i \in \arg_{u_t^i} \min_{u_t^i} \Big\{ \min_{\substack{u_t^{i+1:N} \\ x_{t+1}}} \ell_t^i(x_t, u_t) + V_{t+1}^i(x_{t+1}) \Big\} \\ \text{s.t. } 0 = u_t^j - \pi_t^j(x_t, u_t^{1:j-1}),\ 0 = x_{t+1} - f_t(x_t, u_t) \qquad j \in \mathbf{I}_{i+1}^N \\ 0 = h_t^i(x_t, u_t),\ 0 \leq g_t^i(x_t, u_t)\end{aligned}$$

It follows that $\tilde{u}_t^i \in \arg_{u_t^i} \min_{u_t^i} Z_t^i(\bar{x}_t, \bar{u}_t^{1:i-1}, u_t^i)$. Therefore, the set of strategies $\{\pi_t^i\}_{t=0,i=1}^{T,N}$ constitutes a set of local feedback Stackelberg policies. □

Proof of Theorem 5. For a time $t \in \mathbf{I}_0^T$ and player $i \in \mathbf{I}_1^N$, we set the gradient of $\mathcal{L}_t^i$ with respect to $\{u_\tau^i\}_{\tau=t}^T$ and $\{x_\tau\}_{\tau=t+1}^{T+1}$ to be zero. This constitutes the first two rows of (5.19). In addition, a player $i < N$ considers the feedback interaction constraints $0 = u_t^{j*} - \pi_t^j(x_t^*, u_1^{1:j-1*})$, for $j \in \mathbf{I}_{i+1}^N$. This constraint is implicitly ensured when we include player j's KKT conditions into player i's KKT conditions. Thus, we only need to ensure the gradient $\nabla_{u_t^j} \mathcal{L}_t^i$ to be zero, when synthesizing player i's KKT conditions. This corresponds to the third row of (5.19). Moreover, at a time $t < T$, each player $i \in \mathbf{I}_1^N$ needs to account for the feedback reaction from other players in future steps.

Again this constraint is implicitly ensured when we define player j's KKT conditions. We only need to additionally set the gradient of $\mathcal{L}_t^i$ with respect to u_τ^j to be zero, where $\tau \in \mathbf{I}_{t+1}^T$ and $j \in \mathbf{I}_1^N \setminus \{i\}$. These correspond to the fourth row of (5.19). Finally, we include the dynamics constraints, equality and inequality constraints, and complementary slackness conditions in the last five rows of (5.19). □

Proof of Theorem 6. We can check that the feasible set for the equality constraints of (5.20) is a superset of the critical cone of the problem (5.19). By Theorem 12.6 in [238], the solution$(\mathbf{x}^*, \mathbf{u}^*)$ constitutes a local feedback Stackelberg equilibrium trajectory. □

Proof of Theorem 7. By fundamental theorem of calculus, we have $K_\rho(\mathbf{z} + \alpha\Delta\mathbf{z}) = K_\rho(\mathbf{z}) + \int_0^1 \nabla^* K_\rho(\mathbf{z} + \tau\alpha\Delta\mathbf{z})\alpha\Delta\mathbf{z}d\tau$, and we have

$$\begin{aligned}\|K_\rho(\mathbf{z} + \alpha\Delta\mathbf{z})\|_2 &= \left\|K_\rho(\mathbf{z}) + \int_0^1 \nabla^* K_\rho(\mathbf{z} + \tau\alpha\Delta\mathbf{z})\alpha\Delta\mathbf{z}d\tau\right\|_2 \\ &\leq \|K_\rho(\mathbf{z}) + \alpha\nabla^* K_\rho(\mathbf{z})\Delta\mathbf{z}\|_2 + \left\|\int_0^1 (\nabla^* K_\rho(\mathbf{z} + \tau\alpha\Delta\mathbf{z}) - \nabla^* K_\rho(\mathbf{z}))\alpha\Delta\mathbf{z}d\tau\right\|_2\end{aligned} \tag{5.52}$$

Substituting $\Delta\mathbf{z}$ into $\|K_\rho(\mathbf{z}) + \alpha\nabla^* K_\rho(\mathbf{z})\Delta\mathbf{z}\|_2$, we have

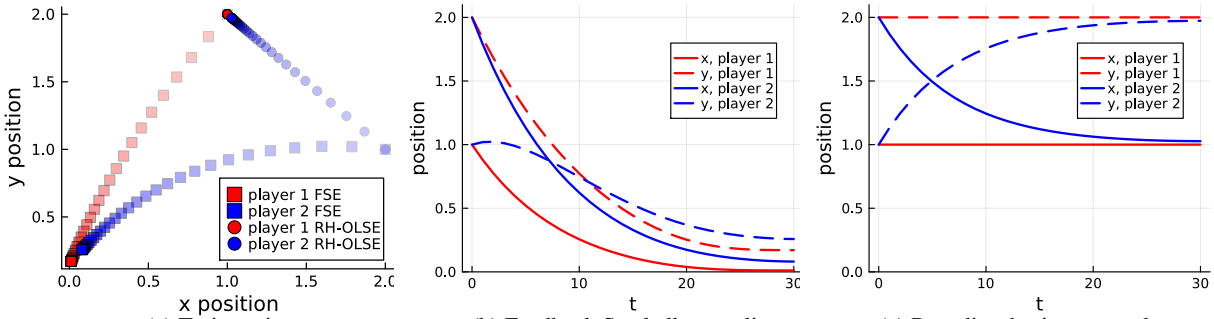
$$\begin{aligned}\|K_\rho(\mathbf{z}) + \alpha\nabla^* K_\rho(\mathbf{z})\Delta\mathbf{z}\|_2 &= \|K_\rho(\mathbf{z}) - \alpha\nabla^* K_\rho(\mathbf{z})(\nabla K_\rho(\mathbf{z}))^{-1}K_\rho(\mathbf{z})\|_2 \\ &\leq (1-\alpha)\|K_\rho(\mathbf{z})\|_2 + \alpha\|\nabla^* K_\rho(\mathbf{z}) - \nabla K_\rho(\mathbf{z})\|_2\|(\nabla K_\rho(\mathbf{z}))^{-1}\|_2\|K_\rho(\mathbf{z})\|_2 \\ &\leq (1-\alpha)\|K_\rho(\mathbf{z})\|_2 + \alpha\delta D\|K_\rho(\mathbf{z})\|_2 = (1-\alpha(1-\delta D))\|K_\rho(\mathbf{z})\|_2\end{aligned} \tag{5.53}$$

Combining (5.53) and (5.52), we have

$$\begin{aligned}&\|K_\rho(\mathbf{z} + \alpha\Delta\mathbf{z})\|_2 \\ &\leq (1-\alpha(1-\delta D))\|K_\rho(\mathbf{z})\|_2 + \|\alpha\Delta\mathbf{z}\|_2\left\|\int_0^1 \|\nabla^* K_\rho(\mathbf{z} + \tau\alpha\Delta\mathbf{z}) - \nabla^* K_\rho(\mathbf{z})\|d\tau\right\|_2 \\ &\leq (1-\alpha(1-\delta D))\|K_\rho(\mathbf{z})\|_2 + \frac{1}{2}\alpha^2 D^2 C\|K_\rho(\mathbf{z})\|_2^2\end{aligned}$$

where the right hand side is minimized when $\alpha^* = \frac{1-D\delta}{D^2C\|K_\rho(\mathbf{z})\|_2}$. Suppose $\|K_\rho(\mathbf{z})\|_2 > \frac{1-D\delta}{D^2C\hat{\alpha}}$, then $\hat{\alpha} > \frac{1-D\delta}{D^2C\|K_\rho(\mathbf{z})\|_2}$ and we have $\|K_\rho(\mathbf{z} + \alpha^*\Delta\mathbf{z})\|_2 \leq \|K_\rho(\mathbf{z})\|_2 - \frac{(1-D\delta)^2}{2D^2C}$.

For the case $\|K_\rho(\mathbf{z})\|_2 \leq \frac{1-D\delta}{D^2C\hat{\alpha}}$, let $\alpha := \hat{\alpha}$. By $\hat{\alpha}D^2C\|K_\rho(\mathbf{z})\|_2 \leq 1 - D\delta$, we have $\|K_\rho(\mathbf{z} + \hat{\alpha}\Delta\mathbf{z})\|_2 \leq (1 - \frac{1}{2}\hat{\alpha}(1-D\delta))\|K_\rho(\mathbf{z})\|_2$. □



(a) Trajectories (b) Feedback Stackelberg policy (c) Receding-horizon open-loop Stackelberg policy

Figure 5.3: The trajectories under the receding horizon open-loop Stackelberg equilibrium (RH-OLSE) policy and those under the FSE policy are quite different, regardless of the initial conditions. For example, in the above case, under the FSE policy, player 1 first moves towards the origin and then player 2 follows. However, under the RH-OLSE policy, player 1 always stays at its initial position, waiting for player 2 to approach.

Proof of Proposition 6. By definition, we have

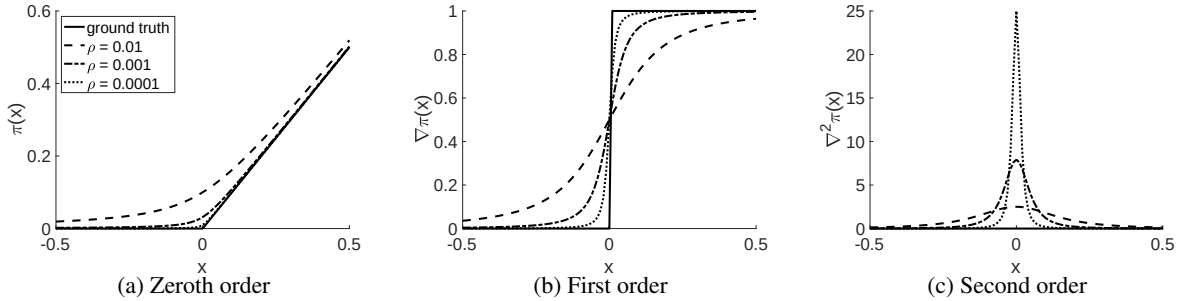
$$
\begin{aligned}
&\|\nabla\tilde{\pi}^i_{t,\rho}(\tilde{\mathbf{z}}) - \nabla\pi^i_{t,\rho}(\mathbf{z})\|_2 = \|\nabla K^i_{t,\rho}(\tilde{\mathbf{z}})^{-1}K^i_{t,\rho}(\tilde{\mathbf{z}}) - \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}K^{i*}_{t,\rho}(\mathbf{z})\|_2 \\
&= \|\nabla K^i_{t,\rho}(\tilde{\mathbf{z}})^{-1}K^i_{t,\rho}(\tilde{\mathbf{z}}) - \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}K^{i*}_{t,\rho}(\mathbf{z}) \\
&\qquad + \nabla K^{i*}_{t,\rho}(\tilde{\mathbf{z}})^{-1}K^i_{t,\rho}(\tilde{\mathbf{z}}) - \nabla K^{i*}_{t,\rho}(\tilde{\mathbf{z}})^{-1}K^i_{t,\rho}(\tilde{\mathbf{z}}) \\
&\qquad + \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}K^i_{t,\rho}(\tilde{\mathbf{z}}) - \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}K^i_{t,\rho}(\tilde{\mathbf{z}}) \\
&\qquad + \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}K^i_{t,\rho}(\mathbf{z}) - \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}K^i_{t,\rho}(\mathbf{z})\|_2 \\
&\leq \Big(\|\nabla K^i_{t,\rho}(\tilde{\mathbf{z}})^{-1} - \nabla K^{i*}_{t,\rho}(\tilde{\mathbf{z}})^{-1}\|_2 + \|\nabla K^{i*}_{t,\rho}(\tilde{\mathbf{z}})^{-1} - \nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}\|_2\Big)\|K^i_{t,\rho}(\tilde{\mathbf{z}})\|_2 \\
&\qquad + \Big(\|K^i_{t,\rho}(\tilde{\mathbf{z}}) - K^i_{t,\rho}(\mathbf{z})\|_2 + \|K^i_{t,\rho}(\mathbf{z}) - K^{i*}_{t,\rho}(\mathbf{z})\|_2\Big)\|\nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}\|_2 \\
&\leq \epsilon_{\mathbf{z},\tilde{\mathbf{z}}}\Big(\|\nabla K^i_{t,\rho}(\tilde{\mathbf{z}})^{-1}\|_2 + \|\nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}\|_2\Big)\|\nabla K^{i*}_{t,\rho}(\tilde{\mathbf{z}})^{-1}\|_2\|K^i_{t,\rho}(\tilde{\mathbf{z}})\|_2 \\
&\qquad + 2\epsilon_{\mathbf{z},\tilde{\mathbf{z}}}\|\nabla K^{i*}_{t,\rho}(\mathbf{z})^{-1}\|_2
\end{aligned}
$$

where the last line follows by applying Lemma 2. □

Lemma 2. *Let K and $\tilde{K}$ be two invertible matrices. Suppose $\|K - \tilde{K}\|_2 \leq \epsilon$, then we have $\|K^{-1} - \tilde{K}^{-1}\|_2 \leq \epsilon\|K^{-1}\|_2 \cdot \|\tilde{K}^{-1}\|_2$.*

Proof of Lemma 2. Define $\bar{K} := K - \tilde{K}$. Applying the Woodbury matrix equality, we have $\tilde{K}^{-1} = K^{-1} + K^{-1} \cdot \bar{K} \cdot \tilde{K}^{-1}$, and this implies $\|\tilde{K}^{-1} - K^{-1}\| \leq \epsilon\|K^{-1}\| \cdot \|\tilde{K}^{-1}\|$. □

Proof of Theorem 8. Observe that the first order-approximation of the KKT conditions for the local LQ game approximations coincides with the one for nonlinear games. By Theorem 7, for each $\rho > 0$, $\lim_{k\to\infty} \|K_\rho(\mathbf{z}^{(k)}_\rho)\|_2 = 0$, and we have exponential convergence when $k \geq \|K_\rho(\mathbf{z}^{(0)}_\rho)\|_2/(\frac{1-D\delta}{D^2C\hat{\alpha}})$. Moreover, by Theorem 6, the solution $\lim_{\rho\to 0}\mathbf{z}^*_\rho$ recovers a local FSE trajectory. □



(a) Zeroth order (b) First order (c) Second order

Figure 5.4: Visualization of the policy gradients of a constrained single-stage Linear Quadratic Regulator problem under different values of ρ. The cost is given by $(u_0 - x_0)^2$. The dynamics is defined as $x_1 = x_0 + u_0$. We consider a constraint $u_0 \geq 0$. The ground truth piecewise linear policy is not differentiable at $x = 0$. As $\rho \to 0$, the policy obtained from PDIP and its first-order gradient closely approximate the ground truth policy and its first-order gradient, for all nonzero x. As shown in Figure 5.4c, the high-order gradient of the PDIP policy decays to zero as $\rho \to 0$, for all nonzero x.

The difference between Stackelberg equilibrium and Nash equilibrium in oligopoly games [298]

Consider a two-player Oligopoly game. We denote by the action u^i the amount of production of player i. Consider three positive constants C_1, C_2, and C_3, where $C_1 > C_3$. The cost of each player $i \in \{1, 2\}$ is modeled as

$$\ell^i(u^1, u^2) := -u^i \cdot (C_1 - C_2(u^1 + u^2) - C_3) \tag{5.54}$$

Both players aim to minimize their respective costs. We compute the Nash equilibrium by solving the KKT conditions

$$\begin{aligned} \frac{\partial}{\partial u^{1*}} \ell^1(u^{1*}, u^{2*}) &= -C_1 + 2C_2 u^{1*} + C_2 u^{2*} + C_3 \\ \frac{\partial}{\partial u^{2*}} \ell^2(u^{1*}, u^{2*}) &= -C_1 + 2C_2 u^{2*} + C_2 u^{1*} + C_3 \end{aligned} \tag{5.55}$$

The Nash equilibrium is $u^{1*}_{Nash} = u^{2*}_{Nash} = \frac{C_1 - C_3}{3C_2}$. In what follows, we compute the Stackelberg equilibrium. By solving player 2's KKT condition, as derived in the second line of (5.55), the optimal reaction control of player 2 is given by

$$u^{2*} = \frac{C_1 - C_3}{2C_2} - \frac{1}{2} u^{1*} \tag{5.56}$$

Substituting this into player 1's decision problem, we have

$$\begin{aligned} &\min_{u^1} \ell^1(u^1, u^2) = -u^1 \cdot (C_1 - C_2(u^1 + u^2) - C_3) \\ &\text{s.t. } u^2 = \frac{C_1 - C_3}{2C_2} - \frac{1}{2} u^1 \end{aligned} \tag{5.57}$$

Solving the above problem, we have that the optimal action of player 1 is $u^{1*}_{Stackelberg} = \frac{C_1 - C_3}{2C_2}$, and the optimal action of player 2 is $u^{2*}_{Stackelberg} = \frac{C_1 - C_3}{4C_2}$. The Stackelberg equilibrium can be arbitrarily different from the Nash equilibrium when the value $\frac{C_1 - C_3}{C_2}$ varies.

A counter example that the receding horizon open-loop Stackelberg equilibrium fails to approximate the FSE well

We consider Example 1 from [180]. We show in Figure 5.3a that the receding-horizon open-loop Stackelberg policy could lead to a trajectory quite different from the one under FSE. Therefore, it is crucial to study the computation of FSE.

The decay of high-order policy gradients when we apply PDIP to solve constrained LQ games

We validate the quasi-policy assumption in LQ games in Proposition 7, and include a simplified example in Figure 5.4.

Proposition 7. *Under the same assumptions of Theorem 7, let $\rho > 0$ and denote by $\mathbf{z}^*_\rho$ a converged solution to an LQ game under Algorithm 3 with considering high-order policy gradients. Let $\{\pi^i_{t,\rho}\}^{T,N}_{t=0,i=1}$ be the converged policies. Suppose that $\lim_{\rho\to 0}\mathbf{z}^*_\rho$ exists and we denote it by $\mathbf{z}^*$. Moreover, suppose that the ground truth FSE policies $\{\pi^i_t\}^{T,N}_{t=0,i=1}$ are differentiable at $(\mathbf{x}^*, \mathbf{u}^*)$. Then, $\lim_{\rho\to 0}\|\nabla\pi^i_{t,\rho} - \nabla\pi^i_t\|_2 = 0$, and $\lim_{\rho\to 0}\|\nabla^j\pi^i_{t,\rho}\|_2 = 0, \forall i \in \mathbf{I}^N_1, t \in \mathbf{I}^T_0$, $j \geq 2$.*

Proof. At time $t = T$, there is no policy gradient term in the N-th player's KKT conditions. Recall that $\nabla\pi^N_{T,\rho} = -[(\nabla K^N_{T,\rho})^{-1}]_{u^N_T}\nabla_{[x_T,u^{1:N-1}_T]}K^N_{T,\rho}$ and $\nabla\pi^N_T = -[(\nabla K^N_T)^{-1}]_{u^N_T}\nabla_{[x_T,u^{1:N-1}_T]}K^N_T$. Since $\lim_{\rho\to 0}\|K^N_{T,\rho}(\mathbf{z}^{N*}_{T,\rho}) - K^N_T(\mathbf{z}^{N*}_T)\|_2 = 0$, we have pointwise convergence $\lim_{\rho\to 0}\|\nabla\pi^N_{T,\rho} - \nabla\pi^N_T\|_2 = 0$ almost everywhere. We characterize those high-order quasi-policy gradients of $\pi^N_{T,\rho}$ as follows. We denote the map from $\mathbf{z}^{N*}_{T,\rho}$ to the j-th order gradient of $\pi^N_{T,\rho}$ by an operator $\mathcal{A}^{N,j}_T : \mathbf{z}^{N*}_{T,\rho} \to \nabla^j\pi^N_{T,\rho}$. Observe that $[(\nabla K^N_{T,\rho})^{-1}]_{u^N_T}$ can be considered as the concatenation of a matrix inverse operator $\mathcal{M} : X \in \mathbb{R}^{n\times n} \to X^{-1} \in \mathbb{R}^{n\times n}$ and a linear operator $\hat{\mathcal{M}} : \mathbf{z}^{N*}_{T,\rho} \to \nabla K^N_{T,\rho}$. Note that the matrix inverse is an infinitely differentiable operator when X is invertible and $\nabla_{[x_T,u^{1:N-1}_T]}K^N_{T,\rho}$ is a constant matrix. Thus, by the chain rule [266], $\pi^N_{T,\rho}$ is infinitely differentiable, which also implies that $\nabla^j\pi^N_{T,\rho}$ is continuous, $\forall j \geq 1$.

Since $\nabla K^N_{T,\rho}(\mathbf{z}^{N*}_{T,\rho})$ is invertible at $\mathbf{z}^{N*}_{T,\rho}$ and $\mathcal{A}^{N,j}_T$, $\forall j \geq 1$, is a continuous operator, there exists a compact set $\mathcal{S}$ containing $\mathbf{z}^{N*}_{T,\rho}$ such that $\nabla K^N_{T,\rho}(\mathbf{z}^N_T)$ is invertible for all $\mathbf{z}^N_T \in \mathcal{S}$. By the compactness of $\mathcal{S}$ and the continuity of $\mathcal{A}^{N,j}_T$, we have that $\mathcal{A}^{N,j}_T$ is a uniformly continuous operator on $\mathcal{S}$. By Theorem 2 in [22], a uniformly continuous operator preserves the pointwise convergence. Thus, $\lim_{\rho\to 0}\|\nabla^j\pi^N_{T,\rho} - \nabla^j\pi^N_T\|_2 = 0$. Since the ground truth policy π^N_T is piecewise linear and the high-order gradients of π^N_T vanish, we have $\lim_{\rho\to 0}\|\nabla^j\pi^N_{T,\rho}\|_2 = 0, \forall j > 1$.

Subsequently, for player $i = N-1$, since $\lim_{\rho\to 0}\|\nabla\pi^N_{T,\rho} - \nabla\pi^N_T\|_2 = 0$, we have $\lim_{\rho\to 0}\|\nabla K^i_T(\mathbf{z}^{i*}_{T,\rho}) - \nabla K^i_T(\mathbf{z}^{i*}_T)\|_2 = 0$, which implies $\lim_{\rho\to 0}\|\nabla\pi^i_{T,\rho} - \nabla\pi^i_T\|_2 = 0$. A similar reasoning as above

yields that $\lim_{\rho\to\infty} \|\nabla^j \pi^i_{T,\rho} - \nabla^j \pi^i_T\|_2 = 0, \forall j > 1$. Moreover, we can show that for all players $i < N-1$, $\lim_{\rho\to 0} \|\nabla^j \pi^i_{T,\rho} - \nabla^j \pi^i_T\|_2 = 0, \forall j \geq 1$. We continue this backward induction proof of $\lim_{\rho\to 0} \|\nabla^j \pi^i_{t,\rho} - \nabla^j \pi^i_t\|_2 = 0, \forall j \geq 1$, for prior stages backwards in players decision order until $t = 0$ and $i = 1$. □

5.10 KKT conditions for two-player LQ games

The KKT conditions $0 = K^2_{T,\rho}(\mathbf{z}^2_T)$ of player 2 at time T are

$$\begin{cases} 0 = \Sigma^2_{j=1} R^{2,2,j}_T u^j_T + S^{2,2}_T x_T + r^{2,2}_T + B^{2\top}_T \lambda^2_T - G^{2\top}_{u^2_T} \gamma^2_T - H^{2\top}_{u^2_T} \mu^2_T \\ 0 = Q^2_{T+1} x_{T+1} + q^2_{T+1} - \lambda^2_T - G^{2\top}_{x_{T+1}} \gamma^2_{T+1} - H^{2\top}_{x_{T+1}} \mu^2_{T+1} \\ 0 = x_{T+1} - A_T x_T - B^1_T u^1_T - B^2_T u^2_T - c_T \\ 0 = H^2_{u^2_T} u^2_T + H^2_{x_T} x_T + H^2_{u^1_T} u^1_T + \bar{h}^2_T \\ 0 = H^2_{x_{T+1}} x_{T+1} + \bar{h}^2_{T+1} \\ 0 = \gamma^2_{T:T+1} \odot s^2_{T:T+1} - \rho \mathbf{1} \\ 0 = G^2_{u^2_T} u^2_T + G^2_{x_T} x_T + G^2_{u^1_T} u^1_T + \bar{g}^2_T - s^2_T \\ 0 = G^2_{x_{T+1}} x_{T+1} + \bar{g}^2_{T+1} - s^2_{T+1} \end{cases}$$

We construct the KKT conditions $0 = K^1_{T,\rho}(\mathbf{z}^1_T)$ of player 1 at time T:

$$\begin{cases} 0 = \Sigma^2_{j=1} R^{1,1,j}_T u^j_T + S^{1,1}_T x_T + r^{1,1}_T + B^{1\top}_T \lambda^1_T - G^{1\top}_{u^1_T} \gamma^1_T - H^{1\top}_{u^1_T} \mu^1_T + (\nabla_{u^1_T} \pi^2_{T,\rho})^\top \psi^{1,2}_T \\ 0 = Q^1_{T+1} x_{T+1} + q^1_{T+1} - \lambda^1_T - G^{1\top}_{x_{T+1}} \gamma^1_{T+1} - H^{1\top}_{x_{T+1}} \mu^1_{T+1} \\ 0 = \Sigma^2_{j=1} R^{1,2,j}_T u^j_T + S^{1,2}_T x_T + r^{1,2}_T + B^{2\top}_T \lambda^1_T - G^{1\top}_{u^2_T} \gamma^1_T - H^{1\top}_{u^2_T} \mu^1_T - \psi^{1,2}_T \\ 0 = H^1_{u^1_T} u^1_T + H^1_{x_T} x_T + H^1_{u^2_T} u^2_T + \bar{h}^1_T \\ 0 = H^1_{x_{T+1}} x_{T+1} + \bar{h}^1_{T+1} \\ 0 = \gamma^1_{T:T+1} \odot s^1_{T:T+1} - \rho \mathbf{1} \\ 0 = G^1_{u^1_T} u^1_T + G^1_{x_T} x_T + G^1_{u^2_T} u^2_T + \bar{g}^1_T - s^1_T \\ 0 = G^1_{x_{T+1}} x_{T+1} + \bar{g}^1_{T+1} - s^1_{T+1} \\ 0 = K^2_{T,\rho}(\mathbf{z}^2_T) \end{cases}$$

We construct the KKT conditions $0 = K^2_{t,\rho}(\mathbf{z}^2_t)$ of player 2 at time $t < T$:

$$\begin{cases} 0 = \Sigma^2_{j=1} R^{2,2,j}_t u^j_t + S^{2,2}_t x_t + r^{2,2}_t + B^{2\top}_t \lambda^2_t - G^{2\top}_{u^2_t} \gamma^2_t - H^{2\top}_{u^2_t} \mu^2_t \\ 0 = Q^2_{t+1} x_{t+1} + q^2_{t+1} - \lambda^2_t - G^{2\top}_{x_{t+1}} \gamma^2_{t+1} - H^{2\top}_{x_{t+1}} \mu^2_{t+1} \\ \quad - A^\top_{t+1} \lambda^2_{t+1} + \Sigma^2_{j=1} S^{2,j}_{t+1} u^j_{t+1} + (\nabla_{x_{t+1}} \pi^1_{t+1,\rho})^\top \eta^{2,1}_t \\ 0 = \Sigma^2_{j=1} R^{2,1,j}_{t+1} u^j_{t+1} + S^{2,1}_{t+1} x_{t+1} + r^{2,1}_{t+1} + B^{1\top}_{t+1} \lambda^2_{t+1} - G^{2\top}_{u^1_{t+1}} \gamma^2_{t+1} - H^{2\top}_{u^1_{t+1}} \mu^2_{t+1} - \eta^{2,1}_t \\ 0 = x_{t+1} - A_t x_t - B^1_t u^1_t - B^2_t u^2_t - c_t \\ 0 = H^2_{u^2_t} u^2_t + H^2_{x_t} x_t + H^2_{u^1_t} u^1_t + \bar{h}^2_t \\ 0 = \gamma^2_t \odot s^2_t - \rho \mathbf{1} \\ 0 = G^2_{u^2_t} u^2_t + G^2_{x_t} x_t + G^2_{u^1_t} u^1_t + \bar{g}^2_t - s^2_t \\ 0 = K^1_{t+1,\rho}(\mathbf{z}^1_{t+1}) \end{cases}$$

We construct the KKT conditions $0 = K^1_{t,\rho}(\mathbf{z}^1_t)$ of player 1 at time $t < T$:

$$\begin{cases} 0 = \Sigma^2_{j=1} R^{1,1,j}_t u^j_t + S^{1,1}_t x_t + r^{1,1}_t + B^{1\top}_t \lambda^1_t - G^{1\top}_{u^1_t} \gamma^1_t - H^{1\top}_{u^1_t} \mu^1_t + (\nabla_{u^1_t} \pi^2_{t,\rho})^\top \psi^{1,2}_t \\ 0 = Q^1_{t+1} x_{t+1} + q^1_{t+1} - \lambda^1_t - G^{1\top}_{x_{t+1}} \gamma^1_{t+1} - H^{1\top}_{x_{t+1}} \mu^1_{t+1} \\ \quad - A^\top_{t+1} \lambda^1_{t+1} + \Sigma^2_{j=1} S^{1,j}_{t+1} u^j_{t+1} + (\nabla_{x_{t+1}} \pi^2_{t+1,\rho})^\top \eta^{1,2}_t \\ 0 = \Sigma^2_{j=1} R^{1,2,j}_t u^j_t + S^{1,2}_t x_t + r^{1,2}_t + B^{2\top}_t \lambda^1_t - G^{1\top}_{u^2_t} \gamma^1_t - H^{1\top}_{u^2_t} \mu^1_t - \psi^{1,2}_t \\ 0 = \Sigma^2_{j=1} R^{1,2,j}_{t+1} u^j_t + S^{1,2}_{t+1} x_{t+1} + r^{1,2}_{t+1} + B^{2\top}_{t+1} \lambda^1_{t+1} - G^{1\top}_{u^2_{t+1}} \gamma^1_{t+1} - H^{1\top}_{u^2_{t+1}} \mu^1_{t+1} - \eta^{1,2}_t \\ 0 = H^1_{u^1_t} u^1_t + H^1_{x_t} x_t + H_{u^2_t} u^2_t + \bar{h}^1_t \\ 0 = \gamma^1_t \odot s^1_t - \rho \mathbf{1} \\ 0 = G^1_{u^1_t} u^1_t + G^1_{x_t} x_t + G^1_{u^2_t} u^2_t + \bar{g}^1_t - s^1_t \\ 0 = K^2_{t,\rho}(\mathbf{z}^2_t) \end{cases}$$

We continue the above construction process until $i = 1$ and $t = 0$.

Chapter 6

Scenario-Game ADMM for Chance-Constrained Stochastic Games

In this chapter, we leverage the game-theoretic KKT conditions for deterministic dynamic games to approximately solve stochastic dynamic games. The challenge is the efficient approximation of the chance constraints under stochastic dynamics. Inspired by the idea of scenario optimization, which is a technique for single-objective stochastic optimization problems, we propose a new approximation for the chance-constrained stochastic games, using a large number of sampled realizations of the original chance-constrained stochastic game, and efficiently solve a consensus safe Nash equilibrium across different realizations via a new Alternating Direction Method of Multipliers (ADMM) algorithm. This chapter is based on the published work [184], co-authored with Chih-Yuan Chiu, Lasse Peters, Fernando Palafox, Mustafa Karabag, Somayeh Sojoudi, Claire J. Tomlin, and David Fridovich-Keil.

6.1 Background

Stochastic game theory [278] provides a principled mathematical foundation for modeling interactions between multiple self-interested players in uncertain environments, and has applications in traffic control [79], multi-robot coordination [319], and human-robot interaction [189]. In this framework, each player selects actions to optimize their own objective, obey a set of constraints, and reason about the strategic response of other players. Uncertainty in the players' potentially conflicting objectives and coupled constraints makes these problems extremely challenging to solve.

Classical results in stochastic games are often derived under strong assumptions regarding the problem structure and the distribution of the underlying random process. In the class of linear quadratic Gaussian games, necessary and sufficient conditions for the existence of Nash equilibria are characterized in [24]. It is also shown that in N-player noncooperative stochastic games, the convexity of player-specific objectives and convex, compact strategy sets are sufficient for the existence of the Nash equilibria [175]. However, for general stochastic games, it is NP-hard to determine the existence of Nash equilibria [66]. Moreover, computing a Nash equilibrium can also

be a hard problem [70], partially due to the complexity of solving the nonlinear equations induced by the Nash equilibrium condition.

Several recent efforts provide computationally efficient, approximate solutions to stochastic games with coupled constraints. Two lines of work provide high probability guarantees of both optimality and feasibility. The first [315, 316, 247] approximates the players' expected value objectives and constraints with sample average approximations. The second [242, 94, 88] follows the idea of scenario programming and approximates the objectives and constraints using worst-case samples. The former approach enjoys low sample complexity under certain distributional assumptions [140]. However, when the sample size is finite, this method may lead to situations in which the optimal solution is infeasible for the original chance constraint. The latter technique does not require strong distributional assumptions and returns conservative feasible solutions with high probability, but may require a large number of samples. In this chapter, we combine the benefits of the two approaches such that we obtain accurate approximations for the objectives and maintain the high probability feasibility guarantee.

Our contributions are threefold: (1) We first propose a new sample-based approximation to the constrained stochastic game problem. In this framework, we approximate the expected objectives using a sample average approximation and ensure the feasibility of the original chance constraints by considering a large number of sampled constraints. We validate this scenario-game approximation by characterizing its sample complexity, and we show how the sample complexity can be improved by using problem structure. (2) To overcome the computational burden induced by the sampled constraints, we decompose the approximated game into smaller games with few constraints per scenario, and propose a decentralized ADMM algorithm to compute the joint Nash equilibrium solution in parallel. (3) We prove the convergence of our method to a generalized Nash equilibrium of the approximated constrained game. Empirical results show that our method can handle a large number of constraints with faster convergence than a state-of-the-art baseline.

6.2 Related Work

Stochastic Games

Originally due to Shapley [278], the field of stochastic game theory has expanded to model uncertainties in players' objectives [123], constraints [48], and in the case of dynamic games, underlying state dynamics [97, 244, 239]. Exact generalized Nash equilibrium solutions to stochastic constrained games can be obtained by solving their equivalent stochastic variational inequality problems [91, 247]. Under an appropriate constraint qualification, the well-known Karush–Kuhn–Tucker (KKT) conditions must be satisfied for all players at a generalized Nash equilibrium [91, 167]. We focus on games with monotone objective pseudogradients [246] and convex constraints, where solutions can be found in polynomial time [156].

ADMM for Games

We are ultimately interested in decentralized methods [54, 276] for identifying generalized Nash equilibria, because they can often exploit computational parallelism for efficiency gains compared with centralized method. In particular, the alternating direction method of multipliers (ADMM) [105, 36] is an appealing approach for efficient decentralized computation. The ADMM enjoys convergence guarantees for convex problems [110, 35], convex-concave saddle point problems [37, 152] and monotone variational inequality problems [125, 322]. Recent work [322] has adopted an interior-point method to ensure constraint feasibility, thereby outperforming projection-based consensus ADMM methods [159, 232, 73].

Our algorithm differs from prior work [34, 269, 169] in that we decompose the objective and constraints *over scenarios*. For each scenario, we solve an N-player game with relatively few constraints, and then synchronize across scenarios via ADMM. Unlike prior methods, we do not require constraint projection or an interior-point method in the consensus step. Moreover, we can handle nonlinear coupled constraints, while prior works [191, 246] consider affine constraints.

Approximation Methods for Stochastic Optimization

The sample average approximation (SAA) method [300] is a well-known technique for solving stochastic optimization problems via Monte Carlo simulation [131]. This method approximates the objectives and constraints of the original problem using sample averages, and has been shown to be able to recover original optimal solutions, as the sample size grows to infinity [315, 316, 247]. Another approach for approximating the stochastic optimization problem is the scenario optimization approach [72, 43], where the original chance constraints are replaced with a large number of sampled constraints [41]. This method has been extensively studied, and subsequent work has characterized its sample complexity and feasibility guarantees [44]. Moreover, it is recently extended to constrained variational inequality problems [242]. Our approach approximates the expected value objective by a sample average, and replaces the chance constraint with a large number of sampled constraints.

6.3 Preliminaries

We begin by introducing a deterministic, general-sum static game played among N players. Concretely, each player i (**P**i) seeks to solve a problem of the form:

$$x_i^* \in \arg\min_{x_i} f_i(\mathbf{x}) \tag{6.1a}$$

$$\text{s.t. } h_i(\mathbf{x}) \leq 0\,, \tag{6.1b}$$

where $x_i \in \mathcal{X}_i \subseteq \mathbb{R}^n$, for each $i \in [N] := \{1, 2, \ldots, N\}$, $\mathcal{X}_i$ is the domain of x_i and $\mathbf{x} := (x_1, \ldots, x_N) \in \mathbb{R}^{Nn}$. Let the joint decision space be denoted by $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, and let each player i's constraint be denoted by $h_i(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^\ell$. Observe that players' problems are *coupled*,

both in the objectives and the constraints. We are interested in finding unilaterally optimal strategies for all players in this setting, i.e., the generalized Nash equilibria.

Definition 2 ([90]). *A point $\mathbf{x}^* \in \mathbb{R}^{Nn}$ is a generalized Nash equilibrium (GNE) if for all $i \in [N]$, $h_i(\mathbf{x}^*) \leq 0$, and $f_i(x_i, \mathbf{x}^*_{-i}) \geq f_i(\mathbf{x}^*)$, for each x_i satisfying $h_i(x_i, \mathbf{x}^*_{-i}) \leq 0$.*

6.4 Scenario Game Problem

In this chapter, we focus our attention on constrained *stochastic* general-sum games, in which both the objective and constraints are subject to uncertainty and parameterized by the random vector θ, i.e. $f_i(\mathbf{x}; \theta)$ and $h_i(\mathbf{x}; \theta)$. Let the random vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^d$ be drawn from a probability distribution p_θ that is unknown to all players. We denote player i's decision problem as:

$$x_i^* \in \arg\min_{x_i} \mathbb{E}\left[f_i(\mathbf{x}; \theta)\right] \tag{6.2a}$$

$$\text{s.t. } \mathbb{P}_\theta\big(h_i(\mathbf{x}; \theta) \leq 0\big) \geq 1 - \epsilon\,. \tag{6.2b}$$

Note that we have replaced $\mathbf{P}i$'s objective with its expectation under distribution p_θ, and likewise we have replaced the deterministic constraint $h_i(\mathbf{x}; \theta) \leq 0$ with the chance constraint $\mathbb{P}_\theta\left(h_i(\mathbf{x}; \theta) \leq 0\right) \geq 1 - \epsilon$, with $\epsilon \in (0, 1)$ as the probability of failure. In full generality—i.e., without making further assumptions about the distribution p_θ, such as normality—it is intractable to find a generalized Nash equilibrium for Eq. (6.2). In the sequel, we will construct a sampled approximation to Eq. (6.2) which is amenable to both theoretical complexity analysis and efficient, parallel implementation.

Drawing upon ideas developed in the stochastic optimization [72, 41] and model predictive control [43, 42, 45] communities, we approximate the stochastic game Eq. (6.2) with the following deterministic problem:

$$x_i^* \in \arg\min_{x_i} \frac{1}{S} \sum_{j=1}^{S} f_i(\mathbf{x}; \theta^j) \tag{6.3a}$$

$$\text{s.t. } h_i(\mathbf{x}; \theta^j) \leq 0,\ \forall j \in \{1, \ldots, S\}\,, \tag{6.3b}$$

in which each so-called *scenario* θ^j is sampled independently from the probability distribution p_θ. In Eq. (6.3), we have replaced the expected value of the objective from Eq. (6.2) with its empirical mean, and enforced the original constraint in Eq. (6.1) for all of the scenarios $\{\theta^j\}_{j=1}^S$. We propose to compute the generalized Nash equilibrium of (6.3), which always exists if the following assumption holds true [101].

Assumption 3. *For each player $i \in [N]$, the constraint $h_i(\mathbf{x}; \theta)$ is convex in $\mathbf{x}$ and satisfies Slater's condition [35]. The objective function of each player is upper bounded, i.e.*

$$\sup_{\theta \in \Theta, \mathbf{x} \in \mathcal{X}} \|f_i(\mathbf{x}; \theta)\|_\infty \leq D, \tag{6.4}$$

for some finite $D \in \mathbb{R}$. The pseudogradient $F(\mathbf{x};\theta) := [\nabla_{x_i} f_i(\mathbf{x};\theta)]_{i=1}^N$, where $\nabla_{x_i} f_i(\mathbf{x};\theta)$ denotes the gradient of f_i with respect to x_i, is a continuous and monotone operator of $\mathbf{x}$, i.e., $(\mathbf{x}-\mathbf{y})^\top (F(\mathbf{x};\theta) - F(\mathbf{y};\theta)) \geq 0, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{Nn}$. □

Assumption 3 implies that the objective of each player is convex with respect to its own decision variable, a standard assumption in variational inequality problems [322]. It is shown in [148] that a convex-concave saddle point problem can be reformulated to satisfy Assumption 3. Note also that Assumption 3 allows *non*convex objectives for each player. An example is a two-player game, with the objectives $f_1(x_1, x_2) = x_1 - x_2^2$ and $f_2(x_1, x_2) = x_2 - x_1^2$.

Running Example: We consider a simplified spacecraft rendezvous problem, where two spacecraft approach each other at a predefined rendezvous point in space. We model this problem as a two player general-sum game with a planning horizon T. At time $t \in \{0, 1, \ldots, T\}$, each spacecraft has a state vector $\xi_i(t) = [\xi_i^x(t), \xi_i^{v_x}(t), \xi_i^y(t), \xi_i^{v_y}(t)] \in \mathbb{R}^4$, where $[\xi_i^x(t), \xi_i^y(t)]$ is the position of the spacecraft in the rendezvous hyperplane and $[\xi_i^{v_x}, \xi_i^{v_y}]$ is the velocity vector. It also has a control vector $u_i(t) = [u_i^x(t), u_i^y(t)] \in \mathbb{R}^2$ representing the x- and y-axis acceleration. The dynamics of each spacecraft is approximated as a double integrator for simplicity [122],

$$\xi_i(t+1) = \underbrace{\begin{bmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{A} \xi_i(t) + \underbrace{\begin{bmatrix} \frac{1}{2}\Delta t^2 & 0 \\ \Delta t & 0 \\ 0 & \frac{1}{2}\Delta t^2 \\ 0 & \Delta t \end{bmatrix}}_{B} u_i(t) \tag{6.5}$$

where $\Delta t > 0$ is the time discretization constant. We assume the initial state $\xi_i(0)$ is drawn from a known distribution p_ξ. We concatenate all the random parameters into a vector $\theta \in \mathbb{R}^d$, and assume it follows a distribution $\theta \sim p_\theta$. As such, the general-sum game that each player considers can be summarized as follows,

$$\begin{aligned} \min_{\{u_i(t)\}_{t=0}^{T-1}} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}_\theta \left[\frac{1}{2} \xi_i(t)^\top Q_i^\theta \xi_i(t) + \frac{1}{2} u_i(t)^\top u_i(t) \right] \\ \text{s.t. } \mathbb{P}_\theta \Big(\begin{bmatrix} \xi_1^x(t) - \xi_2^x(t) \\ \xi_1^y(t) - \xi_2^y(t) \end{bmatrix} - b_i^\theta \leq 0, \|u_i(t)\|_\infty \leq 1, \\ \left\| \begin{bmatrix} \xi_1^x(t) - \xi_2^x(t) \\ \xi_1^y(t) - \xi_2^y(t) \end{bmatrix} \right\|_2^2 \leq 1, \forall t \in [T] \Big) \geq 0.95 \end{aligned} \tag{6.6}$$

where $\xi_i(t+1) = A\xi_i(t) + Bu_i(t)$, $\forall i \in \{1, 2\}$, $\forall t \in \{0, 1, \ldots, T-1\}$, and b_i^θ parameterizes an inequality chance constraint ensuring no hard contact between two spacecraft with high probability. Note that each player's feasible set depends upon the decisions of the other player. Hence, this is a *generalized* Nash equilibrium problem. In the following sections, we will discuss how many samples are required such that we can approximate (6.6) well using (6.3), and develop an efficient method for computing a generalized Nash equilibrium of the sample-approximated game.

6.5 Sample Complexity of Scenario Games

One of the appealing aspects of scenario programming [44, 242] is its generality with respect to the distribution of parameter vector θ. Indeed, one can establish probabilistic guarantees on the feasibility of the original chance constraint without strong assumptions that p_θ be, e.g. sub-Gaussian or sub-exponential. We extend this result to the scenario game problem, and characterize sample complexity as follows:

Proposition 8. *Consider $\epsilon, \delta \in (0,1)$ and $\tilde{\epsilon} > 0$. Let $\{\theta^j\}_{j=1}^S$ be i.i.d. samples of the random variable $\theta \sim p_\theta$. Let S be the sample size. Define $\mathcal{H}_S := \{\mathbf{x} \in \mathbb{R}^{Nn} : h_i(\mathbf{x}; \theta^j) \leq 0, \forall i \in [N], j \in [S]\}$. Suppose that $\mathcal{H}_S$ is non-empty, then under Assumption 3, the following statements hold true simultaneously for each player $i \in [N]$,*

1. $\|\frac{1}{S}\sum_{j=1}^S f_i(\mathbf{x}; \theta^j) - \mathbb{E}_\theta[f_i(\mathbf{x}; \theta)]\| \leq \tilde{\epsilon}$, *for all* $x \in \mathcal{H}_S$
2. $\mathbb{P}_\theta(h_i(\mathbf{x}; \theta) \leq 0) \geq 1 - \epsilon$, *for all* $x \in \mathcal{H}_S$

with probability at least $1 - \delta$, where $\delta := 2Ne^{-\frac{S\tilde{\epsilon}^2}{4D^2}} + \sum_{\ell=0}^{Nn-1} \binom{S}{\ell} \epsilon^\ell (1-\epsilon)^{S-\ell}$.

Proof. The proof can be found in the Appendix. □

Note that we have not made strong assumptions on the distribution p_θ; the bound can be improved if more prior knowledge about the problem structure and distribution p_θ is available. For example, if each player's constraint $h_i(\mathbf{x}; \theta) \leq 0$ only depends on its own decision variable x_i, then the constraint $h_i(\mathbf{x}; \theta)$ can be simplified as $h_i(x_i; \theta) \leq 0$, where the decision variable $x_i \in \mathbb{R}^n$ has a lower dimension than the original decision variable $\mathbf{x} \in \mathbb{R}^{Nn}$. This dimension reduction simplifies the sample complexity for approximating each constraint. By combining this simplification with the union bound, we can improve the sample complexity result of Proposition 8, as shown in the following result.

Proposition 9. *Under the same assumptions of Proposition 8, suppose that $\mathcal{H}_S$ is non-empty and each player's constraint $h_i(\mathbf{x}; \theta^j) \leq 0$ only depends on x_i, $\forall j \in [S]$. Then, the following statements hold true simultaneously for each player $i \in [N]$,*

1. $\|\frac{1}{S}\sum_{j=1}^S f_i(\mathbf{x}; \theta^j) - \mathbb{E}_\theta[f_i(\mathbf{x}; \theta)]\| \leq \tilde{\epsilon}$, *for all* $x \in \mathcal{H}_S$
2. $\mathbb{P}_\theta(h_i(\mathbf{x}; \theta) \leq 0) \geq 1 - \epsilon$, *for all* $x \in \mathcal{H}_S$

with probability at least $1 - \delta$, where $\delta := 2Ne^{-\frac{S\tilde{\epsilon}^2}{4D^2}} + N\sum_{\ell=0}^{n-1} \binom{S}{\ell} \epsilon^\ell (1-\epsilon)^{S-\ell}$.

Proof. The proof can be found in the Appendix. □

The above characterization of sample complexity suggests that a sufficient number of samples leads to an accurate estimation of the objective and ensures the feasibility of the chance constraint with high probability. However, solving a constrained game with a large number of sampled constraints presents a significant computational challenge. This motivates the following algorithmic development.

Algorithm 4: Scenario-Game ADMM (SG-ADMM)

Input: Initialization $\{\mathbf{w}(0), \mathbf{x}(0), \boldsymbol{\lambda}(0)\}$, convergence tolerance $\epsilon > 0$.

for $k = 0, 1, 2, \cdots$ **do**

  for *scenarios* $j = 1, \ldots, S$ *in parallel* **do**

    $w_i^j(k+1) \leftarrow \arg\min_{w_i^j} \mathcal{L}_i^j(\mathbf{w}^j, \mathbf{x}(k), \boldsymbol{\lambda}_i(k))$

    s.t. $h_i(\mathbf{w}^j; \theta^j) \leq 0$

  Update $\{x_i(k+1)\}_{i=1}^N$: $\forall i \in [N]$, $x_i(k+1) \leftarrow \frac{1}{S}\sum_{j=1}^S (\frac{1}{\rho}\lambda_i^j(k) + w_i^j(k+1))$

  Update $\{\lambda_i^j(k+1)\}_{i=1,j=1}^{N,S}$: $\forall i \in [N]$, $j \in [S]$, $\lambda_i^j(k+1) \leftarrow \lambda_i^j(k) + \rho(w_i^j(k+1) - x_i(k+1))$

  If $\|\mathbf{w}(k+1) - M\mathbf{x}(k)\|^2 \leq \epsilon$, **return** $\{x_i(k+1)\}_{i=1}^N$

6.6 Scenario Games via Decentralized ADMM

Decentralized ADMM

In the scenario game Eq. (6.3), both the objective and constraints involve significantly more terms than in Eq. (6.1). When S is large, therefore, it can be computationally demanding to find a generalized Nash equilibrium. Therefore, we propose the following splitting method to enable parallel—and hence more efficient—computation of equilibrium solutions. This technique is an analog of the well-known ADMM algorithm tailored to generalized Nash equilibrium problems, and is summarized in Algorithm 4.

In order to develop this technique, we shall begin by introducing auxiliary decision variables $\{w_i^j\}_{j=1}^S$ for each player $\mathbf{P}i$, and employing the shorthand $\mathbf{w}_i := (w_i^1, \ldots, w_i^S)$ for the decision variables of player i across scenarios $j = 1$ to $j = S$, $\mathbf{w}^j := (w_1^j, \ldots, w_N^j)$ for the decision variables of players $i = 1$ to $i = N$ in the jth scenario, and $\mathbf{w} := (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ for all the decision variables. We will later use the same shorthand $(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}^j, \boldsymbol{\lambda})$ for Lagrange multipliers for the constraints Eq. (6.7c):

$$x_i^*, \mathbf{w}_i^* \in \arg\min_{x_i, \mathbf{w}_i} \frac{1}{S}\sum_{j=1}^S f_i(\mathbf{w}^j; \theta^j) \tag{6.7a}$$

$$\text{s.t. } h_i(\mathbf{w}^j; \theta^j) \leq 0, \ \forall j \in \{1, \ldots, S\} \tag{6.7b}$$

$$w_i^j - x_i = 0, \quad \forall j \in \{1, \ldots, S\}. \tag{6.7c}$$

In Eq. (6.7), $\mathbf{P}i$ evaluates its objective and constraints for scenario j using only the auxiliary variables $\mathbf{w}^j$. However, in the end, each player must select a single decision variable; hence, we also enforce the consensus constraints in Eq. (6.7c). These constraints effectively *couple* S games which would otherwise be entirely independent. To facilitate such a decomposition, we construct a *partial* augmented Lagrangian for each player, in which only Eq. (6.7c) have been dualized:

$$\mathcal{L}_i(\mathbf{w}, \mathbf{x}, \boldsymbol{\lambda}_i) := \sum_{j=1}^S \mathcal{L}_i^j(\mathbf{w}^j, \mathbf{x}, \boldsymbol{\lambda}_i), \tag{6.8}$$

$$\mathcal{L}_i^j(\mathbf{w}^j, \mathbf{x}, \boldsymbol{\lambda}_i) := \frac{f_i(\mathbf{w}^j; \theta^j)}{S} + \lambda_i^{j\top} \delta_i^j + \frac{\rho}{2} \|\delta_i^j\|_2^2 .$$

Here, $\delta_i^j := w_i^j - x_i$, and λ_i^j may be interpreted as an estimate of the Lagrange multiplier corresponding to the j^{th} instance of Eq. (6.7c) in $\mathbf{P}i$'s problem. Thus equipped, we develop the key steps of Algorithm 4, a decentralized technique for solving Eq. (6.3) via Eq. (6.7). To do so, we re-express Eq. (6.7) in terms of the augmented Lagrangians (6.8):

$$x_i^*, \mathbf{w}_i^* \in \arg\min_{x_i, \mathbf{w}_i} \mathcal{L}_i(\mathbf{w}, \mathbf{x}, \boldsymbol{\lambda}_i) \tag{6.9a}$$

$$\text{s.t. } h_i(\mathbf{w}^j; \theta^j) \leq 0, \forall j \in \{1, \ldots, S\} . \tag{6.9b}$$

Solving for auxiliary variable, $\mathbf{w}$

Holding $\mathbf{x}$ and $\boldsymbol{\lambda}$ constant, each player's problem Eq. (6.9) is convex in the decision variable $\mathbf{w}_i$ due to Assumption 3. Thus, we can be assured that any point $\mathbf{w}_i^*$ which satisfies the KKT conditions for all players simultaneously is a generalized Nash equilibrium. Such a point may be identified by, e.g., reformulating the joint KKT conditions as a mixed complementarity program (MCP) [96] and invoking a standard solution method, e.g. PATH [74].

Remark 12. *This equilibrium problem may be separated into S independent problems, involving distinct variables $\{\mathbf{w}^j\}_{j=1}^S$, objectives, and constraints. Consequently, if parallel computation is available, these games may be solved in separate computational threads or on separate computer processors; therefore, Algorithm 4 may still operate efficiently and converge even when many scenarios are required, as shown in Theorem 9.*

Solving for consensus variables, $\mathbf{x}$

Holding $\mathbf{w}$ and $\boldsymbol{\lambda}$ fixed, player i's problem Eq. (6.9) may be simplified to take the following form:

$$x_i = \arg\min_{\tilde{x}_i} \sum_{j=1}^{S} \left(\lambda_i^{j\top} (w_i^j - \tilde{x}_i) + \frac{\rho}{2} \|w_i^j - \tilde{x}_i\|_2^2 \right) . \tag{6.10}$$

Because $\rho > 0$, we readily identify the global solution to (6.10) for each player as

$$x_i \leftarrow \frac{1}{S} \sum_{j=1}^{S} \left(\frac{1}{\rho} \lambda_i^j + w_i^j \right) . \tag{6.11}$$

Updating dual variables, $\boldsymbol{\lambda}$

In order to choose new values of the dual variables which account for the solutions to the previous subproblems, we first examine player i's vanishing gradient condition. We find:

$$0 = \partial_{w_i^j} (\mathcal{L}_i^j(\mathbf{w}^j, \mathbf{x}, \boldsymbol{\lambda}_i) + \mathbb{I}_{h_i(\mathbf{w}^j; \theta^j)}(\mathbf{w}^j)) \tag{6.12}$$

$$= \frac{\nabla_{w_i^j} f_i(\mathbf{w}^j; \theta^j)}{S} + \partial_{w_i^j} \mathbb{I}_{h_i(\mathbf{w}^j;\theta^j)}(\mathbf{w}^j) + \lambda_i^j + \rho(w_i^j - x_i),$$

where $\mathbb{I}_{h(\mathbf{x};\theta)}(\mathbf{x}) : \mathbb{R}^{Nn} \to \{0, \infty\}$ and $\mathbb{I}_{h(\mathbf{x};\theta)}(\mathbf{x}) = 0$ if and only if $h(\mathbf{x}; \theta) \leq 0$. Following well-established reasoning for augmented Lagrangian methods [238, Ch. 17], we recognize the latter two terms as the (unique) value of the Lagrange multiplier for the original constraint Eq. (6.7c) which satisfies the vanishing gradient optimality condition. Therefore, we set:

$$\lambda_i^j \leftarrow \lambda_i^j + \rho(w_i^j - x_i)\,. \tag{6.13}$$

The above update rule is formalized in Algorithm 4.

Convergence of Scenario-Game ADMM

In this section, we first characterize the optimality condition of the general-sum game problem. We then show that the special structure of the consensus constraint allows us to measure convergence by monitoring the residual of the consensus constraint. Building upon this result, we prove the convergence of Algorithm 4.

Similar to standard, single-objective optimization problems, under an appropriate constraint qualification the KKT conditions must be satisfied at solutions to the variational inequality problem [91]. From the KKT conditions, an optimal solution $\mathbf{z}^* := (\mathbf{w}^*, \mathbf{x}^*, \boldsymbol{\lambda}^*)$ should satisfy the following conditions,

$$\begin{cases} (p - w_i^{j*})^\top (\nabla_{w_i^{j*}} f_i(\mathbf{w}^{j*}; \theta^j) + \partial_{w_i^{j*}} \mathbb{I}_{h_i(\mathbf{w}^{j*};\theta^j)}(\mathbf{w}^{j*}) \\ \quad + \boldsymbol{\lambda}^*) \geq 0, \forall p \in \mathbb{R}^n, \forall i \in [N], j \in [S] \\ \mathbf{w}^* - M\mathbf{x} = 0 \\ h_i(\mathbf{w}^{j*}; \theta^j) \leq 0, \forall i \in [N], \forall j \in [S] \end{cases} \tag{6.14}$$

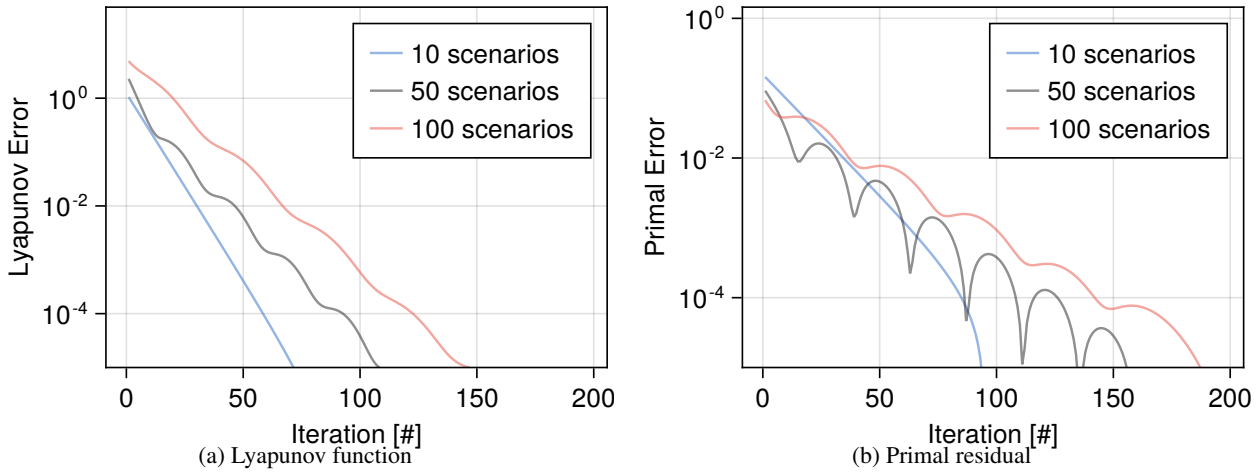
where we represent the consensus constraint (6.7c) compactly as $\mathbf{w} - M\mathbf{x} = 0$ by introducing a constant matrix $M := \mathbf{1}_N \otimes I_n$. Let $F(\mathbf{w}) := [\nabla_{w_i^j} f_i(\mathbf{w}^j; \theta^j)]_{i=1,j=1}^{N,S}$ and $H(\mathbf{w}) := [\partial_{w_i^j} \mathbb{I}_{h_i(\mathbf{w}^j;\theta^j)}(\mathbf{w}^j)]_{i=1,j=1}^{N,S}$. We can also represent the above optimality condition as the variational inequality problem:

$$(\mathbf{z} - \mathbf{z}^*)^\top Q(\mathbf{z}^*) \geq 0, \forall \mathbf{z} \in \mathbb{R}^{SNn} \times \mathbb{R}^{Nn} \times \mathbb{R}^{Nm}, \tag{6.15}$$

$$\mathbf{z} = \begin{bmatrix} \mathbf{w} \\ \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix}, Q(z) = \begin{bmatrix} F(\mathbf{w}) + H(\mathbf{w}) + \boldsymbol{\lambda} \\ -M^\top \boldsymbol{\lambda} \\ \mathbf{w} - M\mathbf{x} \end{bmatrix}. \tag{6.16}$$

Observing that the M matrix in the consensus constraint has full column rank, we see that it must have trivial null space. Consequently, we can show in the following lemma that an optimal solution is reached when the consensus constraint residual is zero.

Lemma 3. *Suppose* $\mathbf{w}(k+1) - M\mathbf{x}(k) = 0$*, then* $(\mathbf{w}(k+1), \mathbf{x}(k+1), \boldsymbol{\lambda}(k+1))$ *is an optimal solution to the VI problem (6.15).*



(a) Lyapunov function (b) Primal residual

Figure 6.1: The convergence of Scenario-Game ADMM under different numbers of sampled scenarios in running example (6.6). With only 10 samples, we have no binding constraint, and we converge exponentially fast. With 50 and 100 samples, we suffer binding constraints, and the primal residual $\rho\|M(\mathbf{x}(k)-\mathbf{x}^*)\|^2$ oscillates. However, the Lyapunov function, which is defined as the sum of primal residual and dual residual $\frac{1}{\rho}\|\boldsymbol{\lambda}(k)-\boldsymbol{\lambda}^*\|^2$, decays monotonically.

Proof. The proof can be found in the Appendix. □

Building upon this result, we show in the following theorem that a Lyapunov function, defined by the Lagrange multiplier error and the consensus constraint's residual, is monotonically decreasing with each iteration of Algorithm 3.

Theorem 9. *Under Assumption 3, let* $\mathbf{z}^* = (\mathbf{w}^*, \mathbf{x}^*, \boldsymbol{\lambda}^*)$ *be an optimal solution of (6.15). Define* $V(k) := (1/\rho)\|\boldsymbol{\lambda}(k) - \boldsymbol{\lambda}^*\|^2 + \rho\|M(\mathbf{x}(k) - \mathbf{x}^*)\|^2$. *We have*

$$V(k+1) \leq V(k) - \rho\|\mathbf{w}(k+1) - M\mathbf{x}(k)\|^2, \tag{6.17}$$

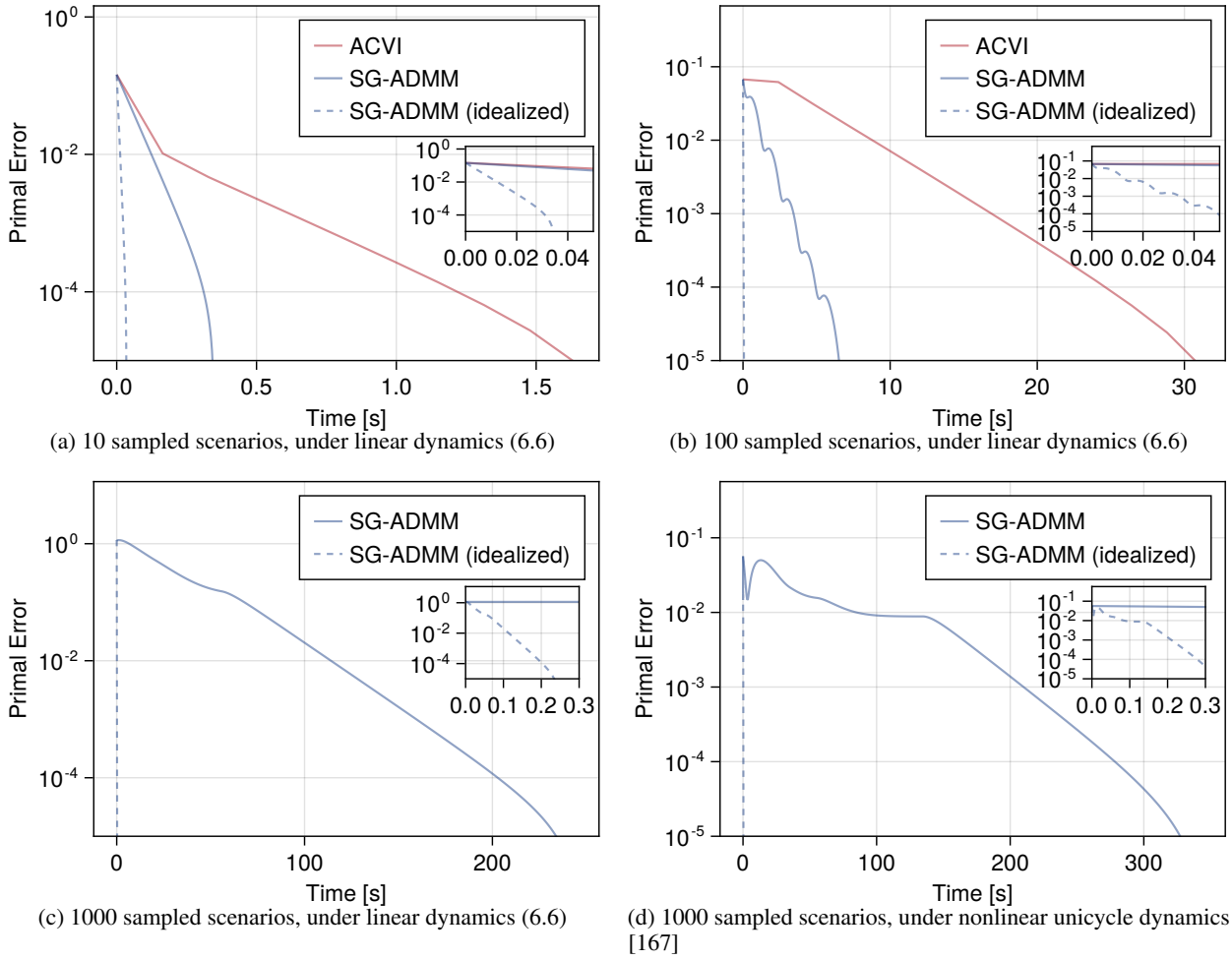
and $\lim_{k\to\infty} V(k) = 0$.

Proof. The proof can be found in the Appendix. □

Theorem 9 establishes the asymptotic convergence of Algorithm 3, by showing that $V(k) \to 0$ as $k \to \infty$; thus, for any convergence tolerance $\epsilon > 0$, there exists some sufficiently large $k > 0$ such that $\|\mathbf{w}(k+1) - M\mathbf{x}(k)\|^2 \leq \epsilon$. When the players' objectives satisfy the following assumption, we can strengthen the convergence result in Theorem 10.

Assumption 4 ([246])**.** *For each player* $i \in [N]$, *the objective* $f_i(\mathbf{x}; \theta)$ *is differentiable. The function* $F(\mathbf{w}) = [\nabla_{w_i^j} f_i(\mathbf{w}^j; \theta^j)]_{i=1,j=1}^{N,S}$ *is L-Lipschitz continuous and is an* m*-strongly monotone operator, i.e.,* $(\mathbf{w} - \tilde{\mathbf{w}})^\top (F(\mathbf{w}) - F(\tilde{\mathbf{w}})) \geq m\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2, \forall \mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^{SNn}$.

Theorem 10. *Under Assumptions 3 and 4, let* $\mathbf{z}^* = (\mathbf{w}^*, \mathbf{x}^*, \boldsymbol{\lambda}^*)$ *be an optimal solution of* (6.15). *Define* $V(k)$ *as in Theorem 9. At the* k*-th iteration, we have:*



(a) 10 sampled scenarios, under linear dynamics (6.6)

(b) 100 sampled scenarios, under linear dynamics (6.6)

(c) 1000 sampled scenarios, under linear dynamics (6.6)

(d) 1000 sampled scenarios, under nonlinear unicycle dynamics [167]

Figure 6.2: Comparison of the CPU time under different numbers of sampled scenarios. The solid blue curves represent the implementation of Scenario-Game ADMM which solves step 4 of Algorithm 4 sequentially, i.e., one scenario by one scenario. The dashed blue curves represent the expected computation time when we implement step 4 of Algorithm 4 in parallel. This expected computation time is derived by dividing the computation time of the blue solid curves by the number of scenarios. In both cases, Scenario-Game ADMM converges faster than ACVI. For each sampled scenario, we have 20 dimensional decision variables, and 35 constraints. When the number of sampled scenarios is 1000, there are $1000 \times 35 = 35000$ constraints. ACVI fails to compile due to the scale of problem. With 1000 samples, our algorithm converges even when we replace the linear dynamics in (6.6) with the nonlinear unicycle dynamics in [167], as shown in Fig.6.2d.

1. *If there is no binding constraint at* $\mathbf{w}(k)$*, i.e.,* $h_i(\mathbf{x}(k); \theta^j) < 0$*,* $\forall i \in [N]$*,* $\forall j \in [S]$*, then:*

$$V(k) \leq \left(1 - \frac{1}{2\kappa_f^{0.5+|\epsilon|}}\right) V(k-1)$$

where $\kappa_f = L/m$ *and* $\epsilon = \log_{\kappa_f}(\rho/\sqrt{mL})$*;*

2. *Otherwise,* $V(k) \leq V(k-1) - \rho\|\mathbf{w}(k) - M\mathbf{x}(k-1)\|^2$.

Proof. The proof can be found in the Appendix. □

6.7 Experiments

In this section, we continue the running example (6.6). We characterize the sample complexity and the empirical performance of the Scenario-Game ADMM. The details of the experiment parameters are included in the Appendix.

By Proposition 8, if the sample size is $S = 1000$, then for each player $i \in \{1, 2\}$,

$$\mathbb{P}(\|\frac{1}{S}\sum_{j=1}^{S} f_i(\mathbf{x}; \theta^j) - \mathbb{E}[f_i(\mathbf{x})]\| \leq 0.5) \geq 1 - 4.0 \times 10^{-3}, \tag{6.18}$$

and

$$\mathbb{P}(\mathbb{P}(h_i(\mathbf{x}; \theta) \leq 0) \geq 0.95) \geq 1 - 2.9 \times 10^{-7}. \tag{6.19}$$

Therefore, by having 1000 sampled scenarios, we are able to obtain a reasonable approximation (6.3) of the stochastic game problem (6.6).

We proceed to apply Scenario-Game ADMM to solve the sample-approximated game problem (6.3). We first validate the convergence of Scenario-Game ADMM in Fig. 6.1. As proven in Theorem 9, the Lyapunov function decays monotonically in Fig. 6.1a. Note that the primal residual $\rho\|M(\mathbf{x}(k) - \mathbf{x}^*)\|^2$ may still oscillate due to the existence of binding constraints, as shown in Theorem 10 and Fig. 6.1b.

We then compare the performance of Scenario-Game ADMM with the baseline method. Since prior works [191, 246] do not consider coupled nonlinear constraints among players, we compare Scenario-Game ADMM with the state-of-the-art ADMM-based constrained variational inequality solver (ACVI) [322]. As shown in Fig. 6.2, Scenario-Game ADMM converges faster than ACVI across different scenario sizes. In particular, when we have 1000 sampled scenarios, Scenario-Game ADMM converges, but ACVI fails to compile due to the scale of the problem, where we have 35000 coupled inequality constraints in total. This experiment suggests that Scenario-Game ADMM can solve game problems with a large number of constraints within a reasonable amount of time.

As an additional ablation, we also compare our method's computation time to the centralized PATH solver that our method uses at the inner loop [74]; c.f. appendix. While PATH is competitive, in particular for small-scale problems, we observe that the parallelized version of our method is still more than 2x faster for scenario sizes $S \in [10, 100]$. Finally, as with ACVI, the scenario-number-dependent compilation overhead of this centralized approach precludes application to larger problems.

6.8 Conclusion and Future Work

In this chapter, we introduced a new sample-based approximation for stochastic games. We characterized the sample complexity and the feasibility guarantees of this approximation scheme.

We proposed a decentralized ADMM solver and characterized its convergence. We empirically validated the performance of this algorithm in a stochastic game with a large number of sampled constraints. Future work should extend our results on sample complexity and analyze how well equilibria of the scenario game approximate solutions to the original chance-constrained stochastic game.

Acknowledgements For This Chapter

This work is supported by the DARPA Assured Autonomy and ANSR programs, the NASA ULI program in Safe Aviation Autonomy, and the ONR Basic Research Challenge in Multibody Control Systems. This work is also supported by the National Science Foundation under Grant Nos. 2211548 and 1652113, and the Army Research Laboratory under Cooperative Agreement Numbers W911NF-23-2-0011 and W911NF-20-1-0140.

Proofs. Before we present the proof of Proposition 8, we first introduce the following lemmas.

Lemma 4 (Thm. 3.26, [303])**.** *Let $\{\theta^j\}_{j=1}^S$ be i.i.d. samples from p_θ. Suppose $\exists D$, s.t.*

$$\sup_{\theta\in\Theta,\mathbf{x}\in\mathcal{X}} \|f(\mathbf{x};\theta)\|_2 \leq D < \infty. \tag{6.20}$$

Then, $\mathbb{P}(\|\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{S}\sum_{j=1}^S f(\mathbf{x};\theta^j) - \mathbb{E}_\theta[f(\mathbf{x};\theta)]\|_2 \geq \tilde{\epsilon}) \leq 2e^{(-\frac{S\tilde{\epsilon}^2}{4D^2})}$.

Lemma 5 ([44])**.** *Let $\{\theta^j\}_{j=1}^S$ be a set of i.i.d. samples of the random variable θ. For all $\mathbf{x}\in\mathbb{R}^{Nn}$, we have* $\mathbb{P}\left(\mathbb{P}\left(h(\mathbf{x};\theta)\leq 0\right)\geq\epsilon\right) \leq \sum_{\ell=0}^{Nn-1}\binom{S}{\ell}\epsilon^\ell(1-\epsilon)^{S-\ell}$.

Proof of Proposition 8. By Assumption 3, $\sup_{\theta\in\Theta,\mathbf{x}\in\mathcal{X}} f_i(\mathbf{x};\theta)\leq D$, for some finite $D\in\mathbb{R}$. Let $h(\mathbf{x};\theta) := [h_i(\mathbf{x};\theta)]_{i=1}^N$. By Lemmas 4 and 5 and the union bound, $\mathbb{P}(\sup_{\mathbf{x}\in\mathcal{X}}\|\frac{1}{S}\sum_{j=1}^S f_i(\mathbf{x};\theta^j) - \mathbb{E}_\theta[f_i(\mathbf{x};\theta)]\| \leq \tilde{\epsilon}$ and $\mathbb{P}\left(h(\mathbf{x};\theta)\leq 0\right)\geq 1-\epsilon, \forall i\in[N]) \geq 1 - 2Ne^{(-\frac{S\tilde{\epsilon}^2}{4D^2})} - \sum_{\ell=0}^{Nn-1}\binom{S}{\ell}\epsilon^\ell(1-\epsilon)^{S-\ell}$. □

Proof of Proposition 9. Under the independent constraint assumption, we have

$$\mathbb{P}(\mathbb{P}(h_i(x_i;\theta)\leq 0)\geq\epsilon) \leq \sum_{\ell=0}^{n-1}\binom{S}{\ell}\epsilon^\ell(1-\epsilon)^{S-\ell}. \tag{6.21}$$

Then, by Lemma 4 and the union bound, we have $\mathbb{P}(\sup_{\mathbf{x}\in\mathcal{X}}\|\frac{1}{S}\sum_{j=1}^S f_i(\mathbf{x};\theta^j) - \mathbb{E}_\theta[f_i(\mathbf{x};\theta)]\| \leq \tilde{\epsilon}$ and $\mathbb{P}\left(h_i(\mathbf{x};\theta)\leq 0\right)\geq 1-\epsilon, \forall i\in[N]) \geq 1 - 2Ne^{(-\frac{S\tilde{\epsilon}^2}{4D^2})} - N\sum_{\ell=0}^{n-1}\binom{S}{\ell}\epsilon^\ell(1-\epsilon)^{S-\ell}$. □

Proof of Lemma 3. From Algorithm 4, we have $(\mathbf{w}-\mathbf{w}(k+1))^\top(F(\mathbf{w}(k+1)) + H(\mathbf{w}(k+1)) + \boldsymbol{\lambda}(k)) \geq 0, \forall\mathbf{w}$, $\mathbf{x}(k+1) = M^\dagger(\mathbf{w}(k+1) + (1/\rho)\boldsymbol{\lambda}(k))$, and $\boldsymbol{\lambda}(k+1) = \boldsymbol{\lambda}(k) + \rho(\mathbf{w}(k+1) - M\mathbf{x}(k+1))$. Since M has full column rank, we have the null space of M is $\{0\}$. Also, by definition, $\sum_{j=1}^S\lambda_i^j(k+1) = \sum_{j=1}^S\lambda_i^j(k) + \rho(\sum_{j=1}^S w_i^j(k+1) - Mx_i(k+1)) = \sum_{j=1}^S\lambda_i^j(k) - \sum_{j=1}^S\lambda_i^j(k) =$

0, and therefore $M^\dagger \boldsymbol{\lambda}(k) = 0$. Thus, $\mathbf{x}(k+1) = M^\dagger \mathbf{w}(k+1) = \mathbf{x}(k)$. This implies that $\mathbf{w}(k+1) - M\mathbf{x}(k+1) = 0$, and $\boldsymbol{\lambda}(k+1) = \boldsymbol{\lambda}(k)$. $(\mathbf{w}(k+1), \mathbf{x}(k+1), \boldsymbol{\lambda}(k+1))$ satisfies the optimality condition (6.15). □

Lemma 6. *Let* $(\mathbf{x}^*, \mathbf{w}^*, \boldsymbol{\lambda}^*)$ *be an optimal solution to* (6.14)*, it holds* $(1/\rho)(\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}^*)^\top (\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}(k)) \leq \rho(\mathbf{w}(k+1) - \mathbf{w}^*)^\top (M\mathbf{x}(k) - M\mathbf{x}(k+1))$.

Proof. From the optimality condition (6.15), we have:

$$(\mathbf{w}(k+1) - \mathbf{w}^*)^\top (F(\mathbf{w}^*) + H(\mathbf{w}^*) + \boldsymbol{\lambda}^*) \geq 0 \tag{6.22}$$

By the optimality condition at the k-th iteration, we have $(\mathbf{w}^* - \mathbf{w}(k+1))^\top (F(\mathbf{w}(k+1)) + H(\mathbf{w}(k+1)) + \boldsymbol{\lambda}(k) + \rho(\mathbf{w}(k+1) - M\mathbf{x}(k))) \geq 0$. Substituting $\boldsymbol{\lambda}(k+1) = \boldsymbol{\lambda}(k) + \rho(\mathbf{w}(k+1) - M\mathbf{x}(k+1))$, we derive:

$$\begin{aligned} &(\mathbf{w}^* - \mathbf{w}(k+1))^\top (F(\mathbf{w}(k+1)) + H(\mathbf{w}(k+1)) \\ &+ \boldsymbol{\lambda}(k+1) + \rho(M(\mathbf{x}(k+1) - M\mathbf{x}(k)))) \geq 0 \end{aligned} \tag{6.23}$$

Adding (6.22) and (6.23), and using monotonicity, we have:

$$\begin{aligned} &(\mathbf{w}(k+1) - \mathbf{w}^*)^\top (\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}^*) \\ &\leq \rho(\mathbf{w}(k+1) - \mathbf{w}^*)^\top M(\mathbf{x}(k) - \mathbf{x}(k+1)) \end{aligned} \tag{6.24}$$

Similarly, by the optimality of $\mathbf{x}^*$ and $\mathbf{x}(k+1)$, we have $(\mathbf{x}(k+1) - \mathbf{x}^*)^\top (-M^\top \boldsymbol{\lambda}^*) \geq 0$ and $(\mathbf{x}^* - \mathbf{x}(k+1))^\top (-M^\top \boldsymbol{\lambda}(k+1)) \geq 0$. Adding these two inequalities:

$$(M\mathbf{x}^* - M\mathbf{x}(k+1))^\top (\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}^*) \leq 0 \tag{6.25}$$

Adding (6.24) and (6.25), and using $\mathbf{w}^* - M\mathbf{x}^* = 0$ and $\mathbf{w}(k+1) - M\mathbf{x}(k+1) = \frac{1}{\rho}(\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}(k))$, we have $\frac{1}{\rho}(\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}(k))^\top (\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}^*) \leq \rho(\mathbf{w}(k+1) - \mathbf{w}^*)^\top (M\mathbf{x}(k) - M\mathbf{x}(k+1))$. □

Proof of Theorem 9. Observe that:

$$\begin{aligned} &(1/\rho)\|\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}^*\|^2 + \rho\|M(\mathbf{x}(k+1) - \mathbf{x}^*)\|^2 \\ &= (1/\rho)\|\boldsymbol{\lambda}(k) - \boldsymbol{\lambda}^*\|^2 + \rho\|M(\mathbf{x}(k) - \mathbf{x}^*)\|^2 \\ &\quad - ((1/\rho)\|\boldsymbol{\lambda}(k+1) - \boldsymbol{\lambda}(k)\|^2 + \rho\|M(\mathbf{x}(k+1) - \mathbf{x}(k))\|^2) \\ &\quad + (2/\rho)(\boldsymbol{\lambda}^* - \boldsymbol{\lambda}(k+1))^\top (\boldsymbol{\lambda}(k) - \boldsymbol{\lambda}(k+1)) \\ &\quad + 2\rho(M\mathbf{x}^* - M\mathbf{x}(k+1))^\top (M\mathbf{x}(k) - M\mathbf{x}(k+1)) \end{aligned} \tag{6.26}$$

The last two terms can be bounded as:

$$\begin{aligned} &(2/\rho)(\boldsymbol{\lambda}^* - \boldsymbol{\lambda}(k+1))^\top (\boldsymbol{\lambda}(k) - \boldsymbol{\lambda}(k+1)) \\ &\quad + 2\rho(M\mathbf{x}^* - M\mathbf{x}(k+1))^\top (M\mathbf{x}(k) - M\mathbf{x}(k+1)) \\ &\leq 2\rho(\mathbf{w}(k+1) - \mathbf{w}^*)^\top M(\mathbf{x}(k) - \mathbf{x}(k+1)) \\ &\quad + 2\rho(M\mathbf{x}^* - M\mathbf{x}(k+1))^\top (M\mathbf{x}(k) - M\mathbf{x}(k+1)) \\ &= 2\rho(\mathbf{w}(k+1) + M\mathbf{x}(k+1))^\top (M\mathbf{x}(k) - M\mathbf{x}(k+1)) \\ &= -2(\boldsymbol{\lambda}(k) - \boldsymbol{\lambda}(k+1))^\top (M(\mathbf{x}(k) - \mathbf{x}(k+1))) \end{aligned} \tag{6.27}$$

where the first inequality follows from Lemma 6, and the first equality is derived by substituting $\mathbf{w}^* - M\mathbf{x}^* = 0$. The last equality holds true because of the update rule of $\boldsymbol{\lambda}(k+1)$.

From (6.26) and (6.27), we have $(1/\rho)\|\boldsymbol{\lambda}(k+1)-\boldsymbol{\lambda}^*\|^2+\rho\|M(\mathbf{x}(k+1)-\mathbf{x}^*)\|^2 \leq (1/\rho)\|\boldsymbol{\lambda}(k)-\boldsymbol{\lambda}^*\|^2 + \rho\|M(\mathbf{x}(k)-\mathbf{x}^*)\|^2 - \rho\|\mathbf{w}(k+1) - M\mathbf{x}(k)\|^2$. □

Before we present the proof of Theorem 10, we first introduce a few preliminaries. Define $\hat{f}_i := \frac{1}{\rho} f_i$ and $\hat{g} := \mathbb{I}_{\text{im } M}$, where $\mathbb{I}_{\text{im } M}$ is the $\{0,\infty\}$-indicator function of the image of M. Additionally, we define $s(k) := M\mathbf{x}(k)$, $u(k) := \lambda(k)/\rho$. Let $\beta(k) := [\nabla_{x_i} \hat{f}_i(\mathbf{w}^j(k);\theta^j)]_{i=1,j=1}^{N,S}$ and $\gamma(k) := [\partial_{x_i} \hat{g}(\mathbf{x})]_{i=1,j=1}^{N,S}$.

Lemma 7. *Under Assumption 4, let* $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^{SNn}$, $E = [\nabla_{w_i^j} f_i(\mathbf{w}^j;\theta^j)]_{i=1,j=1}^{N,S}$ *and* $\tilde{E} = [\nabla_{w_i^j} f_i(\tilde{\mathbf{w}}^j;\theta^j)]_{i=1,j=1}^{N,S}$. *We have* $\begin{bmatrix} \mathbf{w}-\tilde{\mathbf{w}} \\ E-\tilde{E} \end{bmatrix}^\top \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} \otimes I_{SNn} \begin{bmatrix} \mathbf{w}-\tilde{\mathbf{w}} \\ E-\tilde{E} \end{bmatrix} \geq 0$.

Proof. Using the co-coercivity of $F(\mathbf{w})$ and the fact that $F(\mathbf{w}) - m\|\mathbf{w}\|_2^2$ is $L-m$ Lipschitz continuous, we have $(m+L)(\mathbf{w}-\tilde{\mathbf{w}})^\top(E-\tilde{E}) \geq mL\|\mathbf{w}-\tilde{\mathbf{w}}\|_2^2 + \|E-\tilde{E}\|_2^2$. We complete the proof by putting it in matrix form. □

Lemma 8. *Suppose there is no binding constraint at* $\mathbf{w}(k+1)$. *Let* $\eta(k) := [s(k), u(k)]$, $v(k) := [\beta(k+1), \gamma(k+1)]$, $y(k) := [\mathbf{w}(k+1), \beta(k+1)]$ *and* $z(k) := [s(k+1), \gamma(k+1)]$. *We consider* $\eta(k)$, $v(k)$ *and* $[y(k), z(k)]$ *as the state, control input and output of a dynamical system. Define the following matrices,* $\hat{A} := \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, $\hat{B} := \left[\begin{smallmatrix} -1 & -1 \\ 0 & -1 \end{smallmatrix}\right]$, $\hat{C}^1 := \left[\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right]$, $\hat{D}^1 := \left[\begin{smallmatrix} -1 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, $\hat{C}^2 := \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, *and* $\hat{D}^2 := \left[\begin{smallmatrix} -1 & -1 \\ 0 & 1 \end{smallmatrix}\right]$. *Then, we have the dynamics* $\begin{bmatrix} \eta(k+1) \\ y(k) \\ z(k) \end{bmatrix} = \begin{bmatrix} \hat{A} & \hat{B} \\ \hat{C}^1 & \hat{D}^1 \\ \hat{C}^2 & \hat{D}^2 \end{bmatrix} \begin{bmatrix} \eta(k) \\ v(k) \end{bmatrix}$.

Proof. By the KKT condition, we have $\nabla_i f_i(\mathbf{w}^j(k+1);\theta^j) + \lambda_i^j(k) + \rho(\mathbf{w}^j(k+1) - Mx_i(k)) = 0$, which is equivalent to

$$\beta(k+1) + u(k) + \mathbf{w}(k+1) - s(k) = 0 \tag{6.28}$$

Subsequently, when we minimize $\mathbf{x}$ with $\mathbf{w}(k+1)$ and $\lambda(k)$ fixed, we have the problem of minimizing $\mathbf{x}$ is equivalent to, for each $i \in [N]$, $\min_{s_i} \partial_i \hat{g}(s_i) + u_i(k)^\top(\mathbf{w}_i(k+1) - s_i) + \frac{\rho}{2}\|\mathbf{w}_i(k+1) - s_i\|^2$: which has the optimality condition:

$$\gamma(k+1) + s(k+1) - \mathbf{w}(k+1) - u(k) = 0. \tag{6.29}$$

Finally, from the update rule of the Lagrange multiplier, we have $\lambda(k+1) = \lambda(k) + \rho(\mathbf{w}(k+1) - B\mathbf{x}(k+1))$, and this implies:

$$u(k+1) = u(k) + \mathbf{w}(k+1) - s(k+1) = \gamma(k+1) \tag{6.30}$$

where the last equality follows by substituting (6.29). We complete the proof by rearranging terms in (6.28)-(6.30). □

Proof of Theorem 10. The second part has been shown in Theorem 9, we only need to prove the first part. Note that the gradient of $\hat{f}$ is $\frac{\rho}{(mL)^{1/2}}\kappa_f^{-1/2}$-Strongly monotone and $\frac{\rho}{(mL)^{1/2}}\kappa_f^{1/2}$-Lipschitz, and

M is full column rank. We can extend Theorem 6 [237] to variational inequality problem by using Lemma 7, and Lemma 8. Then, by Theorem 7 [237], we have $V(k) \leq (1 - 1/(2\kappa_f^{0.5+|\epsilon|}))V(k-1)$, where $\epsilon = \log_{\kappa_f}(\rho/\sqrt{mL})$. □

Chapter 7

Stochastic Game Theory for Distributed Energy Management

In this chapter, we apply stochastic game theory to design a pricing mechanism that regulates individual users' behavior in power grids such that we can optimize the overall system's performance. This chapter is adapted from the work [38], co-authored with Eli Brock, Javad Lavaei, and Somayeh Sojoudi.

7.1 Background

DER Coordination in Distribution Networks

An additional 217 GW of distributed energy resources (DERs) is expected on the American electric power grid by 2028, a pace of growth similar to that of bulk generation capacity [261]. The widespread introduction of DERs, which include electric vehicles, heat pumps, storage systems, and distributed solar, marks a critical moment for our energy systems. If operated passively the extra load from DERs will necessitate expensive infrastructure upgrades and new carbon-intensive fossil-based dispatchable generation. However, with efficient coordination, DER flexibility can improve efficiency on the grid by shaping electric demand to align with intermittent renewable supply and providing local voltage support [6].

There is extensive literature on market mechanisms and control strategies for DER coordination in distribution networks [120]. Coordinated approaches such as distributed optimal power flow, aggregators, and distribution locational marginal pricing (DLMP) assume that a third party orchestrates groups of DERs, either through direct load control or through (shadow) price incentives, to optimize a single objective function [120, Section 5], [49, 16]. In contrast, peer-to-peer (P2P) markets assume that prosumers control their own devices and optimize their own utility functions; as such, P2P is often analyzed using noncooperative game theory [120, Section 6], [55]. These frameworks make strong assumptions regarding future communication and incentive infrastructures. For example, P2P and distributed optimal power flow methods typically assume that neighboring

devices exchange multiple rounds of communication before reaching a collective decision, while DLMP schemes require prosumers[1] to share device parameters with a coordinator and schedule future consumption in a rolling-horizon fashion.

Achieving any of the aforementioned frameworks at scale would constitute a major departure from the state of most modern power systems. In practice, distribution system operators (DSOs) usually do not have visibility behind the meter of their customers, and self-interested prosumers autonomously dispatch their own devices without communicating with their neighbors. A more practical mechanism, real-time pricing (RTP), coordinates DERs on the transmission scale by exposing them to time-varying substation-level nodal prices and allowing each prosumer's energy management system (EMS) to optimize their personal electricity usage [217]. Most RTP work has focused on optimal control of individual devices as in [304] or considers game-theoretic equilibria on the wholesale market as in [204], without considering the implications for distribution networks.

We propose a new real and reactive nodal pricing structure for minimizing costs and stabilizing voltages on distribution networks. The framework can be understood as a natural extension of RTP to distribution nodes. Unlike traditional DLMPs, the nodal prices are set online, and prosumers do not need to share their device parameters with a DSO. Though the prosumers cannot communicate directly, the nodal pricing scheme allows for indirect coordination through the coupled prices. The resulting networked market is represented as a *stochastic game* (SG) where prosumers attempt to learn closed-loop control policies in the face of uncertain wholesale prices and demand profiles [278]. To the authors' knowledge, this is the first network-aware distribution-level DER coordination scheme that does not require prosumers to share any information, either between themselves or with a central operator. Next, we derive new, generalizable sufficient conditions under which an SG is a Markov potential game [176], allowing us to compute an equilibrium for the proposed market. Finally, we demonstrate on an IEEE test system that the proposed mechanism results in near-socially-optimal equilibrium policies, despite the potential suboptimality associated with the prosumers' market power.

Notation

$G(p)$ is the geometric distribution with success probability p. If a variable or function is defined as y_i, then y refers to the vector or vector-valued function collecting all indices. If a set is defined as $\mathcal{X}_i$, then $\mathcal{X}$ refers to the Cartesian product over all indices. If $\mathcal{I}$ is a collection of indices, $y_{\mathcal{I}}$ refers to the elements of the vector y indexed by $\mathcal{I}$. The subscript $-i$ indexes all components except i. $\mathrm{diag}(v)$ is the square matrix with the vector v along the diagonal and zeros elsewhere.

Stochastic Game Theory Preliminaries

An infinite-horizon stochastic game (SG) [278] with set of agents $\mathcal{N}$ is a tuple

$$\mathcal{G} \coloneqq (\mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \rho_0, T, \{U_i\}_{i\in\mathcal{N}}, \gamma, \{\Pi_i\}_{i\in\mathcal{N}}). \tag{7.1}$$

[1] In active distribution networks, a *prosumer* is a customer who can both consume and produce energy, for example through a rooftop photovoltaic or vehicle-to-grid system.

$\mathcal{S}$ is a (possibly infinite) state space, $\mathcal{A}_i$ is the (possibly infinite) action space of agent i, $\rho_0 : \mathcal{S} \to \mathbb{R}$ is an initial state distribution, $T : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the transition density, $U_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function of agent i, $\gamma \in [0, 1]$ is the discount factor, and Π_i is the set of agent i's available policies. At time $t = 0$, the initial state s_0 is drawn from the initial state distribution ρ_0. At each time t, each agent i chooses an action a_i^t. Based on the current state and actions, each agent i receives a deterministic reward $U_i(s^t, a^t)$ and the game transitions to the next state according to the transition density function: $s^{t+1} \sim T(\cdot|s^t, a^t)$.

Agent i aims to maximize its infinite-horizon discounted rewards, defined by its *value function*

$$V_i^\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t U_i(s^t, a^t) \right]. \tag{7.2}$$

The notation $\mathbb{E}_\pi$ is shorthand for the expectation with respect to $s^0 \sim \rho_0$, $a_i^t \sim \pi_i(\cdot|s^t)$ and $s^{t+1} \sim T(\cdot|s^t, a^t)$, where $\pi_i \in \Pi_i$ is agent i's policy. We are now ready to define *Nash equilibria*, the most common solution concept for SGs.

Definition 3 (Nash Equilibrium [24])**.** *A joint policy profile $\pi \in \Pi$ is called a Nash equilibrium of $\mathcal{G}$ if*

$$V_i^{\pi_i, \pi_{-i}} \geq V_i^{\tilde{\pi}_i, \pi_{-i}}$$

for all $i \in \mathcal{N}$ and $\tilde{\pi} \in \Pi$.

In the remainder of the paper, we will use "equilibrium" and "Nash equilibrium" interchangeably. Intuitively, an equilibrium is a configuration of policies where no agent can unilaterally improve their payoff by changing their policy if all other agents keep their policy fixed. We will often consider *parametric* policy sets of the form $\pi_i = \{\pi_i^{\theta_i}(\cdot|s^t, a^t) : \theta \in \Theta\}$. For compact notation, when Π is parametric, we will often use θ in place of π^θ. For parametric policy sets, we can also define the notion of a *local* Nash equilibrium.

Definition 4 (Local Nash Equilibrium)**.** *If Π is a parametric policy class, a set of policy parameters $\theta \in \Theta$ is called a local Nash equilibrium of $\mathcal{G}$ if there exists a $\nu > 0$ such that*

$$V^{\theta_i, \theta_{-i}} \geq V^{\tilde{\theta}_i, \theta_{-i}}$$

for all $\tilde{\theta} \in \Theta$ such that $\|\tilde{\theta} - \theta\| \leq \nu$.

Going forward, we will assume that the value functions V^{θ_i} of parameterized SGs are continuous in θ_i and differentiable in θ_i almost everywhere. Local Nash equilibria are stationary points under the *gradient play* algorithm, in which each agent applies the update rule

$$\left(\theta_i^{(k+1)} - \theta_i^{(k)}\right) \propto \nabla_\theta V_i^{\theta^{(k)}}. \tag{GP}$$

An SG with a single agent is known as a *Markov Decision Process* (MDP) and a (local) equilibrium of an MDP is called a (locally) *optimal policy*.

Notice that our definition of SGs explicitly includes the set of feasible policies Π. Most existing literature on SGs (and MDPs) omits this specification, implicitly assuming that Π includes all stochastic policies. For large or infinite state and action spaces, policies are often parameterized in order to design tractable solution methods. However, these parameterizations are treated as function approximations that induce some suboptimality due to their limited expressiveness [4]. While we also use Π to encode finite-dimensional parameterizations, in this work, the policy sets are also restricted to capture constraints on the information structure between agents (see Section 7.2). We only seek to find equilibria with respect to the specified policy set without considering the relationship between said equilibria and the equilibria over all stochastic policies.

Computing local equilibria for SGs is hard in general. Unlike in single-objective optimization problems, many natural algorithms such as GP may cycle instead of converging to a stationary point [211]. In the following, we will use the stochastic game framework to model a distribution grid under a new nodal pricing coordination mechanism (Section 7.2). Then, we will develop theory inspired by this application to show that our model belongs to a subclass of SGs for which equilibria can be tractably computed (Section 7.3).

7.2 Model Formulation

We develop a stochastic game model of a distribution grid with dynamic nodal pricing and autonomous DERs. The proposed market is designed to achieve efficient outcomes for the grid without central coordination or prosumer-to-prosumer communication, as discussed in Section 7.1.

Grid Model

A radial distribution grid is modeled as a directed acyclic graph with a root node 0 and a set of non-root nodes $\mathcal{N}$. Denote the full set of nodes as $\mathcal{N}^+ = \{0\} \cup \mathcal{N}$ and the set of lines and transformers (directed edges) $\mathcal{L} \in \mathcal{N}^+ \times \mathcal{N}^+$. By convention, edges are oriented away from the root node. The linear DistFlow model is, for all $i \to j \in \mathcal{L}$,

$$P_{ij} = p_j + \sum_{k:j\to k} P_{jk} \tag{7.3a}$$

$$Q_{ij} = q_j + \sum_{k:j\to k} Q_{jk} \tag{7.3b}$$

$$v_i - v_j = r_{ij}P_{ij} + x_{ij}Q_{ij} \tag{7.3c}$$

where p_i, q_i are the real and reactive power consumption of node i, P_{ij} and Q_{ij} are the real and reactive power flows on line $i \to j$, and v_i is the voltage magnitude at node i [21]. The voltage magnitude at the substation v_0 is fixed and constant. Solving the system of equations (7.3) for P, Q, and v reveal these quantities to be linear functions of (p, q), where p and q are vectors collecting the nodal injections at the non-root nodes $\mathcal{N}$. We define the matrices H, R, and X to represent (7.3) in

the compact form

$$P = Hp \tag{7.4a}$$
$$Q = Hq \tag{7.4b}$$
$$v - v_0 = Rp + Xq. \tag{7.4c}$$

In what follows, we will use P, Q, and v to denote these functions, leaving the dependence on (p, q) implicit. At each time, the load profiles p and q are autonomously determined by the prosumers and the DSO incurs the cost

$$C(p, q, \lambda) = (1 - w)\lambda \left(\sum_{i \in \mathcal{N}} p_i + \sum_{i \to j \in \mathcal{L}} r_{ij} \left(P_{ij}^2 + Q_{ij}^2 \right) \right) + w \sum_{i \in \mathcal{N}} (v_i - v_0)^2 \tag{7.5}$$

where λ is the wholesale locational marginal price (LMP) at the substation node and $w \in [0, 1]$ is a given parameter. The first term in (7.5), given by the price multiplied by the sum of the loads and the approximate real power losses, captures the cost of importing real power from the wholesale market. While the true real power loss on line $i \to j$ is $r_{ij}(P_{ij}^2 + Q_{ij}^2)/v_i^2$, we use the fact that $v_i \approx 1$ under normal operating conditions to approximate the losses in (7.5). Since the losses do not appear in the linear model (7.3), the DSO cost (7.5) does not account for the small fraction of losses incurred on each line due to losses on downstream lines. We also assume that the distribution network is a price-taker, meaning that it is a small enough participant in the wholesale market that it cannot affect λ. The second term penalizes deviations from the nominal voltage as in the voltage control literature [335]. For simplicity, we assume the nominal voltage across the network is equal to the substation voltage v_0. The voltage control weight w controls the trade-off between cost minimization and voltage control, and can be tuned by the DSO until voltages are within an acceptable range.

On distribution networks, the DSO typically handles line ampacity limits through network reconfiguration. Given the complexity introduced by time-varying network topologies, we do not consider line limits; however, this is an important direction for future work.

Prosumer Model

We model a single prosumer at every non-substation bus, so the set of prosumers is also $\mathcal{N}$. Multiple prosumers per bus may also be handled by the model without affecting the theory. Prosumer $i \in \mathcal{N}$ exhibits inelastic real and reactive power demand $(\bar{p}_i, \bar{q}_i)$ representing the sum of inflexible consumption from devices other than DERs, such as kitchen appliances, lighting, and most other plug loads. Prosumer i also owns a set of flexible DERs $\mathcal{N}_i$ and an EMS enabling automatic intelligent control. These DERs may include heat pumps, electric vehicles (EVs), and energy storage devices. Denoting the consumption from DER $j \in \mathcal{N}_i$ belonging to prosumer i as $\tilde{p}_{i,j}$, the load from prosumer i is given by the sum of their inelastic and flexible demand:

$$(p_i, q_i) \coloneqq (\bar{p}_i, \bar{q}_i) + \sum_{j \in \mathcal{N}_i} (\tilde{p}_{i,j}, \tilde{q}_{i,j}). \tag{7.6}$$

The DERs also have temporal state dynamics:

$$d_i^{t+1} = f_i(d_i^t, \tilde{p}_i^t, \tilde{q}_i^t, \omega_i^t) \tag{7.7}$$

where d_i is a vector collecting the states of the DERs belonging to prosumer i and ω_i^t is a random perturbation that will be further discussed in section 7.2. Here we introduce the subscript t to index discrete time. The quantities p, q, P, Q, λ, and v also vary in time—the superscript was previously omitted for clarity. The state dynamics f may represent the state-of-charge of a storage unit/EV or the air/water temperature for a heat pump. The states serve to constrain the DER consumption at each stage:

$$(\tilde{p}_i^t, \tilde{q}_i^t) \in \mathcal{P}_i(d_i^t) \tag{7.8}$$

where we assume $\mathcal{P}_i(\cdot)$ is the feasible set for agent i given d_i. Constraint (7.8) may encode state-of-charge, inverter capacity, or comfort constraints.

Pricing Mechanism

The DSO sets real-time nodal prices for real and reactive power equal to the marginal cost of serving the load at each bus. We assume a *net metering* policy, meaning agents are charged the same rate for net consumption as they are credited for net generation. Specifically, prosumer i's reward function is given by

$$U_i(d_i, \lambda, \tilde{p}, \tilde{q}, \bar{p}, \bar{q}) \coloneqq u_i(d_i, \tilde{p}_i, \tilde{q}_i) - p_i \frac{\partial}{\partial p_i} C(p, q, \lambda) - q_i \frac{\partial}{\partial q_i} C(p, q, \lambda) \tag{7.9}$$

where u_i is prosumer i's utility function. u_i may encode indoor air temperature preferences or battery degradation costs; for purely shiftable loads, u_i may be set to zero. Intuitively, the last two terms in (7.9) are prosumer i's payment to the DSO. The first term is their instantaneous benefit from the state-action configuration of their DERs. The pricing mechanism in (7.9) is inspired by the DLMP literature [16, 243]. In these prior works, the DSO computes the prices over a rolling horizon by solving a multi-period scheduling problem for all the devices in the network. By contrast, the prices in (7.9) are set online for the current time period and only require the DSO to meter the current aggregate consumption at each node.

The nodal prices are composed of three components associated with the three terms in (7.5) after expansion: the energy price, the losses price, and the voltage price. The energy price, given by the substation LMP for real power and zero for reactive power, is constant across nodes. Each prosumer's losses and voltage prices, however, depend on their own consumption as well as that of all agents who share a common ancestor on the network. This coupling introduces strategic interactions and some *market power*, that is, the ability of a participant in a market to manipulate the price. When participants have market power, market equilibria may not maximize social welfare as they do in the fully competitive case. For (7.9) to be an effective coordination mechanism, the gap between the socially optimal outcome and the equilibrium outcome, known as the *price-of-anarchy*, should be small. We empirically verify this in Section 7.4.

Exogenous Quantities

We propose a time-invariant state-space model for the exogenous variables λ, $\bar{p}$, and $\bar{q}$:

$$\alpha_i^{t+1} = g_i(\alpha_i^t, \xi_i^{t+1}) \qquad \forall i \in \{0\} \cup \mathcal{N} \tag{7.10a}$$

$$\lambda^t = m_0(\alpha_0^t) \tag{7.10b}$$

$$(\bar{p}_i^t, \bar{q}_i^t) = m_i(\alpha_i^t) \qquad \forall i \in \mathcal{N} \tag{7.10c}$$

$$(\xi^t, \omega^t) \sim \rho_{\xi,\omega} \tag{7.10d}$$

where g is the transition function, m is the measurement function, and $\rho_{\xi,\omega}$ is the noise distribution. The framework (7.10) encompasses a rich class of models while satisfying the Markov property. Models of the form (7.10) include seasonal autoregressive integrated moving average models, which are common for forecasting time-series econometric data such as electricity demand and prices [78].

While each prosumer's inelastic demand $\bar{p}$ and the LMP λ have separable dynamics in (7.10), they are coupled through the joint distribution $\rho_{\xi,\omega}$, which generates the noise ξ. ξ can represent both independent regressors, such as weather, and decoupled perturbations that may simultaneously affect the wholesale price and the demand at different buses. Moreover, since the DER dynamics noise ω is also generated by $\rho_{\xi,\omega}$, it is correlated with ξ in general. For example, if the DER state is the indoor air temperature, its evolution will be subject to the same perturbations as the ambient outdoor temperature.

Policies

Just as the DSO lacks knowledge of the behind-the-meter devices of its customers, prosumers also cannot see behind the meter of their neighbors. Specifically, we assume that prosumer i observes only their local DER state d_i and has knowledge of their own inelastic demand $(\bar{p}_i, \bar{q}_i)$ and the LMP λ, which we assume is public[2]. By "has knowledge", we mean that the prosumer i's EMS can access all the information it needs to make the best possible prediction of the next quantities $\bar{p}_i^{t+1}$, $\bar{q}_i^{t+1}$ and λ^{t+1}. In practice, this may include recent histories, periodic information related to the time of day, day-ahead forecasts, and correlated data such as weather—all of which would be accessible to a cloud-connected controller. For our model, given the Markovian structure of (7.10), we achieve the desired information structure by simply allowing agent i to condition its policy directly on α_i and α_0, specifically

$$a_i^t = \mu_i^{\theta_i}(d_i^t, \alpha_i^t, \alpha_0^t, \eta_i^t) \tag{7.11}$$

where $a_i^t \in \mathbb{R}^2$ is agent i's action, $\eta_i^t \sim \rho_{\eta_i}$ is the policy noise and $\mu_i^{\theta_i}$ is prosumer i's policy, parameterized by θ^i. The generative model (7.10), combined with the local policies (7.11), abstracts away the choice of specific information on which real EMSs might condition their policies and allows for clean theoretical analysis. Optimizing the performance of DERs in a more realistic context without assuming Markovian states and allowing for possibly incomplete observations is an important direction for future research, but falls outside the scope of this work.

[2]As a real-world justification of this assumption, the California Independent System Operator publishes substation-level real-time nodal prices on its website.

To handle the constraint (7.8), we define the unbounded action spaces $\mathcal{A}_i \coloneqq \mathbb{R}^{2|\mathcal{N}_i|}$, $\forall i \in \mathcal{N}$, and compute the DER consumption by mapping the action onto the feasible set:

$$(\tilde{p}_i^t, \tilde{q}_i^t) = M_i(a_i^t, d_i^t), \tag{7.12}$$

where $M_i : \mathbb{R}^{2|\mathcal{N}_i|} \to \mathbb{R}^{2|\mathcal{N}_i|}$ is an appropriately chosen function satisfying $M_i(a_i, d_i) \in \mathcal{P}_i(d_i)$ for all possible a_i, d_i. Depending on the application, M might be a clipping or projection operation. We assume that M is differentiable almost everywhere and continuous.

As an alternative to the local policies (7.11), one could also consider a partially observable stochastic game paradigm, where policies are conditioned on the history of agent i's observations. However, since any given agent may be unaware of their neighbors, it is unlikely that realistic EMSs would attempt to characterize the state of other prosumers' devices. Therefore, we argue that the local policy parameterization (7.11) is appropriate for this setting.

Stochastic Game Formulation

We are now ready to formally express the proposed model as an SG, which we call $\mathcal{D}$. The state vector is $s = (x, \alpha)$. The transition density T of $\mathcal{D}$ is characterized by (7.7), (7.10a), and (7.10d). The reward functions U_i, $i \in \mathcal{N}$ are given in (7.9). The (parametric) joint policy set Π is defined according to (7.11). The discount factor γ and initial state distribution ρ_0 are case-specific.

7.3 Markov Potential Games

In this section, we introduce a class of well-behaved SGs known as Markov Potential Games (MPGs) that admit tractable algorithms for computing equilibria [176]. We then present new, generalizable sufficient conditions under which an SG is an MPG and show that the game $\mathcal{D}$ introduced in Section 7.2 satisfies these conditions.

MPGs generalize the notion of *potential games* in static game theory to SGs [224].

Definition 5 (Markov Potential Game [329])**.** *An SG $\mathcal{G}$ is a Markov potential game if there exists a potential function $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that*

$$V_i^{\pi_i, \pi_{-i}} - V_i^{\tilde{\pi}_i, \pi_{-i}} = \Phi^{\pi_i, \pi_{-i}} - \Phi^{\tilde{\pi}_i, \pi_{-i}}$$

where

$$\Phi^{\pi} \coloneqq \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s^t, a^t)\right]$$

for all $i \in \mathcal{N}$ and $\pi, \tilde{\pi} \in \Pi$.

In other words, an MPG is an SG where each agent's value function is characterized by a single *potential value function* Φ^{π}, given by the discounted sum of the potential function ϕ. An immediate consequence of Definition 5 is that joint policy profiles that (locally) optimize the potential function Φ^{π} are also (local) Nash equilibria.

If a game is an MPG and the potential function ϕ is known, then finding a (local) equilibrium is reduced to (locally) solving a single-agent MDP for the optimal joint policy with rewards given by ϕ. Unlike multi-agent SGs, MDPs can be reliably solved by methods from reinforcement learning and dynamic programming. In particular, for parameterized policy classes, a local equilibrium of an MPG can be found by the gradient ascent algorithm

$$\left(\theta^{(k+1)} - \theta^{(k)}\right) \propto \nabla_\theta \Phi^{\theta^{(k)}}. \tag{Φ-GA}$$

Definition 5 implies that, for MPGs, Φ-GA is equivalent to the decentralized gradient play algorithm GP; see [176, Proposition B.1] for details.

Definition 5 is challenging to verify. It may be natural to suspect that an SG is an MPG if it is potential at each stage, that is, if there exists ϕ such that

$$U_i(s, a_i, a_{-i}) - U_i(s, \tilde{a}_i, a_{-i}) = \phi(s, a_i, a_{-i}) - \phi(s, \tilde{a}_i, a_{-i}) \tag{7.13}$$

for all $i \in \mathcal{N}$, $a, \tilde{a} \in \mathcal{A}$, and $s \in \mathcal{S}$. Unfortunately, this is not true; see [329, Proposition 2] for a counterexample. SGs satisfying (7.13) are, however, the most promising candidates for MPGs since static game theory techniques may be used to find a candidate potential function ϕ given the rewards U_i. As such, the existing literature has focused on obtaining sufficient conditions under which SGs satisfying (7.13) are MPGs [176, 329, 213].

[213] claims to establish a sufficient condition for MPGs under an assumption, called "state transitivity", that the reward functions are also potential with respect to the state. This assumption is quite restrictive and is not satisfied by the game of interest $\mathcal{D}$. [176, Prop 3.2, C1] and [329, Lemma 8] both introduce what we call the *action-independent transitions* (AIT) sufficient condition requiring that the transition density is independent of the agents' actions. However, this restriction precludes any controllable state dynamics and is only satisfied by repeated one-shot games. [176, Prop 3.2, C2] generalizes AIT, but it is difficult to check and is specific to SGs with finite state and action spaces.

[329, Lemma 8] introduces what we call the *local states and policies* (LSP) sufficient condition. LSP holds when 1) each state is owned by a certain agent, 2) the rewards are potential in states as well as actions, 3) each local state space has its own conditionally independent transition density, and 4) each agent's policy is conditioned only on its local state. LSP does not apply to $\mathcal{D}$ because the state space includes the component α_0 that does not belong to any agent, and the agents' local state transitions are coupled through the joint distribution $\rho_{\xi,\omega}$.

While neither AIT nor LSP applies directly to $\mathcal{D}$, $\mathcal{D}$ exhibits features of each. Similar to AIT, the exogenous states α evolve independently of the agents' actions. As in LSP, the agents condition their policies on private local states with transition densities that are independent of other agents' actions. We now present a new, verifiable sufficient condition generalizing both AIT and LS that applies to the proposed distribution system game $\mathcal{D}$.

Theorem 11. *An SG $\mathcal{G}$ is an MPG with potential function ϕ if, for each agent $i \in \mathcal{N}$, there exists a local state space $\mathcal{S}_i$ such that $\mathcal{S} = \mathcal{S}_i \times \mathcal{S}_{-i}$ and*

1. *ϕ is a stagewise potential function for both the states and actions, that is*

$$\begin{aligned} &U_i(s_i, s_{-i}, a_i, a_{-i}) - U_i(\tilde{s}_i, s_{-i}, \tilde{a}_i, a_{-i}) \\ &= \phi(s_i, s_{-i}, a_i, a_{-i}) - \phi(\tilde{s}_i, s_{-i}, \tilde{a}_i, a_{-i}) \end{aligned} \tag{7.14}$$

   for all $s_i, \tilde{s}_i \in \mathcal{S}_i$, $s_{-i} \in \mathcal{S}_{-i}$, and $a, \tilde{a} \in \mathcal{A}$.

2. *Other agents' policies do not depend on the local states, that is*

$$\pi_{-i}(a_{-i}|s_i, s_{-i}) = \pi_{-i}(a_{-i}|s_{-i}). \tag{7.15}$$

3. *The marginal transition density of the non-local states, defined as*

$$T_{-i}(s'_{-i}|s, a) \coloneqq \int_{s'_i \in \mathcal{S}_i} T(s'_i, s'_{-i}|s, a) ds'_i$$

   does not depend on agent i's action or local state, that is

$$T_{-i}(s'_{-i}|s_i, s_{-i}, a_i, a_{-i}) = T_{-i}(s'_{-i}|s_{-i}, a_{-i}). \tag{7.16}$$

Proof. See appendix. □

Theorem 1 generalizes LSP by allowing the agents' non-local state spaces to overlap and by relaxing the requirement that the local state transition probabilities be conditionally independent. While Theorem 11 is motivated by the power systems setting considered here, it is quite general and may extend to other applications. The challenge of the proof is to show that the difference between the potential value function and agent i's reward is independent of its policy. At a high level, this is achieved by combining and generalizing the fundamental ideas of AIT and LSP.

Referring back to Section 7.2, we apply Theorem 11 to show that $\mathcal{D}$ is an MPG. From (7.4), define the matrix

$$\begin{aligned} L(\lambda) = \lambda(1-w) &\begin{bmatrix} H^T \operatorname{diag}(r) H & \mathbf{0} \\ \mathbf{0} & H^T \operatorname{diag}(r) H \end{bmatrix} \\ + w &\begin{bmatrix} R^T \\ X^T \end{bmatrix} \begin{bmatrix} R & X \end{bmatrix} \end{aligned}$$

Corollary 1. *$\mathcal{D}$ is an MPG with potential function*

$$\phi_{\mathcal{D}}(d, a, \lambda) = \sum_{i \in \mathcal{N}} u_i(d_i, \tilde{p}_i, \tilde{q}_i) - \tilde{C}(p, q, \lambda)$$

where

$$\tilde{C}(p, q, \lambda) = C(p, q, \lambda) + \sum_{i \in \mathcal{N}} \begin{bmatrix} p_i \\ q_i \end{bmatrix}^T L(\lambda)_{\mathcal{I}_i \mathcal{I}_i} \begin{bmatrix} p_i \\ q_i \end{bmatrix} \tag{7.17}$$

and $\mathcal{I}_i = \{i, i + |\mathcal{N}|\}$ for all $i \in \mathcal{N}$.

Proof. See appendix. □

Corollary 1 is instructive. Ideally, the agents would cooperate to maximize the discounted sum of the social benefit $\sum_{i\in\mathcal{N}} u_i(d_i, \tilde{p}_i) - C(p, q, \lambda)$. Instead, due to the market power associated with the losses and voltage prices, they minimize the discounted sum of $\phi_{\mathcal{D}}$, which includes the the additional second term in (7.17). In Section 7.4, we will demonstrate that the effect of this additional term is small in practice.

7.4 Experiment

We evaluate the efficiency of the proposed market on a benchmark IEEE network populated with autonomous storage units. The equilibrium policies are compared with a more naive non-nodal RTP pricing structure as well as the theoretical socially optimal policies. Our analysis is from a mechanism design perspective, meaning that we are interested in evaluating the social benefit of an equilibrium found through centralized computation, assuming that, in practice, prosumers will find it in the process of optimizing their local policies.

Algorithm

For MPGs, the potential value function is simply the discounted sum of the potential function. Therefore, the problem of finding a local equilibrium reduces to the problem of computing a locally optimal policy for an MDP. Given that we have access to the transition density and reward functions, we choose to apply a simple *value gradient* algorithm inspired by [127] instead of more popular model-free policy gradient algorithms. Value gradient algorithms depend on the reparameterization trick, where the policy and transition densities are expressed as deterministic functions of independent random variables:

$$a^t = \mu_\theta(s^t, a^t, \eta^t) \tag{7.18a}$$
$$s^{t+1} = h(s^t, a^t, \zeta^{t+1}) \tag{7.18b}$$
$$\eta^t \sim \rho_\eta,\ \zeta^t \sim \rho_\zeta \tag{7.18c}$$

For the game of interest $\mathcal{D}$, the transition density defined in (7.7), (7.10a), and (7.10d) as well as the policies (7.11) are already written in reparameterized form; we reuse μ and η given the one-to-one correspondence between their use in the general case and their use in $\mathcal{D}$. Given an MPG with potential function ϕ, define the estimated potential value function

$$\hat{\Phi}^\theta(s^0, H, \zeta, \eta) \coloneqq \sum_{t=0}^{H} \phi(s^t, a^t) \tag{7.19}$$

where s^t, $t \geq 1$ and a^t are generated by (7.18). $\hat{\Phi}^\theta$ is an unbiased estimate of the potential value function

$$\mathbb{E}\left[\hat{\Phi}^\theta(s^0, H, \zeta, \eta)\right] = \Phi^\theta$$

Algorithm 5: SGA

Require: S-MPG $\mathcal{G}$, γ, ρ_ξ, ρ_η, ρ_0, γ, β, N_{train}, N_{batch}

1: Define $\hat{\Phi}^\theta$ from $\mathcal{G}$
2: Arbitrarily initialize θ
3: **for** 1 **to** N_{train} **do**
4: **for** 1 **to** N_{batch} **do**
5: Sample $s^0 \sim \rho_0$, $H \sim G(1-\gamma)$, $\zeta_t \sim \rho_\zeta$, $\eta_t \sim \rho_\eta$
6: $\hat{\nabla} \leftarrow \hat{\nabla} + \frac{1}{N_{\text{batch}}} \nabla_\theta \hat{\Phi}^\theta(s_0, H, \zeta, \eta)$
7: **end for**
8: $\theta \leftarrow \theta + \beta \hat{\nabla}$
9: **end for**
10: **return** θ

where the expectation is taken with respect to (7.18c), $s^0 \sim \rho_0$, and $H \sim G(1-\gamma)$. By extension, the gradient of the estimator is an unbiased estimate of the gradient of the potential value function:

$$\mathbb{E}\left[\nabla_\theta \hat{\Phi}^\theta(s^0, H, \zeta, \eta)\right] = \nabla_\theta \Phi^\theta.$$

This leads to Algorithm 5 for computing local equilibria, which is equivalent to stochastic gradient ascent (SGA) on the potential value function. Step 6 can be computed by backpropogation through the Bellman equation as in [127, Sec. 4.1]; automatic differentiation software such as PyTorch can perform this computation off-the-shelf.

The update rule in Step 6 is equivalent to the stochastic gradient play update rule. As $N_{\text{batch}} \to \infty$, Algorithm 5 recovers the deterministic gradient ascent algorithm (Φ-GA), which is in turn equivalent to the gradient play algorithm (GP) as discussed in Section 7.3. The scope of this work is limited to computing a local equilibrium from a mechanism design perspective. However, consider a more practical multi-agent reinforcement learning setting where agents compute unbiased estimates of their policy gradients in a decentralized, model-free manner using the policy gradient theorem [287]. In expectation (with large batch sizes), this multi-agent policy gradient algorithm is equivalent to Algorithm 5. Therefore, Algorithm 5 may mirror the learning process of multi-agent systems when the sufficient conditions in Theorem 11 hold. See [328] for a convergence analysis of policy gradient algorithms for continuous state and action spaces (single-agent policy gradient results suffice in the MPG setting).

Setup

We demonstrate the proposed distribution grid market on the IEEE 18-bus test case from [114] with 1-hour timesteps. Each prosumer owns a single battery energy storage unit, such as a Tesla Powerwall, with its internal state being the current state-of-charge. A storage unit's charging is constrained by its maximum state-of-charge as well as its inverter capacity. Customer utility functions are set to zero, so they focus only on price arbitrage. The substation LMP and inelastic

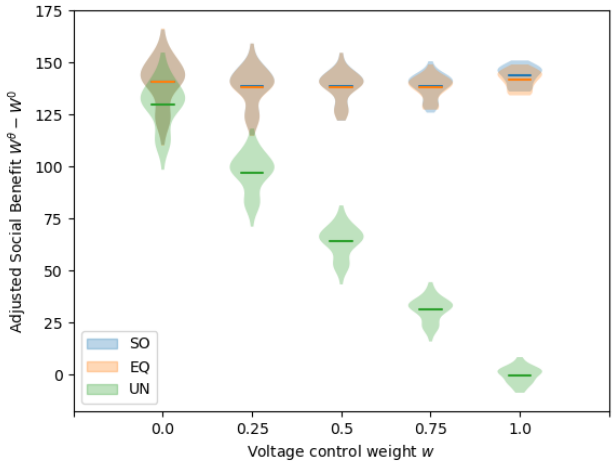


Figure 7.1: Relative performance of the EQ, SO, and UN policies. EQ is almost equivalent to SO and is more socially beneficial than UN pricing regardless of the tradeoff between import costs and voltage control.

demand profiles are simple noisy sinusoids with 1-day periods, with the demand profiles staggered such that they peak (on average) six hours before the LMP. Loads are scaled until losses reach approximately 10% of the total power flow, consistent with a heavily loaded distribution network. The local policies (7.11) are Gaussian with the mean parameterized as a simple affine function of the states. The full experimental setup is detailed in the appendix.

Performance

We train three joint policies for comparison. All three policies are trained by Algorithm 5 with $N_{\text{batch}} = 1$, $N_{\text{train}} = 500$, $\beta = .001$, and the policy parameters θ initialized to the zero vector. The trained policies are:

1. Equilibrium (EQ) policies, trained to minimize the potential value function by setting $\phi = \phi_{\mathcal{D}}$ in (7.19).

2. Socially optimal (SO) policies, computed by setting ϕ in (7.19) to the social welfare value function
$$W^\theta := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(\sum_{i \in \mathcal{N}} u_i(d_i^t, \tilde{p}_i^t) - C(p^t, \lambda^t) \right) \right].$$

3. Policies under uniform pricing (UN), trained identically to the EQ policies but with the network removed by setting $r = x = 0$.

Fig. 7.1 shows the approximate adjusted social welfare value function $W^\theta - W^0$ for each policy across five values of the voltage control weight w[3]. W^0 is the social welfare value function when

[3]Because lower values of w deprioritize voltage control, the losses approximation in (7.5) may only be well-justified for sufficiently high w. The low w cases in Fig. 7.1 are meant to explore the strategic implications of the game, rather than to serve as realistic simulations.

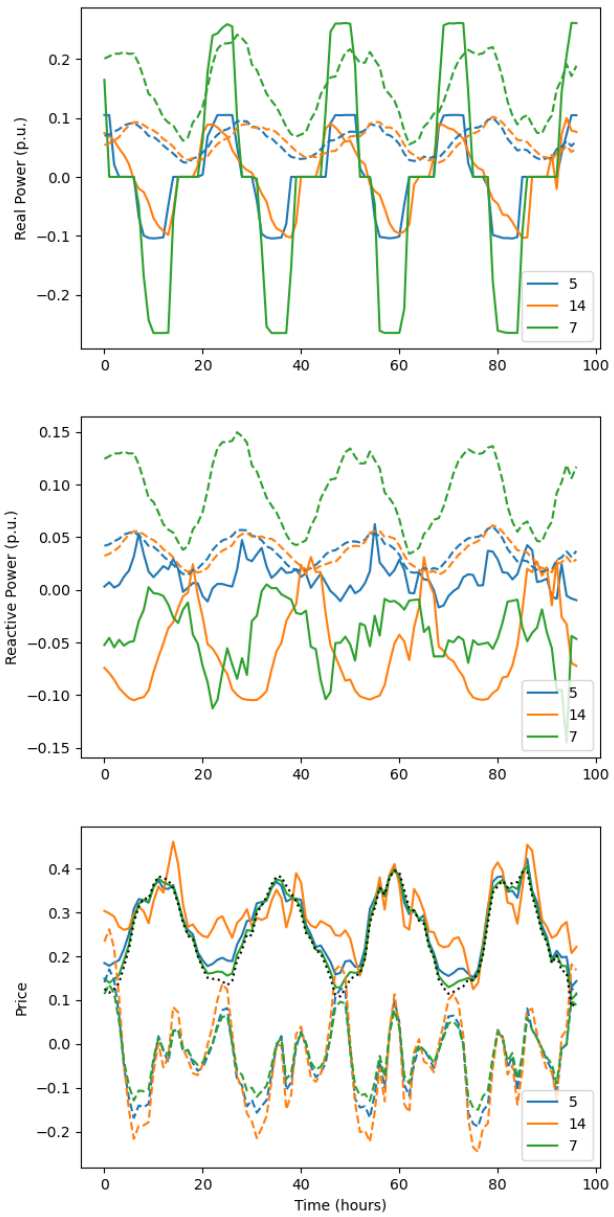


Figure 7.2: Equilibrium behavior of three random agents over four-day rollout.

the prosumers idle their storage units at every timestep. The adjustment is necessary because W^0 appears as a constant term in the expanded form of W^θ, so including it would distort the relative performance of the policies. The approximate social welfare value function is the empirical average over 50 discounted rollouts of length 500 ($.99^{500} < .01$). To allow for a one-to-one comparison, the same 50 trajectories of the random variables ξ and ω were used for all 15 evaluations. Fig. 7.1 visualizes the distributions of the 50 rollouts and their means, indicated by horizontal lines. All policies converged by the end of the 500 training rollouts.

UN is the equilibrium outcome when the DSO ignores the network and simply forwards the

LMP to each prosumer. In this aggregated context, the prices no longer vary by node, so there is no market power. The EQ policies learn to respond to nodal prices reflecting network-related social costs. However, in theory, they may exploit their market power to benefit themselves at the expense of their neighbors, thereby failing to meaningfully increase social welfare compared to the network-blind UN policies. The results in Fig. 7.1 indicate that this is not the case: the EQ policies outperform the UN policies by at least 8% across all five voltage control weights.

The SO policies capture the ficticious best-case-scenario where all devices cooperatively maximize social welfare instead of responding to prices. Fig. 7.1 shows that the price-of-anarchy, given by the gap between EQ and SO, is negligible relative to the gap between EQ and UN [4]. For this benchmark network, the takeaway is that the benefits of the proposed nodal pricing mechanism outweigh the costs—in fact, it is almost as efficient as the best possible coordination mechanism given the local policy structure.

Demonstration

To build intuition, we include a four-day simulation of the deterministic EQ policies with $w = 0.75$ in Fig. 7.2. Three randomly chosen agents are presented. The first and second subplots show the real and reactive power load, with inelastic demand as a dashed line and storage consumption as a solid line. The third subplot shows the nodal prices for real and reactive power as solid and dashed lines, respectively, with the substation LMP included as a dotted black line. The nodal pricing mechanism creates different incentive structures and, by extension, different behaviors between the prosumers. Fundamentally, the agents must balance arbitraging over the substation LMP, minimizing losses, and stabilizing voltage magnitudes, given the apparent power capacity of their inverters.

During the four-day simulation, voltage magnitudes fluctuated between ± 0.10 p.u. and network losses ranged from about 2% and 12% of the total power flow through the substation.

7.5 Conclusion

We propose a practical pricing scheme for DER coordination in distribution networks, accounting for both import costs and voltage stability. In order to compute equilibria for the resulting model, we characterize new generalizable sufficient conditions under which a stochastic game is a Markov potential game and prove that our application satisfies these new conditions. Finally, the proposed mechanism is shown to be efficient on a benchmark distribution network. Interesting directions for future research include accounting for the affect of the aggregate distribution network load on the LMP and rigorously bounding the price-of-anarchy.

[4] While it is difficult to see from the figure, SO did indeed slightly outperform EQ for all five voltage control weights, as expected.

Additional Notation

I_k is the identity matrix of dimension k. $\mathbf{1}$, $\mathbf{0}$, and e_k are the matrix of all ones, the matrix of all zeros, and the kth standard basis vector, respectively, with dimensions inferred from context. $U(\underline{x}, \overline{x})$ is the uniform distribution between its $\underline{x}$ and $\overline{x}$. $N(\mu, \Sigma)$ is the multivariate normal distribution with mean μ and covariance Σ.

Proofs

Proof of Theorem 11

For each agent $i \in \mathcal{N}$, define the *dummy function*

$$\psi_i(s, a) = U_i(s, a) - \phi(s, a). \tag{7.20}$$

Sufficient condition 1 implies that the dummy function does not depend on the local state or action:

$$\psi_i(s_i, s_{-i}, a_i, a_{-i}) = \psi_i(s_{-i}, a_{-i}). \tag{7.21}$$

Combining (7.2), (7.20), and (7.21), agent i's value function can be decomposed as follows

$$V_i^\pi = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \phi(s^t, a^t) \right] + \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \psi_i(s_{-i}^t, a_{-i}^t) \right]. \tag{7.22}$$

Notice that the first term is the desired potential value function Φ^π from Definition 5. To satisfy Definition 5, we need to show that the second term in (7.22) does not depend on π_i. First, bring the expectation inside the summation:

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \psi(s_{-i}^t, a_{-i}^t) \right] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\pi \left[\psi(s_{-i}^t, a_{-i}^t) \right].$$

Clearly, it suffices to show that $\mathbb{E}_\pi \left[\psi(s_{-i}^t, a_{-i}^t) \right]$ is independent of π_i for all t. For compactness, we write $\int_{s_i \in \mathcal{S}_i} \cdot ds_i$ simply as $\int_{s_i} \cdot ds_i$, and likewise for the integral over the nonlocal state space $\mathcal{S}_{-i}$ and relevant subsets of the action space.

$$\begin{aligned}
&\mathbb{E}_\pi \left[\psi(s_{-i}^t, a_{-i}^t) \right] \\
&= \int_s \Pr_\pi(s^t = s) \int_{a_{-i}} \pi_{-i}(a_{-i}|s) \psi_i(s_{-i}, a_{-i}) da_{-i} ds \\
&= \int_{s_{-i}} \Pr_\pi(s_{-i}^t = s_{-i}) \int_{s_i} \Pr_\pi(s_i^t = s_i | s_{-i}^t = s_{-i}) \\
&\quad \int_{a_{-i}} \pi_{-i}(a_{-i}|s) \psi_i(s_{-i}, a_{-i}) da_{-i} ds_i ds_{-i}
\end{aligned}$$

where $\Pr_\pi(\cdot)$ is shorthand for the probability with respect to where $s^0 \sim \rho_0$, $a^t \sim \pi(\cdot|s^t)$, and $s^{t+1} \sim T(\cdot|s^t, a^t)$. We now apply sufficient condition 2 to rewrite π_{-i} without the dependence on s_i:

$$
\begin{aligned}
&= \int_{s_{-i}} \Pr_\pi(s^t_{-i} = s_{-i}) \int_{s_i} \Pr_\pi(s^t_i = s_i | s^t_{-i} = s_{-i}) \\
&\quad \int_{a_{-i}} \pi_{-i}(a_{-i}|s_{-i}) \psi_i(s_{-i}, a_{-i}) da_{-i} ds_i ds_{-i} \\
&= \int_{s^t_{-i}} \Pr_\pi(s^t_{-i} = s_{-i}) \int_{a^t_{-i}} \pi_{-i}(a^t_{-i}|s^t_{-i}) \psi_i(s^t_{-i}, a^t_{-i}) \\
&\quad \int_{s^t_i} \Pr_\pi(s^t_i | s^t_{-i}) ds^t_i da^t_{-i} ds^t_{-i} \\
&= \int_{s^t_{-i}} \Pr_\pi(s^t_{-i} = s_{-i}) \\
&\quad \int_{a^t_{-i}} \pi_{-i}(a^t_{-i}|s^t_{-i}) \psi_i(s^t_{-i}, a^t_{-i}) da^t_{-i} ds^t_{-i}
\end{aligned}
$$

The only remaining nominal dependence on π is through the term $\Pr_\pi(s^t_{-i} = s_{-i})$. We show that this term also does not depend on π by induction. Suppose that $\Pr_\pi(s^t_{-i} = s_{-i})$ is independent of π for some t. In general, we have

$$
\begin{aligned}
\Pr_\pi(s^{t+1}_{-i} = s'_{-i}) = \int_s \Pr_\pi(s^t = s) \int_a \pi(a^t|s^t) \\
\int_{s'_i} T(s'_i, s'_{-i}|s, a) ds'_i da ds
\end{aligned}
$$

Applying sufficient condition 3 gives

$$
= \int_s \Pr_\pi(s^t = s) \int_a \pi(a|s) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) da_i da_{-i} ds
$$

Decomposing the inner integral,

$$
\begin{aligned}
&= \int_s \Pr_\pi(s^t = s) \int_{a_{-i}} \pi_{-i}(a_{-i}|s) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) \\
&\quad \int_{a_i} \pi_i(a_i|s^t) da_i da_{-i} ds \\
&= \int_s \Pr_\pi(s^t = s) \int_{a_{-i}} \pi_{-i}(a_{-i}|s) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) da_{-i} ds
\end{aligned}
$$

Decomposing the outer integral,

$$= \int_{s_{-i}} \Pr_{\pi}(s^t_{-i} = s_{-i}) \int_{s_i} \Pr_{\pi}(s^t_i = s_i | s^t_{-i} = s_{-i}) \int_{a_{-i}} \pi_{-i}(a_{-i}|s) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) da_{-i} ds_i ds_{-i}$$

Applying sufficient condition 2,

$$\begin{aligned}
&= \int_{s_{-i}} \Pr_{\pi}(s^t_{-i} = s_{-i}) \int_{s_i} \Pr_{\pi}(s^t_i = s_i | s^t_{-i} = s_{-i}) \\
&\quad \int_{a_{-i}} \pi_{-i}(a_{-i}|s_{-i}) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) da_{-i} ds_i ds_{-i} \\
&= \int_{s_{-i}} \Pr_{\pi}(s^t_{-i} = s_{-i}) \int_{a_{-i}} \pi_{-i}(a_{-i}|s_{-i}) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) \\
&\quad \int_{s_i} \Pr_{\pi}(s^t_i = s_i | s^t_{-i} = s_{-i}) ds_i da_{-i} ds_{-i} \\
&= \int_{s_{-i}} \Pr_{\pi}(s^t_{-i} = s_{-i}) \\
&\quad \int_{a_{-i}} \pi_{-i}(a_{-i}|s_{-i}) T_{-i}(s'_{-i}|s_{-i}, a_{-i}) da_{-i} ds_{-i}
\end{aligned}$$

The only remaining nominal dependence on π_i is through the term $\Pr_{\pi}(s^t_{-i} = s_{-i})$, which is independent of π_i by the inductive hypothesis. For the base case $t = 0$, note that $\Pr_{\pi}(s^0_{-i} = s_{-i})$ does not depend on π since the initial state is drawn directly from the distribution ρ_0.

Proof of Corollary 1

We will use the following lemma.

Lemma 9. *For any vector $v \in \mathbb{R}^\ell$, matrix $Q \in \mathbb{R}^{\ell \times \ell}$, and partition of indices $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_J\}$ such that $\bigcup_{k=1}^J \mathcal{I}_k = [\ell]$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ for all $i \neq j$,*

$$\nabla_{v_{\mathcal{I}_i}} \left[v^T_{\mathcal{I}_i} \nabla_{v_{\mathcal{I}_i}} (v^T Q v) \right] = \nabla_{v_{\mathcal{I}_i}} \left(v^T Q v + \sum_{j=1}^N v^T_{\mathcal{I}_j} Q_{\mathcal{I}_j \mathcal{I}_j} v_{\mathcal{I}_j} \right).$$

Proof. The expressions are shown to be equivalent through expansion. To avoid clutter, we will write the subscript $\mathcal{I}_i$ as simply i. We will use the E_i to denote the matrix composed of the rows of

the identiy matrix indexed by $\mathcal{I}_i$. Beginning with the left-hand side:

$$\begin{aligned}
&\nabla_{v_i}\left[v_i^T\nabla_{v_i}(v^TQv)\right]\\
&= E_i\nabla\left[(E_iv)^TE_i\nabla(v^TQv)\right]\\
&= E_i\nabla\left[v^TE_i^TE_i(Q+Q^T)v\right]\\
&= E_i\left[E_i^TE_i(Q+Q^T)+(Q+Q^T)E_i^TE_i\right]v\\
&= E_i(Q+Q^T)(I+E_i^TE_i)v
\end{aligned}$$

where we use the fact that $E_iE_i^T = I$. For the right-hand side:

$$\begin{aligned}
&\nabla_{v_i}\left(v^TQv+\sum_{j\in\mathcal{I}}v_j^TQ_{jj}v_j\right)\\
&= E_i\nabla\left[v^T\left(Q+\sum_{j\in\mathcal{I}}E_j^TE_jQE_j^TE_j\right)v\right]\\
&= E_i\left[(Q+Q^T)+\sum_{j\in\mathcal{I}}E_j^TE_j(Q+Q^T)E_j^TE_j\right]v\\
&= E_i(Q+Q^T)(I+E_i^TE_i)v
\end{aligned}$$

where we use the fact that $E_iE_j = 0$ when $i \neq j$. □

We now prove Corollary 1 by checking the sufficient conditions from Theorem 11 in sequence, starting with condition 1. For each agent $i \in \mathcal{N}$, define the local state $s_i = d_i$. The property (7.14) holds if and only if $\phi_{\mathcal{D}}(d,\lambda,a) - U_i(d,\lambda,a)$ does not depend on the local state d_i or the local action a_i. Beginning with the local state, we have

$$\begin{aligned}
\phi_{\mathcal{D}}(d,\lambda,a) - U_i(d,\lambda,a) &= \sum_{j\in\mathcal{N}}u_j(d_j,\tilde{p}_j,\tilde{q}_j) - \tilde{C}(p,q,\lambda)\\
&- \left(u_i(d_i,\tilde{p}_i,\tilde{q}_i) - p_i\frac{\partial}{\partial p_i}C(p,q,\lambda) - q_i\frac{\partial}{\partial q_i}C(p,q,\lambda)\right)\\
&= \sum_{j\neq i}u_j(d_j,\tilde{p}_j,\tilde{q}_j) - \tilde{C}(p,q,\lambda)\\
&- \left(-p_i\frac{\partial}{\partial p_i}C(p,q,\lambda) - q_i\frac{\partial}{\partial q_i}C(p,q,\lambda)\right)
\end{aligned}$$

which does not include d_i as desired. We now want to show the final expression does not depend on the action a_i by way of $(\tilde{p}_i,\tilde{q}_i)$ or, by extension, (p_i,q_i). Clearly, the summation of the other agents' utilities does not depend on a_i, so we need only consider the other terms where (p_i,q_i) appears.

Checking the gradient of these terms:

$$\begin{aligned}
&\nabla_{(p_i,q_i)}\left(p_i\frac{\partial}{\partial p_i}C(p,q,\lambda)+q_i\frac{\partial}{\partial q_i}C(p,q,\lambda)-\tilde{C}(p,q,\lambda)\right)\\
&=\nabla_{(p_i,q_i)}\left[\begin{bmatrix}p_i\\q_i\end{bmatrix}^T\nabla_{(p_i,q_i)}\left(\begin{bmatrix}p\\q\end{bmatrix}^T L(\lambda)\begin{bmatrix}p\\q\end{bmatrix}\right)\right]\\
&-\nabla_{(p_i,q_i)}\left[\begin{bmatrix}p\\q\end{bmatrix}^T L(\lambda)\begin{bmatrix}p\\q\end{bmatrix}+\sum_{j\in\mathcal{N}}\begin{bmatrix}p_j\\q_j\end{bmatrix}^T L_{\mathcal{I}_j\mathcal{I}_j}(\lambda)\begin{bmatrix}p_j\\q_j\end{bmatrix}\right]\\
&=0.
\end{aligned}$$

The first step uses the fact that

$$C(p,q,\lambda)=(w-1)\lambda\sum_{i\in\mathcal{N}}p_i+\begin{bmatrix}p\\q\end{bmatrix}^T L(\lambda)\begin{bmatrix}p\\q\end{bmatrix},$$

which can be checked from (7.4) and (7.5). The linear term cancels trivially and the second step applies Lemma (9). We can conclude from the resulting equality that $\phi_{\mathcal{D}}(d,\lambda,a)-U_i(d,\lambda,a)$ does not depend on a_i, satisfying condition 1.

The local policy of agent j is given by

$$\pi_j^{\theta_j}(a_j|s)=\rho_{\eta_j}\left(\left\{\eta_j : a_j=\mu_j^{\theta_j}(d_j,\alpha_j,\alpha_0,\eta_j)\right\}\right)$$

When $j\neq i$, d_i does not appear in this expression, so condition 2 is satisfied.

By construction of the local state space, we have $s_{-i}=(d_{-i},\alpha)$. The marginal transition density of the nonlocal states is given by

$$\begin{aligned}
T_{-i}(s'_{-i}|s,a)=\rho_{\xi,\omega}(\{(\xi,\omega):&\\
&d'_j=f_j(d_j,\tilde{p}_j,\tilde{q}_j,\omega_j)\quad\forall j\neq i,\\
&\alpha'_i=g_j(\alpha_j,\xi_j)\quad\forall j\in\mathcal{N}\cup\{0\}\})
\end{aligned}$$

Since neither the local state d_i nor the local action a_i (by way of $(\tilde{p}_i,\tilde{q}_i)$ or (p_i,q_i)) appear in this expression, condition 3 is satisfied and the proof is complete.

Experimental Setup

A scalar state variable d_i represents storage unit i's state-of-charge. The state dynamics (7.7) are given by

$$f_i(d_i^t,\tilde{p}_i^t)=d_i^t+\tilde{p}_i^t$$

where $\tilde{p}_i^t$ is also a scalar, since the prosumer owns only a single DER. Note that the storage dynamics include no stochastic component (represented in the general game $\mathcal{G}$ by ω). The feasible set (7.8) is given by

$$\mathcal{P}_i(d_i)=\{(\tilde{p}_i,\tilde{q}_i):-d_i\leq\tilde{p}_i\leq\overline{d_i}-d_i,{\tilde{p}_i}^2+\tilde{q}_i^2\leq b_i^2\}$$

where b_i is the apparent power inverter capacity and $\overline{d_i}$ is the maximum state-of-charge. We choose M_i in (7.12) to be a lazy projection operation that first clips the real component fo a_i into the range $[-d_i, \overline{d_i}]$ and then projects the resulting action onto the 2-norm ball of radius b_i. It is also possible to perform a true projection, but doing so requires solving a convex optimization problem at every timestep and therefore takes much longer to train. Since the authors observed nearly identical results using the true and lazy projections, we choose the latter for ease of reproducibility.

The agents' utility functions u_i are set to zero, so they will only seek to minimize their utility bills given their fixed inelastic demand. We do not include round-trip inefficiencies or battery degradation since the purpose of the example is to demonstrate the fundamental strategic features of the proposed model, but such details can easily be incorporated under the general framework in Section 7.2.

The agent policies are Gaussian with fixed variance where the mean is an affine function of the states:

$$\mu_i^{\theta_i}(d_i^t, \alpha_i^t, \alpha_0^t, \eta_i^t) = \theta_i^{d_i} d_i^t + \theta_i^{\alpha_i} \alpha_i^t + \theta_i^{\alpha_0} \alpha_0^t + \theta_i^0 + \eta_i^t \tag{7.23a}$$

$$\eta_i^t \sim N(\mathbf{0}, \Sigma_{\eta_i}). \tag{7.23b}$$

While more complex neural-network-based policies could also be employed, we found that affine policies are sufficiently expressive for this simple application and exhibited more reliable training.

The inelastic demand and LMP profiles are sinusoids with periods of one day perturbed by normally-distributed random noise. Specifically, the dynamics (7.10a) are given by

$$\begin{aligned} g_i(\alpha_i^t, \xi_i^{t+1}) &= A\alpha_i^t + B\xi_i^t \\ \lambda^t &= \lambda^*(1 + \kappa m^T \alpha_0^t) \\ (\bar{p}_i^t, \bar{q}_i^t) &= (\bar{p}_i^*, \bar{q}_i^*)(1 + \kappa m^T \alpha_0^t) \\ \xi_i^t &\sim N(0, \Sigma_\xi) \end{aligned}$$

where

$$\begin{aligned} A &= \begin{bmatrix} \cos\frac{\pi}{12} & -\sin\frac{\pi}{12} & & \mathbf{0} & \\ \sin\frac{\pi}{12} & \cos\frac{\pi}{12} & & & \\ & & & \mathbf{0} & 0 \\ & \mathbf{0} & & I_\tau & \mathbf{0} \end{bmatrix} \\ B &= \begin{bmatrix} 0 & 0 & e_1^T \end{bmatrix}^T \\ m &= \begin{bmatrix} 1 & 0 & \sigma_\xi \mathbf{1}^T \end{bmatrix} \\ \Sigma_\xi &= z\mathbf{1}\mathbf{1}^T + \mathrm{diag}((1-z)\mathbf{1}). \end{aligned}$$

κ is the amplitude factor of the sinusoids. λ^* and $\tilde{p}_i^*$ are the mean LMP and agent inelastic demand, respectively. τ is the noise duration, σ_ξ is the standard deviation factor for the exogenous parameters, and z is the correlation coefficient.

The initial state distribution is defined by

$$\begin{aligned} d_i^0 &\sim U(0, \overline{d_i}) \\ \alpha_i^0 &= \begin{bmatrix} \cos\frac{2\pi}{24}(t_0 + \delta_i) & \sin\frac{2\pi}{24}(t_0 + \delta_i) & 0 & 0 & 0 \end{bmatrix} \\ t_0 &\sim U(0, 23) \\ \delta_i &\sim U(\underline{\delta}, \overline{\delta}) \quad \forall i \in \mathcal{N} \\ \delta_0 &= 0 \end{aligned}$$

where t_0 represents the starting hour of the day of the simulation and δ_i is the phase difference between agent i's inelastic demand and the LMP, randomized between agents to introduce heterogeneity.

The network model is the 18-bus radial distribution system from [114], one of seven distribution benchmark cases distributed with MATPOWER. One storage-equipped agent is located at each of the 15 load buses. To approximate a heavily-loaded distribution network with losses on the order of 10%, the nominal loads are tripled.

$\bar{p}^*$ and $\bar{q}^*$ are set to the (tripled) real and reactive nominal load values from the MATPOWER case file and λ^* is set to 1. Branch parameters r and x are also taken from the MATPOWER case file. We further set $\kappa_0 = 1/2$, $\sigma_\xi = 1/10$, $z = .9$, $\tau = 3$, $\underline{\delta} = 3$, and $\overline{\delta} = 9$. For the storage units, we set $b_i = 1.5\sqrt{\bar{p}_i^{*2} + \bar{q}_i^{*2}}$, $\overline{d_i} = 6\widetilde{\bar{p}_i}$, and $\Sigma_{\eta_i} = \mathrm{diag}((\bar{p}_i^{*2}, \bar{q}_i^{*2}))$. Finally, we set the discount factor $\gamma = 0.99$.

Part III

Strategic Information Alignment

Chapter 8

Inferring Agents' Objectives in Feedback Dynamic Games

In the preceding chapters, we introduced efficient computational methods for determining various game-theoretic equilibria under the assumption of complete information, wherein agents possess full knowledge of each other's private information, such as objectives and constraints. In this chapter, we relax this complete-information assumption and focus on scenarios involving incomplete or asymmetric information. Specifically, we propose an efficient approach grounded in differentiable optimization techniques to infer maximum-likelihood estimates of agents' objectives from observed data in feedback dynamic games. This chapter is based on the paper [180], co-authored with Chih-Yuan Chiu, Lasse Peters, Somayeh Sojoudi, Claire J. Tomlin, and David Fridovich-Keil.

8.1 Background

The safety and efficiency of urban traffic rely heavily on the ability of each participant to predict the effects of their actions on others' decisions [218, 275]. For example, drivers on a highway may wish to halt an overtaking maneuver if they believe the other drivers are aggressive, and some drivers may decelerate their cars to avoid collision if they believe that another driver wishes to merge.

A powerful paradigm for modeling the interdependence of decisions in multi-agent settings is provided *general-sum dynamic games* [24, 144]. A Nash equilibrium solution of a game-theoretic model can be used to *simultaneously* predict the actions of all agents in the scene. This equilibrium solution is particularly expressive when the game possesses a feedback information structure. In this case, each equilibrium strategy explicitly accounts for the dynamically evolving information available to each player over time.

Despite the theoretical attractiveness of this modeling paradigm, in reality, autonomous agents often have only limited information available about the world around them. For example, in urban traffic an autonomous agent typically has incomplete knowledge of the objectives of other players. To address this challenge, recent works on *inverse dynamic game theory* [265, 251, 221] recover these objectives from past trajectory data. Moreover, in realistic applications, only noisy

sensor measurements of agents' states are available. This partial observability further complicates the inverse game problem, and existing work [251] treats this case in the open-loop information structure.

In this chapter, we present a gradient-based solver for inverse dynamic games, under the state feedback information structure. Our solver can recover objectives from partial state observations of incomplete trajectories. Both of these effects are common in robotics due to noisy perception and occlusions. We show that our algorithm converges reliably in practice, and demonstrate the superior robustness and generalization performance as compared with a baseline method which learns cost functions under the open-loop assumption [251], when the observation data is from a group of players pursuing a feedback Nash equilibrium strategy.

Our contributions are threefold. Firstly, we characterize the solution set of the inverse feedback dynamic game problem. In particular, we show that the set of the global minima could be nonconvex and disconnected, and discuss regularization schemes to mitigate this problem. Secondly, we show the differentiability of the loss function in linear quadratic games and propose a computationally efficient procedure to approximate the gradient for nonlinear games. Finally, we propose an efficient first-order coordinate-descent solver for the inverse feedback game problem, using noisy partial observations of an incomplete expert state trajectory. Experimental results show that our method reliably converges for inverse feedback games with nonlinear dynamics and is able to learn nonconvex costs. Moreover, the converged cost function can accurately predict the feedback Nash equilibrium state trajectories even for unseen initial states.

8.2 Related works

Non-cooperative Dynamic Games

Non-cooperative dynamic game theory [24, 144] provides a formal framework for analyzing strategic interaction in a multi-agent setting [69, 24, 173]. In non-cooperative games, each player minimizes its own individual cost function; since players' costs may not be mutually aligned, the resulting equilibrium behavior is generally competitive. Among different equilibrium concepts, the Nash equilibrium has been extensively studied because of its representative power of capturing many non-cooperative behaviors arising in real-world multi-agent systems [106, 274].

Recent advances in the literature aim to develop efficient solutions to Nash equilibrium problems in dynamic games. Though the solutions for the open-loop and feedback Nash equilibrium in linear quadratic (LQ) games are well understood [24], for nonlinear games there is no closed-form solution in general. The work [259] characterizes the local Nash solution concept for open-loop Nash equilibrium. In the feedback setting, numerous approaches have been proposed under various special cases [290, 161]. A value iteration based approach for computing feedback Nash equilibria of nonlinear games without constraints is introduced in [128]. Recently, a set of KKT conditions for feedback Nash equilibria in constrained nonlinear games is derived in [167]. Computing a feedback Nash equilibrium is challenging due to the nested KKT conditions in different time steps.

Our work draws upon the ILQGames [104] framework, which at each iteration solves a linear-quadratic game that approximates the original game. The construction of the approximate game parallels the iterative linearization and quadraticization methods of iterative LQR [188], and the dynamic programming equations that characterize equilibrium strategies in linear quadratic dynamic games [24]. This approach differs from the ALGames [65] method, which computes an open-loop Nash equilibrium strategy.

Inverse Non-cooperative Dynamic Games

In contrast to the forward game problem of computing a strategy in dynamic games, an inverse game problem amounts to finding objectives for all agents such that the corresponding strategic (e.g., Nash equilibrium) interactions reproduce expert demonstrations. The inverse game problem is important because it paves the way for an agent to understand the preferences which explain other agents' behavior, which may facilitate more efficient multi-agent interaction and coordination.

The problem of inverse infinite-horizon LQ games is considered in [143], where the set of cost functions whose feedback Nash equilibrium strategies coincide with an expert strategy is derived. In [265, 326], the two-player inverse LQ game is solved by transforming the problem to an inverse optimal control under the assumption that the control input data of one player is known. Two methods based on the KKT conditions of an open-loop Nash equilibrium are proposed for open-loop general-sum differential games in [220]. Several necessary conditions for open-loop Nash equilibria are proposed in [222] and used for developing an inverse game solution for some classes of open-loop games.

Recently, an efficient bilevel optimization framework [251] based on the open-loop Nash equilibrium KKT conditions was proposed for solving inverse games with an open-loop Nash assumption. Another line of work on inferring costs in open-loop games [15, 142, 81] proposes to minimize the residual violation of the KKT conditions. This KKT residual framework assumes the knowledge of complete trajectory data and is a convex problem. Given the difficulty of evaluating KKT conditions for feedback Nash equilibria in nonlinear games [167], the extension of the KKT residual method to feedback nonlinear games may be subject to numerical difficulty.

A bilevel optimization approach for inverse feedback game problem is proposed in [219], with the assumption that both the expert state and control trajectories are observed without noise. In addition, an inverse game solver is proposed in [212] where they infer the players' cost functions with the assumption that the expert strategy follows a new concept called Maximum Entropy Nash Equilibrium. To the best of the authors' knowledge, there is no work on inferring cost functions of nonlinear dynamic games under feedback Nash equilibrium condition, from noisy partial state observation and incomplete trajectory data.

8.3 Preliminaries

Consider an N-player, T-stage, deterministic, discrete-time dynamic game, with a state $x_t^i \in \mathbb{R}^{n_i}$ and control input $u_t^i \in \mathbb{R}^{m_i}$ for each player $i \in [N] := \{1, \cdots, N\}$, $t \in [T]$. Let the dimension

of the joint state and control input be $n := \sum_{i=1}^{N} n_i$ and $m := \sum_{i=1}^{N} m_i$, respectively. We denote by $x_t := [x_t^1, \ldots, x_t^N] \in \mathbb{R}^n$ and $u_t := [u_t^1, \ldots, u_t^N] \in \mathbb{R}^m$ the joint state and joint control at time $t \in [T]$, respectively. The joint dynamics for the system is given by the differentiable dynamics map $f_t(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$:

$$x_{t+1} = f_t(x_t, u_t), \qquad \forall t = 1, \cdots, T. \tag{8.1}$$

We denote by $\mathbf{f} := \{f_t\}_{t=1}^T$ the set of dynamics across all the time instances within horizon T. We define $\mathbf{x} := \{x_t\}_{t=1}^T$ and $\mathbf{u} := \{u_t\}_{t=1}^T$ to be a state trajectory and control trajectory, respectively, if $x_{t+1} = f(x_t, u_t)$, for each $t \in [T]$. The objective of each agent i is to minimize its overall cost, given by the sum of its running costs $g_t^i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ over the time horizon:

$$J^i(\mathbf{x}, \mathbf{u}) := \sum_{t=1}^{T} g_t^i(x_t, u_t) \tag{8.2}$$

Define $g_t := \{g_t^1, g_t^2, \cdots, g_t^N\}$, $t \in [T]$. We denote by $\mathbf{g} := \{g_t\}_{t=1}^T$ the set of cost functions for all the agents within horizon T.

To minimize (8.2), each player uses their observations of the environment to design a sequence of control inputs to deploy during the discrete time interval $[T]$. The information available to each player at each time characterizes the *information pattern* of the dynamic game, and plays a major role in shaping the optimal responses of each player [24]. Below, we explore two such information patterns—*feedback* and *open-loop*.

Nash Solutions in Feedback Strategies

Under the state feedback information pattern, each player observes the state x_t at each time t, and uses this information to design a *feedback strategy* $\gamma_t^i : \mathbb{R}^n \to \mathbb{R}^{m_i}$, given by: $u_t^i := \gamma_t^i(x_t)$, for each $i \in [N]$ and $t \in [T]$. Let $\gamma_t(x_t) := [\gamma_t^1(x_t), \gamma_t^2(x_t), \ldots, \gamma_t^N(x_t)] \in \mathbb{R}^m$.

Following the notation of [24], we denote by Γ_t^i the set of all state feedback strategies of player i, for each $i \in [N]$. Under this *feedback* information pattern, the Nash equilibrium of the dynamic game is as defined below.

Definition 6 (Feedback Nash Equilibrium (FBNE) [24, Ch. 6]**).** *The set of control strategies $\{\gamma_t^{1*}, \cdots, \gamma_t^{N*}\}_{t=1}^T$ is called a feedback Nash equilibrium if no player is incentivized to unilaterally alter its strategy. Formally:*

$$\begin{aligned} &W_t^{i*}\big(x_t, [\gamma_t^{1*}(x_t), \ldots, \gamma_t^{i*}(x_t), \ldots, \gamma_t^{N*}(x_t)]\big) \\ &\leq W_t^{i*}\big(x_t, [\gamma_t^{1*}(x_t), \ldots, \gamma_t^{i}(x_t), \ldots, \gamma_t^{N*}(x_t)]\big), \forall \gamma_t^i \in \Gamma_t^i, \forall t \in [T]. \end{aligned} \tag{8.3}$$

where $W_t^{i*}(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $t \in [T]$ is the optimal state-action function defined as follows,

$$\begin{aligned} W_T^{i*}(x_T, u_T) &:= g_T^i(x_T, u_T) \\ W_t^{i*}(x_t, u_t) &:= g_t^i(x_t, u_t) + V_{t+1}^{i*}(x_{t+1}), \forall t \in [T-1], \\ V_t^{i*}(x_t) &:= W_t^{i*}(x_t, [\gamma_t^{1^*}(x_t), \ldots, \gamma_t^{N^*}(x_t)]), \forall t \in [T]. \end{aligned} \tag{8.4}$$

We define $\mathbf{x}$ and $\mathbf{u}$ to be a FBNE state trajectory and a FBNE control trajectory, respectively, if $u_t^i = \gamma_t^{i*}(x_t)$, for each $i \in [N]$ and $t \in [T]$. We denote by $\xi(\mathbf{f}, \mathbf{g})$ the set of all FBNE state trajectories in the game defined by the dynamics $\mathbf{f}$ and cost functions $\mathbf{g}$.

Remark 13 (Strong Time Consistency)**.** *The FBNE conditions of* (8.3) *implicitly enforce strong time-consistency [24, Def. 5.14] of the equilibrium strategies. That is, FBNE does not admit arbitrary feedback strategies, but imposes the additional condition that those strategies must also be in equilibrium for any subgame starting at a later stage from an arbitrary state.*

Nash Solutions in Open-loop Strategies

In contrast, under the open-loop information pattern, each player only observes the initial state x_1. In this case, the strategy for each player $i \in [N]$ is a map from x_1 to $\{u_1^i, u_2^i, \cdots, u_T^i\}$, which we denote by $\phi^i(\cdot) : \mathbb{R}^n \to \underbrace{\mathbb{R}^{m_i} \times \cdots \times \mathbb{R}^{m_i}}_{T}$. Let Φ^i be the set of all open-loop strategies of the player i, $i \in [N]$. The corresponding *open-loop Nash equilibrium* is defined as follows.

Definition 7 (**Open-Loop Nash Equilibrium (OLNE)** [24, Ch. 6])**.** *The tuple of control strategies $\{\phi_1^*, \cdots, \phi_N^*\}$ is called an open-loop Nash equilibrium if no player is incentivized to unilaterally alter its sequence of control inputs. Formally:*

$$\begin{aligned} &J^i\big(\mathbf{x}, [\phi^{1^*}(x_1), \cdots, \phi^{i^*}(x_1), \cdots, \phi^{N^*}(x_1)]\big) \\ \leq &J^i\big(\mathbf{x}, [\phi^{1^*}(x_1), \cdots, \phi^{i}(x_1), \cdots, \phi^{N^*}(x_1)]\big), \forall \phi^i \in \Phi^i, \forall x_1 \in \mathbb{R}^n. \end{aligned} \tag{8.5}$$

Remark 14. *The OLNE definition does not imply the strong time-consistence as in the feedback counterpart [24].*

Feedback vs. Open-loop Nash Equilibria

In this subsection, we demonstrate the difference between open-loop and feedback Nash equilibria and show the necessity of developing specific solutions for cost inference problems with the feedback information pattern, instead of applying existing work with the open-loop assumption [250]. To this end, we introduce below several linear-quadratic (LQ) games where the open-loop Nash equilibrium (OLNE) and feedback Nash equilibrium (FBNE) state trajectories differ substantially. LQ games are a class of dynamic games with dynamics and player objectives of the form in (8.6) and (8.7), respectively,

$$x_{t+1} = A_t x_t + \sum_{i \in [N]} B_t^i u_t^i, \ \forall t \in [T], \tag{8.6}$$

$$g_t^i(x_t, u_t) = \frac{1}{2}(x_t^\top Q_t^i x_t + \sum_{j \in [N]} {u_t^j}^\top R_t^{ij} u_t^j), \forall t \in [T], \forall i \in [N], \tag{8.7}$$

where matrices $\{A_t, B_t^i\}$, positive semidefinite matrix Q_t^i and positive definite matrix R_t^{ij} are defined with appropriate dimensions, for each $i, j \in [N]$ and $t \in [T]$.

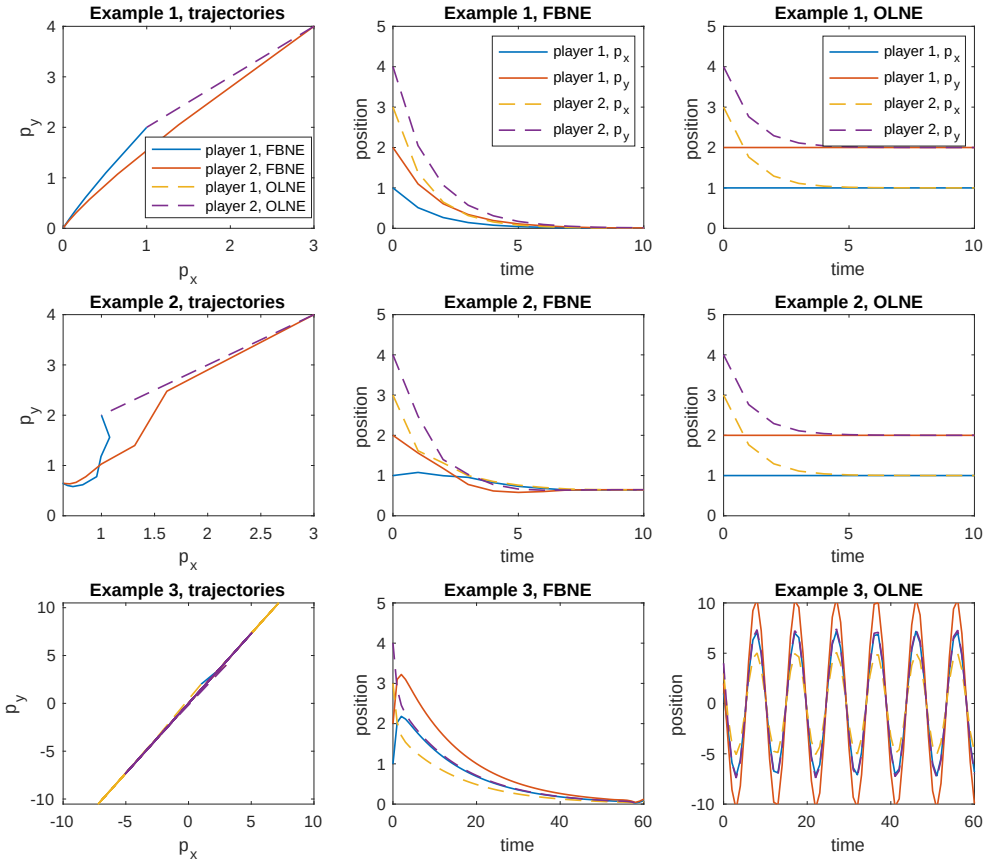


Figure 8.1: Examples of cost functions that yield trajectories that are different under the OLNE and FBNE assumptions.

Case Study: We consider a two-player LQ game with a state vector $x_t = [p^1_{x,t}, p^1_{y,t}, p^2_{x,t}, p^2_{y,t}]$, where $p^i_{x,t}$ and $p^i_{y,t}$ are the x- and y-coordinates of agent $i \in \{1, 2\}$, respectively. Let $u^i_t \in \mathbb{R}^2$ be the control input for the i-th agent, $i \in \{1, 2\}$. In this setting, we consider a class of games in which the first agent wants to drive the second agent to the origin, while the second agent wants to catch the first agent. The agents' joint dynamics and costs at time $t \in [T]$ are specified as follows:

$$\begin{aligned} x_{t+1} &= \begin{bmatrix} I_2 & 0 \\ 0 & I_2 \end{bmatrix} x_t + \begin{bmatrix} I_2 \\ 0 \end{bmatrix} u^1_t + \begin{bmatrix} 0 \\ I_2 \end{bmatrix} u^2_t, \\ g^1_t(x_t, u_t) &= \|p^2_{x,t}\|^2_2 + \|p^2_{y,t}\|^2_2 + \|u^1_t\|^2_2, \\ g^2_t(x_t, u_t) &= \|p^2_{x,t} - p^1_{x,t}\|^2_2 + \|p^2_{y,t} - p^1_{y,t}\|^2_2 + \|u^2_t\|^2_2, \end{aligned} \tag{8.8}$$

where I_2 is the 2×2 identity matrix. We visualize the unique FBNE and OLNE state trajectories of this example in the first row in Fig. 8.1. If we modify the cost function of the first player such that it wants to lead the x- and y-position of the second player to be aligned with each other, i.e.,

$$\hat{g}^1_t(x_t, u_t) := \|p^2_{x,t} - p^2_{y,t}\|^2_2 + \|u^1_t\|^2_2, \tag{8.9}$$

then, the unique FBNE and OLNE state trajectories are still different, as shown in the second row of Fig. 8.1. Moreover, observations of players may be noisy in practice. To illustrate this, we consider a task where the two agents want to catch each other, but the first player's observation of the second player's position is inaccurate. We modify the first player's cost in (8.8) as follows:

$$\hat{\hat{g}}_t^1(x_t, u_t) := \|p_{x,t}^1 - 2p_{x,t}^2\|_2^2 + \|p_{y,t}^1 - 2p_{y,t}^2\|_2^2 + \|u_t^1\|_2^2. \tag{8.10}$$

The third row of Fig. 8.1 reveals that the FBNE state trajectory is robust to inaccurate observations, but the unique OLNE state trajectory is not.

Thus, it is readily apparent that the OLNE and FBNE state strategies can be substantially different even for fixed cost functions. This difference in expressive power may be understood as a consequence of the strong time consistency property, which is enforced in the feedback information structure but not in the open-loop setting, per Remarks 13 and 14. A similar problem arises in the cost inference problem, where the existing OLNE cost inference algorithms may fail to infer the correct cost function in feedback games.

8.4 Problem Statement

Let $\mathbf{x}$ be an expert FBNE state trajectory under the nonlinear dynamics $\mathbf{f}$ but unknown cost functions $\{g_t^i\}_{t=1,i=1}^{T,N}$. Let $\mathcal{T} \subseteq [T]$ be the set of observed time indices of the trajectory $\mathbf{x}$. We denote by $\mathbf{y}_\mathcal{T} := \{y_t\}_{t\in\mathcal{T}}$ the observation data of $\mathbf{x}$, where $y_t \in \mathbb{R}^\ell$ is a partial observation of the state, composed of certain coordinates of x_t corrupted by noise. The task is to infer the cost function of each player such that those inferred costs jointly yield a FBNE state trajectory that is as close as possible to the observed trajectory. We parameterize the cost of the player i by a vector $\theta^i \in \mathbb{R}^{d_i}$, and let $\theta := [\theta^1, \theta^2, \ldots, \theta^N] \in \mathbb{R}^d$. Denote by $g_{t,\theta}^i(x_t, u_t) = \sum_{j=1}^{d_i} \theta_j^i b_{t,j}^i(x_t, u_t)$ player i's parameterized cost at time $t \in [T]$, for some basis functions $\{\{b_{t,j}^i\}_{j=1}^{d_i}\}_{t=1,i=1}^{T,N}$. Define $\mathbf{g}_\theta := \{g_{t,\theta}^i\}_{t=1,i=1}^{T,N}$. Formally, this problem is of the form:

$$\begin{aligned} \min_{\theta, x_1, \hat{\mathbf{x}}} \quad & -p(\mathbf{y}_\mathcal{T}|\hat{\mathbf{x}}) \\ \text{s.t.} \quad & \hat{\mathbf{x}} \in \xi(\mathbf{f}, \mathbf{g}_\theta, x_1), \end{aligned} \tag{8.11}$$

where $p(\cdot|\cdot)$ is the likelihood function corresponding to a known sensor model and $\xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ represents the set of state trajectories from the initial condition $x_1 \in \mathbb{R}^n$ following a FBNE strategy, under the cost set $\mathbf{g}_\theta$. Due to the noisy partial observation, x_1 is not assumed to be known and instead needs to be inferred as well in (8.11). Note that the above formulation can also be extended to the cases where multiple partially observed incomplete trajectories from different initial conditions are available.

Running example: We consider a highway platooning scenario where player 1 wants to guide player 2 to a particular lane of the road. The joint state vector is $x_t = [p_{x,t}^1, p_{y,t}^1, \beta_t^1, v_t^1, p_{x,t}^2, p_{y,t}^2, \beta_t^2, v_t^2]$.

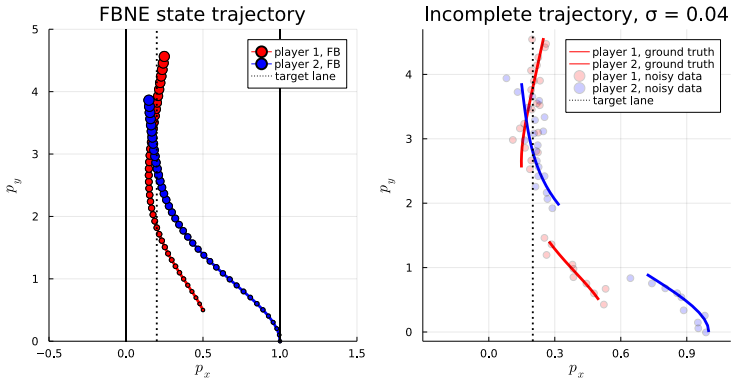


Figure 8.2: Visualization of the running example.

The time horizon $T = 40$. The dynamics model for the player i is:

$$\begin{bmatrix} p^i_{x,t+1} \\ p^i_{y,t+1} \\ \beta^i_{t+1} \\ v^i_{t+1} \end{bmatrix} = \begin{bmatrix} p^i_{x,t} \\ p^i_{y,t} \\ \beta^i_t \\ v^i_t \end{bmatrix} + \Delta T \begin{bmatrix} v^i_t \cos(\beta^i_t) \\ v^i_t \sin(\beta^i_t) \\ \omega^i_t \\ a^i_t \end{bmatrix} \tag{8.12}$$

where ΔT is a time discretization constant and $u^i_t = [\omega^i_t, a^i_t] \in \mathbb{R}^2$ is the control input for player $i \in [N]$. Let p^*_x be the target lane that player 1 wants to guide player 2 to. We parameterize the cost function of the player i by $\theta^i \in \mathbb{R}^2$,

$$\begin{aligned} g^1_{t,\theta}(x_t, u_t) &= \theta^1_1 \|p^1_{x,t}\|^2_2 + \theta^1_2 \|p^2_{x,t} - p^*_x\|^2_2 + \|u^1_t\|^2_2 \\ g^2_{t,\theta}(x_t, u_t) &= \theta^2_1 \|p^2_{x,t} - p^1_{x,t}\|^2_2 + \theta^2_2 \|v^2_t - 1\|^2_2 + \|u^2_t\|^2_2, \forall t \in [T]. \end{aligned} \tag{8.13}$$

The ground truth solution is $\theta^* = [0, 8, 4, 4]$. We assume that there is a period of occlusion happening from the time index $t = 11$ to $t = 19$, and the observed time index set is $\mathcal{T} = \{1, 2, \ldots, 10, 20, 21, \ldots, 40\}$. Also, it may be difficult for a human driver to measure other vehicles' velocity accurately, and therefore we assume that partial observation data $\mathbf{y}_\mathcal{T}$ excludes the velocity of both cars in the data set, and is further subject to Gaussian noise of standard deviation σ. The initial condition x_1 is not known and needs to be inferred. We visualize the ground truth solution in the first subplot of Fig. 8.2 and the noisy incomplete trajectory data in the second subplot of Fig. 8.2.

The many challenges of the above problem include: (a) partial observation; (b) noisy and incomplete expert trajectory data; and (c) the difficulty of evaluating and differentiating the objective in (8.11), due to the challenge of computing a FBNE strategy in nonlinear games [167]. In the following sections, we will characterize the complexity of this inverse feedback game problem and propose an efficient solution.

8.5 Results: From Characterization to Computation

In this section, we first characterize the complexity of the inverse feedback game problem (8.11). In particular, we will show the nonconvexity of the loss function and the existence of multiple isolated *global* minima. Based on this observation, we discuss regularization schemes that can mitigate this issue. Our main contribution is to characterize the differentiability of the inverse feedback game loss function in (8.11). Finally, we present a gradient approximation scheme that can be used in a first-order optimization formulation.

Characterization of the Inverse Feedback Dynamic Game Problem

The inverse feedback dynamic game problem (8.11) is a constrained optimization problem, which is hard to solve due to the nonconvexity of the set $\xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$. With a slight abuse of notation, we denote by $\hat{\mathbf{x}}(\mathbf{f}, \mathbf{g}_\theta, x_1) \in \xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ a FBNE state trajectory. To simplify the problem, we transform (8.11) to an unconstrained problem by substituting a forward game solution $\hat{\mathbf{x}}(\mathbf{f}, \mathbf{g}_\theta, x_1)$ into the likelihood function $p(\mathbf{y}_\mathcal{T}|\hat{\mathbf{x}})$, as follows:

$$\hat{L}(\theta, x_1) := -p(\mathbf{y}_\mathcal{T}|\hat{\mathbf{x}}(\mathbf{f}, \mathbf{g}_\theta, x_1)). \tag{8.14}$$

The minimizer of (8.14) is a local optimum to the original problem (8.11) and becomes global when $\xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ contains only a single element.

Before we dive into the nonlinear setting, let us first consider a simplified LQ case to highlight the main challenges associated with the optimization of this loss. In the LQ case, the *evaluation* of the loss (8.14) is straightforward if there exists a closed-form expression for $p(\mathbf{y}_\mathcal{T}|\hat{\mathbf{x}})$, e.g., under a Gaussian observation model. Even in that setting, however, it is important to realize that the problem remains nonconvex, as shown in Fig. 8.3. The following proposition makes this challenge explicit, and the proof can be found in the Appendix.

Proposition 10. *There exists an inverse LQ game problem* (8.11)*: (a) whose global minima are isolated, and (b) for which there exist multiple cost functions that exactly match expert data from any initial condition, when there is no observation noise.*

Remark 15. *Proposition 10 does not imply that any inverse LQ game problem will suffer from the multiple global minima issue. Instead, Proposition 10 suggests that simply normalizing the cost vector does not rule out the possibility of having multiple global solutions. That is, there exist two cost parameter vectors which are linearly independent, but generate the same FBNE state trajectories for any given initial state. This non-injective mapping from the cost parameter space to the FBNE state trajectory space is a fundamental problem in inverse feedback games, and is not particular to the formulation* (8.11)*. In practice, this multiple global minima issue could be mitigated by adding L_2 regularization, as visualized in Fig. 8.3.*

Though being nonconvex, the loss function $\hat{L}(\theta, x_1)$ is differentiable with respect to both θ and x_1 under the condition of Theorem 3.2 in [167], which follows from the implicit function theorem [162]. Inspired by the success of gradient-based methods in non-convex optimization with

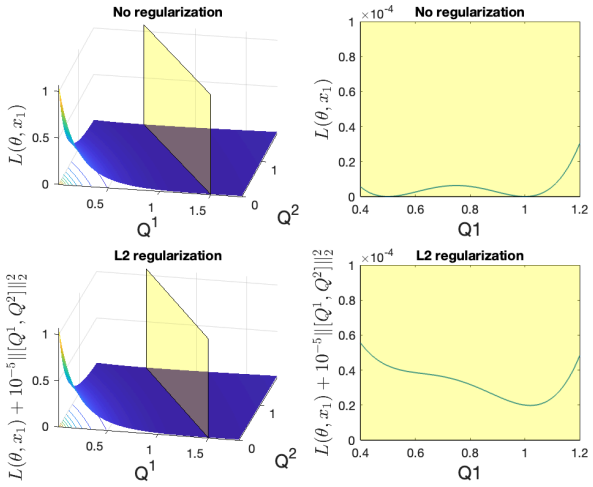


Figure 8.3: Visualization of the loss function $L(\theta, x_1)$ of the LQ game specified in (8.16) and (8.17), and its L_2 regularization, with an initial condition $x_1 = 1$. We adopt Gaussian likelihood function. The yellow hyperplane is drawn according to $2Q^1 + Q^2 = 3$. With L_2 regularization, the number of global minima is reduced.

differentiable objective functions [233, 35, 285], one natural idea is to apply gradient descent to minimize $\hat{L}(\theta, x_1)$. In what follows, we discuss efficient ways to evaluate and differentiate $\hat{L}(\theta, x_1)$ in nonlinear games.

Efficient Computation for a FBNE State Trajectory in Nonlinear Games

It is easy to evaluate $\hat{L}(\theta, x_1)$ for LQ games, but when dynamics are nonlinear or objectives are non-quadratic, the problem becomes more challenging [167]. In forward games, this challenge can be addressed by using the ILQGames algorithm [104], which finds approximate local FBNE solutions in smooth non-LQ dynamic games. Given the effectiveness of this approximation scheme in those domains, we also adopt it as a submodule for evaluating the loss $\hat{L}(\theta, x_1)$. Akin to the ILQR method [209, 188], in each step of the ILQGames algorithm, the system dynamics $x_{t+1} = f(x_t, u_t)$ and the costs $\{g_t^i(x, u)\}_{t=1,i=1}^{T,N}$ are linearized and quadraticized, respectively, around a state trajectory $\mathbf{x}$ and a control trajectory $\mathbf{u}$. A FBNE strategy for each player of the derived LQ game is then used to update the state and control trajectories. This iteration continues until a convergence criterion is satisfied.

To be more specific, we approximate $\hat{L}(\theta, x_1)$ by a new loss function $\tilde{L}(\theta, x_1)$ defined as,

$$\hat{L}(\theta, x_1) \simeq \tilde{L}(\theta, x_1) := -p\big(\mathbf{y}_{\mathcal{T}} | \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)\big) \tag{8.15}$$

where $\mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)$ represents a FBNE state trajectory from initial condition x_1, for the LQ game defined by the linearized dynamics $\tilde{\mathbf{f}}_\theta$, quadraticized cost set $\tilde{\mathbf{g}}_\theta := \{\tilde{\mathbf{g}}_{t,\theta}^i\}_{t=1,i=1}^{T,N}$ at the converged solution returned by ILQGames solver. Note that the linearized dynamics $\tilde{\mathbf{f}}_\theta$ depend upon θ via the state trajectory about which $\mathbf{f}$ is linearized; this trajectory is simulated under the feedback policy returned by ILQGames, where the policy depends upon costs $\mathbf{g}_\theta$.

Differentiating the Loss in the Inverse Feedback Game Problem

The challenge of computing a feedback Nash equilibrium strategy not only makes the evaluation of the loss function $\hat{L}(\theta, x_1)$ hard, but also renders differentiation difficult. In this work, we approximate the gradient of $\hat{L}(\theta, x_1)$ using a similar idea as the ILQGames algorithm in the previous section. In other words, we propose to use the LQ approximation of the nonlinear game specified by $\tilde{\mathbf{f}}_\theta$ and $\tilde{\mathbf{g}}_\theta$ to derive an approximation to the gradient of $\hat{L}(\theta, x_1)$. Note that $\tilde{g}^i_{t,\theta}(x,u) = \sum_{j=1}^{d_i} \theta^i_j \tilde{b}^i_{t,j,\theta}(x,u)$, where $\tilde{b}^i_{t,j,\theta}(x,u) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the j-th quadraticized cost basis function. By the chain rule, we have

$$\frac{\partial \tilde{L}(\theta, x_1)}{\partial \theta^i_j} = -\nabla_{\mathbf{x}} p(\mathbf{y}_{\mathcal{T}}|\mathbf{x})\Big|_{\mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)} \cdot \frac{\partial \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)}{\partial \theta^i_j},$$
$$\frac{\partial \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)}{\partial \theta^i_j} = \Big(\nabla_{\tilde{\mathbf{f}}_\theta} \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1) \frac{\partial \tilde{\mathbf{f}}_\theta}{\partial \theta^i_j} + \nabla_{\tilde{\mathbf{g}}_\theta} \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1) \frac{\partial \tilde{\mathbf{g}}_\theta}{\partial \theta^i_j}\Big).$$

The complexity of differentiating $\tilde{L}(\theta, x_1)$ comes from the fact that the linearized dynamics and the quadraticized costs are functions of θ implicitly, which makes the total derivative hard to compute. We propose to approximate the above gradient by treating the linearized $\tilde{\mathbf{f}}_\theta$ and each quadraticized cost basis function $\tilde{b}^i_{t,j,\theta}$ as constants with respect to θ, denoted by $\tilde{\mathbf{f}}$ and $\tilde{b}^i_{t,j}$, and only compute the partial derivative with respect to θ, rather than the total derivative:

$$\frac{\partial \tilde{L}(\theta, x_1)}{\partial \theta^i_j} \simeq -\nabla_{\mathbf{x}} p(\mathbf{y}_{\mathcal{T}}|\mathbf{x})\Big|_{\mathbf{x}(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}_\theta, x_1)} \cdot \frac{\partial \mathbf{x}(\tilde{\mathbf{f}}, \{\sum_{j=1}^{d_i} \theta^i_j \tilde{b}^i_{t,j}\}_{t=1,i=1}^{T,N}, x_1)}{\partial \theta^i_j}.$$

This is based on the observation that at the convergence of the forward ILQGames solver, the linearized dynamics are a good approximation of the full nonlinear dynamics $\mathbf{f}$, so long as the cost parameter being perturbed remains sufficiently small. We adopt a similar approximation for the gradient $\nabla_{x_1} \tilde{L}(\theta, x_1)$ by fixing the linearized dynamics and quadraticized costs and obtaining the partial derivative with respect to x_1.

In summary, we approximate $\nabla \hat{L}(\theta, x_1)$ by $\nabla \tilde{L}(\theta, x_1)$. In practice, $\nabla \tilde{L}(\theta, x_1)$ can be efficiently computed by automatic differentiation [238, Ch. 8]. As exemplified in Fig. 8.4, the proposed gradient approximation is virtually always a descent direction and therefore aligns well with the true gradient of $\hat{L}(\theta, x_1)$.

An Inverse Feedback Game Solver

In this subsection, we present a solver for the inverse feedback game problem (8.11). In what follows, we first discuss how the three challenges mentioned in Section 8.4 are handled in our solver. We then introduce the proposed solver in Algorithm 6.

The first two challenges on noisy partial observation and incomplete trajectory data are handled by maintaining an estimate of the full initial condition and a noise-free state-input trajectory.

Algorithm 6: Inverse Iterative LQ (i^{2}LQ) Games

Data: Horizon $T > 0$, initial solution $\theta^{(0)} \in \mathbb{R}^d$, observed time index set $\mathcal{T} \subseteq [T]$, observation data $\mathbf{y}_{\mathcal{T}}$, max iteration number K , tolerance ϵ.

Result: Inferred cost parameter $\hat{\theta}$ and $\hat{x}_1$

for $k = 0, 1, \ldots, K$ **do**

$(\tilde{\mathbf{x}}^{(k)}, \{\tilde{\gamma}_t^i\}_{t=1,i=1}^{T,N}, \tilde{\mathbf{f}}_{\theta^{(k)}}, \tilde{\mathbf{g}}_{\theta^{(k)}}) \leftarrow \text{ILQGames}(\mathbf{f}, \mathbf{g}_{\theta^{(k)}}, x_1^{(k)})$

$\nabla_{x_1} \hat{L}(\theta^{(k)}, x_1^{(k)}) \leftarrow$ evaluated using $\tilde{\mathbf{f}}_{\theta^{(k)}}$ and $\tilde{\mathbf{g}}_{\theta^{(k)}}$ via Gradient Approximation in Section 8.5

$x_1^{(k+1)} \leftarrow x_1^{(k)} - \eta \nabla_{x_1} \hat{L}(\theta^{(k)}, x_1^{(k)})$ with line search over η

$(\check{x}^{(k)}, \{\check{\gamma}_t^i\}_{t=1,i=1}^{T,N}, \check{\mathbf{f}}_{\theta^{(k)}}, \check{\mathbf{g}}_{\theta^{(k)}}) \leftarrow \text{ILQGames}\big(\mathbf{f}, \mathbf{g}_{\theta^{(k)}}, x_1^{(k+1)}\big)$

$\nabla_{\theta} \hat{L}(\theta^{(k)}, x_1^{(k+1)}) \leftarrow$ evaluated using $\check{\mathbf{f}}_{\theta^{(k)}}$ and $\check{\mathbf{g}}_{\theta^{(k)}}$ via Gradient Approximation in Section 8.5

$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta' \nabla_{\theta} \hat{L}(\theta^{(k)}, x_1^{(k+1)})$ with line search over η'

Return $(\theta^{(k+1)}, x_1^{(k+1)})$ if $\|\theta^{(k)} - \theta^{(k-1)}\|_2 \leq \epsilon$ or **Return** $(\theta^{(k')}, x_1^{(k')})$, where $k' \leftarrow \arg\min_k \tilde{L}(\theta^{(k)}, x_1^{(k)})$, if iteration number k reaches K.

As shown in Section 8.6, this procedure of joint reconstruction and filtering enables our solver to reliably recover player costs even in scenarios of substantial partial observability. The third difficulty of evaluating and differentiating the objective function in the inverse *feedback* game problem is mitigated by the efficient approximation outlined in Section 8.5. To jointly infer the initial condition, the cost and the state-input trajectory, we first adopt the coordinate gradient descent method, where gradient descent steps are first taken over the initial condition $\hat{x}_1$, and then taken over the cost parameter. We update the estimate of the noise-free full state-input trajectory by computing a FBNE state trajectory from the inferred initial condition and the cost.

We summarize our proposed solver in Algorithm 6. At the k-th iteration, we first compute an approximate FBNE state trajectory $\tilde{x}^{(k)}$ and the associated LQ approximation via the ILQGames algorithm of [104]. Using this LQ approximation, we estimate $\nabla_{x_1} \hat{L}(\theta, x_1^{(k)})$ using the procedure outlined in Section 8.5. We then update the initial condition $x_1^{(k)}$ by a step of gradient descent, where the stepsize is chosen by a suitable linesearch technique [238, Ch. 3] such that the loss $\hat{L}(\theta, x_1)$ is sufficiently decreased. Given the updated initial condition $x_1^{(k+1)}$, we find a new approximate FBNE state trajectory via the ILQGames algorithm again, which is then used to estimate $\nabla_{\theta} \hat{L}(\theta^{(k)}, x_1^{(k+1)})$ via the procedure in Section 8.5. With this gradient, we update $\theta^{(k)}$ by one step of gradient descent with linesearch. We repeat this procedure until, at convergence, we find a locally optimal solution $(\hat{\theta}, \hat{x}_1)$.

8.6 Experiments

In this section, we adopt the open-loop solution method of [251] as the baseline method and compare it to Algorithm 1. The experiment codes can be found in https://github.com/jamesjingqili/inverse-iLQGames.git. In particular, we evaluate Algorithm 6 in several Monte Carlo studies

which aim to justify the following claims.

- The proposed gradient approximation often aligns with a descent direction in the loss function.
- Algorithm 1 is more robust than the open-loop baseline method [251] with respect to noise in, and incomplete observations of, the expert demonstration trajectory.
- The cost functions inferred by Algorithm 6 can be generalized to predict trajectories from unseen initial conditions.
- Algorithm 1 can infer nonconvex costs in nonlinear games.

Gradient Approximation Quality

We continue the 2-vehicle platooning example defined in (8.12) and (8.13). We measure the performance of Algorithm 6 in two settings, incomplete expert trajectory data with noisy partial state observation, and complete expert trajectory data with noisy full observation. In the first case, each player's partial observation only contains its x-position, y-position and heading angle. The time index set of the incomplete trajectory is $\mathcal{T} = [T] \setminus \{11, 12, \ldots, 19\}$. In the second case, the expert data includes the noisy observation of all the states of both players at all $t \in [T]$. The ground truth expert state trajectory follows a FBNE strategy from the initial condition $x_1 = [0, 0.5, \frac{\pi}{2}, 1, 1, 0, \frac{\pi}{2}, 1]$ and the target lane is $p_x^* = 0.0$. At each variance level $\sigma \in \{0.004, 0.008, \ldots, 0.04\}$, we generate 10 noisy observations of the ground truth expert trajectory, with isotropic zero-mean Gaussian noise. For each noisy expert data set $\mathbf{y}_\mathcal{T}$, we minimize the negative log-likelihood objective in (1), i.e., $\sum_{t\in\mathcal{T}} \|y_t - h(x_t)\|_2^2$, where $h(\cdot) : \mathbb{R}^n \to \mathbb{R}^\ell$ maps a state x_t to its partial observation.

As shown in Fig. 8.4, the loss decreases monotonically on the average. This indicates that the gradient approximation proposed in Section 8.5 provides a reliable descent direction. The inverse feedback game problem becomes challenging when there is only partial state observation and incomplete trajectory data, and the quality of inferred costs may degrade when the observation noise is high.

Robustness, Generalization and the Ability to Infer Nonconvex Costs

We continue the previous 2-vehicle example and compare Algorithm 6 and the baseline in a Monte Carlo study, where we infer the costs under 10 different levels of Gaussian noise with increasing variance. In particular, we evaluate three metrics in Fig. 8.5: (a) the distance between the noisy expert data and the FBNE state trajectory which results from players' inferred costs; (b) the distance between the computed FBNE state trajectory (under the players' inferred costs) and the ground truth expert data. An example of such a comparison is shown in Fig. 8.6. Finally, we evaluate (c) the distance between the inferred FBNE state trajectories and the FBNE state trajectory under the ground truth costs for some randomly sampled initial conditions, which is also visualized in Fig. 8.7. Collectively, the results demonstrate that *Algorithm 6 has better robustness and generalization performance than the open-loop baseline when the expert data follows the FBNE assumption.*

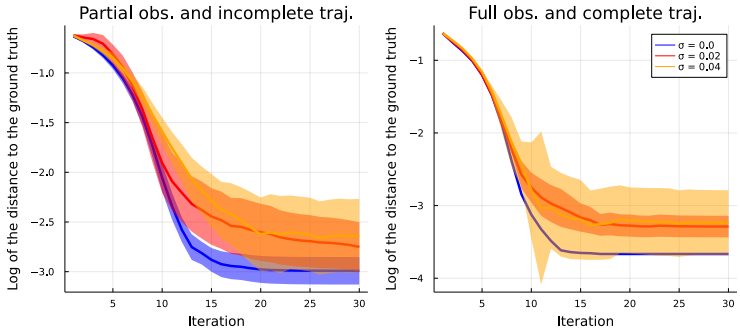


Figure 8.4: Convergence of Algorithm 1 with the Gradient Approximation proposed in Section 8.5. The loss decreases monotonically on the average. The bold lines and shaded areas represent the mean values and their standard error, i.e., the variance divided by the square root of the sample size, respectively.

To show that Algorithm 1 can infer nonconvex cost functions, we extend the previous 2-vehicle platooning example and assume that the 2-vehicle team encounters a third vehicle and the follower wants to stay close to the leader without colliding with the third vehicle. We model this scenario as a 3-vehicle game with a 12 dimensional state space and a horizon $T = 30$. The dynamics for each vehicle is the same as (8.12) and the costs are as follows,

$$\begin{aligned}
g^1_{t,\theta}(x_t, u_t) =& \theta^1_1 \|p^1_{x,t}\|^2_2 + \theta^1_2 \|p^2_{x,t} - p^*_x\|^2_2 + \|v^1_t - 2\|^2_2 \\
&+ \|\beta^1_t - \frac{\pi}{2}\|^2_2 + \|u^1_t\|^2_2 \\
g^2_{t,\theta}(x_t, u_t) =& \theta^2_1 \|p^2_{x,t}\|^2_2 + \|\beta^2_t - \frac{\pi}{2}\|^2_2 + \theta^2_2 \|p^2_{x,t} - p^1_{x,t}\|^2_2 + \|v^2_t - 2\|^2_2 \\
&- \frac{1}{2}\log(\|p^2_{x,t} - p^3_{x,t}\|^2_2 + \|p^2_{y,t} - p^3_{y,t}\|^2_2) + \|u^2_t\|^2_2 \\
g^3_{t,\theta}(x_t, u_t) =& \theta^3_1 \|p^3_{x,t} - \frac{1}{2}\|^2_2 + \|u^3_t\|^2_2
\end{aligned}$$

where the ground truth $\theta^* \in \mathbb{R}^5$ is $[0, 4, 0, 4, 2]$. The ground truth expert state trajectory follows a FBNE strategy from the initial condition $x_1 = [0, 1, \frac{\pi}{2}, 2, 0.3, 0, \frac{\pi}{2}, 2, 0.5, 0.5, \frac{\pi}{2}, 2]$, where the last four elements encode the state of the third vehicle. The target lane in the expert data is $p^*_x = 0.2$.

Similar to the 2-vehicle experiment, we consider two settings, incomplete trajectory data with partial state observation and complete trajectory data with full state observation. The partial state observation includes all the states of each vehicle except for the velocity of all the vehicles, and the time indices set of the incomplete trajectory is $\mathcal{T} = [T] \setminus \{11, 12, \ldots, 19\}$. The nonconvex cost of player 2 causes numerical problems in the baseline KKT OLNE solver [251]. Thus, we add an L_2 regularization $10^{-4}\|\theta\|^2_2$ to the loss $\hat{L}(\theta, x_1)$ and summarize the Monte Carlo study in Fig. 8.8, where we see Algorithm 6 is also able to learn better cost functions reflecting the true intentions of each vehicle in feedback games, even with only partial state observations and incomplete trajectory data.

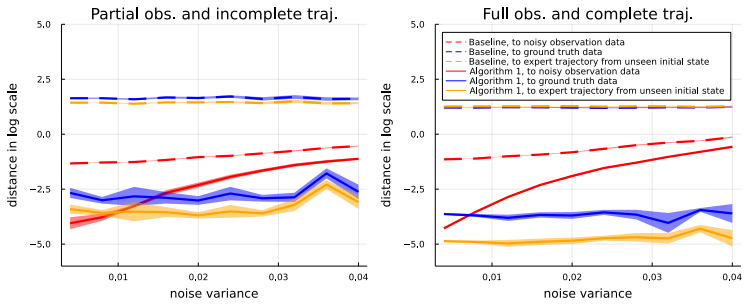


Figure 8.5: 2-vehicle platooning scenario. The bold lines and shaded areas represent the mean values and their standard error, i.e., the variance divided by the square root of the sample size, respectively. As the noise variance growing, the converged loss value increases, as shown in the red curves. However, Algorithm 6 is still able to learn a more accurate cost and has less generalization error than the baseline, as shown in the blue and yellow curves, respectively.

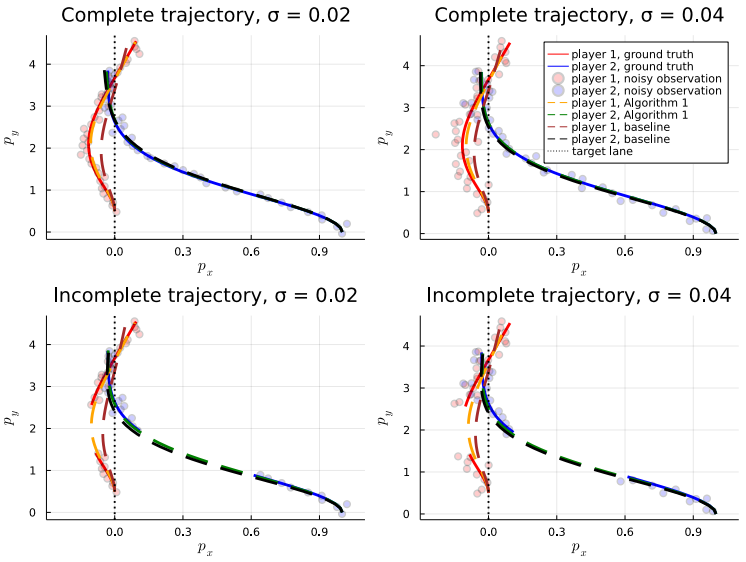


Figure 8.6: Full and partial, noisy observation of the expert trajectories. Dashed lines represent predicted trajectories which result from inferred costs, and solid lines are ground truth. The trajectories predicted by Algorithm 6 are closer to the ground truth than the baseline.

8.7 Conclusion

In this chapter, we propose an efficient cost inference algorithm for inverse feedback nonlinear games, with only partial state observation and incomplete trajectory data. Empirical results show that the proposed solver converges reliably for inverse games with nonconvex costs and has superior generalization performance than a state-of-the-art open-loop baseline method when the expert demonstration reflects a group of agents acting in a dynamic feedback game. There are many future directions. We can investigate under what conditions the cost can be inferred exactly in feedback games. The active and online inference are also promising directions. In addition, we are eager to extend this work to settings of closed-loop interaction. In such an extension, rather than merely inferring the objectives of observed players, this information would be used to guide the

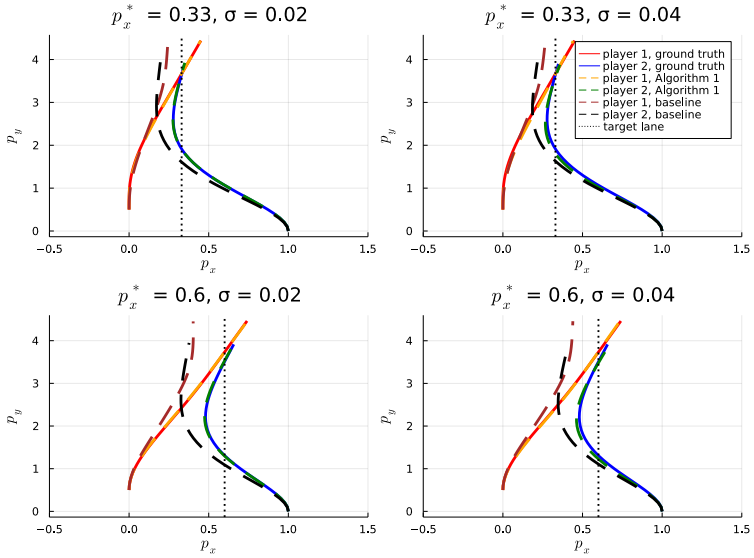


Figure 8.7: Generalization performance comparison. p_x^* is the target lane position that player 1 wants to guide player 2 toward. All the costs are inferred from partial observations and incomplete trajectory data, with different noise variance specified in each of the subplot. The trajectories predicted by Algorithm 6 are closer to the ground truth than the baseline.

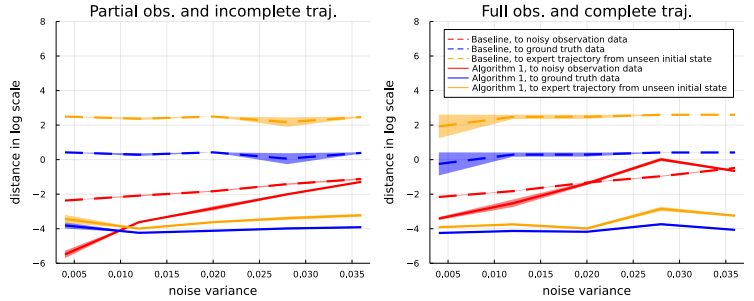


Figure 8.8: 3-vehicle platooning scenario. The bold lines and shaded areas represent the mean values and their standard error, i.e., the variance divided by the square root of the sample size, respectively. As the noise variance growing, the converged loss value increases on the average, as shown in the red curves. However, Algorithm 6 is still able to learn a more accurate cost and has less generalization error than the baseline, as shown in the blue and yellow curves, respectively.

decision-making of an autonomous agent in that scene.

8.8 Acknowledgements For This Chapter

This work is supported by the National Science Foundation under Grant No. 2211548.

Appendix

Proof of Proposition 10. Proposition 1 claims that there exists an inverse LQ game, which has isolated global minima and the induced FBNE state trajectories of those solutions match the expert demonstration. Here, we show such a counterexample, which supports the claim. Consider a 2-player horizon-3 LQ game with the linear dynamics

$$x_{t+1} = x_t + u_t^1 + u_t^2, \;\; t \in \{1, 2, 3\}, \tag{8.16}$$

and the cost

$$\begin{aligned} g_t^1(x_t, u_t) &= \frac{1}{2}(Q^1\|x_t\|_2^2 + \|u_t^1\|_2^2), \;\; t \in \{1, 2\}, \\ g_t^2(x_t, u_t) &= \frac{1}{2}(Q^2\|x_t\|_2^2 + 2\|u_t^2\|_2^2), \;\; t \in \{1, 2\}, \\ g_3^1(x_3, u_3) &= \frac{1}{2}Q^1\|x_3\|_2^2, \; g_3^2(x_3, u_3) = \frac{1}{2}Q^2\|x_3\|_2^2. \end{aligned} \tag{8.17}$$

We assume that the ground truth solutions are $Q^1 = 1$, $Q^2 = 1$. We will show there is also one extra solution $\hat{Q}^1 = \frac{1}{2}$ and $\hat{Q}^2 = 2$, which yields the same FBNE state trajectory as the ground truth for any initial condition. We follow the same definition of the variable $\{Z_t^i\}_{t=1,i=1}^{3,2}$ as in [24]. By definition, we have $Z_t^i \geq Q^i > 0$, when $Q^1 \in \mathbb{R}_+$ and $Q^2 \in \mathbb{R}_+$. Following the notations in FBNE condition in Corollary 6.1 of [24], we consider the feedback matrices $\{P_t^i\}_{t=1,i=1}^{2,2}$,

$$\begin{bmatrix} P_t^1 \\ P_t^2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 + Z_{t+1}^1 & Z_{t+1}^1 \\ Z_{t+1}^2 & 2 + Z_{t+1}^2 \end{bmatrix}}_{G_t^i} \begin{bmatrix} Z_{t+1}^1 \\ Z_{t+1}^2 \end{bmatrix}, \;\; \forall t \in \{1, 2\}, \tag{8.18}$$

where the matrix G_t^i is invertible because $\det(G_t^i) = 2 + Z_{t+1}^2 + 2Z_{t+1}^1 > 0$. The above analysis suggests that the FBNE state trajectory for all $Q^1 > 0$ and $Q^2 > 0$ are uniquely determined. We consider the time instant $t = 2$, and observe

$$\begin{bmatrix} P_2^1 \\ P_2^2 \end{bmatrix} = \begin{bmatrix} 1 + Q^1 & Q^1 \\ Q^2 & 2 + Q^2 \end{bmatrix}^{-1} \begin{bmatrix} Q^1 \\ Q^2 \end{bmatrix} = \frac{1}{2 + 2Q^1 + Q^2} \begin{bmatrix} 2Q^1 \\ Q^2 \end{bmatrix}. \tag{8.19}$$

We then have the closed-loop dynamics $x_3 = (1 - P_2^1 - P_2^2)x_2 = \frac{2}{2+2Q^1+Q^2}x_2$, which yields that for two pairs of positive variables (Q^1, Q^2) and $(\hat{Q}^1, \hat{Q}^2)$, a necessary condition for them to have the same FBNE trajectory is that $2Q^1 + Q^2 = 2\hat{Q}^1 + \hat{Q}^2$. We have $Z_2^1 = Q^1 + \frac{Q^1 + (2Q^1)^2}{(2+2Q^1+Q^2)^2}$, $Z_2^2 = Q^2 + \frac{Q^2 + 2(Q^2)^2}{(2+2Q^1+Q^2)^2}$. Similarly, for the time instant $t = 1$, we have $x_2 = (1 - P_1^1 - P_1^2)x_1 = \frac{2}{2+2Z_2^1+Z_2^2}x_1$. A necessary condition for $(\hat{Q}^1, \hat{Q}^2)$ to have the same FBNE state trajectory as (Q^1, Q^2) is that the

following 2 equations are satisfied,

$$\begin{aligned} &2Q^1 + Q^2 = 2\hat{Q}^1 + \hat{Q}^2 \\ &2\big(Q^1 + \frac{Q^1 + (2Q^1)^2}{(2 + 2Q^1 + Q^2)^2}\big) + Q^2 + \frac{Q^2 + 2(Q^2)^2}{(2 + 2Q^1 + Q^2)^2} \\ &\quad = 2\big(\hat{Q}^1 + \frac{\hat{Q}^1 + (2\hat{Q}^1)^2}{(2 + 2\hat{Q}^1 + \hat{Q}^2)^2}\big) + \hat{Q}^2 + \frac{\hat{Q}^2 + 2(\hat{Q}^2)^2}{(2 + 2\hat{Q}^1 + \hat{Q}^2)^2}. \end{aligned} \tag{8.20}$$

We substitute $Q^1 = 1$, $Q^2 = 1$ and $\hat{Q}^2 = 3 - 2\hat{Q}^1$ into the second row of (8.20), which is reduced to a 2-degree polynomial of $\hat{Q}^2$. By the fundamental theorem of algebra [46], there exist at most 2 solutions for $\hat{Q}^2$. The two pairs of $(\hat{Q}^1, \hat{Q}^2)$ satisfying (8.20) are $(1, 1)$ and $(\frac{1}{2}, 2)$. The two global minima are isolated. Since the dimension of the state x_t is 1, for all initial states $x_1 \in \mathbb{R}$, the FBNE state trajectories under the costs specified by the two pairs cost parameters $(1, 1)$ and $(\frac{1}{2}, 2)$ coincide with each other. □

Chapter 9

Beyond Alignment: Exploiting Information Asymmetry in Multi-Agent Coordination

In this chapter, we investigate the possibility of strategically influencing agents with incomplete information and shaping their beliefs for enhanced overall multi-agent coordination task performance. This chapter is adapted from the work [183], co-authored with Anand Siththaranjan, Somayeh Sojoudi, Claire J. Tomlin, and Andrea Bajcsy.

9.1 Background

General-sum dynamic games—wherein agents may have competing (but not opposing) objectives—are a powerful mathematical framework that can model a range of multi-agent behaviors, such as autonomous vehicle coordination [274] and human-robot interaction [225]. When these models are put into practice, an outstanding challenge is accounting for the fact that all agents' objectives (i.e., intents) may not be known *a priori*. For example, when a car is merging onto the highway, the highway drivers typically pay attention to see if the new car is aggressively merging in front of them, or passively yielding to them.

Prior game-theoretic planners predominantly handle intent uncertainty from the perspective of the agents that are uncertain about the behavior of another agent. We call these the *uncertain agents*. These works propose that the uncertain agent plays the game under point estimates of the other agent's intents [274, 212] or plan in expectation under the average of all opponent strategies parameterized by their intents (e.g., aggressive and passive merging driver) [166, 171]. Other works focus on how the uncertain agent can take information-gathering actions to probe at the opponent's intent [267, 136, 327], thus improving the long-term performance. However, both of these models miss out on the fact that the other agent, here called the *certain agent*, can also *demonstrate* their intent to the uncertain agent. For example, the merging driver may speed up more aggressively when entering the highway, conveying its intent in a more exaggerated way to the highway vehicles behind. The key here is acknowledging that the agent with certainty can *plan to influence the belief of the uncertain agents* through its own actions.

In this chapter, we study *strategic* intent demonstration in dynamic games, where a certain agent interacts with multiple uncertain agents. Our core idea is to model the certain agent as planning over both the evolution of the joint physical state and the dynamics of the uncertain agents' beliefs. With this, we can design objectives enabling the certain agent to trade off between *demonstrating its intent* (i.e., aligning the uncertain agents' beliefs with its true intent) and *pursuing its own task performance*, while the uncertain agents respond through belief updates and the rational physical actions.

Our primary contribution is a scalable continuous state-action algorithm for solving nonlinear intent demonstration games via iterative linear-quadratic approximations. Our algorithm consists of two sub-optimizations: first solving for all agents' game-theoretic feedback policies parameterized by *any* intent, and then solving the certain agent's optimization over the joint physical and estimate dynamics. We theoretically characterize the convergence of the uncertain agents' beliefs and the certain agent's ability to balance intent demonstration with task performance. We also evaluate our method in a suite of multi-agent settings such as decentralized bi-manual robot manipulation, three-vehicle platooning, and shared control. We find that when agents can strategically demonstrate their (dynamically changing) intents to others, they can achieve superior task performance and coordination.

9.2 Related Works

Efficient Solutions to General-Sum Dynamic Games. Even without intent uncertainty, solving general-sum dynamic games over continuous state and action spaces is challenging. Most of these games have no analytic solution, and classical dynamic programming approach for finding Nash equilibria of these games suffers from the "curse of dimensionality" [255]. However, under linear dynamics and quadratic costs, there exist efficient numerical solutions for solving these linear-quadratic (LQ) games [24]. Recent works propose to solve nonlinear games by iteratively approximating them via LQ games [104, 248]. In this work, we leverage these fast and approximate iterative LQ game solvers as a submodule in our intent demonstration algorithm.

Incomplete Information Games: From Theory to Algorithms. Prior dynamic programming solutions to incomplete information games [123, 241, 299, 141, 268] do not scale to high-dimensional nonlinear games with continuous state, action and intent spaces. Thus, recent works focus on scalable approximations. One overarching approximation is assuming that some agents have complete information and others do inference. These approaches model the uncertain agents as planning in expectation [275, 166], planning with the most likely estimate and recovering a complete-information game [171, 47], doing intent inference from an offline dataset [251, 180, 212], planning multiple contingencies based on discrete intent hypotheses [249], and modeling incentives for uncertain agents to take information-gathering actions [267, 136]. While prior works focus on how *uncertain agents* should tractably plan under their beliefs, we focus on how the *certain agent* can demonstrate their intent by exploiting the learning dynamics of other agents.

Intent Demonstration in Multi-Agent Interactions. Prior works on intent demonstration, such as legibility in robot motion planning around humans [76], typically model uncertain agents as

passive observers. However, in scenarios like multi-agent highway driving [313] or collaborative manipulation [200], all agents actively interact while some simultaneously learn the missing information of the games. Unlike previous multi-agent intent demonstration frameworks [208, 26], our model explicitly accounts for rational feedback from uncertain agents within general-sum dynamic games. Moreover, rather than simply aligning uncertain agents' beliefs with the certain agent's true intent, our approach allows the certain agent to strategically shape these beliefs, thereby guiding uncertain agents' actions to enhance overall task performance beyond conventional belief-alignment methods [76, 171].

9.3 Background: General-sum Games and Nash Equilibrium

In this section, we present the necessary background on general-sum dynamic games. For narrative simplicity, we will use the terms "players" and "agents" interchangeably.

Notation. We consider general-sum games played over the finite time horizon T. We consider N players in the game, each of whose control action is denoted by $u_t^i \in \mathbb{R}^m$ for $i \in \{1, 2, \dots, N\}$. Let the set of times $\{0, 1, \dots, T\}$ be denoted by $\mathbf{T}$ and the set of player indicies $\{1, 2, \dots, N\}$ be denoted by $\mathbf{N}$. We denote $x_t \in \mathbb{R}^n$ to be the joint physical states of all players (e.g., positions, velocities) which evolves via the deterministic discrete-time dynamics, $x_{t+1} = f_t(x_t, u_t^1, \dots, u_t^N)\ \forall t \in \mathbf{T}$, where $f_t(\cdot) : \mathbb{R}^n \times \mathbb{R}^m \times \cdots \times \mathbb{R}^m \to \mathbb{R}^n$ is assumed to be a differentiable function. For notational convenience, we denote the vector of all N agents' actions at time t to be $u_t := [u_t^1, \dots, u_t^N]$.

Player Objectives. Let each player $i \in \mathbf{N}$ seek to minimize their own cost function, $c_t^i(x_t, u_t)$. Note that in general this cost function depends on *both* the joint physical state of all players and also the actions of all players. It is precisely this coupling that induces a dynamic game between all players. The Nash equilibrium defines a scenario wherein no player wants to deviate from their current state-action profile under their respective cost functions. Specifically, in our work, we consider *feedback Nash equilibrium* (FNE) [24], wherein each player $i \in \mathbf{N}$ solves for a policy $\pi_t^i(x_t) : \mathbb{R}^n \to \mathbb{R}^m$ which gets access to the current joint physical state, x_t, at any time, and outputs an action. When the cost functions for all agents were assumed to be known a priori, such games are called *complete information games*. However, when players have uncertainty over other players' objectives, these are *incomplete information games*, which is what we study here.

9.4 Problem Formulation: Intent Demonstration in General-Sum Dynamic Games

In this chapter, we study the problem of intent demonstration—wherein one agent can express their intent to uncertain agents—in general-sum game-theoretic interactions. Similar to prior work [47, 171], we consider incomplete information *asymmetry* between the players: one player (e.g., player 1) has complete information, i.e., they know the cost functions of all players, but players 2 through N have incomplete information about player 1's cost function. Moreover, we assume that each

agent is aware of its status as either certain or uncertain, and that this information is shared among all agents. For example, from our introductory example, the driver merging in from an on-ramp has certainty over their own driving style, but all other road agents on the highway do not. However, players 2 through N have the ability to *estimate* or learn about player 1's cost function during game-theoretic interaction. This problem cannot be reformulated as another complete information dynamic game with deterministic dynamics because players 2 through N are not aware of player 1's cost function and there can be an infinite number of possible cost functions for player 1. We formalize these ideas below.

Certain Player: Cost Parameterization. Without loss of generality, let player 1 be the agent with complete information of the game, including the cost functions of other players. We model player 1's task-centric cost function, $c_t^1(x_t, u_t; \theta^*)$, as parameterized by a low-dimensional parameter, $\theta^* \in \Theta$, which could in theory be discrete (e.g., aggressive or passive driving style) or continuous (e.g., weights on a linear feature basis).

Uncertain Players: Estimation and Cost Functions. All agents, except for player 1, are uncertain about player 1's cost function parameter. They maintain *estimates* of this parameter via $\hat{\theta}$, which in general can be a full Bayesian belief or a point estimate. All uncertain agents possess the ability to learn, based on the joint physical states (x_t) and the action of player 1 (u_t^1) observed during interaction. Mathematically, for any uncertain player $j \in \{2, 3 \ldots, N\}$ and their associated estimate $\hat{\theta}_t^j$ at time t, let $\hat{\theta}_{t+1}^j = g_t(\hat{\theta}_t^j, x_t, u_t^1)$ be the updated estimate via update rule g_t. Ultimately, each uncertain player aims to minimize their own cost function $c_t^j(x_t, u_t)$.

Intent Demonstration Formulation. We can now formulate the intent demonstration problem in general-sum games. One of our core ideas is to augment player 1's state space with the estimates of all uncertain agent's beliefs. Let the vector of all uncertain agent's current estimates be denoted by $\hat{\theta}_t := [\hat{\theta}_t^2, \ldots, \hat{\theta}_t^N]$. We model the certain agent's cost as a combination of their task-centric cost, $c_t^1(x_t, u_t; \theta^*)$, (e.g., for an autonomous car this could be lane-keeping and smoothness of motion), and the "error" between the uncertain agent's estimates and the true intent, $c^{\text{demo}}(\hat{\theta}_t, \theta^*)$, (e.g., expressing that they are aggressive or in a rush):

$$\bar{c}_t^1(x_t, \hat{\theta}_t, u_t; \theta^*) := \rho_1 \cdot c_t^1(x_t, u_t; \theta^*) + \rho_2 \cdot c^{\text{demo}}(\hat{\theta}_t, \theta^*),$$

where $\rho_1, \rho_2 \geq 0$ are hyper-parameters. Intuitively, this enables player 1 to synthesize a range of behaviors, from prioritizing task-cost and only influencing the uncertain agent's beliefs when beneficial for minimizing task cost (i.e., $\rho_1 > 0, \rho_2 \equiv 0$), to encouraging player 1 to actively express their intent (i.e., $\rho_1 \equiv 0, \rho_2 > 0$). Ultimately, player 1's intent demonstration problem optimizes

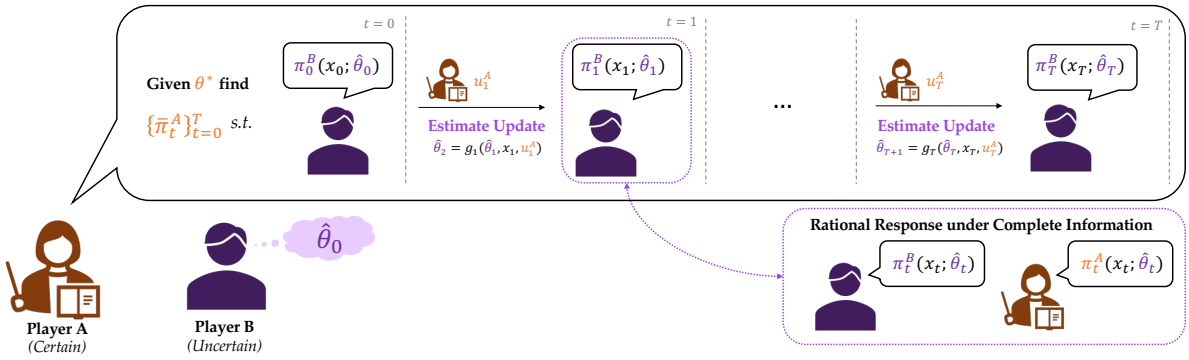


Figure 9.1: **Intent Demonstration Problem in General-Sum Games.** The *certain* player A optimizes $u_t^A = \bar{\pi}_t^A(x_t, \hat{\theta}_t; \theta^*)$, which trades off its own task cost and demonstrating their intent. The *uncertain* player B engages with player A through rational actions $u_t^B = \pi_t^B(x_t; \hat{\theta}_t)$ and updates their estimate $\hat{\theta}_t$ of player A's intent θ^* by observing A's actions. This enables player A to choose to influence player B's estimate.

their augmented cost function subject to several key constraints:

$$\min_{\{u_t^1\}_{t=0}^T} \sum_{t=0}^{T} \bar{c}_t^1(x_t, \hat{\theta}_t, u_t; \theta^*) \tag{9.1a}$$

$$\text{s.t. } x_{t+1} = f_t(x_t, u_t), \forall t \in \mathbf{T} \tag{9.1b}$$

$$\hat{\theta}_{t+1}^j = g_t(\hat{\theta}_t^j, x_t, u_t^1), \forall j \in \mathbf{N} \setminus \{1\}, \forall t \in \mathbf{T} \tag{9.1c}$$

$$u_t^j = \pi_t^j(x_t; \hat{\theta}_t^j), \forall j \in \mathbf{N} \setminus \{1\}, \forall t \in \mathbf{T} \tag{9.1d}$$

$$x_0 = x^{\text{init}}, \ \hat{\theta}_0 = \hat{\theta}_{\text{init}}. \tag{9.1e}$$

Here, Equations (9.1b) and (9.1c) constrain the solution to abide by the physical dynamics of the joint system and ensure that the estimates of the uncertain players follow their update rules. Given any player's current estimate $\hat{\theta}_t^i$, Equation (9.1d) models the uncertain players as rationally responding under their current FNE strategy[1] $\pi_t^i(x_t; \hat{\theta}_t^i)$, *assuming that all agents also play under the player i's current intent estimate, $\hat{\theta}_t^i$*. Note that this is simply a virtual game model in the mind of each uncertain player (see purple dashed box in Figure 9.1). In reality, player 1 can behave differently than the current estimate $\hat{\theta}_t^i$; however, this is not a problem for player 2 since they will update their intent estimate at the next timestep. Finally, similar to prior first-order belief assumptions [275], in Equation (9.1e) we assume that the initial estimates, $\hat{\theta}_0^j$, of each uncertain player $j \in \{2, \ldots, N\}$ are common knowledge. An illustrative diagram of our interaction model between two players is visualized in Figure 9.1.

[1]We assume that there exists a unique FNE. When multiple FNEs exist, we can align the FNE strategies of players by taking the technique in [250].

9.5 Theoretical and Algorithmic Results

In this section, we investigate both the theoretical and algorithmic properties of the proposed intent demonstration formulation, involving a certain player (player 1) and uncertain players (players 2 to N). We begin by analyzing linear-quadratic games, for which we establish rigorous theoretical guarantees concerning the efficiency of intent teaching. Subsequently, we extend our framework algorithmically to address intent demonstration problems within multi-player nonlinear games, such as those incorporating nonlinear Bayesian estimation rules.

Case 1: LQ Games with Linear Estimation Dynamics

LQ Setup. We consider settings where player 1's true intent parameter is a continuous goal parameter (i.e., part of their cost). Each player $j \in \{2, \dots, N\}$ maintains a point estimate $\hat{\theta}_t^j$ of θ^*. Let the joint physical dynamics in optimization problem (9.1) be a time-varying linear system, $f_t := A_t x_t + \sum_{i=1}^N B_t^i u_t^i, \quad t \in \mathbf{T}$, with $A_t \in \mathbb{R}^{n \times n}$ and $B_t^i \in \mathbb{R}^{n \times m}$. Let player 1's task and intent-demonstration costs be quadratic in physical state and control: $c_t^1(x_t, u_t; \theta^*) := x_t^\top Q_t^1 x_t + u_t^{1\top} R_t^1 u_t^1 + x_t^\top \theta^*$ and $c^{\text{demo}}(\hat{\theta}_t, \theta^*) := \sum_{j=2}^N \|\hat{\theta}_t^j - \theta^*\|_2^2$. Similarly, for each $i \in \{1, \dots, N\}$, let player i's quadratic cost be $c_t^i(x_t, u_t) := x_t^\top Q_t^i x_t + u_t^{i\top} R_t^i u_t^i$ where $Q_t^i \in \mathbb{R}^{n \times n}$ and $R_t^i \in \mathbb{R}^{m \times m}$ are positive semi-definite matrices.

Uncertain Player's Feedback Policy. Given their current point estimate, $\hat{\theta}_t^j$, the uncertain player j rationally responds under their current FNE policy $\pi_t^j(x_t; \hat{\theta}_t^j)$, assuming a complete information game where player 1 also acts rationally under player j's estimate, $u_t^1 = \pi_t^1(x_t; \hat{\theta}_t^j)$. Importantly, since we are in the LQ setting, all players' complete information game FNE policies are linear feedback policies [24].

Linear Estimation Dynamics. Finally, let the estimate dynamics of an uncertain player $j \in \{2 \dots, N\}$ to be linear in state and estimate. Specifically, we consider a constant step size $\alpha > 0$, and we study a gradient descent-based maximum likelihood estimation (MLE) update rule [199], $g_t(\hat{\theta}_t^j, x_t, u_t^1)$, as

$$g_t := \hat{\theta}_t^j - \alpha \nabla_{\hat{\theta}_t^j} \|u_t^1 - \pi_t^1(x_t; \hat{\theta}_t^j)\|_2^2. \tag{9.2}$$

Player j updates their estimate based on the difference between the action they expected player 1 to take under their estimate, $\pi_t^1(x_t; \hat{\theta}_t^j)$, and player 1's observed action, u_t^1.

Bellman Equation and Algorithm. When an uncertain player learns via a linear MLE update rule, intent demonstration is an LQR problem in the joint physical state x_t, the estimate $\hat{\theta}_t$, and the true cost parameter θ^*. The Bellman equation for player 1's intent demonstration problem specified in Equation (9.1) is defined as:

$$\begin{aligned} V_t^1(x_t, \hat{\theta}_t; \theta^*) = &\min_{u_t^1} \bar{c}_t^1(x_t, \hat{\theta}_t, u_t^1, \{\pi_t^j(x_t; \hat{\theta}_t^j)\}_{j=2}^N; \theta^*) \\ &+ V_{t+1}^1(x_{t+1}, \hat{\theta}_{t+1}; \theta^*). \end{aligned} \tag{9.3}$$

Algorithm 7: Strategic Intent Demonstration Games

Require: dynamics f, player 1's task cost $c_t^1(x, u; \theta^*)$ and demonstration cost $c^{\text{demo}}(\hat{\theta}, \theta^*)$, $\rho_1, \rho_2 \geq 0$, player j's cost $c_t^j(x, u)$, initial estimate $\hat{\theta}_0^j$, for each $j \in \{2, \ldots, N\}$, and estimation dynamics g

`// Solve complete information game for all potential intents`

$\{\pi_t^i(x; \theta)\}_{i=1,t=0}^{N,T} \leftarrow$ `FeedbackGame`$(\{c_t^i\}_{i=1}^N, f)$

$\Pi = \{\pi_t^i(x; \theta)\}_{i=1,t=0}^{N,T}$

`// Compute intent demonstration policy`

$\{\bar{\pi}_t^1(x, \hat{\theta}; \theta^*)\}_{t=0}^T \leftarrow$ `OptimalControl`$(\hat{\theta}_0, \bar{c}_t^1, f, g)$

$\Pi \leftarrow \{\bar{\pi}_t^1(x, \hat{\theta}; \theta^*)\}_{t=0}^T \cup \Pi$

return Π

With this Bellman equation in hand, we can now pose our intent demonstration Algorithm 7 and leverage a suite of off-the-shelf numerical techniques for each component of our algorithm. Specifically, in Algorithm 7, we first solve a complete information linear-quadratic game for all players under each possible intent parameter $\theta \in \Theta$. Importantly, here we can obtain feedback policies, $\{\pi_t^i\}_{i=1,t=0}^{N,T}$, for all agents with efficient (polynomial time) off-the-shelf algorithms. These feedback policies are re-used by all players. Player $j \in \{2, \ldots, N\}$ uses $\pi_t^1(x_t; \hat{\theta}_t^j)$ to predict player 1's actions under their current estimate, $\hat{\theta}_t^j$, and then update the estimate. Player 1 plans over the estimation dynamics of players $j \in \{2, \ldots, N\}$ when it solves the LQR problem leveraging the value function specified in Equation (9.3). Once again, this yields a feedback control law for player 1 in the joint physical and estimate state space, $\bar{\pi}_t^1(x_t, \hat{\theta}_t; \theta^*)$, and enjoys the benefits of off-the-shelf LQR solvers. We note that the active intent demonstration policy computed by Algorithm 7 is guaranteed to converge to the optimal one when the associated LQ games and the LQR problems are well-defined and admit valid solutions.

Theoretical Results. Finally, in the LQ setting, we prove a sufficient condition for the existence of an intent demonstration policy for player 1 which guarantees to drive player j's estimate to the true parameter exponentially fast, for all $j \in \{2, \ldots, N\}$. Our proof operates under player 1's cost, $\bar{c}_t^1$, with $\rho_1 = 0$ and $\rho_2 > 0$, meaning that player 1 only considers demonstrating their intent.

Proposition 11 (Effective Intent Demonstration)**.** *Consider a two-player LQ game. Suppose that the linear policy $\pi_t^1(x_t; \theta)$ takes the form $\pi_t^1(x_t; \theta) = K_{t,x}^1 x_t + K_{t,\theta}^1 \theta$, $\forall t \in \mathbf{T}$ and $K_{t,\theta}^{1\top} K_{t,\theta}^1 > 0$. Moreover, let player $j \in \{2, \ldots, N\}$ learn via linear estimate dynamics $\hat{\theta}_{t+1}^j = g_t(\hat{\theta}_t^j, x_t, u_t^1)$. Pick a step size $\alpha \in (0, 1)$ such that the largest singular value of $(I - \alpha K_{t,\theta}^{1\top} K_{t,\theta}^1)$ is less than 1, $\forall t \leq T$. Then, there exists a linear intent demonstration policy $u_t^1 = \bar{\pi}_t^1(x_t, \hat{\theta}_t; \theta^*)$ such that $\|\hat{\theta}_{t+1}^j - \theta^*\|_2 < c\|\hat{\theta}_t^j - \theta^*\|_2$, $\forall t \in \mathbf{T}$, $\forall j \in \{2, \ldots, N\}$, where $0 < c < 1$ is a constant dependent on $\bar{\pi}_t^1$.*

Proof. The proof can be found in the Appendix. □

Proposition 11 is a feasibility result, and the strong assumption on the form of the policy π_t^1 is not necessary for the existence of active intent demonstration policies. Moreover, always actively demonstrating the intent to other uncertain agents could be excessive and may impair the certain agent's task performance. We show in the following result that the active teaching policy can trade-off between the certain agent's task completion and intent demonstration such that it can achieve a task performance even higher than in the complete information game, when setting $\rho_1 > 0$ and $\rho_2 = 0$.

Proposition 12 (Strategic Intent Demonstration)**.** *Let $\rho_1 = 1$ and $\rho_2 = 0$. Suppose that g_t is a linear estimate dynamics and each player's cost is convex with respect to the state x_t and the control u_t. Let $\{\tilde{u}_t^i\}_{i=1,t=0}^{N,T}$ be the controls of all players corresponding to the Nash equilibrium in the complete information game, and denote by $\{\tilde{x}_t\}_{t=0}^{T+1}$ the resulted Nash equilibrium state trajectory. Moreover, suppose that there exists a stage $t < T$ such that the Jacobian of the cost-to-go function $\tilde{c}_{t:T}^1$, defined in* (9.4)*, with respect to the control $\tilde{u}_{t:T}^1 := [\tilde{u}_t^1, \tilde{u}_{t+1}^1, \dots, \tilde{u}_T^1]$ is nonzero,*

$$\begin{aligned} \tilde{c}_{t:T}^1(\tilde{x}_t, u_{t:T}^1) := &\sum_{\tau=t}^{T} c_\tau^1(x_\tau, u_\tau^1, \{\pi_\tau^j(x_\tau; \hat{\theta}_\tau^j)\}_{j=2}^N; \theta^*) \\ &\textit{s.t. } x_{\tau+1} = f_\tau(x_\tau, u_\tau^1, \{\pi_\tau^j(x_\tau; \hat{\theta}_\tau^j)\}_{j=2}^N), \\ &\quad \hat{\theta}_{\tau+1}^j = g_\tau(\hat{\theta}_\tau^j, x_\tau, u_\tau^1), j \in \mathbf{N} \setminus \{1\} \\ &\quad \tau \in [t, T], x_t = \tilde{x}_t, \hat{\theta}_t^j = \theta^*, j \in \mathbf{N} \setminus \{1\} \end{aligned} \tag{9.4}$$

then, the optimal cost of player 1 in (9.1) *is strictly lower than its optimal cost in the complete information game.*

Proof. The proof can be found in the Appendix. □

Proposition 12 suggests that the ability of influencing the uncertain agent's belief enables the certain agent to achieve a higher task performance. In practice, we can replace the estimation dynamics in (9.2) with other types of estimation dynamics, e.g., Bayesian inference or general maximum likelihood estimation. We will explore this in the next subsection.

Case 2: Nonlinear Games with Nonlinear Estimation Dynamics

With small modifications, we can adapt Algorithm 7 to non-quadratic costs and for nonlinear dynamics. This is particularly important as many estimation update rules, such as the Bayesian belief update, are nonlinear in the estimate.

The algorithm takes inspiration from iterative LQ games (`iLQGames`) and iterative LQR (`iLQR`). Similar to the first phase in Algorithm 7, we first approximately solve the complete-information FNE equilibrium policies by calling an `iLQGames` solver [104]. At each iteration, the solver linearizes the dynamics and approximates the costs quadratically around the current trajectory–a critical procedure of `iLQgames` that affects its convergence [104]. Subsequently, it

computes an optimal control for this local LQ game, updates the trajectory, and repeats these steps until convergence.

Similar to Section 9.5, after we compute the complete information `iLQGames` policies $\{\pi_t^i(x_t; \theta)\}_{i=1}^N$, we use these policies, once again, in both the uncertain player's estimation dynamics and for the certain player's intent demonstration. For example, if player $j \in \{2, \ldots, N\}$ maintains a Gaussian belief $\hat{\theta}_t^j := b_t^j(\theta) = \mathcal{N}(\mu_t^j, \Sigma_t^j)$ over the intent parameter and learns via a nonlinear belief update rule like Bayesian inference, they use π_t^1 computed from `iLQGames` to construct their (Gaussian) likelihood function and obtain the posterior:

$$b_{t+1}^j(\theta) \propto p(u_t^1 | x_t, \theta) \cdot b_t^j(\theta) \tag{9.5}$$

Assuming that the likelihood model $p(u_t^1 | x_t, \theta)$ follows a Gaussian distribution $\mathcal{N}(\pi_t^1(x_t; \theta), I)$, and the initial belief is also a Gaussian distribution $\mathcal{N}(\mu_0^j, \Sigma_0^j)$, we can simplify the belief update by substituting the policy $\pi_t^1(x_t; \theta)$ and obtain the update rule of the belief distribution parameters:

$$\begin{aligned}
\mu_{t+1}^j =& \mu_t^j + \Sigma_t^j \cdot \nabla_\theta \pi_t^{1\top} \cdot (I + \nabla_\theta \pi_t^1 \cdot \Sigma_t^j \cdot \nabla_\theta \pi_t^{1\top})^{-1} \cdot \\
& (u_t^1 - \pi_t^1(x_t; \mu_t^j)) \\
\Sigma_{t+1}^j =& \Sigma_t^j - \Sigma_t^j \cdot \nabla_\theta \pi_t^{1\top} \cdot (I + \nabla_\theta \pi_t^1 \cdot \Sigma_t^j \cdot \nabla_\theta \pi_t^{1\top})^{-1} \cdot \\
& \nabla_\theta \pi_t^1 \cdot \Sigma_t^j
\end{aligned}$$

To optimize $c^{demo}(\cdot, \cdot)$, the certain agent can, for example, minimize the error between the average intent under the other agent's belief, $\tilde{\theta}_t^j := \mathbb{E}_{\theta \sim b_t^j(\theta)}[\theta]$, and θ^*. From player 1's perspective, instead of solving an LQR problem as in Section 9.5, it solves an `iLQR` problem to obtain the intent demonstration policy $\bar{\pi}_t^1$ in the joint physical-estimate space.

Remark 16. *We can enhance Algorithm 7 by integrating deep reinforcement learning to compute policies for intent demonstration problems in general-sum nonlinear dynamic games. For instance, multi-agent reinforcement learning [201] can be applied in step 1 of Algorithm 7 to compute complete-information FNE policies, while deep reinforcement learning can be used in step 3 of Algorithm 7 to derive a strategic intent demonstration policy.*

9.6 Experiments

In this section, we evaluate our algorithm[2] in four multi-agent scenarios shown in Figure 9.2 and study the benefits of strategic intent demonstration over alternative game-theoretic interaction models.

Bi-Manual Robot Manipulation. In the robosuite simulation environment [337], we consider a bi-manual robot manipulation problem, where two robot arms must coordinate in a decentralized

[2]The source code and additional details of the experiments are available at https://github.com/jamesjingqili/Active-Intent-Demonstration-in-Games.git.

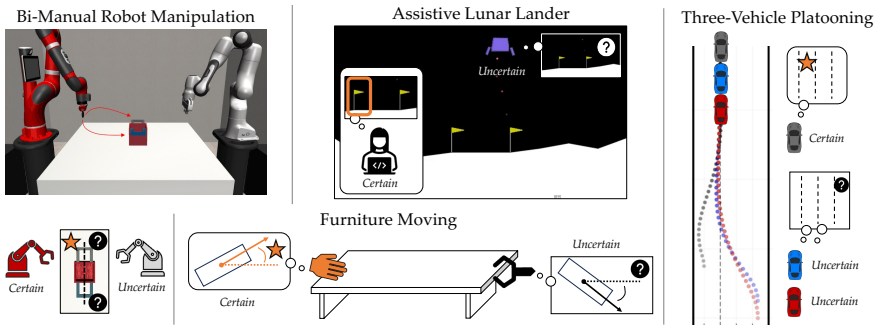


Figure 9.2: **Environments.** Four incomplete information general-sum games considered in this work.

way to pick up a pot (top left, Figure 9.2). The certain agent (red robot) wants to grab one of the handles, but the uncertain agent (silver robot) does not know this. Let $\theta^* \in \mathbb{R}$ be the red robot's preferred y-goal location (right handle) and $\hat{\theta} \in \mathbb{R}$ the silver robot's point estimate of the red robot's desired goal position which evolves via the update rule in Equation (9.2). Let $x_t = [p^1_{x,t}, p^1_{y,t}, p^2_{x,t}, p^2_{y,t}]$ be the joint physical state consisting of the (x, y)-positions of the i-th robot's end-effector. Players control linear velocity of their end-effector, $u^i_t = [v^i_{x,t}, v^i_{y,t}]$, $i \in \{1, 2\}$, and the physical system evolves via double integrator dynamics. The certain robot's quadratic task-cost minimizes distance to the target handle, avoids agent collisions, and minimizes velocity. The uncertain agent has a similar objective but is incentivized to pick up the opposite side of the pot.

Assistive Lunar Lander. A lunar lander autopilot shares control with a human pilot. The human pilot controls horizontal thrust and wants to land at their preferred destination on the x-axis (top center Figure 9.2), $\theta^* \in \mathbb{R}$, which is unknown to the autopilot. The autopilot controls both the vertical and horizontal thrust, aiming to avoid crashing on the ground while conserving fuel. For the convenience of analysis, we focus on its horizontal and vertical movements, excluding the rotation dynamics, and model this interaction as a two-player linear-quadratic game. The autopilot maintains a point estimate $\hat{\theta} \in \mathbb{R}$ and learns via linear estimate update rule (e.g., as in Equation (9.2)).

Furniture Moving. A human and robot must move table to a known destination together. The human's task cost is parameterized by their desired furniture moving angle, θ^*, and they seek to minimize their effort. The robot maintains a Bayesian belief $\hat{\theta} := b(\theta)$ over the human's preferred orientation angle (bottom, Figure 9.2). The joint physical state is position and current table angle $x_t = [p^H_{t,x}, p^H_{t,y}, p^R_{t,x}, p^R_{t,y}, \theta_t]$ and players control their x and y velocity. The dynamics of the furniture moving follows a simple kinematics model. The robot learns via a Bayesian belief update.

Three-Vehicle Platooning. A human driver guides two autonomous vehicles (AV) towards a target lane, unknown to the autonomous vehicles. Each vehicle has a unicycle dynamics with a state vector $x^i_t := [p^i_{t,x}, p^i_{t,y}, \psi^i_t, v^i_t]$ and control inputs are acceleration a^i_t and turning rate w^i_t (12-D joint state vector). Each AV optimizes 1) following the human driver's lane, 2) maintaining a forward orientation, 3) minimizing control effort and 4) avoiding collisions. Each AV has a separate Gaussian belief over the human driver's target lane, $\hat{\theta}^i := b^i(\theta)$, and updates via Bayesian estimation.

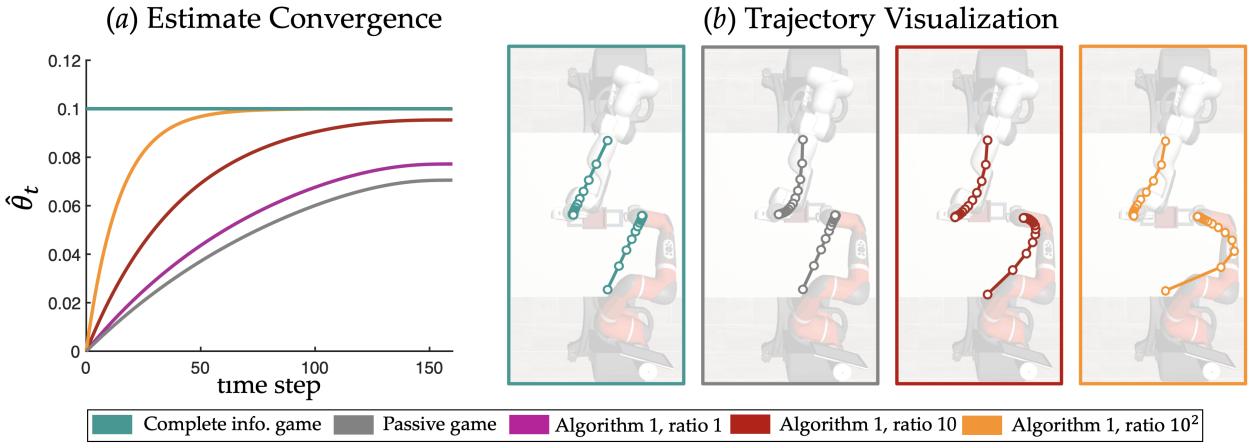


Figure 9.3: **Results: H1.** Algorithm 1 enables the uncertain agent to learn fast (left) and generate behavior qualitatively similar to complete information game when ratio = 100 (right).

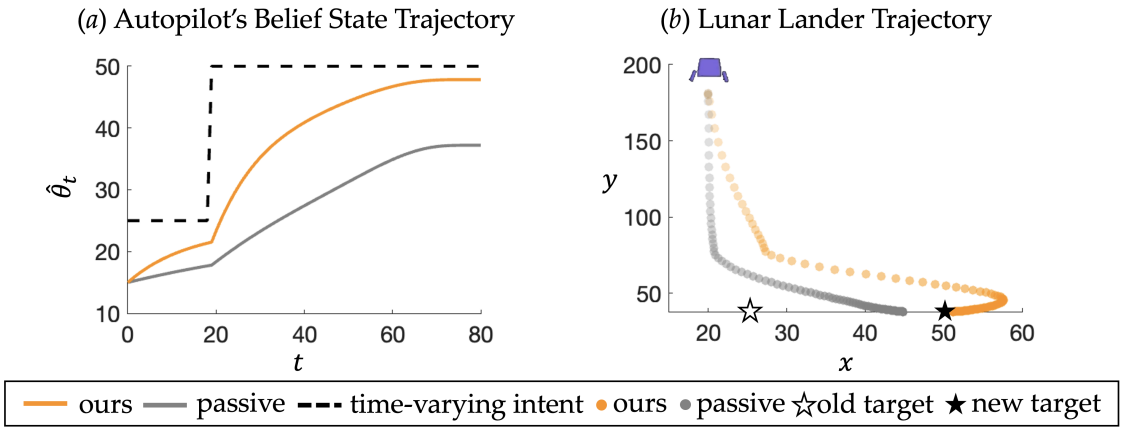


Figure 9.4: **Results: H2.** The human pilot changes their target landing position θ^* from 25 to 50 at time $t = 20$. The strategic intent demonstration policy $\bar{\pi}_t^1$, computed without anticipating this change, efficiently conveys the unforeseen dynamic intent, enabling the autopilot's belief to converge faster than in the passive game, without the need of recomputing $\bar{\pi}_t^1$.

Simulation Results

We compare our game-theoretic intent demonstration algorithm (Algorithm 1) with two other models. One is a state-of-the-art incomplete information game solver [171] where uncertain agents infer intent via a Kalman filter and the certain player acts under a FNE in a complete information setting. We call this method passive game since any learning on the part of the uncertain agents is not explicitly planned for by the certain agent. We also compare with a complete information game model where all players have complete information about each others' intent. We study four hypotheses described in detail below.

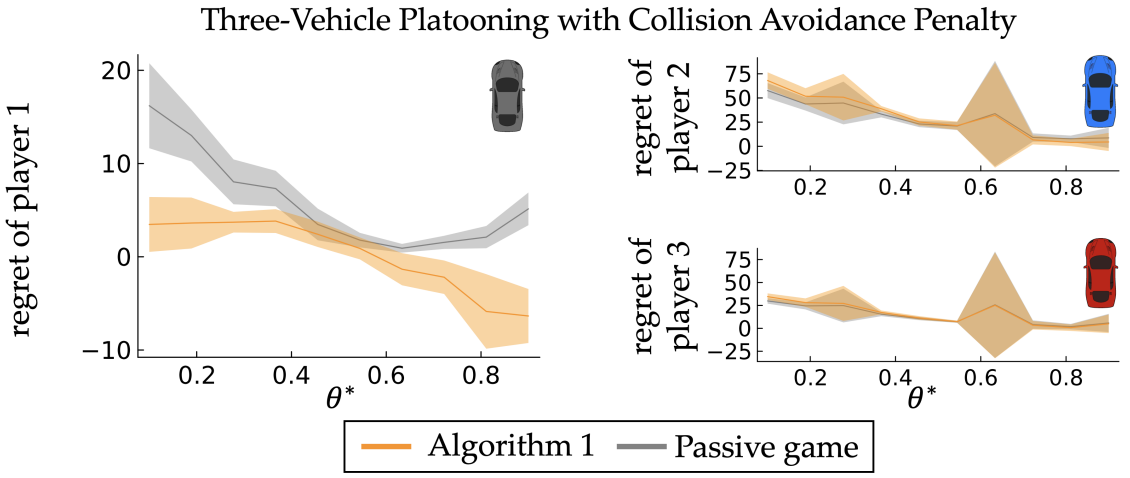


Figure 9.5: **Results: H3.** The regrets of the certain player (player 1) under the active teaching strategy are consistently lower compared to those under the passive teaching strategy, across different ground truth intents of the certain player. This empirically validates the claim in Proposition 12.

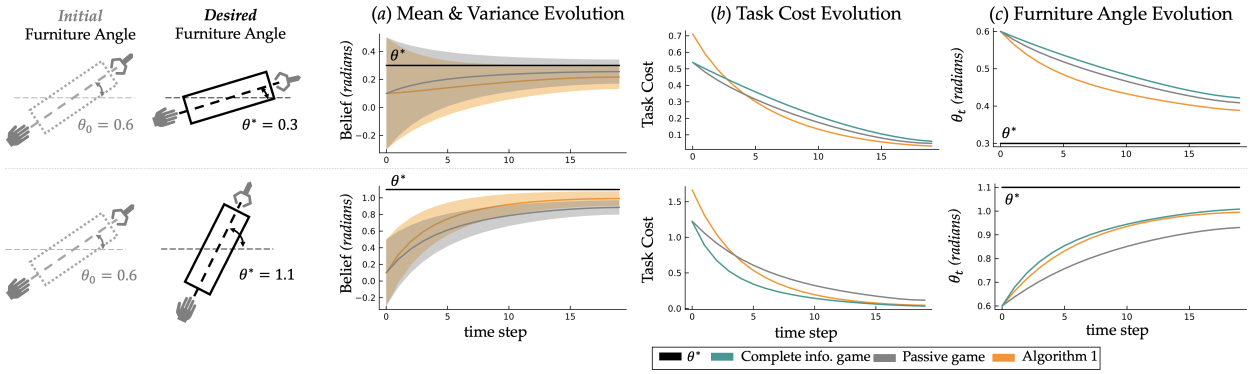


Figure 9.6: **Results: H4**. Even without explicit incentives to express intent, Algorithm 1 influences the uncertain agent's belief in a way that improves task cost over passive game (plot (b)). However, intent demonstration is strategic: if the state is already sufficiently good, our method pauses its influence (top left, plot (a)) but still achieves better task performance.

H1. *Uncertain agents coordinating under Algorithm 1 reduce uncertainty faster than passive game-theoretic models that do not account for agent learning.*

Setup and Metrics. We focus on the **Bi-Manual Robot Manipulation** environment where the uncertain agent maintains a point estimate. The uncertain silver robot initially believes that the red robot wants to grab the center of the pot, $\hat{\theta}_0 = 0$. We measure the convergence of $\hat{\theta}_t$ to θ^* under passive game and Algorithm 1. For our method, we also vary the hyperparameters ρ_1 and ρ_2, denoted by ratio $= \frac{\rho_2}{\rho_1}$, to study how different weight ratios between belief alignment and task performance affect the uncertain agents' learning.

Results. Figure 9.3 shows both quantitative and qualitative results. In complete information game, when all agents know what part of the pot they want to grab, then they coordinate seamlessly (left in Figure 9.3 (b)). However, under a passive game interaction model, the uncertain silver robot first moves towards the center of the pot and then moves to the correct handle at the last minute. The left plot in Figure 9.3 shows how with the passive game algorithm, the red robot doesn't take advantage of the silver agent's learning dynamics and thus its behavior doesn't enable fast learning. On the other hand, even when Algorithm 1 has low weight on intent demonstration, the red robot still actively influences the silver robot's estimation dynamics, enabling faster convergence than passive game. As more weight is put on intent demonstration, the red robot automatically *exaggerates* its motion towards the right handle, and the planned response of the silver robot is a more direct movement towards the complementary pot handle (right, Figure 9.3 (b)), supporting **H1**.

H2. *The pre-computed intent demonstration feedback policy $\bar{\pi}_t^1$ can efficiently convey the certain player's unforeseen changes in its intent without the need of recomputing $\bar{\pi}_t^1$.*

Setup and Metrics. We focus on the **Assistive Lunar Lander** environment. We measure the belief state and the physical state trajectories of the lunar lander under passive game and Algorithm 7. For our method, we set $\rho_1 = 1$ and $\rho_2 = 4$, to ensure the task performance as well as belief alignment.

Results. Although the feedback policy $\bar{\pi}_t^1$ is computed under the assumption that the certain agent's intent remains stationary, it can still effectively convey unforeseen changes in certain agent's intent during deployment. Figure 9.4 demonstrates that the autopilot's belief rapidly converges toward the initial target $\theta^* = 25$ during the interval $t \in [0, 20]$, and then adjusts efficiently to the updated human-preferred destination $\theta^* = 50$ when $t \geq 20$. This highlights the robustness of the feedback policy $\bar{\pi}_t^1$ in realistic scenarios where task objectives shift unexpectedly during deployment—situations not explicitly considered when computing $\bar{\pi}_t^1$, yet handled effectively due to the feedback policy's strong generalization to dynamically changing intents.

H3. *The certain agent can improve its task performance by teaching agents with uncertainty.*

Setup and Metrics. We focus on the **Three-Vehicle Platooning** environment. We measure each player's task regret by comparing the optimal state-action trajectory ξ^* under the complete information game with the executed state-action trajectory ξ under one of the incomplete information models: $\text{Regret}^i(\xi, \xi^*) := \sum_{t=0}^{T}[c^i(x_t, u_t) - c^i(x_t^*, u_t^*)]$. Lower regret indicates better performance. We set $\rho_1 = 1$ and $\rho_2 = 0$ to evaluate whether the policy $\bar{\pi}_t^1$ can strategically reduce the certain agent's task cost when prioritizing task performance.

Results. Figure 9.5 shows the regret of each player (y-axis) under all possible true intents of the certain player θ^* (x-axis) in both environments. Across both environments, Algorithm 1 achieves lower regret for the certain player 1 than the passive game approach. This indicates that the certain agent can exploit the estimation dynamics of the other players to improve its task performance, bringing the task regret down, supporting **H3**.

H4. *When $\rho_2 \equiv 0$, Algorithm 1 balances task performance and intent demonstration for the certain agent.*

Setup and Metrics. We evaluate our method in the **Furniture Moving** environment, focusing on (1) the uncertain agent's belief convergence and (2) the certain agent's task cost. To test whether the

certain agent can strategically influence belief without explicitly optimizing for intent demonstration, we set the intent demonstration hyperparameter to $\rho_2 = 0$, thereby prioritizing task performance. We consider two true furniture angle preferences: $\theta^* = 0.3$ rad ($\sim 17°$) and $\theta^* = 1.1$ rad ($\sim 63°$). The furniture's initial angle is always set to $\theta_0 = 0.6$ rad ($\sim 35°$), and the uncertain agent's initial belief is Gaussian, with mean 0.1 and variance 0.4.

Results. Even without explicit intent demonstration in the cost, Algorithm 1 enables the certain agent to influence the uncertain agent's belief to improve task cost compared to passive game (Figure 9.6 (a) and (b)). While the real furniture angle is always moved towards θ^* faster with Algorithm 1 than with passive game (plot (c)), we notice that when $\theta^* = 0.3$, the uncertain player's *belief* converges *slower* with ours than with the baseline (top plot (a)). This arises as a function of initial conditions. Since the initial furniture angle $\theta_0 = 0.6$ is quite close to the desired one $\theta^* = 0.3$, the certain agent minimizes their effort by focusing on task completion rather than correcting the uncertain agent's belief. However, when the initial and desired angles are very different, then it is worth the certain agent to correct the uncertain agent's belief to improve overall task performance, supporting **H4**.

9.7 Discussion

Conclusion. In this chapter, we studied intent demonstration in multi-agent general-sum games, a problem commonly encountered in game-theoretic control applications such as autonomous driving, multi-robot manipulation, shared control systems, and human-robot interactions. Theoretically, we proved a sufficient condition for the convergence of an uncertain agent's beliefs to the ground truth certain agent's intent. Additionally, we showed that the certain agent could achieve a higher task performance by strategically demonstrating its intent to the uncertain agents. Algorithmically, we proposed an efficient method to solve linear and nonlinear intent demonstration problems via iterative linear-quadratic approximations. Our empirical results show that intent demonstration accelerates the learning of uncertain agents, reduces task regret for players, and enables the certain agent to balance task performance with intent expression.

Limitations and Future Work. One modeling limitation of our framework is the assumption of a shared initial estimate. While this assumption may be reasonable based on the context (e.g., a strong prior on expected maneuvers at a four-way intersection in driving scenario), it remains an assumption that could be relaxed in future work. Additionally, future work could relax the assumption of knowing the estimate dynamics of uncertain agents. This can be achieved by having the certain agent first infer the estimate dynamics of the uncertain agents and then compute its optimal intent demonstration policies.

Appendix

Proof of Proposition 1: We approach the proof by showing that there exists a teaching policy under which the belief $\hat{\theta}_t^j$, where $j \in \{2, \dots, N\}$, converges to the ground truth parameter exponentially

fast. Substituting $u_t^1 = \pi_t^1(x_t; \theta)$ into player j's estimate dynamics, we have

$$\begin{aligned} \hat{\theta}_{t+1}^j &= \hat{\theta}_t^j - \alpha \nabla_{\hat{\theta}_t^j} \|u_t^1 - \pi_t^1(x_t; \hat{\theta}_t^j)\|_2^2 \\ &= \hat{\theta}_t^j + \alpha K_{t,\theta}^{1\top} K_{t,\theta}^1 (\theta - \hat{\theta}_t^j). \end{aligned} \tag{9.6}$$

Subtracting θ from both sides, we have $\hat{\theta}_{t+1}^j - \theta = (I - \alpha K_{t,\theta}^{1\top} K_{t,\theta}^1)(\hat{\theta}_t^j - \theta)$. Since the largest singular value of $(I - \alpha K_{t,\theta}^{1\top} K_{t,\theta}^1), \forall t \leq T$, denoted as c, is strict less than 1, we have $\|\hat{\theta}_{t+1}^j - \theta\|_2 \leq c\|\hat{\theta}_t^j - \theta\|_2$. Thus, there exists an active teaching policy, defined as $\bar{\pi}_t^1(x_t, \hat{\theta}_t; \theta) := \pi_t^1(x_t; \theta)$, that guarantees the exponential convergence of $\hat{\theta}_t^j$ towards θ^*. □

Proof of Proposition 2: First of all, we observe that $\{\tilde{u}_t^1\}_{t=0}^T$ and its resulted state trajectory $\{\tilde{x}_t\}_{t=0}^{T+1}$ is a feasible solution to (9.1). Thus, the optimal solution (9.1) leads to a cost value not greater than player 1's cost in complete information game. Moreover, when the Jacobian of $\tilde{c}_{t:T}^1$ with respect to $\tilde{u}_{t:T}^1$ is nonzero, by convexity of the cost c_t^1 [35, Section 4.2.3], for some $\epsilon > 0$, there exists a solution $\breve{u}_{t:T}^1 \in \{u_{t:T}^1 : \|u_{t:T}^1 - \tilde{u}_{t:T}^i\|_2 \leq \epsilon\}$ such that player 1's control $\breve{u}_{t:T}^1$ achieves a lower task cost value $\breve{c}_{t:T}^1$ than under the control $\tilde{u}_{t:T}^1$. This completes the proof. □

Chapter 10

Conclusion and Future Directions

This dissertation addresses key challenges in multi-agent autonomous systems by focusing on three interconnected research areas: safe learning-based control, efficient computation for dynamic game-theoretic decision-making, and methods that alleviate information asymmetry. By leveraging advanced methodologies from reachability analysis, differentiable optimization, and deep reinforcement learning (RL), we have developed solutions enabling safe, efficient, and intelligent interactions among decentralized autonomous agents operating in uncertain, dynamic environments.

In addressing multi-agent safety, we introduced a new reachability learning framework that integrates deep reinforcement learning with formal verification techniques. This framework provides certifiable safety guarantees, even under bounded disturbances, and was empirically validated through real-world drone racing experiments, demonstrating its practical effectiveness. Additionally, we extended classical optimization methods, particularly the augmented Lagrangian approach, into reinforcement learning contexts. These advancements effectively manage discrete action spaces and non-differentiable objectives and were further adapted to multi-agent reinforcement learning scenarios relevant to air mobility applications, highlighting their robustness and practical relevance.

The second chapter explored efficient computational methods for dynamic game-theoretic decision-making. Recognizing the complexity inherent in computing equilibrium solutions such as Nash and Stackelberg equilibria, we developed new computational frameworks specifically designed for dynamic feedback Stackelberg games. By formulating nested KKT conditions, we ensured recursive feasibility and consistent equilibrium solutions across sequential decision stages. Additionally, addressing the scalability limitations prevalent in large-scale networked systems such as power grids, air traffic management, and transportation networks, we exploited structural properties, such as the existence of a potential function in potential game frameworks, to develop scalable policy gradient algorithms. These contributions enhance computational efficiency, enabling practical deployment in large-scale multi-agent systems.

The third theme focused on the critical issue of information asymmetry in decentralized multi-agent decision-making processes. Traditional inference methodologies typically focus on static scenarios and fail to capture the complexities of dynamic, feedback-driven interactions. We advanced this area by developing robust methods capable of accurately inferring agent objectives from dynamically observed interaction data. These methods distinguish immediate strategic actions

from forward-looking strategic considerations. Further extending these inference capabilities, we explored strategic intent demonstrations, empowering agents to proactively influence other agents' belief updates and subsequent decisions. Thus, we demonstrated how leveraging information asymmetry strategically could significantly enhance overall system performance.

Future Research Directions

Looking ahead, the next generation of multi-agent autonomy must support strategic decision-making that approaches human-level reasoning. As multi-agent systems become increasingly central to robotics, autonomous infrastructure, and large-scale societal applications, the need for agents to act safely, strategically, and independently under uncertainty is more urgent than ever. Traditional centralized control paradigms face scalability and robustness limitations, motivating the development of decentralized, game-theoretic approaches that enable agents to reason locally and coordinate effectively at scale.

To move beyond global planners and full observability, agents must be equipped to reason about their own objectives, anticipate others' behavior, and adapt under uncertainty. This shift enables scalable deployment in real-world settings such as collaborative robotics, drone swarms, mixed-autonomy traffic, and smart grids. It also raises a central scientific question:

> *What information structures, learning mechanisms, and interaction protocols are necessary to ensure safe, efficient, and strategic decision-making in decentralized multi-agent systems, especially when centralized oversight is infeasible or undesirable?*

The following research directions build toward answering this question by integrating ideas from control, game theory, and learning into a cohesive roadmap for robust and socially intelligent autonomy.

10.1 Intent Inferability and Active Social Reasoning

A core enabler of strategic autonomy is understanding the *inferability of intent*: under what conditions can agents reliably infer others' goals, strategies, or preferences from observed behavior? Future research should formalize observability conditions in dynamic games, characterize the identifiability of latent objectives, and develop uncertainty-aware inference metrics. Beyond passive observation, agents may also engage in *active intent probing*—deliberately perturbing their behavior to elicit informative responses from others, mirroring human-like social learning.

Integrating inverse game-theoretic inference with real-time planning opens the door to intent-aware policies that proactively adapt in the presence of others. These ideas are particularly relevant in domains such as socially compliant driving, ad hoc human-agent collaboration, and multi-robot coordination in unknown or adversarial environments.

10.2 Theory of Mind and Hierarchical Belief Modeling

Building on intent inference, the next step is incorporating a *theory of mind*—the recursive capacity to reason about what others believe, know, or intend. Real-world agents must operate under deep informational asymmetries, far beyond the near-complete information typically assumed in classical models. Level-2 and level-K reasoning in reinforcement learning provide promising frameworks for capturing this recursive structure.

An exciting future direction is the development of *Bayesian games with latent beliefs*, where agents reason not only about others' goals but also about their understanding of the environment and the game structure itself—such as hidden constraints, reward functions, or roles. These approaches are especially critical for high-stakes applications like disaster response, competitive planning, or multi-agent negotiation under limited communication.

10.3 Reasoning under Partial Observability and Epistemic Uncertainty

Autonomous agents must often operate with incomplete knowledge of system states, dynamics, or other agents' behaviors. Strategic decision-making under such *epistemic uncertainty* remains a fundamental challenge. One promising direction is the study of *value-of-information games*, where agents actively balance task objectives with the need to acquire informative observations.

Extending reachability and viability analysis to multi-agent systems under partial observability can provide formal safety guarantees even with limited information. Meanwhile, cooperative exploration and distributed sensing strategies enable agents to reduce uncertainty collectively, supporting applications such as decentralized monitoring, collaborative mapping, and search-and-rescue missions.

10.4 Safe Coordination in Dynamic, Heterogeneous Environments

In realistic deployments, multi-agent systems must withstand dynamic environments with communication dropouts, unexpected agent failures, or conflicting instructions. This calls for robust coordination protocols that adapt to changes while preserving global safety and performance.

Potential research directions include *contract-based planning*, where agents make local commitments with fallback guarantees, and *emergent social institutions*, where coordination norms and roles evolve through repeated decentralized interactions. Such frameworks are critical for long-term collaboration in applications like warehouse automation, household robotics, and human-robot teaming.

10.5 Safe Decentralized Learning under Incomplete Information

To realize scalable autonomy, agents must learn and adapt in real time under structural, informational, and safety constraints. Extending reachability analysis and game-theoretic equilibria into multi-agent reinforcement learning is a promising path forward. These tools allow agents to optimize constrained objectives in a distributed manner while respecting local safety and coordination requirements.

Future research should explore how to incorporate communication topology, temporal dependencies, and delayed feedback into decentralized learning algorithms. This may enable *multi-agent policy inference*, where agents estimate others' gradients or value functions from partial observations. Additionally, *curriculum learning* and hierarchical task decomposition can accelerate training and improve generalization in complex domains like urban air mobility, distributed energy control, and collaborative manipulation.

10.6 Toward Socially Intelligent Multi-Agent Systems

The research directions described above aim to contribute toward the development of autonomous systems that are not only capable and efficient, but also socially and strategically aware. These systems should be able to reason about fairness, shared goals, and social norms; collaborate effectively with humans; and make decisions transparently in the presence of uncertainty and disagreement. By integrating tools from control theory, game theory, and reinforcement learning, this work seeks to explore how safety, strategic reasoning, and belief modeling can jointly support more robust and intelligent multi-agent decision-making.

While the results presented in this dissertation offer several contributions, many challenges remain. Progress toward truly socially intelligent systems will likely require insights from multiple disciplines, including cognitive science, communication, and economics. As autonomous agents become more present in human environments, important questions arise: How should agents balance individual goals with collective outcomes? How should they communicate uncertainty and intent? And how can we ensure their decisions are interpretable, fair, and aligned with human values?

This dissertation represents a small step in addressing these broader challenges. The approaches developed across safety assurance, intent inference, equilibrium computation, and decentralized learning are intended as foundational tools that others may build upon. It is my hope that this work will contribute to the continued advancement of reliable and responsible multi-agent autonomy in domains such as robotics, transportation, logistics, and human-robot collaboration.

Bibliography

[1] J. Achiam et al. 'Constrained policy optimization'. In: *International Conference on Machine Learning*. PMLR. 2017.

[2] B. Acikmese and S. R. Ploen. 'Convex programming approach to powered descent guidance for mars landing'. In: *Journal of Guidance, Control, and Dynamics* 30.5 (2007), pp. 1353–1366.

[3] F. A. Administration. *Urban Air Mobility Concept of Operations 2.0*. Technical report. Federal Aviation Administration, 2023.

[4] A. Agarwal et al. 'On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift'. In: *Journal of Machine Learning Research* 22.98 (2021), pp. 1–76. ISSN: 1533-7928. URL: http://jmlr.org/papers/v22/19-736.html (visited on 14/01/2025).

[5] M. Alshiekh et al. 'Safe reinforcement learning via shielding'. In: *Proceedings of the AAAI conference on artificial intelligence*. 2018.

[6] P. Alstone et al. *2025 California Demand Response Potential Study - Charting California's Demand Response Future. Final Report on Phase 2 Results*. en. Tech. rep. 1421800. Mar. 2017, p. 1421800. DOI: 10.2172/1421800. URL: http://www.osti.gov/servlets/purl/1421800/ (visited on 25/03/2025).

[7] E. Altman. *Constrained Markov decision processes*. Vol. 7. CRC Press, 1999.

[8] L. E. Alvarez et al. 'ACAS sXu: Robust decentralized detect and avoid for small unmanned aircraft systems'. In: *IEEE/AIAA Digital Avionics Systems Conference (DASC)*. 2019.

[9] A. D. Ames et al. 'Control barrier function based quadratic programs for safety critical systems'. In: *IEEE Transactions on Automatic Control* 62.8 (2016), pp. 3861–3876.

[10] P.-A. Andersen, M. Goodwin and O.-C. Granmo. 'The dreaming variational autoencoder for reinforcement learning environments'. In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer. 2018, pp. 143–155.

[11] J. D. Anderson. 'Introduction to Flight'. In: 8th. New York, NY: McGraw-Hill Education, 2016.

[12] Archer Aviation. *Archer Aviation - Our Aircraft*. Accessed: 2025-01-31. 2025. URL: https://archer.com/aircraft.

[13] S. Asayesh et al. 'Least-restrictive multi-agent collision avoidance via deep meta reinforcement learning and optimal control'. In: *International Conference on Robot Intelligence Technology and Applications*. Springer. 2022.

[14] P. Auer. 'Using confidence bounds for exploitation-exploration trade-offs'. In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 397–422.

[15] C. Awasthi and A. Lamperski. 'Inverse differential games with mixed inequality constraints'. In: *2020 American control conference (ACC)*. IEEE. 2020, pp. 2182–2187.

[16] L. Bai et al. 'Distribution Locational Marginal Pricing (DLMP) for Congestion Management and Voltage Support'. In: *IEEE Transactions on Power Systems* 33.4 (July 2018), pp. 4061–4073. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2017.2767632. URL: https://ieeexplore.ieee.org/document/8089425 (visited on 04/09/2024).

[17] Y. Bai et al. 'Sample-efficient learning of Stackelberg equilibria in general-sum games'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25799–25811.

[18] L. Bakule and M. Straškraba. 'On structural control strategies in aquatic ecosystems'. In: *Ecological Modelling* 39.1-2 (1987), pp. 171–180.

[19] S. Bansal and C. J. Tomlin. 'Deepreach: A deep learning approach to high-dimensional reachability'. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2021.

[20] S. Bansal et al. 'Hamilton-Jacobi Reachability: A Brief Overview and Recent Advances'. In: IEEE. 2017, pp. 2242–2253.

[21] M. Baran and F. Wu. 'Network reconfiguration in distribution systems for loss reduction and load balancing'. In: *IEEE Transactions on Power Delivery* 4.2 (Apr. 1989), pp. 1401–1407. ISSN: 1937-4208. DOI: 10.1109/61.25627. URL: https://ieeexplore.ieee.org/abstract/document/25627 (visited on 29/03/2025).

[22] R. G. Bartle and J. T. Joichi. 'The preservation of convergence of measurable functions under composition'. In: *Proceedings of the American Mathematical Society* 12.1 (1961), pp. 122–126.

[23] T. Basar. 'On the uniqueness of the Nash solution in linear-quadratic differential games'. In: *International Journal of Game Theory* 5 (1976), pp. 65–90.

[24] T. Başar and G. J. Olsder. *Dynamic noncooperative game theory*. SIAM, 1999.

[25] O. Bastani, Y. Pu and A. Solar-Lezama. 'Verifiable reinforcement learning via policy extraction'. In: *Advances in neural information processing systems*. 2018, pp. 2494–2504.

[26] J.-L. Bastarache, C. Nielsen and S. L. Smith. 'On Legible and Predictable Robot Navigation in Multi-Agent Environments'. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 5508–5514.

[27] F. P. Bejarano, L. Brunke and A. P. Schoellig. 'Safety Filtering While Training: Improving the Performance and Sample Efficiency of Reinforcement Learning Agents'. In: *IEEE Robotics and Automation Letters* (2024).

[28] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[29] A. Bensoussan et al. 'Feedback Stackelberg–Nash Equilibria in Mixed Leadership Games with an Application to Cooperative Advertising'. In: *SIAM Journal on Control and Optimization* 57.5 (2019), pp. 3413–3444.

[30] J. Bertram and P. Wei. 'Distributed computational guidance for high-density urban air mobility with cooperative and non-cooperative collision avoidance'. In: *AIAA Scitech Forum*. 2020.

[31] D. Bertsekas. *Dynamic programming and optimal control: Volume I*. Vol. 1. Athena scientific, 2012.

[32] T. Besselmann, J. Lofberg and M. Morari. 'Explicit MPC for LPV systems: Stability and optimality'. In: *IEEE Transactions on Automatic Control* 57.9 (2012), pp. 2322–2332.

[33] M. Bhatt, Y. Jia and N. Mehr. 'Efficient constrained multi-agent trajectory optimization using dynamic potential games'. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2023.

[34] E. Börgens and C. Kanzow. 'ADMM-Type methods for Generalized Nash Equilibrium Problems in Hilbert Spaces'. In: *SIAM Journal on Optimization* 31.1 (2021), pp. 377–403.

[35] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[36] S. Boyd et al. 'Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers'. In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.

[37] K. Bredies and H. Sun. 'Preconditioned Douglas-Rachford Splitting Methods for Convex-concave Saddle-point Problems'. In: *SIAM Journal on Numerical Analysis* 53.1 (2015), pp. 421–444.

[38] E. Brock et al. 'Coordinating Distributed Energy Resources with Nodal Pricing in Distribution Networks: a Game-Theoretic Approach'. In: *arXiv preprint arXiv:2503.24342* (2025).

[39] G. Brockman et al. 'Openai gym'. In: *arXiv preprint arXiv:1606.01540* (2016).

[40] V. Bulusu et al. 'A Traffic Demand Analysis Method for Urban Air Mobility'. In: *IEEE Transactions on Intelligent Transportation Systems* (2021).

[41] G. Calafiore and M. C. Campi. 'Uncertain Convex Programs: Randomized Solutions and Confidence Levels'. In: *Mathematical Programming* 102.1 (2005), pp. 25–46.

[42] G. C. Calafiore and L. Fagiano. 'Robust Model Predictive Control via Scenario Optimization'. In: *IEEE Transactions on Automatic Control* 58.1 (2012), pp. 219–224.

[43] G. C. Calafiore and M. C. Campi. 'The Scenario Approach to Robust Control Design'. In: *IEEE Transactions on automatic control* 51.5 (2006), pp. 742–753.

[44] M. C. Campi and S. Garatti. 'The Exact Feasibility of Randomized Solutions of Uncertain Convex Programs'. In: *SIAM Journal on Optimization* 19.3 (2008), pp. 1211–1230.

[45] M. C. Campi, S. Garatti and F. A. Ramponi. 'A General Scenario Theory for Nonconvex Optimization and Decision Making'. In: *IEEE Transactions on Automatic Control* 63.12 (2018), pp. 4067–4078.

[46] A.-L. Cauchy. 'Cours d'analyse de l'Ecole royale polytechnique, 1re partie'. In: *Analyse algébrique. Debure freres, Paris* (1821).

[47] M. Chahine et al. 'Intention Communication and Hypothesis Likelihood in Game-Theoretic Motion Planning'. In: *IEEE Robotics and Automation Letters* 8.3 (2023), pp. 1223–1230.

[48] A. Charnes and D. Granot. *Prior Solutions: Extensions of Convex Nucleus Solutions to Chance-Constrained Games.* Tech. rep. Texas Univ Austin Center For Cybernetic Studies, 1973.

[49] C. Chen et al. 'Wholesale Market Participation of DERAs: DSO-DERA-ISO Coordination'. In: *IEEE Transactions on Power Systems* 39.5 (Sept. 2024), pp. 6605–6614. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2024.3352003. URL: https://ieeexplore.ieee.org/abstract/document/10398499 (visited on 13/03/2025).

[50] M. Chen, J. C. Shih and C. J. Tomlin. 'Multi-vehicle collision avoidance via hamilton-jacobi reachability and mixed integer programming'. In: *IEEE Conference on Decision and Control (CDC)*. 2016.

[51] M. Chen et al. 'Reachability-based safety and goal satisfaction of unmanned aerial platoons on air highways'. In: *Journal of Guidance, Control, and Dynamics* (2017).

[52] O. Chen and M. Ben-Akiva. 'Game-theoretic formulations of interaction between dynamic traffic control and dynamic traffic assignment'. In: *Transportation Research Record* 1617.1 (1998), pp. 179–188.

[53] S. Chen et al. 'Integrated Conflict Management for UAM With Strategic Demand Capacity Balancing and Learning-Based Tactical Deconfliction'. In: *IEEE Transactions on Intelligent Transportation Systems* (2024).

[54] Y. Chen and T. Li. 'Decentralized Policy Gradient for Nash Equilibria Learning of General-sum Stochastic Games'. In: *arXiv preprint arXiv:2210.07651* (2022).

[55] Y. Chen et al. 'An Energy Sharing Mechanism Considering Network Constraints and Market Power Limitation'. In: *IEEE Transactions on Smart Grid* 14.2 (Mar. 2023), pp. 1027–1041. ISSN: 1949-3061. DOI: 10.1109/TSG.2022.3198721. URL: https://ieeexplore.ieee.org/abstract/document/9857667 (visited on 22/11/2024).

[56] V. K. Chilakamarri, Z. Feng and S. Bansal. 'Reachability analysis for black-box dynamical systems'. In: *arXiv preprint arXiv:2410.07796* (2024).

[57] R. Chinchilla, G. Yang and J. P. Hespanha. 'Newton and interior-point methods for (constrained) nonconvex–nonconcave minmax optimization with stability and instability guarantees'. In: *Mathematics of Control, Signals, and Systems* (2023), pp. 1–41.

[58] J. J. Choi et al. 'A forward reachability perspective on robust control invariance and discount factors in reachability analysis'. In: *arXiv preprint arXiv:2310.17180* (2023).

[59] J. J. Choi et al. 'Robust control barrier–value functions for safety-critical control'. In: *IEEE Conference on Decision and Control (CDC)*. 2021.

[60] J. J. Choi et al. 'Resolving Conflicting Constraints in Multi-Agent Reinforcement Learning with Layered Safety'. In: ().

[61] Y. Chow et al. 'A Lyapunov-based Approach to Safe Reinforcement Learning'. In: *Advances in neural information processing systems*. 2018.

[62] Y. Chow et al. 'Lyapunov-based safe policy optimization for continuous control'. In: *arXiv preprint arXiv:1901.10031* (2019).

[63] Y. Chow et al. 'Risk-constrained reinforcement learning with percentile risk criteria'. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6070–6120.

[64] K. Chu, M. Lee and M. Sunwoo. 'Local path planning for off-road autonomous driving with avoidance of static obstacles'. In: *IEEE Transactions on Intelligent Transportation Systems* (2012).

[65] S. L. Cleac'h, M. Schwager and Z. Manchester. 'ALGAMES: A fast solver for constrained dynamic games'. In: *arXiv preprint arXiv:1910.09713* (2019).

[66] V. Conitzer and T. Sandholm. 'New Complexity Results About Nash Equilibria'. In: *Games and Economic Behavior* 63.2 (2008), pp. 621–641.

[67] S. Coogan. 'Mixed monotonicity for reachability and safety in dynamical systems'. In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 5074–5085.

[68] E. Cruck and J. Lygeros. 'Subliminal air traffic control: Human friendly control of a multi-agent system'. In: *American Control Conference*. 2007.

[69] J. Cruz Jr. 'Survey of nash and stackelberg equilibrim strategies in dynamic games'. In: *Annals of Economic and Social Measurement, Volume 4, number 2*. NBER, 1975, pp. 339–344.

[70] C. Daskalakis, P. W. Goldberg and C. H. Papadimitriou. 'The Complexity of Computing a Nash Equilibrium'. In: *SIAM Journal on Computing* 39.1 (2009), pp. 195–259.

[71] M. Dawood et al. 'Safe Multi-Agent Reinforcement Learning for Behavior-Based Cooperative Navigation'. In: *arXiv preprint arXiv:2312.12861* (2023).

[72] R. S. Dembo. 'Scenario Optimization'. In: *Annals of Operations Research* 30.1 (1991), pp. 63–80.

[73] J. Diakonikolas. 'Halpern Iteration for Near-Optimal and Parameter-Free Monotone Inclusion and Strong Solutions to Variational Inequalities'. In: *Conference on Learning Theory*. PMLR. 2020, pp. 1428–1451.

[74] S. P. Dirkse and M. C. Ferris. 'The PATH Solver: a Non-monotone Stabilization Scheme for Mixed Complementarity Problems'. In: *Optimization methods and software* 5.2 (1995), pp. 123–156.

[75] E. Dockner. *Differential games in economics and management science*. Cambridge University Press, 2000.

[76] A. D. Dragan, K. C. Lee and S. S. Srinivasa. 'Legibility and predictability of robot motion'. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 301–308.

[77] Y. Duan et al. 'Benchmarking deep reinforcement learning for continuous control'. In: *International conference on machine learning*. PMLR. 2016, pp. 1329–1338.

[78] J. Durbin and S. J. Koopman, eds. *Time Series Analysis by State Space Methods*. Oxford University Press, May 2012. ISBN: 978-0-19-964117-8. DOI: 10.1093/acprof:oso/9780199641178.003.0010. URL: https://doi.org/10.1093/acprof:oso/9780199641178.003.0010.

[79] M. Elhenawy et al. 'An Intersection Game-theory-based Traffic Control Algorithm in a Connected Vehicle Environment'. In: *itsc*. 2015.

[80] I. ElSayed-Aly et al. 'Safe Multi-Agent Reinforcement Learning via Shielding'. In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2021.

[81] P. Englert, N. A. Vien and M. Toussaint. 'Inverse KKT: Learning cost functions of manipulation tasks from demonstrations'. In: *The International Journal of Robotics Research* 36.13-14 (2017), pp. 1474–1488.

[82] U. Eren et al. 'Model predictive control in aerospace systems: Current state and opportunities'. In: *Journal of Guidance, Control, and Dynamics* (2017).

[83] H. Erzberger and K. Heere. 'Algorithm and operational concept for resolving short-range conflicts'. In: *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering* (2010).

[84] H. Erzberger, T. A. Lauderdale and Y.-C. Chu. 'Automated conflict resolution, arrival management, and weather avoidance for air traffic management'. In: *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of aerospace engineering* (2012).

[85] A. Esteva et al. 'A guide to deep learning in healthcare'. In: *Nature medicine* 25.1 (2019), pp. 24–29.

[86] A. Evans, V. Vaze and C. Barnhart. 'Airline-driven performance-based air traffic management: Game theoretic models and multicriteria evaluation'. In: *Transportation Science* (2016).

[87] M. Everett, G. Habibi and J. P. How. 'Efficient reachability analysis of closed-loop systems with neural network controllers'. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 4384–4390.

[88] F. Fabiani, K. Margellos and P. J. Goulart. 'Probabilistic Feasibility Guarantees for Solution Sets to Uncertain Variational Inequalities'. In: *Automatica* 137 (2022), p. 110120.

[89] F. Fabiani et al. 'Local Stackelberg equilibrium seeking in generalized aggregative games'. In: *IEEE Transactions on Automatic Control* 67.2 (2021), pp. 965–970.

[90] F. Facchinei and C. Kanzow. 'Generalized Nash Equilibrium Problems'. In: *Annals of Operations Research* 175.1 (2010), pp. 177–211.

[91] F. Facchinei and J.-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

[92] Federal Aviation Administration. *Aeronautical Information Manual (AIM), Chapter 7: Safety of Flight, Section 7: Near Midair Collision Reporting*. `https://www.faa.gov/air_traffic/publications/atpubs/aim_html/chap7_section_7.html`. Accessed: 26 January 2025.

[93] Federal Aviation Administration. *Pilot's Handbook of Aeronautical Knowledge*. U.S. Department of Transportation, Federal Aviation Administration, 2023.

[94] F. Fele and K. Margellos. 'Probabilistic Sensitivity of Nash Equilibria in Multi-Agent Games: a Wait-and-Judge Approach'. In: *cdc*. IEEE. 2019, pp. 5026–5031.

[95] S. El-Ferik, B. A. Siddiqui and F. L. Lewis. 'Distributed nonlinear MPC of multi-agent systems with data compression and random delays'. In: *IEEE Transactions on Automatic Control* (2015).

[96] M. C. Ferris, S. P. Dirkse and A. Meeraus. 'Mathematical Programs with Equilibrium Constraints: Automatic Reformulation and Solution via Constrained Optimization'. In: *Frontiers in applied general equilibrium modeling* (2005), pp. 67–93.

[97] A. M. Fink. 'Equilibrium in a Stochastic n-Person Game'. In: *Journal of science of the hiroshima university, series ai (mathematics)* 28.1 (1964), pp. 89–93.

[98] J. F. Fisac et al. 'Bridging hamilton-jacobi safety analysis and reinforcement learning'. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2019.

[99] J. F. Fisac et al. 'Reach-Avoid Problems with Time-Varying Dynamics, Targets and Constraints'. In: 2015, pp. 11–20.

[100] J. F. Fisac et al. 'A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems'. In: *IEEE Transactions on Automatic Control* (2019).

[101] S. D. Flåm. 'Paths to Constrained Nash Equilibria'. In: *Applied Mathematics and Optimization* 27 (1993), pp. 275–289.

[102] A. A. Flem et al. 'Experimental Characterization of Propeller-Induced Flow (PIF) below a Multi-Rotor UAV'. In: *Atmosphere* 15.3 (2024), p. 242.

[103] P. Franceschi, N. Pedrocchi and M. Beschi. 'Human–Robot Role Arbitration via Differential Game Theory'. In: *IEEE Transactions on Automation Science and Engineering* (2023).

[104] D. Fridovich-Keil et al. 'Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games'. In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 1475–1481.

[105] D. Gabay and B. Mercier. 'A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximation'. In: *Computers & mathematics with applications* 2.1 (1976), pp. 17–40.

[106] V. Gabler et al. 'A game-theoretic approach for adaptive action selection in close proximity human-robot-collaboration'. In: *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2017, pp. 2897–2903.

[107] B. Gardner and J. Cruz. 'Feedback Stackelberg strategy for M-level hierarchical games'. In: *IEEE Transactions on Automatic Control* 23.3 (1978), pp. 489–491.

[108] R. Ghosh and C. Tomlin. 'Maneuver design for multiple aircraft conflict resolution'. In: *American Control Conference*. 2000.

[109] W. Giernacki et al. 'Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering'. In: *the 22nd International Conference on Methods and Models in Automation and Robotics*. 2017, pp. 37–42. DOI: 10.1109/MMAR.2017.8046794.

[110] R. Glowinski and A. Marroco. 'Sur l'approximation, par Éléments Finis d'Ordre un, et la Résolution, par Pénalisation-dualité d'une Classe de Problèmes de Dirichlet non Linéaires'. In: *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique* 9.R2 (1975), pp. 41–76.

[111] M. Goarin et al. 'Decentralized nonlinear model predictive control for safe collision avoidance in quadrotor teams with limited detection range'. In: *arXiv preprint arXiv:2409.17379* (2024).

[112] N. Goswami, S. K. Mondal and S. Paruya. 'A comparative study of dual active-set and primal-dual interior-point method'. In: *IFAC Proceedings Volumes* 45.15 (2012), pp. 620–625.

[113] H. Gouk et al. 'Regularisation of neural networks by enforcing lipschitz continuity'. In: *Machine Learning* 110.2 (2021), pp. 393–416.

[114] W. Grady, M. Samotyj and A. Noyola. 'The application of network objective functions for actively minimizing the impact of voltage harmonics in power systems'. In: *IEEE Transactions on Power Delivery* 7.3 (July 1992), pp. 1379–1386. ISSN: 1937-4208. DOI: 10.1109/61.141855. URL: https://ieeexplore.ieee.org/document/141855 (visited on 14/02/2025).

[115] D. Graves, K. Rezaee and S. Scheideman. 'Perception as prediction using general value functions in autonomous driving applications'. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 1202–1209.

[116] S. Gros, M. Zanon and A. Bemporad. 'Safe reinforcement learning via projection on a safe set: How to achieve optimality?' In: *arXiv preprint arXiv:2004.00915* (2020).

[117] S. Gu et al. 'Multi-agent constrained policy optimisation'. In: *arXiv preprint arXiv:2110.02793* (2021).

[118] S. Gu et al. 'Continuous deep Q-learning with model-based acceleration'. In: *International Conference on Machine Learning*. 2016, pp. 2829–2838.

[119] S. Gu et al. 'Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates'. In: *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2017, pp. 3389–3396.

[120] J. Guerrero et al. 'Towards a transactive energy system for integration of distributed energy resources: Home energy management, distributed optimal power flow, and peer-to-peer energy trading'. In: *Renewable and Sustainable Energy Reviews* 132 (Oct. 2020), p. 110000. ISSN: 1364-0321. DOI: 10.1016/j.rser.2020.110000. URL: https://www.sciencedirect.com/science/article/pii/S1364032120302914 (visited on 25/03/2025).

[121] T. Haarnoja et al. 'Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor'. In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.

[122] M. W. Harris and B. Açıkmeşe. 'Minimum Time Rendezvous of Multiple Spacecraft using Differential Drag'. In: *Journal of Guidance, Control, and Dynamics* 37.2 (2014), pp. 365–373.

[123] J. C. Harsanyi. 'Games with Incomplete Information Played by "Bayesian" players, I–III Part I. The Basic Model'. In: *Management science* 14.3 (1967), pp. 159–182.

[124] E. Hazan et al. 'Provably efficient maximum entropy exploration'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2681–2691.

[125] B. He et al. 'A New Inexact Alternating Directions Method for Monotone Variational Inequalities'. In: *Mathematical Programming* 92 (2002), pp. 103–118.

[126] X. He et al. 'A survey of Stackelberg differential game models in supply and marketing channels'. In: *Journal of Systems Science and Systems Engineering* 16 (2007), pp. 385–413.

[127] N. Heess et al. 'Learning Continuous Control Policies by Stochastic Value Gradients'. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/hash/148510031349642de5ca0c544f31b2ef-Abstract.html (visited on 05/03/2025).

[128] J. Herrera de la Cruz, B. Ivorra and Á. M. Ramos. 'An Algorithm for Solving a Class of Multiplayer Feedback-Nash Differential Games'. In: *Mathematical Problems in Engineering* 2019 (2019).

[129] Y.-C. Ho, P. Luh and R. Muralidharan. 'Information structure, Stackelberg games, and incentive controllability'. In: *IEEE Transactions on Automatic Control* 26.2 (1981), pp. 454–460.

[130] J. E. Holland, M. J. Kochenderfer and W. A. Olson. 'Optimizing the next generation collision avoidance system for safe, suitable, and acceptable operational performance'. In: *Air Traffic Control Quarterly* (2013).

[131] T. Homem-de-Mello and G. Bayraksan. 'Monte Carlo Sampling-based Methods for Stochastic Optimization'. In: *Surveys in Operations Research and Management Science* 19.1 (2014), pp. 56–85.

[132] K.-C. Hsu, H. Hu and J. F. Fisac. 'The safety filter: A unified view of safety-critical control in autonomous systems'. In: *Annual Review of Control, Robotics, and Autonomous Systems* (2023).

[133] K.-C. Hsu, D. P. Nguyen and J. F. Fisac. 'Isaacs: Iterative soft adversarial actor-critic for safety'. In: *Learning for Dynamics and Control Conference*. PMLR. 2023.

[134] K.-C. Hsu et al. 'Safety and Liveness Guarantees through Reach-Avoid Reinforcement Learning'. In: *Proceedings of Robotics: Science and Systems*. 2021.

[135] K.-C. Hsu et al. 'Sim-to-lab-to-real: Safe reinforcement learning with shielding and generalization guarantees'. In: *Artificial Intelligence* 314 (2023), p. 103811.

[136] H. Hu and J. F. Fisac. 'Active uncertainty learning for human-robot interaction: An implicit dual control approach'. In: *arXiv preprint arXiv:2202.07720* (2022).

[137] H. Hu et al. 'Emergent coordination through game-induced nonlinear opinion dynamics'. In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023, pp. 8122–8129.

[138] H. Hu et al. 'Who Plays First? Optimizing the Order of Play in Stackelberg Games with Many Robots'. In: *arXiv preprint arXiv:2402.09246* (2024).

[139] H. Hu et al. 'Reach-sdp: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming'. In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 5929–5934.

[140] Y. Hu, X. Chen and N. He. 'Sample Complexity of Sample Average Approximation for Conditional Stochastic Optimization'. In: *SIAM Journal on Optimization* 30.3 (2020), pp. 2103–2133.

[141] L. Huang and Q. Zhu. 'Dynamic bayesian games for adversarial and defensive cyber deception'. In: *Autonomous Cyber Deception: Reasoning, Adaptive Planning, and Evaluation of HoneyThings* (2019), pp. 75–97.

[142] J. Inga et al. 'Inverse dynamic games based on maximum entropy inverse reinforcement learning'. In: *arXiv preprint arXiv:1911.07503* (2019).

[143] J. Inga et al. 'Solution Sets for Inverse Non-Cooperative Linear-Quadratic Differential Games'. In: *IEEE Control Systems Letters* 3.4 (2019), pp. 871–876.

[144] R. Isaacs. *Differential games: a mathematical theory with applications to warfare and pursuit, control and optimization*. Courier Corporation, 1999.

[145] R. Janin. *Directional derivative of the marginal function in nonlinear programming*. Springer, 1984.

[146] M. Jankovic, M. Santillo and Y. Wang. 'Multiagent systems with CBF-based controllers: Collision avoidance and liveness from instability'. In: *IEEE Transactions on Control Systems Technology* (2023).

[147] F. J. Jiang et al. 'Guaranteed Completion of Complex Tasks via Temporal Logic Trees and Hamilton-Jacobi Reachability'. In: *arXiv preprint arXiv:2404.08334* (2024).

[148] R. Jiang and A. Mokhtari. 'Generalized Optimistic Methods for Convex-concave Saddle Point Problems'. In: *arXiv preprint arXiv:2202.09674* (2022).

[149] Joby Aviation. *Joby Aviation - Official Website*. Accessed: 2025-01-31. 2025. URL: https://www.jobyaviation.com/.

[150] M. Jungers. 'Feedback strategies for discrete-time linear-quadratic two-player descriptor games'. In: *Linear Algebra and its Applications* 440 (2014), pp. 1–23.

[151] G. Kalweit et al. *Deep constrained Q-learning*. 2020. arXiv: 2003.09398 [cs.LG].

[152] M. O. Karabag, D. Fridovich-Keil and U. Topcu. 'Alternating Direction Method of Multipliers for Decomposable Saddle-Point Problems'. In: *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2022.

[153] S. Karaman and E. Frazzoli. 'Sampling-based algorithms for optimal motion planning'. In: *The international journal of robotics research* 30.7 (2011), pp. 846–894.

[154] H. Khan and D. Fridovich-Keil. 'Leadership Inference for Multi-Agent Interactions'. In: *arXiv preprint arXiv:2310.18171* (2023).

[155] M. Kimmel and S. Hirche. 'Invariance control for safe human–robot interaction in dynamic environments'. In: *IEEE Transactions on Robotics* 33.6 (2017), pp. 1327–1342.

[156] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. SIAM, 2000.

[157] P. Kopardekar et al. 'Unmanned Aircraft System Traffic Management (UTM) Concept of Operations'. In: 2016.

[158] M. Korda. 'Stability and performance verification of dynamical systems controlled by neural networks: algorithms and complexity'. In: *IEEE Control Systems Letters* 6 (2022), pp. 3265–3270.

[159] G. M. Korpelevich. 'The Extragradient Method for Finding Saddle Points and Other Problems'. In: *Matecon* 12 (1976), pp. 747–756.

[160] D. Korzhyk et al. 'Stackelberg vs. Nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness'. In: *Journal of Artificial Intelligence Research* 41 (2011), pp. 297–327.

[161] G. Kossioris et al. 'Feedback Nash equilibria for non-linear differential games in pollution control'. In: *Journal of Economic Dynamics and Control* 32.4 (2008), pp. 1312–1331.

[162] S. G. Krantz and H. R. Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

[163] J. Kuchar et al. 'A safety analysis process for the traffic alert and collision avoidance system (TCAS) and see-and-avoid systems on remotely piloted vehicles'. In: *AIAA 3rd "Unmanned Unlimited" Technical Conference, Workshop and Exhibit*. 2004.

[164] Y. Kwon et al. 'Conformalized Reachable Sets for Obstacle Avoidance With Spheres'. In: *CoRL Workshop on SAFE-ROL*.

[165] S. A. S. Lab. *hj_reachability: Hamilton-Jacobi reachability analysis in JAX*. Accessed: 2025-01-28. 2025.

[166] F. Laine et al. 'Multi-hypothesis interactions in game-theoretic motion planning'. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 8016–8023.

[167] F. Laine et al. 'The computation of approximate generalized feedback nash equilibria'. In: *SIAM Journal on Optimization* 33.1 (2023), pp. 294–318.

[168] S. M. LaValle and S. Hutchinson. 'Game theory as a unifying structure for a variety of robot tasks'. In: *Proceedings of 8th IEEE international symposium on intelligent control*. IEEE. 1993, pp. 429–434.

[169] H. Le Cadre, Y. Mou and H. Höschle. 'Parametrized Inexact-ADMM to Span the Set of Generalized Nash Equilibria: A Normalized Equilibrium Approach'. In: (2020).

[170] S. Le Cleac'h, M. Schwager and Z. Manchester. 'ALGAMES: a fast augmented Lagrangian solver for constrained dynamic games'. In: *Autonomous Robots* 46.1 (2022), pp. 201–215.

[171] S. Le Cleac'h, M. Schwager and Z. Manchester. 'LUCIDGames: Online unscented inverse dynamic games for adaptive trajectory prediction and planning'. In: *IEEE Robotics and Automation Letters* 6.3 (2021), pp. 5485–5492.

[172] D. Lee, M. Chen and C. J. Tomlin. 'Removing Leaking Corners to Reduce Dimensionality in Hamilton-Jacobi Reachability'. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2019.

[173] K. W. Lee and J.-H. Hwang. 'Human–robot interaction as a cooperative game'. In: *Trends in Intelligent Systems and Computer Engineering* (2008), pp. 91–103.

[174] S. Lee et al. 'Preliminary Analysis of Separation Standards for Urban Air Mobility using Unmitigated Fast-Time Simulation'. In: *IEEE/AIAA Digital Avionics Systems Conference (DASC)*. 2022.

[175] J. Lei and U. V. Shanbhag. 'Stochastic Nash Equilibrium Problems: Models, Analysis, and Algorithms'. In: *IEEE Control Systems Magazine* 42.4 (2022), pp. 103–124.

[176] S. Leonardos et al. *Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games*. arXiv:2106.01969 [cs]. Sept. 2021. DOI: 10.48550/arXiv.2106.01969. URL: http://arxiv.org/abs/2106.01969 (visited on 09/01/2025).

[177] S. Levine et al. 'End-to-end training of deep visuomotor policies'. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[178] T. Lew and M. Pavone. 'Sampling-based reachability analysis: A random set theory approach with adversarial sampling'. In: *Conference on robot learning*. PMLR. 2021, pp. 2055–2070.

[179] J. Li et al. 'Augmented lagrangian method for instantaneously constrained reinforcement learning problems'. In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 2982–2989.

[180] J. Li et al. 'Cost Inference for Feedback Dynamic Games from Noisy Partial State Observations and Incomplete Trajectories'. In: *AAMAS*. 2023.

[181] J. Li et al. 'The Computation of Approximate Feedback Stackelberg Equilibria in Multiplayer Nonlinear Constrained Dynamic Games'. In: *SIAM Journal on Optimization* (2024).

[182] J. Li et al. 'Certifiable Reachability Learning Using a New Lipschitz Continuous Value Function'. In: *IEEE Robotics and Automation Letters* (2025).

[183] J. Li et al. 'Intent demonstration in general-sum dynamic games via iterative linear-quadratic approximations'. In: *arXiv preprint arXiv:2402.10182* (2024).

[184] J. Li et al. 'Scenario-Game ADMM: A Parallelized Scenario-Based Solver for Stochastic Noncooperative Games'. In: *IEEE CDC* (2023).

[185] M. Li, J. Qin and L. Ding. 'Two-player stackelberg game for linear system via value iteration algorithm'. In: *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*. IEEE. 2019, pp. 2289–2293.

[186] S. Li et al. 'Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33 Issue 1. 2019, pp. 4213–4220.

[187] T. Li and S. P. Sethi. 'A review of dynamic Stackelberg game models'. In: *Discrete & Continuous Dynamical Systems-B* 22.1 (2017), p. 125.

[188] W. Li and E. Todorov. 'Iterative Linear Quadratic Regulator Design for Nonlinear Biological Movement Systems'. In: *ICINCO* (2004), pp. 222–229.

[189] Y. Li et al. 'Differential Game Theory for Versatile Physical Human-robot Interaction'. In: *Nature Machine Intelligence* 1.1 (2019), pp. 36–43.

[190] Z. Li et al. 'Learning predictive safety filter via decomposition of robust invariant set'. In: *arXiv preprint arXiv:2311.06769* (2023).

[191] S. Liang, P. Yi and Y. Hong. 'Distributed Nash Equilibrium Seeking for Aggregative Games with Coupled Constraints'. In: *Automatica* 85 (2017), pp. 179–185.

[192] T. P. Lillicrap et al. 'Continuous control with deep reinforcement learning'. In: *arXiv preprint arXiv:1509.02971* (2015).

[193] A. Lin and S. Bansal. 'Generating formal safety assurances for high-dimensional reachability'. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 10525–10531.

[194] A. Lin and S. Bansal. 'Verification of neural reachable tubes via scenario optimization and conformal prediction'. In: *6th Annual Learning for Dynamics & Control Conference*. PMLR. 2024, pp. 719–731.

[195] C. Liu and M. Tomizuka. 'Designing the robot behavior for safe human–robot interactions'. In: *Trends in Control and Decision-Making for Human–Robot Collaboration Systems* (2017), pp. 241–270.

[196] M. Liu et al. 'Task and path planning for multi-agent pickup and delivery'. In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2019.

[197] Y. Liu, J. Ding and X. Liu. 'IPO: Interior-point policy optimization under constraints'. In: *arXiv preprint arXiv:1910.09615* (2019).

[198] B. T. Lopez, J.-J. E. Slotine and J. P. How. 'Dynamic Tube MPC for Nonlinear Systems'. In: *2019 American Control Conference (ACC)*. 2019, pp. 1655–1662. DOI: 10.23919/ACC.2019.8814758.

[199] D. P. Losey and M. K. O'Malley. 'Learning the correct robot trajectory in real-time from physical human interactions'. In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.1 (2019), pp. 1–19.

[200] D. P. Losey et al. 'A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction'. In: *Applied Mechanics Reviews* 70.1 (2018), p. 010804.

[201] R. Lowe et al. 'Multi-agent actor-critic for mixed cooperative-competitive environments'. In: *Advances in neural information processing systems* 30 (2017).

[202] R. Lucchetti, F. Mignanego and G. Pieri. 'Existence theorems of equilibrium points in Stackelberg'. In: *Optimization* 18.6 (1987), pp. 857–866.

[203] X. Lyu and M. Chen. 'Ttr-based reward for reinforcement learning with implicit model priors'. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020.

[204] Z. Ma, D. S. Callaway and I. A. Hiskens. 'Decentralized Charging Control of Large Populations of Plug-in Electric Vehicles'. In: *IEEE Transactions on Control Systems Technology* 21.1 (Jan. 2013), pp. 67–78. ISSN: 1558-0865. DOI: 10.1109/TCST.2011.2174059. (Visited on 22/02/2025).

[205] M. Maljkovic, G. Nilsson and N. Geroliminis. 'On Finding the Leader's Strategy in Quadratic Aggregative Stackelberg Pricing Games'. In: *2023 European Control Conference (ECC)*. 2023, pp. 1–6. DOI: 10.23919/ECC57647.2023.10178392.

[206] K. Margellos and J. Lygeros. 'Hamilton–Jacobi Formulation for Reach–Avoid Differential Games'. In: *IEEE Transactions on Automatic Control* 56.8 (2011), pp. 1849–1861. DOI: 10.1109/TAC.2011.2105730.

[207] G. Martín-Herrán and S. J. Rubio. 'On coincidence of feedback and global Stackelberg equilibria in a class of differential games'. In: *European Journal of Operational Research* 293.2 (2021), pp. 761–772.

[208] C. I. Mavrogiannis, W. B. Thomason and R. A. Knepper. 'Social momentum: A framework for legible navigation in dynamic multi-agent environments'. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, pp. 361–369.

[209] D. Mayne. 'A second-order gradient method for determining optimal trajectories of non-linear discrete-time systems'. In: *International Journal of Control* 3.1 (1966), pp. 85–95.

[210] R. Mazouz et al. 'Safety guarantees for neural network dynamic systems via stochastic barrier functions'. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 9672–9686.

[211] E. Mazumdar et al. 'Policy-Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games'. en. In: *New Zealand* (2020).

[212] N. Mehr et al. 'Maximum-entropy multi-agent dynamic games: Forward and inverse solutions'. In: *IEEE Transactions on Robotics* (2023).

[213] D. H. Mguni et al. 'Learning in Nonzero-Sum Stochastic Games with Potentials'. en. In: *Proceedings of the 38th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2021, pp. 7688–7699. URL: https://proceedings.mlr.press/v139/mguni21a.html (visited on 20/12/2024).

[214] I. Mitchell, A. Bayen and C. Tomlin. 'A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games'. In: *IEEE Transactions on Automatic Control* (2005).

[215] I. M. Mitchell. 'Scalable calculation of reach sets and tubes for nonlinear systems with terminal integrators: a mixed implicit explicit formulation'. In: *Proceedings of the International Conference on Hybrid Systems: Computation and Control*. 2011.

[216] V. Mnih et al. 'Human-level control through deep reinforcement learning'. In: *nature* 518.7540 (2015), pp. 529–533.

[217] A.-H. Mohsenian-Rad and A. Leon-Garcia. 'Optimal Residential Load Control With Price Prediction in Real-Time Electricity Pricing Environments'. In: *IEEE Transactions on Smart Grid* 1.2 (Sept. 2010), pp. 120–133. ISSN: 1949-3061. DOI: 10.1109/TSG.2010.2055903. URL: https://ieeexplore.ieee.org/abstract/document/5540263 (visited on 05/12/2024).

[218] T. L. Molloy et al. 'An inverse differential game approach to modelling bird mid-air collision avoidance behaviours'. In: *IFAC-PapersOnLine* 51.15 (2018), pp. 754–759.

[219] T. L. Molloy et al. 'Inverse noncooperative differential games'. In: *Inverse Optimal Control and Inverse Noncooperative Dynamic Game Theory*. Springer, 2022, pp. 189–226.

[220] T. L. Molloy et al. 'Inverse open-loop noncooperative differential games and inverse optimal control'. In: *IEEE Transactions on Automatic Control* 65.2 (2019), pp. 897–904.

[221] T. L. Molloy et al. *Inverse Optimal Control and Inverse Noncooperative Dynamic Game Theory*. 2022.

[222] T. L. Molloy, J. J. Ford and T. Perez. 'Inverse Noncooperative Dynamic Games'. In: *IFAC-PapersOnLine* 50.1 (2017). 20th IFAC World Congress, pp. 11788–11793. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2017.08.1989. URL: https://www.sciencedirect.com/science/article/pii/S2405896317326277.

[223] S. Mondal and P. V. Reddy. 'Linear quadratic Stackelberg difference games with constraints'. In: *2019 18th European Control Conference (ECC)*. IEEE. 2019, pp. 3408–3413.

[224] D. Monderer and L. S. Shapley. 'Potential Games'. In: *Games and Economic Behavior* 14.1 (May 1996), pp. 124–143. ISSN: 0899-8256. DOI: 10.1006/game.1996.0044. URL: https://www.sciencedirect.com/science/article/pii/S0899825696900445 (visited on 22/11/2024).

[225] S. Musić and S. Hirche. 'Haptic shared control for human-robot collaboration: a game-theoretical approach'. In: *IFAC-PapersOnLine* 53.2 (2020), pp. 10216–10222.

[226] T. Mylvaganam and A. Astolfi. 'Approximate solutions to a class of nonlinear stackelberg differential games'. In: *53rd IEEE Conference on Decision and Control*. IEEE. 2014, pp. 420–425.

[227] T. Mylvaganam, M. Sassano and A. Astolfi. 'A differential game approach to multi-agent collision avoidance'. In: *IEEE Transactions on Automatic Control* (2017).

[228] S. Narvekar et al. 'Curriculum learning for reinforcement learning domains: A framework and survey'. In: *Journal of Machine Learning Research* (2020).

[229] J. Nash. 'Non-cooperative games'. In: *Annals of mathematics* (1951), pp. 286–295.

[230] B. Navarro et al. 'An ISO10218-compliant adaptive damping controller for safe physical human-robot interaction'. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 3043–3048.

[231] S. Nayak et al. 'Scalable Multi-Agent Reinforcement Learning through Intelligent Information Aggregation'. In: *International Conference on Machine Learning*. Ed. by A. Krause et al. PMLR. 2023.

[232] A. Nemirovski. 'Prox-method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-concave Saddle Point Problems'. In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251.

[233] Y. Nesterov. 'A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2)'. In: *Doklady an ussr*. Vol. 269. 1983, pp. 543–547.

[234] D. P. Nguyen et al. 'Gameplay Filters: Robust Zero-Shot Safety through Adversarial Imagination'. In: *8th Annual Conference on Robot Learning*. 2024. URL: https://openreview.net/forum?id=Ke5xrnBFAR.

[235] H. H. Nguyen et al. 'Stability certificates for neural network learning-based controllers using robust control theory'. In: *2021 American Control Conference (ACC)*. IEEE. 2021, pp. 3564–3569.

[236] Q. Nguyen and K. Sreenath. 'Exponential control barrier functions for enforcing high relative-degree safety-critical constraints'. In: *2016 American Control Conference (ACC)*. IEEE. 2016, pp. 322–328.

[237] R. Nishihara et al. 'A General Analysis of the Convergence of ADMM'. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 343–352.

[238] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[239] A. S. Nowak. 'On a New Class of Nonzero-sum Discounted Stochastic Games Having Stationary Nash Equilibrium Points'. In: *International Journal of Game Theory* 32.1 (2003), p. 121.

[240] OpenAI. *ChatGPT-4*. [Software]. 2023. URL: https://openai.com/chatgpt.

[241] Y. Ouyang, H. Tavafoghi and D. Teneketzis. 'Dynamic games with asymmetric information: Common information based perfect bayesian equilibria and sequential decomposition'. In: *IEEE Transactions on Automatic Control* 62.1 (2016), pp. 222–237.

[242] D. Paccagnan and M. C. Campi. 'The Scenario Approach Meets Uncertain Game Theory and Variational Inequalities'. In: *cdc*. IEEE. 2019, pp. 6124–6129.

[243] A. Papavasiliou. 'Analysis of Distribution Locational Marginal Prices'. In: *IEEE Transactions on Smart Grid* 9.5 (Sept. 2018), pp. 4872–4882. ISSN: 1949-3053, 1949-3061. DOI: 10.1109/TSG.2017.2673860. URL: https://ieeexplore.ieee.org/document/7862921/ (visited on 07/05/2024).

[244] T. Parthasarathy and S. Sinha. 'Existence of Stationary Equilibrium Strategies in Non-zero-sum Discounted Stochastic Games with Uncountable State Space and State-independent Transitions'. In: *International Journal of Game Theory* 18 (1989), pp. 189–194.

[245] S. Paternain et al. 'Constrained reinforcement learning has zero duality gap'. In: *Advances in Neural Information Processing Systems*. 2019, pp. 7555–7565.

[246] L. Pavel. 'Distributed GNE Seeking Under Partial-Decision Information Over Networks via a Doubly-augmented Operator Splitting Approach'. In: *IEEE Transactions on Automatic Control* 65.4 (2020), pp. 1584–1597.

[247] S. Peng and J. Jiang. 'Stochastic Mathematical Programs with Probabilistic Complementarity Constraints: SAA and Distributionally Robust Approaches'. In: *Computational Optimization and Applications* 80.1 (2021), pp. 153–184.

[248] L. Peters and Z. N. Sunberg. 'iLQGames. jl: Rapidly Designing and Solving Differential Games in Julia'. In: *arXiv preprint arXiv:2002.10185* (2020).

[249] L. Peters et al. 'Contingency Games for Multi-Agent Interaction'. In: *arXiv preprint arXiv:2304.05483* (2023).

[250] L. Peters et al. 'Inference-based strategy alignment for general-sum differential games'. In: *AAMAS*. 2020.

[251] L. Peters et al. 'Inferring objectives in continuous dynamic games from noise-corrupted partial state observations'. In: *RSS*. 2021.

[252] T.-H. Pham, G. De Magistris and R. Tachibana. 'Optlayer-practical constrained optimization for deep reinforcement learning in the real world'. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 6236–6243.

[253] L. Pichierri, A. Testa and G. Notarstefano. 'Crazychoir: Flying swarms of crazyflie quadrotors in ros 2'. In: *IEEE Robotics and Automation Letters* 8.8 (2023), pp. 4713–4720.

[254] H. A. Poonawala et al. 'Collision-free formation control with decentralized connectivity preservation for nonholonomic-wheeled mobile robots'. In: *IEEE Transactions on Control of Network Systems* (2014).

[255] W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*. Vol. 703. John Wiley & Sons, 2007.

[256] S. Prajna and A. Jadbabaie. 'Safety verification of hybrid systems using barrier certificates'. In: *International Workshop on Hybrid Systems: Computation and Control*. Springer. 2004, pp. 477–492.

[257] Z. Qin et al. 'Learning Safe Multi-agent Control with Decentralized Neural Barrier Certificates'. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021.

[258] P. Ramachandran, B. Zoph and Q. V. Le. 'Searching for activation functions'. In: *arXiv preprint arXiv:1710.05941* (2017).

[259] L. J. Ratliff, S. A. Burden and S. S. Sastry. 'On the characterization of local Nash equilibria in continuous games'. In: *IEEE transactions on automatic control* 61.8 (2016), pp. 2301–2307.

[260] A. Ray, J. Achiam and D. Amodei. 'Benchmarking safe exploration in deep reinforcement learning'. In: *arXiv preprint arXiv:1910.01708* (2019).

[261] S. Razdan, J. Downing and L. White. *Pathways to Commercial Liftoff: Virtual Power Plants 2025 Update*. en. Tech. rep. U.S. Department of Energy, Jan. 2025.

[262] P. V. Reddy and G. Zaccour. 'Feedback Nash equilibria in linear-quadratic difference games with constraints'. In: *IEEE Transactions on Automatic Control* 62.2 (2016), pp. 590–604.

[263] A. Reuther et al. 'Interactive supercomputing on 40,000 cores for machine learning and data analysis'. In: *IEEE High Performance extreme Computing Conference (HPEC)*. 2018.

[264] J. Richalet. 'Algorithmic control of industrial processes'. In: *Proc. of the 4ˆ th IFAC Sympo. on Identification and System Parameter Estimation* (1976), pp. 1119–1167.

[265] S. Rothfuß et al. 'Inverse Optimal Control for Identification in Non-Cooperative Differential Games'. In: *IFAC-PapersOnLine* 50.1 (2017), pp. 14909–14915.

[266] W. Rudin et al. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1976.

[267] D. Sadigh et al. 'Information gathering actions over human internal state'. In: (2016), pp. 66–73.

[268] S. Sagheb, S. Gandhi and D. P. Losey. 'Should Collaborative Robots be Transparent?' In: *arXiv preprint arXiv:2304.11753* (2023).

[269] F. Salehisadaghiani and L. Pavel. 'Distributed Nash Equilibrium Seeking via the Alternating Direction Method of Multipliers'. In: *IFAC-PapersOnLine* 50.1 (2017), pp. 6166–6171.

[270] P. A. Scholten et al. 'Variable stability in-flight simulation system based on existing autopilot hardware'. In: *Journal of Guidance, Control, and Dynamics* 43.12 (2020), pp. 2275–2288.

[271] J. Schulman et al. 'Proximal policy optimization algorithms'. In: (2017).

[272] J. Schulman et al. 'Trust region policy optimization'. In: *International Conference on Machine Learning*. 2015, pp. 1889–1897.

[273] R. Schwan, C. N. Jones and D. Kuhn. 'Stability verification of neural network controllers using mixed-integer programming'. In: *IEEE Transactions on Automatic Control* (2023).

[274] W. Schwarting et al. 'Social behavior for autonomous vehicles'. In: *Proceedings of the National Academy of Sciences* 116.50 (2019), pp. 24972–24978.

[275] W. Schwarting et al. 'Stochastic dynamic games in belief space'. In: *IEEE Transactions on Robotics* 37.6 (2021), pp. 2157–2172.

[276] G. Scutari et al. 'Convex Optimization, Game Theory, and Variational Inequality Theory'. In: *IEEE Signal Processing Magazine* 27.3 (2010), pp. 35–49.

[277] S. Shalev-Shwartz, S. Shammah and A. Shashua. 'Safe, multi-agent, reinforcement learning for autonomous driving'. In: *arXiv preprint arXiv:1610.03295* (2016).

[278] L. S. Shapley. 'Stochastic Games'. In: *Proceedings of the national academy of sciences* 39.10 (1953), pp. 1095–1100.

[279] K. H. Shish et al. 'Survey of Capabilities and Gaps in External Perception Sensors for Autonomous Urban Air Mobility Applications'. In: *AIAA Scitech Forum*. 2021.

[280] D. Silver, R. S. Sutton and M. Müller. 'Reinforcement learning of local shape in the game of Go.' In: *IJCAI*. Vol. 7. 2007, pp. 1053–1058.

[281] M. Simaan and J. B. Cruz Jr. 'Additional aspects of the Stackelberg strategy in nonzero-sum games'. In: *Journal of Optimization Theory and Applications* 11.6 (1973), pp. 613–626.

[282] A. Singh, Z. Feng and S. Bansal. 'Imposing Exact Safety Specifications in Neural Reachable Tubes'. In: *arXiv preprint arXiv:2404.00814* (2024).

[283] N. Slegers, J. Kyle and M. Costello. 'Nonlinear model predictive control technique for unmanned air vehicles'. In: *Journal of Guidance, Control, and Dynamics* (2006).

[284] Z. Sun et al. 'Feasibility and coordination of multiple mobile vehicles with mixed equality and inequality constraints'. In: *arXiv preprint arXiv:1809.05509* (2018).

[285] I. Sutskever et al. 'On the importance of initialization and momentum in deep learning'. In: *International conference on machine learning*. PMLR. 2013, pp. 1139–1147.

[286] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[287] R. S. Sutton et al. 'Policy Gradient Methods for Reinforcement Learning with Function Approximation'. In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press, 1999. URL: https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html (visited on 05/02/2025).

[288] A. Talebpour, H. S. Mahmassani and S. H. Hamdar. 'Modeling lane-changing behavior in a connected environment: A game theory approach'. In: *Transportation Research Procedia* 7 (2015), pp. 420–440.

[289] H. Tang et al. 'Tactical Separation Algorithms and Their Interaction with Conflict Avoidance Systems'. In: *AIAA Guidance, Navigation and Control Conference and Exhibit*. 2008.

[290] A. Tanwani and Q. Zhu. 'Feedback Nash Equilibrium for Randomly Switching Differential–Algebraic Games'. In: *IEEE Transactions on Automatic Control* 65.8 (2019), pp. 3286–3301.

[291] A. G. Taye et al. 'Reachability based online safety verification for high-density urban air mobility trajectory planning'. In: *AIAA Aviation Forum*. 2022.

[292] A. Taylor et al. 'Learning for safety-critical control with control barrier functions'. In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 708–717.

[293] B. Tolwinski. 'A Stackelberg solution of dynamic games'. In: *IEEE Transactions on Automatic Control* 28.1 (1983), pp. 85–93.

[294] C. Tomlin, J. Lygeros and S. Sastry. 'Computing controllers for nonlinear hybrid systems'. In: *Hybrid Systems: Computation and Control: Second International Workshop, HSCC'99 Berg en Dal, The Netherlands, March 29–31, 1999 Proceedings 2*. Springer. 1999, pp. 238–255.

[295] 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.

[296] C. Tomlin et al. 'Hybrid Control in Air Traffic Management Systems'. In: *IFAC Proceedings Volumes* (1996).

[297] K. G. Vamvoudakis et al. 'Online learning algorithm for Stackelberg games in problems with hierarchy'. In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE. 2012, pp. 1883–1889.

[298] H. R. Varian. *Intermediate microeconomics with calculus: a modern approach*. WW norton & company, 2014.

[299] D. Vasal, A. Sinha and A. Anastasopoulos. 'A systematic process for evaluating structured perfect Bayesian equilibria in dynamic games with asymmetric information'. In: *IEEE Transactions on Automatic Control* 64.1 (2018), pp. 81–96.

[300] S. Vogel. 'Stability Results for Stochastic Programming Problems'. In: *Optimization* 19.2 (1988), pp. 269–288.

[301] H. Von Stackelberg. 'The theory of the market economy: Oxford University Press, 1952'. In: (1952).

[302] K. P. Wabersich et al. 'Data-Driven Safety Filters: Hamilton-Jacobi Reachability, Control Barrier Functions, and Predictive Methods for Uncertain Systems'. In: *IEEE Control Systems Magazine* (2023).

[303] M. J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Vol. 48. Cambridge university press, 2019.

[304] Z. Wan et al. 'Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning'. In: *IEEE Transactions on Smart Grid* 10.5 (Sept. 2019), pp. 5246–5257. ISSN: 1949-3061. DOI: 10.1109/TSG.2018.2879572. (Visited on 25/03/2025).

[305] K. Wang et al. 'Coordinating followers to reach better equilibria: End-to-end gradient descent for stackelberg games'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 5. 2022, pp. 5219–5227.

[306] L. Wang, A. D. Ames and M. Egerstedt. 'Safety Barrier Certificates for Collisions-Free Multirobot Systems'. In: *IEEE Transactions on Robotics* (2017).

[307] X. Wang, K. Leung and M. Pavone. 'Infusing Reachability-Based Safety into Planning and Control for Multi-agent Interactions'. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020.

[308] E. R. Weintraub. 'On the existence of a competitive equilibrium: 1930-1954'. In: *Journal of Economic Literature* 21.1 (1983), pp. 1–39.

[309] Wisk Aero. *Wisk Aero - Our Aircraft*. Accessed: 2025-01-31. 2025. URL: https://wisk.aero/aircraft/.

[310] C. Wu et al. 'Flow: Architecture and benchmarking for reinforcement learning in traffic control'. In: *arXiv preprint arXiv:1710.05465* (2017), p. 10.

[311] W. Xiao et al. 'Barriernet: Differentiable control barrier functions for learning of safe robot control'. In: *IEEE Transactions on Robotics* (2023).

[312] D. Xie. 'On time inconsistency: a technical issue in Stackelberg differential games'. In: *Journal of Economic Theory* 76.2 (1997), pp. 412–430.

[313] Y. Xing et al. 'Driver lane change intention inference for intelligent vehicles: Framework, survey, and challenges'. In: *IEEE Transactions on Vehicular Technology* 68.5 (2019), pp. 4377–4390.

[314] H. Xiong et al. 'Deterministic policy gradient: Convergence analysis'. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 2159–2169.

[315] H. Xu. 'Sample Average Approximation Methods for a class of Stochastic Variational Inequality Problems'. In: *Asia-Pacific Journal of Operational Research* 27.01 (2010), pp. 103–119.

[316] H. Xu and D. Zhang. 'Stochastic Nash Equilibrium Problems: Sample Average Approximation and Applications'. In: *Computational Optimization and Applications* 55 (2013), pp. 597–645.

[317] K. Xu, X. Zhao and X. Han. 'Adaptive Dynamic Programming for a Class of Two-player Stackelberg Differential Games'. In: *2020 International Conference on System Science and Engineering (ICSSE)*. IEEE. 2020, pp. 1–6.

[318] M. Xue, A. K. Ishihara and P. U. Lee. 'Negotiation Model for Cooperative Operations in Upper Class E Airspace'. In: *IEEE/AIAA Digital Avionics Systems Conference (DASC)*. 2022.

[319] Z. Yan, N. Jouandeau and A. A. Cherif. 'A Survey and Analysis of Multi-Robot Coordination'. In: *International Journal of Advanced Robotic Systems* 10.12 (2013), p. 399.

[320] I. Yang et al. 'One-shot computation of reachable sets for differential games'. In: *Proceedings of the International Conference on Hybrid Systems: Computation and Control*. 2013.

[321] Q. Yang et al. 'WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35 Issue 12. 2021, pp. 10639–10646.

[322] T. Yang, M. Jordan and T. Chavdarova. 'Solving Constrained Variational Inequalities via a First-order Interior Point-based Method'. In: *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*.

[323] Y. Yang et al. 'Scalable synthesis of formally verified neural value function for hamilton-jacobi reachability analysis'. In: *arXiv preprint arXiv:2407.20532* (2024).

[324] J. H. Yoo and R. Langari. 'A stackelberg game theoretic driver model for merging'. In: *Dynamic Systems and Control Conference*. Vol. 56130. American Society of Mechanical Engineers. 2013, V002T30A003.

[325] C. Yu et al. 'The surprising effectiveness of ppo in cooperative multi-agent games'. In: *Advances in Neural Information Processing Systems* (2022).

[326] C. Yu et al. 'Inverse linear quadratic dynamic games using partial state observations'. In: *Automatica* 145 (2022), p. 110534.

[327] Y. Yu et al. 'Active Inverse Learning in Stackelberg Trajectory Games'. In: *arXiv preprint arXiv:2308.08017* (2023).

[328] K. Zhang et al. 'Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies'. In: *SIAM Journal on Control and Optimization* 58.6 (Jan. 2020). Publisher: Society for Industrial and Applied Mathematics, pp. 3586–3612. ISSN: 0363-0129. DOI: 10.1137/19M1288012. URL: https://epubs-siam-org.libproxy.berkeley.edu/doi/10.1137/19M1288012 (visited on 31/03/2025).

[329] R. Zhang, Z. Ren and N. Li. 'Gradient Play in Stochastic Games: Stationary Points, Convergence, and Sample Complexity'. In: *IEEE Transactions on Automatic Control* 69.10 (Oct. 2024), pp. 6499–6514. ISSN: 1558-2523. DOI: 10.1109/TAC.2024.3387208. (Visited on 08/01/2025).

[330] S. Zhang, K. Garg and C. Fan. 'Neural graph control barrier functions guided distributed collision-avoidance multi-agent control'. In: *Conference on Robot Learning*. PMLR. 2023.

[331] S. Zhang et al. 'Gcbf+: A neural graph control barrier function framework for distributed safe multi-agent control'. In: *IEEE Transactions on Robotics* (2025).

[332] S. Zhang et al. 'Discrete GCBF Proximal Policy Optimization for Multi-agent Safe Optimal Control'. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2025.

[333] Y. Zhao and Q. Zhu. 'Stackelberg Game-Theoretic Trajectory Guidance for Multi-Robot Systems with Koopman Operator'. In: *arXiv preprint arXiv:2309.16098* (2023).

[334] D. Zhou et al. 'Fast, on-line collision avoidance for dynamic vehicles using buffered voronoi cells'. In: *IEEE Robotics and Automation Letters* (2017).

[335] H. Zhu and H. J. Liu. 'Fast Local Voltage Control Under Limited Reactive Power: Optimality and Stability Analysis'. In: *IEEE Transactions on Power Systems* 31.5 (Sept. 2016). Conference Name: IEEE Transactions on Power Systems, pp. 3794–3803. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2015.2504419. URL: https://ieeexplore.ieee.org/abstract/document/7361761 (visited on 29/03/2025).

[336] K. Zhu et al. 'Safe Multi-Agent Reinforcement Learning via Approximate Hamilton-Jacobi Reachability'. In: *Journal of Intelligent & Robotic Systems* (2024).

[337] Y. Zhu et al. 'robosuite: A Modular Simulation Framework and Benchmark for Robot Learning'. In: *arXiv preprint arXiv:2009.12293*. 2020.