Exploiting Equilibrium Structure for Scalable Solutions to Zero-Sum Differential Games with One-Sided Information

by

Mukesh Ghimire

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2026 by the
Graduate Supervisory Committee:

Yi Ren, Co-Chair
Zhe Xu, Co-Chair
Jennie Si
Spring Berman
Wenlong Zhang

ARIZONA STATE UNIVERSITY

May 2026

ABSTRACT

Many real-world adversarial planning problems in security and robotics are best modeled as two-player zero-sum (2p0s) differential games, but they remain difficult to solve in practice because they involve continuous-time dynamics with continuous state and action spaces, strategic interaction, and often information asymmetry and state constraints. In contrast to the current state of the art in imperfect-information game solving — largely driven by discrete domains (e.g., poker-style solvers) — these settings demand methods that scale in continuous spaces while preserving the structure induced by asymmetric information.

This dissertation develops a unified theory-and-algorithms for two-player zero-sum one-sided incomplete-information differential games (2p0s1) with and without state constraints, where the informed player "P1" has full information of the game while the uninformed player "P2" only holds partial information along with a public belief, i.e., a probability distribution over the set of possible payoffs/states. Key contributions of this work include the characterization of the value function in 2p0s1 with state constraints, development of scalable plug-and-play algorithms by leveraging the unique equilibrium structure inherent to 2p0s1 and enabling near real-time computation of strategies in 2p0s1 with linear dynamics and quadratic costs via a differentiable tree-Riccati methodology. Together, these contributions bridge rigorous foundations and scalable computation of Nash equilibrium strategies in 2p0s1.

DEDICATION

"Fill your life with great verbs. That will surround you with the right nouns. Do that and you will live a life that merits the best adjectives."

To my family.

ACKNOWLEDGMENTS

I express my deepest gratitude to my co-advisor, Dr. Yi Ren, for his continuous support throughout my Ph.D. journey. His dedication to superior research output has always kept me motivated to pursue this mission. I am equally thankful to my co-advisor Dr. Xu for his guidance and mentorship during this journey. I am also thankful to Dr. Wenlong Zhang for providing opportunities to collaborate with his research team, and to the rest of my committee members — Dr. Spring Berman, and Dr. Jennie Si for their invaluable feedback.

I would also like to thank my collaborators — Dr. Lei Zhang, Dr. Sunny Amatya, who were also my mentors during the early stages of my Ph.D. I am also thankful to Dr. Jennifer Zhu, and Dr. Aosong Feng, who were invaluable mentors to me during my internship at AWS.

Additionally, I am also grateful to wonderful friends that I've made throughout my time at ASU. I will forever cherish the moments spent with Sajal Regmi, Avinaya Tripathi, Sangeet Adhikari, Deepesh Karmacharya, Ujjwal Poudel, Siddhartha Pudasaini, Sadhikshya Rijal, Rajesh Shrestha, Rosha Gumanju, and Som Sagar.

A sincere thank you goes to my family — my parents and my brother Nootan Ghimire — who have always rooted for me. This Ph.D. wouldn't have been possible without their unwavering support and unconditional love. Equally thankful to my dear friend, Amrit Poudel, who has always made himself available in times of need, whether for a research discussion or for much needed reassurances.

Finally, a special thank you goes to my partner, Anushka Subedi for being my cheerleader, and a beacon of light whenever I doubted myself; for reminding me to celebrate the small moments. I will be forever grateful for your support and cherish the memories that we have made.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION 1

1.1 Introduction 1

1.1.1 Motivation and Contributions 2

2 VALUE AND POLICY CHARACTERIZATION IN THE PRESENCE OF STATE CONSTRAINTS 5

2.1 Introduction 5

2.2 Related Work 9

2.2.1 Incomplete-information repeated and differential games 9

2.2.2 Imperfect-information dynamic games 11

2.3 State Constrained Zero-Sum Differential Games with One-Sided Information 12

2.3.1 Preliminaries 12

2.3.1.1 Behavioral strategy 13

2.3.1.2 Backward reachable set 15

2.3.1.3 Payoff and value 15

2.3.1.4 The existence of the value 16

2.3.1.5 Characterization of the value 16

2.4 Bellman Backup and Behavioral Strategies 18

2.4.1 Behavioral strategy for Player 1 and the belief dynamics 19

2.4.2 The dual game and behavioral strategy for Player 2 20

CHAPTER Page

2.5 Numerical Methods 22
2.5.1 Primal and dual value approximation 22
2.5.1.1 Value discontinuity 22
2.5.1.2 Partially convex values 23
2.5.1.3 Convexification error. 24
2.5.1.4 Approximation of the conjugate value. 24
2.5.2 Synthesis of strategies 25
2.6 Case Study 25
2.7 Proofs and Additional Results 29
2.7.1 Proof of Theorem 2.1 29
2.7.2 Proof of Theorem 2.2 41
2.7.3 Proof of Proposition 2.1 42
2.7.4 Bellman Backup of the Conjugate Value with the Presence of Instantaneous Loss 43
2.8 Conclusion and Future Work 45
3 EQUILIBRIUM STRUCTURE EXPLOITATION FOR EFFICIENT GAME SOLVING 46
3.1 Introduction 46
3.2 Related Work 49
3.2.1 2p0s games with incomplete information. 49
3.2.2 IIEFGs. 50
3.2.3 Multigrid for value approximation. 50
3.3 Problem Statement 51
3.4 A Primal-Dual Reformulation of the Game 52

CHAPTER Page

3.4.1 The primal game G_τ. 53

3.4.2 The dual game G_τ^*. 53

3.4.3 Equilibrium strategies of G_τ and G_τ^* are atomic. 54

3.5 Incorporating the Atomic Structure into MARL and Control 58

3.5.1 Value approximation 58

3.5.2 CAMS for G_τ 58

3.5.3 Computational challenge. 59

3.5.4 Multigrid 59

3.5.5 Model-free multiagent reinforcement learning 61

3.5.6 Model predictive control 63

3.6 Empirical Validation 63

3.6.1 Effect of atomic NE on value approximation 65

3.6.2 Effect of atomic NE on model-free MARL 67

3.6.3 Effect of atomic NE on MPC 68

3.6.4 Hexner's game 68

3.6.5 Football game 69

3.6.5.1 Making Sense of Exploitability 70

3.7 Proofs and Additional Results 71

3.7.1 Proof of Theorem 3.1 71

3.7.2 Proof of Theorem 3.2 74

3.7.3 Settings 74

3.7.4 Preliminary lemmas 75

3.7.5 Sample Trajectories 82

3.8 Conclusion 83

CHAPTER Page

4 ENABLING ONLINE SOLVING 84

4.1 INTRODUCTION 84

4.2 PROBLEM FORMULATION 85

4.2.1 Game Setup and Information 85

4.2.2 Strategy and Payoff Structures 86

4.2.3 Reparameterization of $\{\eta_i\}$ 88

4.3 DIFFERENTIABLE RICCATI SOLVER 90

4.3.1 Quadratic Value Representation on Nodes 90

4.3.2 The Solver 94

4.4 Case Studies 96

4.4.1 Case 1: Hexner's Game 96

4.4.2 Case 2: Drone Landing Under Attack 98

4.4.2.1 Optimal policy under nominal conditions 99

4.4.2.2 Optimal policy under stochastic disturbance 99

4.5 Conclusion 100

5 CONCLUSION AND FUTURE WORK 101

5.1 Conclusion 101

5.2 Future Directions and Broader Outlook 103

REFERENCES 106

APPENDIX

A EXAMPLES OF ZERO-SUM GAMES WITH ONE-SIDED INFORMATION 113

B PREDICTION ERROR OF VALUE APPROXIMATION 125

C IMPLEMENTATION DETAILS 129

CHAPTER Page

D PUBLICATIONS ASSOCIATED WITH THE DISSERTATION 142

LIST OF TABLES

Table Page

3.1. Runtime comparison of CAMS with and without multigrid 66

4.1. Δcost of P1 and mean subgame solving time 99

A.1. Payoff table for a zero-sum beer-quiche game 114

C.1. Hyperparameters for DeepCFR Training 135

LIST OF FIGURES

Figure Page

2.1. Value along belief (p) and time (t) in Hexner's game 8

2.2. Trajectories of informed and uninformed players in 8D Hexner's game 10

2.3. Schematics of a simplified football game 26

2.4. Information Delay and Splitting Advantage 27

2.5. Primal-Dual Trajectories 29

3.1. Game tree structure and sample trajectories in football 47

3.2. Value convexification causes NE to be atomic. 56

3.3. Illustration of 2-level multigrid method. 61

3.4. Computational complexity and quality of solutions 64

3.5. Exploitability comparison in the normal-form Hexner's game. 64

3.6. Trajectories comparison 67

3.7. Comparisons b/w CAMS-DRL and standard PG methods 68

3.8. Comparison between equilibrium policy and BR policies 71

3.9. Comparison between CAMS, DeepCFR, and JPSPG Trajectories 82

4.1. Bilevel structure used in the solver 86

4.2. Comparision between GT NEs and solution from 2p0s1 LQ solver 97

4.3. Sample trajectories for the drone landing under attack case 98

A.1. Zero-Sum Variant of the Beer-Quiche Game 115

A.2. Primal value for beer-quiche game 118

A.3. Conjugate value for beer-quiche game 120

A.4. d_1-d_2 plot 122

C.1. P2's trajectory comparison 131

C.2. Visualization of backward reachable set 132

Figure Page

C.3. Hyperparameter sweep for PG baselines 138

C.4. 2D Hexner's game solved by MPC. 139

Chapter 1

INTRODUCTION

1.1 Introduction

Differential games model a wide range of multi-agent interactions such as autonomous driving (Zhang *et al.*, 2023), human-robot interactions (Li *et al.*, 2019), and even safety-critical scenarios arising in defense (Garcia *et al.*, 2018, 2021; Liang *et al.*, 2019) and cybersecurity (Durkota *et al.*, 2017; Mc Carthy *et al.*, 2016; Horák and Bošanskỳ, 2016; Van Dijk *et al.*, 2013). These interactions occur in continuous time and more commonly in continuous state and action spaces. Most of the existing studies on differential games assume complete-information, i.e., all agents have complete understanding of the costs involved in the interaction. However, this is rarely the case in the real-world — agents (or players) often have hidden objectives unknown to others. Hence, it becomes necessary to model the information asymmetry in order to compute the optimal strategies.

While studies on computation of optimal strategies in incomplete-information differential games are limited, there is an abundance of existing work on imperfect-information dynamic games, representing interactions defined on discrete time, discrete state, and action spaces. Applying them to solve incomplete-information differential games would entail discretizing the entire domain, which leads to an explosion in computational complexity for safety-critical interactions that require high frequency control inputs. Nevertheless, these studies provide the necessary building blocks for studying incomplete-information differential games. As we progress, we will observe

that the theory of incomplete-information differential games closely mimics that of the *repeated games with incomplete-information* by Aumann *et al.* (1995), for which the authors won the 2005 Nobel Prize in Economic Sciences.

1.1.1 Motivation and Contributions

As hinted earlier, the primary motivation for this dissertation arises from the computational intractability of existing dynamic game solvers when applied to differential games. This dissertation aims to develop theoretically sound algorithms for efficiently approximating optimal strategy in incomplete-information differential games without the need for discretizing the action space (and subsequently the state space), which often requires domain-specific discretization. To do so, it focuses on a subset of 2p0s differential games with incomplete information, where one player (P1) has full information on the cost associated with the game, while the other player (P2) only knows about the probability distribution from which a true cost is sampled in the beginning of the game. We call this a 2p0s differential game with one-sided information (2p0s1), in which P1 is the informed player and P2 is the uninformed player.

The next three chapters discuss the major contributions of this dissertation that are outlined below.

1. **Existence of Value and Strategies in State-Constrained 2p0s1 Games:** Central to differential games are safety-critical applications such as autonomous driving, where the control actions have to be adhere to constraints to ensure safety. Motivated by this, this dissertation first attempts to extend the theory of 2p0s1 to handle state-constraints. To that end, it provides value and strategy

existence proofs when there exist state-constraints in the system. It leverages reachability analysis to prove that the value exists in such cases, and to develop backward induction for both value and strategy approximation. Using a toy-case it showcases how the strategy changes when the state-constraints are introduced in an interaction. This contribution is discussed in Chapter 2.

2. **Family of Efficient Algorithms for 2p0s1 Games:** This is the core of the dissertation, which takes a step back to address the computational challenge of applying existing state-of-the-art (SOTA) algorithms for dynamic games, also referred to as Imperfect Information Extensive Form Games (IIEFGs) to solving differential games. It first highlights the exponential growth of computational complexity of IIEFG solvers with respect to the size of the action space, making them intractable when the action space is infinite. This chapter further solidifies the special equilibrium structure prevalent in 2p0s1 games — both the informed (P1) and the uninformed (P2) players' equilibrium strategy, at any time-step, consists of only a finite number of actions to randomize from (i.e., equilibrium has finite support). As a result, the underlying complexity of the game reduces by a great extent. Finally, it provides a family of algorithmic implementations: value approximation, model-free multiagent reinforcement learning, and model predictive control, showcasing how this special equilibrium structure can be leveraged to efficiently solve 2p0s1, otherwise unreachable to SOTA IIEFG solvers. It demonstrates the efficacy by solving a 22-player american football game with continuous action spaces and 10 time steps, where the offense needs to strategically conceal their play until a critical moment in order to exploit its information advantage. This contribution is discussed in detail in Chapter 3
3. **Near Real-time 2p0s1 Game Solving for Linear Quadratic Systems:**

In Chapter 5.2, this dissertation focuses on yet another special class of 2p0s1 in which the game dynamics is linear and the costs are quadratic (LQ). It also leverages the equilibrium structure of the game to construct a tree-structured problem, and proposes a bi-level optimization problem to solve the equilibrium strategy of the informed player. The proposed algorithm enables not only fast offline solving of 2p0s1 LQ games, but also enables rapid online strategy-adaptation when there's noise in the state during deployment.

Chapter 2

VALUE AND POLICY CHARACTERIZATION IN THE PRESENCE OF STATE CONSTRAINTS

2.1 Introduction

We study fixed-time zero-sum differential games with state constraints and one-sided information, where Player 1 holds a private type (e.g., an intent or preference) that defines the payoffs of the game. The goal of Player 1 (resp. Player 2) is to minimize (resp. maximize) the cost. Since violation of the state constraint results in $+\infty$ penalty to Player 1, Player 2 resorts to violating the constraints when possible; and Player 1 resigns when state violation is inevitable. At the beginning of the game, Nature draws a type from a distribution known to both players and assigns the type only to Player 1. Initialized as Nature's distribution, the common belief about Player 1's type is updated dynamically during the game based on observations, and shared between the players. A stochastic state trajectory is produced based on the initial state and belief, the deterministic system dynamics, and the behavioral strategies of the two players. The value of the game, when exists, follows a Hamilton-Jacobi (HJ) equation and is a function of time, state, and belief. Importantly, Player 1 may

This chapter is based on the work described in the paper "State-Constrained Zero-Sum Differential Games with One-Sided Information" published in the **Proceedings of Machine Learning Research** for the 41st International Conference on Machine Learning.

control the release of information at the equilibrium to manipulate the common belief and take advantage of information asymmetry.

We use Hexner's game Hexner (1979) as a minimal example to demonstrate information control by Player 1: Consider two players with linear dynamics

$$\dot{x}_i = A_i x_i + B_i u_i,$$

for $i = 1, 2$, where $x_i(t) \in \mathbb{R}^{d_x}$ are system states, $u_i(t) \in \mathcal{U}$ are control inputs belonging to the admissible set $\mathcal{U}$, $A_i, B_i \in \mathbb{R}^{d_x \times d_x}$. Let $\theta \in \{-1, 1\}$ be Player 1's type unknown to Player 2[1]. Let p_θ be Nature's probability distribution of θ. Consider that the game is to be played infinite many times, the payoff is an expectation over θ:

$$J(u_1, u_2) = \mathbb{E}_\theta\Big[\int_0^T \Big(\|u_1\|_{R_1}^2 - \|u_2\|_{R_2}^2\Big)dt + \|x_1(T) - z\theta\|_{K_1(T)}^2 - \|x_2(T) - z\theta\|_{K_2(T)}^2\Big], \tag{2.1}$$

where $z \in \mathbb{R}^{d_x}$, R_1 and R_2 are positive-definite, continuous matrix functions, and $K_1(T)$ and $K_2(T)$ are positive semi-definite matrices. All parameters are common knowledge except θ. Essentially Player 1's goal is to get closer to $z\theta$ than Player 2. Since Player 2 can infer the target based on Player 1's control, Player 1 may play a non-revealing strategy for some time, i.e., as if he also only knows p_θ rather than the actual θ, before he eventually reveals.

The equilibrium of this game is as follows: First, it can be shown that players' control has a 1D representation, denoted by $\tilde{\theta}_i \in \mathbb{R}$, through:

$$u_i = -R_i^{-1} B_i^T K_i x_i + R_i^{-1} B_i^T K_i \Phi_i z \tilde{\theta}_i,$$

for $i = 1, 2$, where $\dot{\Phi}_i = A_i \Phi_i$ with boundary condition $\Phi_i(T) = I$, and

$$\dot{K}_i = -A_i^T K_i - K_i A_i + K_i^T B_i R_i^{-1} B_i^T K_i.$$

[1]Hexner's analysis is applicable to $\theta \in \mathbb{R}^{d_x}$, but is not generalizable to games with arbitrary dynamics and payoff functions. Here we adopt Cardaliaguet's setting where types are categorical Cardaliaguet (2007).

Then by introducing

$$d_i = z^T \Phi_i^T K_i B_i R_i^{-1} B_i^T K_i^T \Phi_i z, \tag{2.2}$$

and defining the critical time as

$$t_r = \arg\min_t \int_0^t (d_1(s) - d_2(s)) ds,$$

one can derive Player 1's strategy as $\tilde{\theta}_1(t) = 0$ for $t \in [0, t_r]$ and $\tilde{\theta}_1(t) = \theta$ for $t \in (t_r, T]$, i.e., Player 1 reveals its type at t_r. Player 2's strategy turns out to be to strictly follow Player 1: $\tilde{\theta}_2(t) = \tilde{\theta}_1(t)$. The original analysis by Hexner exploits the fact that both players solve linear-quadratic regulators parameterized by θ. We will revisit this game after introducing the differential game theory for one-sided information games (Cardaliaguet, 2007, 2009), which arrives at Hexner's solution but can also solve games with arbitrary dynamics and payoff functions, subject to continuity assumptions. This paper extends the unconstrained settings in Cardaliaguet (Cardaliaguet, 2009) and Souquiere (Souquière, 2010): We prove that value exists for differential games *with state constraints* and one-sided information, and derive the primal-dual HJ equations necessary for computing player strategies.

Different from existing works that focus on scalable no-regret learning on imperfect-information games with discrete dynamics (Brown *et al.*, 2020a; Perolat *et al.*, 2022a), this paper builds on top of repeated games and incomplete-information differential games (Cardaliaguet, 2007, 2009) to reveal the underlying mechanism of belief manipulation resulted from information asymmetry and state constraints. Specifically, we show that in any subgame, Player 1 plays a behavioral strategy (i.e., probability distributions over the action space for all his types) that convexifies his value with respect to the common belief. As a result, the common belief "splits" to vertices of the value convex hull with probabilities that are optimal for Player 1. See Fig. 2.1 for an illustration using Hexner's game. Importantly, the number of splits for Player 1

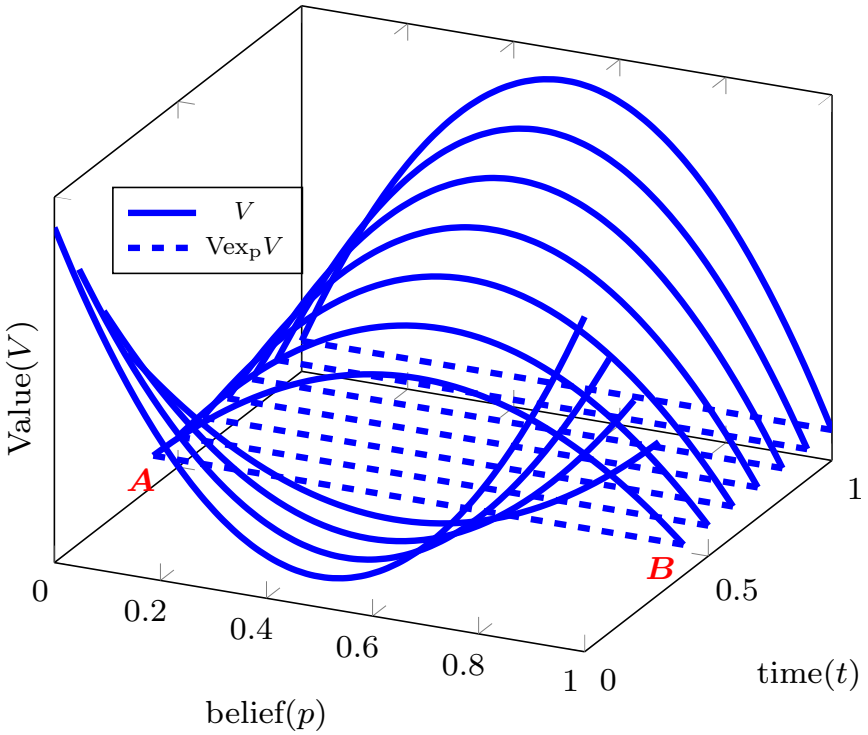


Figure 2.1. Value along belief (p) and time (t) in Hexner's game. Belief splits to A ($p = 0$) and B ($p = 1$) depending on the true type of Player 1, when the value becomes concave should Player 1 play a non-revealing strategy. In other words, Player 1 delays the release of his type until a critical time. In more general cases, belief splitting may not fully reveal Player 1's type, leading to belief manipulation.

is no more than the number of possible player types. On the other hand, Player 2 counters Player 1 by playing a dual game where her behavioral strategy is determined by the convexification of the conjugate value. See Sec. 2.3 and 2.4 for details.

Within this context, it becomes clear that understanding whether and how belief should be manipulated relies on knowing the value landscape over the belief space at any time and state. In addition to the curse of dimensionality (CoD) commonly present for games with non-trivial state/action/belief spaces and time horizons, we also experience computational challenges due to value discontinuity and the need for convexification and splitting. We discuss in Sec. 2.5 a set of solutions, including using

physics-informed neural network to characterize the backward reachable set to smooth value approximation, and using an input convex architecture Amos *et al.* (2017) for predicting convex values.

To summarize, we claim the following contributions:

- We extend the theory of zero-sum differential games with one-sided information to games with state constraints by proving value existence of such games and deriving the primal and dual subdynamic principles;
- We elucidate, with detailed examples, how the subdynamic principles lead to the derivation of behavioral strategies;
- We develop numerical tools to alleviate CoD in value approximation and to infer behavioral strategies from values. In Sec. 2.6, we solve an 8D man-to-man matchup game and reveal player positions in which the attacker can take advantage of information asymmetry by playing specific deceptive moves, and to derive the defender's best response in the lack of information. See Fig. 2.2.

2.2 Related Work

2.2.1 Incomplete-information repeated and differential games

Harsanyi (1967) first formalized information asymmetry in a stage game by introducing a private player types. Aumann *et al.* (1995) provided a framework to study repeated games with incomplete information on one side. De Meyer (1996) introduced dual games from where strategies of the uninformed player can be derived from a recursive structure of the conjugate value. Extending these results to differential

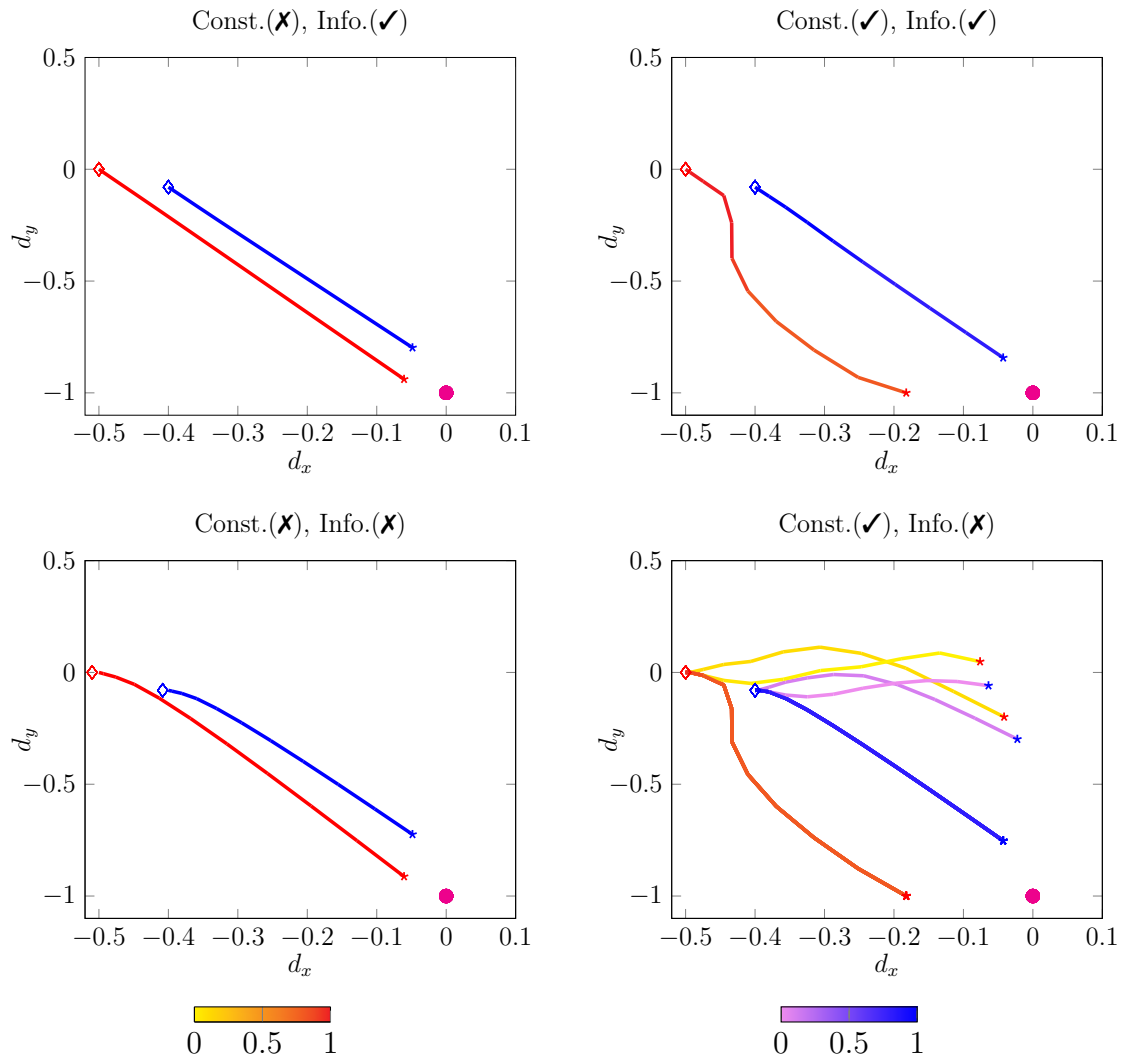


Figure 2.2. Trajectories of informed Player 1 (red) and uninformed Player 2 (blue) in an 8D Hexner's game w/ and w/o a state constraint or information asymmetry. Color shades indicate probabilities. When constrained, Player 1 stays away from Player 2 while trying to be closer to the target (the circle) than Player 2. Diamonds indicate initial states and stars indicate final states. See Sec. 2.6 for details.

games with Markov rewards, Cardaliaguet (2007) and Souquière (2010) confirmed the structures of incomplete-information games with one sided information on player type: (1) the game enjoys a primal-dual decomposition so that the informed player does not need to know the equilibrium of the uninformed player to compute his own; and (2) the value is convexified by belief splitting at the equilibrium. Recently, Hu *et al.* (2023)

proposed independently a belief-space HJ formulation for zero-sum differential games with one-sided information. While their framework can incorporate state constraints, it does not reveal the above structure of the equilibrium strategies of such games.

2.2.2 Imperfect-information dynamic games

Since player types can be considered as static private states, our work belongs to the category of imperfect-information dynamic games, where more general dynamics and information structures (e.g., disturbances, partial observability, and delayed information sharing) are considered. Nayyar *et al.* (2013) showed that the game can be reformulated as perfect-information by introducing a common belief state, provided that the belief is strategy independent. This strategy-independence assumption is relaxed in Kartik and Nayyar (2021) for zero-sum dynamic games by introducing past strategies as part of the players' information state. The general setting of Kartik and Nayyar (2021), however, does not facilitate a value existence proof. A significant amount of recent work build on top of common belief to approximate values of imperfect-information dynamic games (e.g., ReBeL (Brown *et al.*, 2020a), DeepNash (Perolat *et al.*, 2022a), and SoG (Schmid *et al.*, 2023a)). Following Nayyar *et al.* (2013), these algorithms model behavioral strategies as *prescriptions*, i.e., belief-conditioned action distributions. In addition, by taking advantage of the equivalence between local regret matching and Nash equilibrium in two-player zero-sum games (Zinkevich *et al.*, 2007), no-regret algorithms (Brown and Sandholm, 2018, 2019; Brown *et al.*, 2020a) have been developed for more scalably solving games with large action spaces and long time horizons than linear programming based methods (Koller and Pfeffer, 1995). It should be noted that these algorithms scale linearly to the

square-root of the action space, and thus induce high costs as the action space grows. While often disconnected, the studies on imperfect-information dynamic (or extensive-form) games and those on incomplete-information differential games are consistent in theory. Specifically, regret matching, i.e., solving subgame minimax problems with respect to behavioral strategies in the former, leads to strategies that satisfy the subdynamic programming principles stated in Cardaliaguet (2007), due to the fact that the behavioral strategies intrinsically convexify values. The key difference, however, is that regret matching algorithms do not enforce belief splitting. In practice, this means that the resultant strategy, often as a result of manually chosen action discretization, does not explicitly *explain* whether a certain random action is to be taken in a given belief state in order to delay information release or to manipulate the belief in a specific way.

2.3 State Constrained Zero-Sum Differential Games with One-Sided Information

2.3.1 Preliminaries

We consider a time-invariant deterministic dynamical system that defines the dynamics of the combined state x of Players 1 and 2, whose control inputs are u and v, respectively:

$$\begin{cases} \dot{x}(t) = f(x(t), u(t), v(t)), \quad u(t) \in \mathcal{U},\ v(t) \in \mathcal{V} \\ x(t_0) = x_0 \end{cases} \tag{2.3}$$

The game starts at $t_0 \in [0, T]$ with an initial state $x_0 \in \mathbb{R}^{d_x}$. Denote $g_i : \mathbb{R}^{d_x} \to \mathbb{R}$ the terminal payoff functions for $i \in [I]$, each corresponding to a Player 1 type drawn from Nature's distribution $p = \{p_1, ..., p_I\} \in \Delta(I)$, where $\Delta(I)$ is an I-dimensional

simplex; denote $\mathcal{C} = \{x \in \mathbb{R}^{d_x} | c(x) \leq 0\}$ the set of feasible states. The goal of Player 1 is to minimize the expected payoff while keeping the state in $\mathcal{C}$. Player 1 receives $+\infty$ if state violation occurs; the goal of Player 2 is to maximize the expected payoff and hence may resort to violating the state constraint. We omit instantaneous payoffs (e.g., effort losses due to control) for conciseness, and discuss in Sec. 2.4 modifications to the Bellman backup when common-knowledge instantaneous payoffs exist.

The following assumptions will be used:

1. $\mathcal{U}$ and $\mathcal{V}$ are compact and finite-dimensional sets;
2. $f : \mathbb{R}^{d_x} \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}^{d_x}$ is bounded, continuous, and uniformly Lipschitz continuous with respect to x;
3. $g_i : \mathbb{R}^{d_x} \to \mathbb{R}$ for $i = 1, ..., I$ and $c : \mathbb{R}^{d_x} \to \mathbb{R}$ are Lipschitz continuous and bounded.
4. Isaacs' condition holds for the Hamiltonian $H : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$:

$$\begin{aligned} H(x, \xi) &:= \min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} f(x, u, v)^T \xi \\ &= \max_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} f(x, u, v)^T \xi. \end{aligned} \tag{2.4}$$

5. Control inputs and states of both players are fully observable by all. The dynamics, the payoff set, and the equilibrium strategies are common knowledge to all.

2.3.1.1 Behavioral strategy

Let $\mathcal{A}(t)$ (resp. $\mathcal{D}(t)$) be the set of open-loop controls for Player 1 (resp. Player 2):

$$\mathcal{A}(t) := \{\alpha : [t, T] \to \mathcal{U} \mid \text{Lebesgue measurable}\},$$
$$\mathcal{D}(t) := \{\delta : [t, T] \to \mathcal{V} \mid \text{Lebesgue measurable}\}.$$

Following Cardaliaguet (2007), we introduce $\mathcal{H}(t)$ (resp. $\mathcal{Z}(t)$) as the set of non-anticipative pure strategies with delay for Player 1 (resp. Player 2) Elliott and Kalton (1972):

$$\begin{aligned}\mathcal{H}(t) := \{\eta : \mathcal{D}(t) \to \mathcal{A}(t) \mid \exists \tau > 0 \text{ such that } \forall s \in (t, T-\tau) \text{ and } \delta, \bar{\delta} \in \mathcal{D}(t),\\ \text{if } \delta = \bar{\delta} \text{ a.e. in } [t,s], \text{then } \eta(\delta) = \eta(\bar{\delta}) \text{ a.e. in } [t, s+\tau]\}.\\ \mathcal{Z}(t) := \{\zeta : \mathcal{A}(t) \to \mathcal{D}(t) \mid \exists \tau > 0 \text{ such that } \forall s \in (t, T-\tau) \text{ and } \alpha, \bar{\alpha} \in \mathcal{A}(t),\\ \text{if } \alpha = \bar{\alpha} \text{ a.e. in } [t,s], \text{then } \zeta(\alpha) = \zeta(\bar{\alpha}) \text{ a.e. in } [t, s+\tau]\}\end{aligned}$$

A behavioral (mixed) strategy for Player 1 is defined by a pair $((\Omega_\eta, \mathcal{F}_\eta, \mathbf{P}_\eta), \eta)$, where $(\Omega_\eta, \mathcal{F}_\eta, \mathbf{P}_\eta)$ is a probability space and the strategy $\eta : \Omega_\eta \times \mathcal{D}(t) \to \mathcal{A}(t)$ is measurable and non-anticipative with delay, i.e., there is some $\tau > 0$ such that, for any $s \in (t, T-\tau)$ and δ, $\bar{\delta} \in \mathcal{D}(t)$, if $\delta = \bar{\delta}$ a.e. in $[t,s]$ then $\eta(\omega, \delta) = \eta(\omega, \bar{\delta})$ a.e. in $[t, s+\tau]$ for any $\omega \in \Omega_\eta$. We denote the sets of behavioral strategies of Player 1 by $(\mathcal{H}_r(t))^I$ and the behavioral strategy of Player 2 by $\mathcal{Z}_r(t)$. With mild notational abuse, we will denote by $(\eta_i) \in (\mathcal{H}_r(t))^I$ a particular set of behavioral strategies of Player 1, and by $\zeta \in \mathcal{Z}_r(t)$ a particular behavioral strategy of Player 2. Lastly, we assume that η_i for $i = 1, ..., I$ are defined on the same probability space.

Remarks. Nonanticipative strategies *with delay* are used, as opposed to ones without delay that are often used in complete-information games (Elliott and Kalton, 1972), in order to enable Lemma 2.1 that associates random strategies with open-loop controls. This association will become useful in proving the existence of value of incomplete-information differential games and in value characterization (see discussions in Cardaliaguet (2007)):

Lemma 2.1. *(Lemma 2.2 of Cardaliaguet (2007)) For any pair* $(\eta, \zeta) \in \mathcal{H}_r(t) \times \mathcal{Z}_r(t)$ *and any* $\omega := (\omega_1, \omega_2) \in \Omega_\eta \times \Omega_\zeta$*, there is a unique pair* $(\alpha_\omega, \delta_\omega) \in \mathcal{A}(t) \times \mathcal{D}(t)$ *such*

that

$$\eta(\omega_1, \delta_\omega) = \alpha_\omega \text{ and } \zeta(\omega_2, \alpha_\omega) = \delta_\omega. \tag{2.5}$$

Furthermore the map $\omega \to (\alpha_\omega, \delta_\omega)$ *is measurable from* $\Omega_\eta \times \Omega_\zeta$ *endowed with* $\mathcal{F}_\eta \otimes \mathcal{F}_\zeta$ *into* $\mathcal{A}(t) \times \mathcal{D}(t)$ *endowed with the Borel* σ*-field associated with the* L^1 *distance.*

2.3.1.2 Backward reachable set

Let $X_\tau^{t_0,x_0,\alpha,\delta}$ be the solution of Eq. (4.1) at $t = \tau$ when starting at (t_0, x_0) and following (α, δ). With behavioral strategies (η, ζ) and initials (t_0, x_0), we denote by $\mathcal{X}_\tau^{t_0,x_0,\alpha,\delta}$ the trajectory of states reachable by (α, δ) within $[t_0, \tau]$, and $\mathcal{X}_\tau^{t_0,x_0,\eta,\zeta}$ as states reachable by (η, ζ) within $[t_0, \tau]$:

$$\mathcal{X}_\tau^{t_0,x_0,\eta,\zeta} := \bigcup_{\omega \in \Omega_\eta \times \Omega_\zeta} \mathcal{X}_\tau^{t_0,x_0,\alpha_\omega,\delta_\omega}$$

where $(\alpha_\omega, \delta_\omega)$ is defined by Eq. (2.5). Introduce $\rho(\mathcal{S}) = 1$ if $\mathcal{S} \subseteq \mathcal{C}$, and otherwise $\rho(\mathcal{S}) = +\infty$; and the backward reachable (infeasible) set as

$$\bar{\mathcal{Q}}(t) := \{x \in \mathbb{R}^{d_x} \mid \forall \eta \in \mathcal{H}_r(t), \exists \zeta \in \mathcal{Z}_r(t), \tau \in (t, T], s.t., \rho\left(\mathcal{X}_\tau^{t,x,\eta,\zeta}\right) = +\infty\}.$$

$\mathcal{Q}(t) := \mathbb{R}^{d_x} \setminus \bar{\mathcal{Q}}(t)$ is the set of *feasible states*. $\mathcal{Q}(T) = \mathcal{C}$. Lastly, we use $\bar{\rho}(t, x) = 1$ if $x \in \mathcal{Q}(t)$ and otherwise $\bar{\rho}(t, x) = +\infty$.

2.3.1.3 Payoff and value

We define the expected payoff of player type i for taking behavioral strategies (η, ζ) as

$$\begin{aligned} G_i(t_0, x_0, \eta, \zeta) &:= \mathbb{E}_{\eta,\zeta}\left[g_i(X_T^{t_0,x_0,\eta,\zeta})\rho(\mathcal{X}_\tau^{t_0,x_0,\eta,\zeta})\right] \\ &= \int_{\Omega_\eta \times \Omega_\zeta} g_i\left(X_T^{t_0,x_0,\alpha_\omega,\delta_\omega}\right)\rho(\mathcal{X}_\tau^{t_0,x_0,\alpha_\omega,\delta_\omega})d\mathbf{P}_\eta \otimes \mathbf{P}_\zeta(\omega). \end{aligned}$$

The payoff of Player 1 is $\sum_{i=1}^{I} p_i G_i(t_0, x_0, \eta_i, \zeta)$.

With strategies $(\eta_i) \in (\mathcal{H}_r(t_0))^I$ and $\zeta \in \mathcal{Z}_r(t_0)$, the upper value function is defined by

$$V^+(t_0, x_0, p) := \inf_{(\eta_i)} \sup_{\zeta} \sum_{i=1}^{I} p_i G_i(t_0, x_0, \eta_i, \zeta),$$

and the lower value function is given by

$$V^-(t_0, x_0, p) := \sup_{\zeta} \inf_{(\eta_i)} \sum_{i=1}^{I} p_i G_i(t_0, x_0, \eta_i, \zeta).$$

2.3.1.4 The existence of the value

While the existence of value is proven for both zero-sum complete-information state-constrained differential games Lee (2022) and zero-sum differential games with one-sided information Cardaliaguet (2009), the proof for games with both one-sided information and state constraints is missing. Our main theoretical result fills in this gap (see Appendix 2.7.1 for the proof):

Theorem 2.1. *If assumptions 1-5 hold, we have* $V^+(t, x, p) = V^-(t, x, p)$ *for all* $(t, x, p) \in [0, T] \times \mathbb{R}^{d_x} \times \Delta(I)$.

2.3.1.5 Characterization of the value

We need to first characterize the value of the unconstrained game since this value will later appear in that of the state-constrained game.

Let $U : [0, T] \times \mathbb{R}^{d_x} \times \Delta(I) \to \mathbb{R}$ be the value of the unconstrained version of the

game, and $U^* : [0,T] \times \mathbb{R}^{d_x} \times \mathbb{R}^I \to \mathbb{R}$ its convex conjugate:

$$U^*(t,x,\hat{p}) := \sup_{p \in \Delta(I)} \hat{p}^T p - U(t,x,p)$$
$$\forall (t,x,\hat{p}) \in [0,T] \times \mathbb{R}^{d_x} \times \mathbb{R}^I.$$

We have the following properties for U and U^*:

1. U is Lipschitz continuous in (t,x,p) and convex to p. $U(T,x,p) = \sum_{i=1}^I p_i g_i(x)$, $\forall (x,p) \in \mathbb{R}^{d_x} \times \Delta(I)$. U^* is Lipschitz continuous in $(t,x,\hat{p})$ and convex to $\hat{p}$. $U^*(T,x,\hat{p}) = \max_{i \in [I]} \hat{p}_i - g_i(x)$, $\forall (x,\hat{p}) \in \mathbb{R}^{d_x} \times \mathbb{R}^I$.
2. For any $p \in \Delta(I)$, $(t,x) \to U(t,x,p)$ is a viscosity subsolution to the primal HJ equation
$$w_t + H(x, Dw) = 0,$$
where H is defined by Eq. (2.4).
3. For any $\hat{p} \in \mathbb{R}^I$, $(t,x) \to U^*(t,x,p)$ is a viscosity subsolution to the dual HJ equation
$$w_t + H^*(x, Dw) = 0,$$
where $H^*(x,\xi) = -H(x,-\xi)\ \forall\ (x,\xi) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_x}$.

The conjugate U^* defines the value of a *dual game* where Player 2 minimizes her payment, $U^*(T,x,\hat{p})$, to Player 1 where $\hat{p}$ is common knowledge. We note that by definition (see De Meyer (1996)), the dual variables $\hat{p}$ are the info-state values defined in Brown et al. Brown *et al.* (2020a), i.e., $\hat{p}[i]$ captures the value of Player 1 if he is of type i and plays the best response to Player 2's equilibrium strategy. De Meyer De Meyer (1996) showed that when $\hat{p} \in \partial_p U(0,x,p)$, Player 2's strategy in the dual game is her equilibrium in the primal game. We note that Player 2's strategy cannot be derived from the primal subdynamic principle because her best response is

dependent on Player 1's type which is unknown to her. The introduction of the dual game allows us to derive a subdynamic principle of the conjugate value from where her equilibrium strategy can be computed.

For the state-constrained game, the following corollary is a result of the subdynamic principles derived from Theorem 2.1, and will guide the numerical approximation of values for the state-constrained game (Sec. 2.4):

Corollary 2.1.1. *Under assumptions 1-5, the value function $V := V^+ = V^-$ (resp. V^*) is a unique function defined on $[0,T] \times \mathbb{R}^{d_x} \times \Delta(I)$ (resp. $[0,T] \times \mathbb{R}^{d_x} \times \mathbb{R}^I$) such that:*

1. *V is convex respect to p and*

$$V(T,x,p) = \rho(x)U(T,x,p) \; \forall (x,p) \in \mathbb{R}^{d_x} \times \Delta(I); \tag{2.6}$$

V^* is convex respect to $\hat{p}$ and

$$V^*(T,x,\hat{p}) = \max_{i \in [I]} \hat{p}_i - \rho(x) g_i(x) \; \forall (x,\hat{p}) \in \mathbb{R}^{d_x} \times \mathbb{R}^I \tag{2.7}$$

2. *For all $p \in \Delta(I)$, $(t,x) \to V(t,x,p)$ is a viscosity subsolution to the primal HJ equation*

$$\min\{\rho(x)U(t,x,p) - w, w_t + H(x,Dw)\} = 0. \tag{2.8}$$

3. *For all $\hat{p} \in \mathbb{R}^I$, $(t,x,z) \to V^*(t,x,z,p)$ is a viscosity subsolution to the dual HJ equation*

$$\min\{\rho(x)U^*(t,x,\hat{p}/\rho(x)), w_t + H^*(x,Dw)\} = 0. \tag{2.9}$$

2.4 Bellman Backup and Behavioral Strategies

Discrete-time Bellman backup computes an approximated value $V_\tau(t_k,\cdot,\cdot) : \mathbb{R}^{d_x} \times \Delta(I) \to \mathbb{R}$, with time step $\tau = T/L$ for some large L and $t_k = k\tau$ for $k = 0, ..., L$:

(i) At the terminal time, set

$$V_\tau(T, x, p) = \rho(x) \sum_i p_i g_i(x)$$

.

(ii) At $k \in \{0, ..., L-1\}$

$$V_\tau(t_k, x, p) = \rho(x) \text{Vex}_p \left(\min_u \max_v V_\tau(t_{k+1}, x', p) \right), \tag{2.10}$$

where $x' = x + \tau f(x, u, v)$ and $\text{Vex}_p(\cdot)$ is the convex hull with respect to p.

Let $l : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be a Lipschitz continuous and bounded function that represents the instantaneous payoff of the game. To incorporate l, Eq. (2.10) becomes

$$V_\tau(t_k, x, p) = \rho(x) \text{Vex}_p \left(\min_u \max_v V_\tau(t_{k+1}, x', p) + \tau l(u, v) \right) \tag{2.11}$$

Theorem 2.2 states that V_τ uniformly converges to V as $\tau \to 0^+$ (see proof in Appendix 2.7.2):

Theorem 2.2. *If assumptions 1-5 hold, V_τ converges uniformly to V on compact subsets of $[0, T] \times \mathbb{R}^{d_x} \times \Delta(I)$:*

$$\lim_{\substack{\tau \to 0^+,\ t_k \to t, \\ x' \to x, p' \to p}} V_\tau(t_k, x', p') = V(t, x, p) \quad \forall (t, x, p) \in [0, T] \times \mathbb{R}^{d_x} \times \Delta(I).$$

2.4.1 Behavioral strategy for Player 1 and the belief dynamics

Player 1's behavioral strategy is a probability distribution over $\mathcal{U}$ conditioned on (t, x, p). At time t_k, Player 1 resigns if $x_k \in \bar{\mathcal{Q}}(t_k)$; otherwise, he determines his

strategy using the following steps: First he finds $\lambda = \{\lambda_1, ..., \lambda_I\} \in \Delta(I)$ and $p^j \in \Delta(I)$ for $j = 1, ..., I$, such that

$$\begin{aligned} V_\tau(t_k, x_k, p_k) &= \sum_{j=1}^{I} \lambda_j \Big(\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} V_\tau(t_{k+1}, x'_k, p^j) \Big), \\ \sum_{j=1}^{I} \lambda_j p^j &= p_k. \end{aligned} \tag{2.12}$$

Then he computes u^j as the minimax solution corresponding to p^j, and chooses $u_k = u^j$ with probability $\lambda_j p^j[i]/p_k[i]$, where i is its true type. It is proved that this behavioral strategy of Player 1 is ϵ-optimal for small enough τ Cardaliaguet (2009). Importantly, $\{p^j\}_{j=1}^I$ are vertices of the value convex hull. Thus by announcing his strategy, and by assuming that players use the same Bayes belief update, Player 1 controls the belief dynamics to follow a martingale that optimizes his gain, i.e., $p_{k+1} = p^j$ if u^j is chosen. Note that the introduction of state constraints changes the minimax solutions, the value convex hulls, and thus the behavioral strategies. Lastly, Eq. (2.12) will be modified according to Eq. (2.11) when instantaneous loss is present.

2.4.2 The dual game and behavioral strategy for Player 2

Player 2's strategy is determined by a dual game for which the conjugate value is approximated by $V_\tau^*(t_k, \cdot, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^I \to \mathbb{R}$. Specifically,

(i) At the terminal time, set $V_\tau^*(T, x, \hat{p}) = \max_i \{\hat{p}_i - \rho(x) g_i(x)\}$.

(ii) At $k \in \{0, ..., L-1\}$

$$V_\tau^*(t_k, x, \hat{p}) = \mathrm{Vex}_{\hat{p}} \left(\min_v \max_u V_\tau^*(t_{k+1}, x', \hat{p}) \right), \tag{2.13}$$

if $\bar{\rho}(t_k, x) = 1$; otherwise $V_\tau^*(t_k, x, \hat{p}) = -\infty$.

Similar to Theorem 2.2, Theorem 2.3 proves that V_τ^* uniformly converges to V^* as $\tau \to 0^+$ (proof omitted).

Theorem 2.3. *If assumptions 1-5 holds, V_τ^* converges uniformly to V^* on compact subsets of $[0,T] \times \mathbb{R}^{d_x} \times \mathbb{R}^I$:*

$$\lim_{\substack{\tau \to 0^+,\ t_k \to t, \\ x' \to x, \hat{p}' \to \hat{p}}} V_\tau^*(t_k, x', \hat{p}') = V^*(t, x, \hat{p}) \quad \forall (t, x, \hat{p}) \in [0,T] \times \mathbb{R}^{d_x} \times \mathbb{R}^I.$$

With instantaneous loss l, the Bellman backup in Eq. (2.13) becomes

$$V_\tau^*(t_k, x, \hat{p}) = \mathrm{Vex}_{\hat{p}} \left(\min_v \max_u V_\tau^*(t_{k+1}, x', \hat{p} - \tau l(u, v)) \right) \tag{2.14}$$

We explain this modification in detail in Sec. 2.7.4. An intuitive explanation is as follows: Recall that each element of $\hat{p}$ represents Player 1's value for the corresponding type in the primal game. Hence $\hat{p}$ at the next time step should discount the common instantaneous loss incurred at the current time step.

The behavioral strategy of Player 2 defines a probability distribution over $\mathcal{V}$ conditioned on $(t, x(t), \hat{p}(t))$, with the dual variable $\hat{p}(t_0) \in \partial_p V(t_0, x_0, p(t_0))$. At time t_k, if $x_k \in \bar{\mathcal{Q}}(t_k)$, Player 2 plays according to a pursuit-evasion game since she can always catch Player 1 according to the definition of $\bar{\mathcal{Q}}(t_k)$; otherwise, Player 2 determines her strategy using the following steps: First she finds $\lambda = \{\lambda_1, ..., \lambda_{I+1}\} \in \Delta(I+1)$ and $\hat{p}^j \in \mathbb{R}^I$ for $j = 1, ..., I+1$, such that

$$\begin{aligned} V_\tau^*(t_k, x_k, \hat{p}) &= \sum_{j=1}^{I+1} \lambda_j \left(\min_{v \in \mathcal{V}} \max_{u \in \mathcal{U}} V_\tau^*(t_{k+1}, x_k', \hat{p}^j) \right), \\ \sum_{j=1}^{I+1} \lambda_j \hat{p}^j &= \hat{p}_k. \end{aligned} \tag{2.15}$$

Then she computes the minimax solution v^j, and chooses $v_k = v^j$ with probability λ_j. It is proved that this behavioral strategy of Player 2 is ϵ-optimal for small enough

τ Cardaliaguet (2009). $\hat{p}$ follows a martingale $\hat{p}_{k+1} = \hat{p}^j$ if v^j is chosen by Player 2, or $\hat{p}_{k+1} = \hat{p}^j - \tau l(u^j, v^j)$ if l is present, where u^j is the best response to v^j in the dual game. Notice that due to her lack of information, Player 2 solves harder value approximation and control synthesis problems of belief dimension $I + 1$ rather than I.

To help readers better understand the mechanisms described in this section, we provide detailed derivation of behavioral strategies for two sample problems in Appendix A (A.1 for a zero-sum version of the beer-quiche game and A.2 for Hexner's game).

2.5 Numerical Methods

2.5.1 Primal and dual value approximation

We use backward induction to solve Eq. (2.10) and (2.15), and discuss treatments that alleviate error propagation. We focus the discussion on the primal problem for brevity.

2.5.1.1 Value discontinuity

At each time step t, $V_\tau(t, \cdot, \cdot)$ (resp. $V^*_\tau(t, \cdot, \cdot)$) can be approximated separately in $\bar{\mathcal{Q}}(t)$ and $\mathcal{Q}(t)$: the primal (resp. dual) value in the former is set to $+\infty$ (resp. $-\infty$) and value in the latter will be approximated using a neural network $\hat{V}_\tau(t, \cdot, \cdot)$ (resp. $\hat{V}^*_\tau(t, \cdot, \cdot)$). This avoids fitting the value networks to functions with large Lipschitz constants during numerical implementation. $\bar{\mathcal{Q}}(t)$ for $t \in [0, T]$ can be approximated by a physics-informed neural network (PINN) solver Bansal and Tomlin

(2021) (see details in Appendix C.2), by recognizing that $\bar{\mathcal{Q}}(t)$ can be defined by pure strategies instead of behavioral ones using Lemma 2.1. PINN alleviates CoD in solving HJ equations with Lipschitz continuous solutions Shin *et al.* (2020), and here it results in a separate value network $\tilde{V}(\cdot,\cdot) : [0,T] \times \mathbb{R}^{d_x}$ that approximates $\bar{\mathcal{Q}}(t)$ as $\{x \in \mathbb{R}^{d_x} | \tilde{V}(t,x) \leq 0\}$.

2.5.1.2 Partially convex values

At each t_k and for uniformly sampled $\mathcal{S}(t) \subset \mathcal{Q}(t)$, we scan a lattice $\mathcal{P} \subset \Delta(I)$ to obtain the minimax solution of the RHS of Eq. (2.10) (denoted by $\vartheta^0(t,x,p)$ for $(x,p) \in \mathcal{S}(t) \times \mathcal{P}$), resulting in a dataset $\{(p, \vartheta^0(t,x,p))\}_{p\in\mathcal{P}}$. Value convexification is then obtained by applying the Monotone Chain Convex Hull algorithm to this dataset for each $x \in \mathcal{S}(t)$ and taking the lower hull of the resulting convex hull. Let $\vartheta(t,x,p)_{\mathcal{S}(t)}$ be the resultant value on the convex hull. A value network $\hat{V}_\tau(t,\cdot,\cdot)$ is then trained using data $\{(x, \vartheta_{\mathcal{S}(t)}(t,x,p) | (x,p) \in \mathcal{S}(t) \times \mathcal{P}\}$ so that during the Bellman backup at $t-1$, we can predict convexified values at previously unvisited states at t. We use a Partially Input Convex Neural Network (PICNN) Amos *et al.* (2017) to ensure that $\hat{V}_\tau(t,\cdot,\cdot)$ is convex in p. Alg. 1 summarizes the value approximation algorithm. Optionally, we also train a separate value network to predict the minimax values using $\{(p, \vartheta^0(t,x,p))\}_{p\in\mathcal{P}}$. This network helps remove the nested minimax problem during control synthesis.

2.5.1.3 Convexification error.

Backward induction suffers from error propagation, where errors at each time step are originated from (i) value approximation through neural networks, (ii) backward reachable set approximation, (iii) convex hull approximation, and (iv) finite time discretization (and Euler method for ODE). Here we discuss control of the error resulted from convex hull approximation, which is unique to incomplete-information games. We leave a full analysis of data complexity for error control to future studies. At each $t \in [0, T]$ and $x \in \mathcal{Q}(t)$, let $\vartheta(t, x, \cdot)$ be the RHS of Eq. (2.10) after convexification, and the convexification error be $\varepsilon_{vex}(t, x) := \max_{p \in \Delta(I)} \|\vartheta(t, x, p) - \vartheta_{\mathcal{S}(t)}(t, x, p)\|_\infty$. Proposition 2.1 shows that $\varepsilon_{vex}(t, x)$ can be controlled by refining $\mathcal{P}$ (see proof in Appendix 2.7.3):

Proposition 2.1. *For given (t, x), let the Lipschitz constant of $\vartheta(t, x, \cdot)$ be L, and $d_\mathcal{P}$ be the minimum distance between two neighboring nodes of the lattice $\mathcal{P}$. $\varepsilon_{vex}(t, x) \leq 2 d_\mathcal{P} L$.*

2.5.1.4 Approximation of the conjugate value.

Recall that the dual game is initialized by the dual variables $\hat{p} \in \partial_p V(0, x, p)$ when the primal game starts at (x, p). Since $\hat{V}(0, x, \cdot)$ is a differentiable neural network defined on a simplex, subgradients can be found using $\hat{p}^T p = V(0, x, p)$ and $\hat{p}^T q \leq V(0, x, q)$ for all $q \in \Delta(I)$ and $q \neq p$. Specific to the case study where $I = 2$ and $\hat{V} := \hat{V}_\tau(0, x, p[1])$ is modeled to be a function of the first element of p to reduce dimensionality, we have $\hat{p} = (\hat{V} + \nabla_{p[1]} \hat{V}(1 - p[1]), \hat{V} - \nabla_{p[1]} \hat{V} p[1])^T$.

2.5.2 Synthesis of strategies

Given $(t, x, p) \in [0, T] \times \mathbb{R}^{d_x} \times \Delta(I)$, Player 1 computes his behavioral strategy by finding $\lambda \in \Delta(I)$ and splitting beliefs $\{p^j \in \Delta(I)\}_{j=1}^{I}$ that best satisfy Eq. (2.12) in L^2, if $x \in \mathcal{Q}(t)$ (otherwise he surrenders). Given $(t, x, \hat{p}) \in [0, T] \times \mathbb{R}^{d_x} \times \mathbb{R}^I$, Player 2 finds $\lambda \in \Delta(I+1)$ and $\{\hat{p}^j \in \mathbb{R}^I\}_{j=1}^{I+1}$ that best satisfy Eq. (2.15) in L^2. When $I = 2$ as in the case study, the splitting beliefs and resultant strategies for Player 1 can be approximated through sweeping $p[1] \in [0, 1]$. For Player 2, we use gradient descent to solve a 6D optimization problem with the initial guess $\lambda = [1/3, 1/3, 1/3]^T$ and $\hat{p}^j = \hat{p}$ for $j = [3]$.

Algorithm 1 Value Approximation

Inputs: current time step t, time discretization τ, sample size N, admissible action spaces $(\mathcal{U}(t), \mathcal{V}(t))$, value approximation at $t + \tau$: $\hat{V}_{\text{next}}(\cdot, \cdot) := \hat{V}_\tau(t + \tau, \cdot, \cdot)$, feasible state set $\mathcal{Q}(t)$, instantaneous loss $l(\cdot, \cdot)$, terminal loss in Eq. (2.6)
Initialize: $\hat{V}_\tau(t, \cdot, \cdot)$, $\vartheta^0 = \emptyset$

```
S(t) ← sample N states from Q(t)
for x in S do
    for p in P do
        v(x, p) ← min_{u∈U(t)} max_{v∈V} V̂_next(x', p) + τ l(U, V) {if t + τ = T, V̂_next is given by Eq. (2.6)}
        append v(x, p) to ϑ^0
    end for
    ϑ_{S(t)}(t, x, ·) ← compute Vex_p(ϑ^0(x, ·)) via Eq. (2.10)
end for
Update V̂_τ to match ϑ_{S(t)}
```

2.6 Case Study

Setup. We study a state-constrained version of Hexner's game that represents a simplified football game: Player 1 (P1)'s goal is to move closer to one of the two targets than P2 without being caught during the interaction (see Fig. 2.3); P2's goal is to catch P1 if possible, or otherwise move close to P1's target. Each player has

4 state variables: x- and y- position and velocity; and their actions encode x- and y-acceleration.

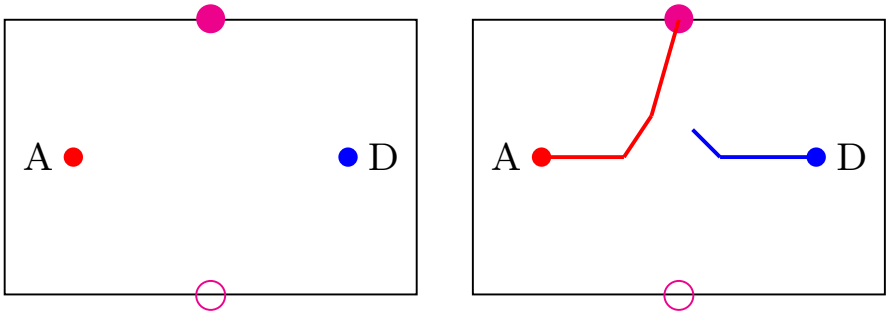


Figure 2.3. Schematics of a simplified football game with Player 1 (red) and Player 2 (blue). Left: the initial configuration. Right: equilibrium trajectory. Magenta circles: two goals. The filled is the current type private to Player 1. Players move in a 2D space bounded by $[-1, 1] \times [-1, 1]$.

The parameters $R_A = diag([0.05, 0.025])$ and $R_D = diag([0.05, 0.1])$ are chosen so that P1 can afford to accelerate faster in the y-direction than P2. The state constraint is $c(x_1, x_2) = r - \|(d_{x_1}, d_{y_1}) - (d_{x_2}, d_{y_2})\|_2$, where $r = 0.05$. We note that due to the introduction of an (instantaneous) effort loss, the backward induction is modified as: $V_\tau(t_k, x, p) = \mathrm{Vex}_p(\min_u \max_v V_\tau(t_{k+1}, x + \tau(f, x, u, v), p) + \tau l(u, v))$, where $l(u, v)$ is the integral term in $[t_k, t_{k+1}]$ in Eq. (2.1).

Value network architecture and training. The value network uses PICNN with 5 hidden layers and 256 neurons each and has 9-dimensional inputs (state and belief). We train 10 separate networks for each time step starting from $t = 0.9$ with $\tau = 0.1$, each being trained for 10 epochs. For each epoch, $\mathcal{S}(t)$ includes 5000 states sampled from $\mathcal{Q}(t)$. Since $I = 2$, value networks can be considered as functions of $p[1]$ and thus we set $\mathcal{P} = \{p[1] = 0, 0.01, \ldots, 0.99, 1\}$. $\hat{V}_\tau(t, \cdot, \cdot)$ is trained on data collected from $\mathcal{S}(t) \times \mathcal{P}$. For the conjugate $\hat{V}^*_\tau$, we set $\hat{\mathcal{P}} = \{\hat{p}[1] = \{-14, \ldots, 14\}, \hat{p}[2] = \{-14, \ldots, 14]\}$. More details can be found in Appendix C.1.

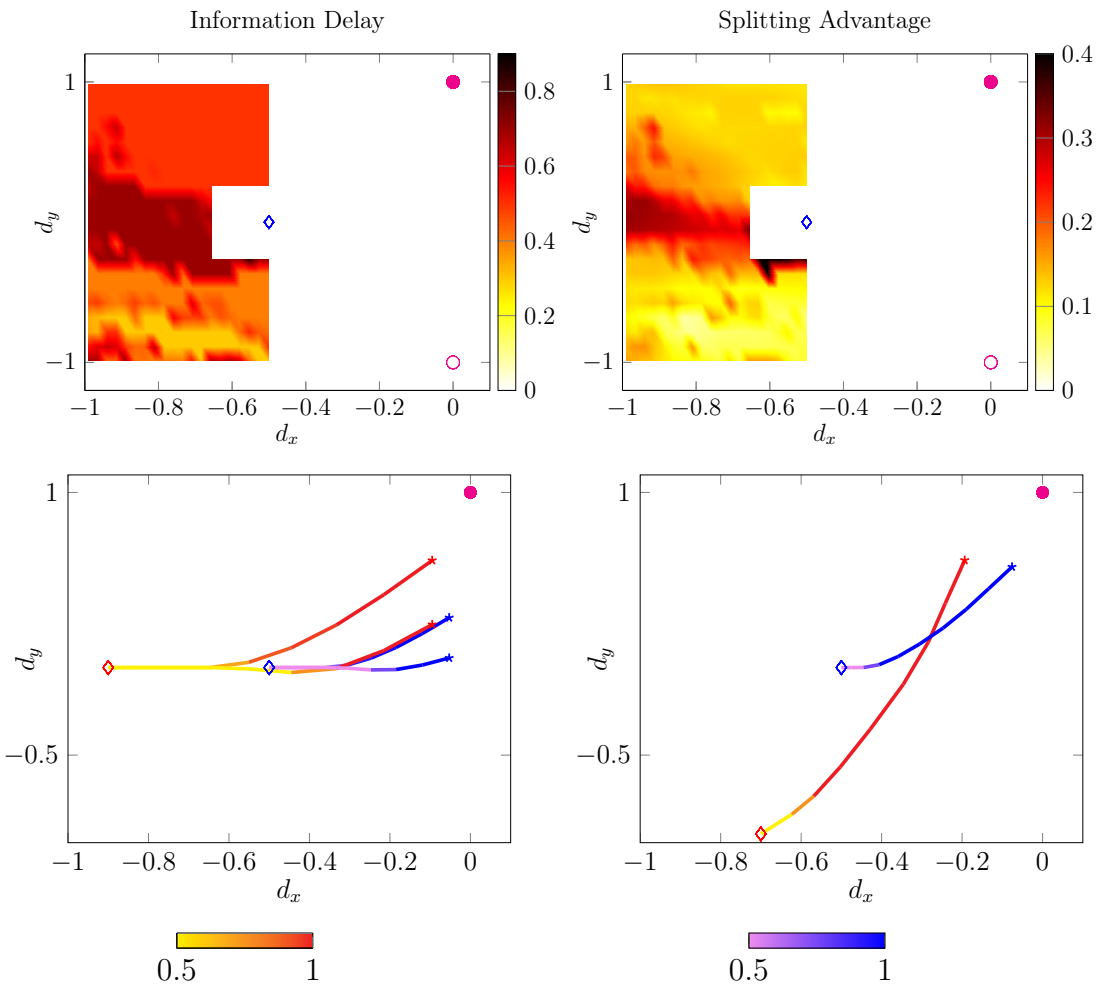


Figure 2.4. **Top:** Average delay ($\mathcal{T}$) in information reveal (left) and average maximum advantage of playing the revealing strategy (right), keeping P2's location fixed at (-0.5, 0) and changing P1's location. **Bottom:** Trajectory with high delay and advantage (left) and with low delay and advantage (right). Color shades indicate current belief.

Constrained vs. unconstrained strategies. Fig. 2.2 compares strategies with and without the state constraint, visualizing the equilibrium strategies of P1 and the best responses of P2 given P1's strategies. Note that the best responses of P2 give P2 an advantage since she does not know actions to be taken by P1 in reality. The analytical solution to the unconstrained game is given by Hexner's analysis, where P1's strategy is to start moving to the goal after the critical time $t_r = 0.4$. This strategy no longer holds in the constrained case as it violates the state constraint. Instead, P1 actively tries to stay clear of P2 while pursuing the goal (see 2nd col. of Fig. 2.2). Note that in this case, P1 resorts to a random strategy with the presence of incomplete information and state constraints, as the two contribute to value nonconvexity with respect to the belief.

Information delay and advantage of random strategies To understand how P1 uses information asymmetry, we examine the delay in information reveal, measured by the time at which the belief converges to the true type, i.e. $\mathcal{T} = \inf\{k \in [L] : p_k = \{0, 1\}\}$. We then take average of $\mathcal{T}$ for each initial state over 10 simulations. In Fig. 2.4, we visualize $\mathcal{T}$ over the space of P1's starting positions, while fixing P2's starting position and setting both players' initial speed to 0. The trajectories represent cases where P1 delays (bottom left) and does not delay (bottom right) the release of information. We also compute the advantage of following a belief manipulating strategy (that convexifies the value) as opposed to taking the non-revealing strategy (i.e. never split), expressed as $[\min_u \max_v V(t+\tau, x', p) - V(t, x, p)]$. Overall, P1 tends to conceal and deceive when it has equal distances to the possible targets.

Equilibrium strategy of Player 2. Fig. 2.5 shows sample trajectories when both players play their equilibrium strategies. Note that compared with P2's best responses to P1 in previous examples, P2's equilibrium strategy is more conservative,

due to her lack of knowledge about P1's type. We also note that P2's dual game is one dimension higher than P1's primal game, and thus encounters higher numerical errors in value and strategy approximation (see Appendix C.1).

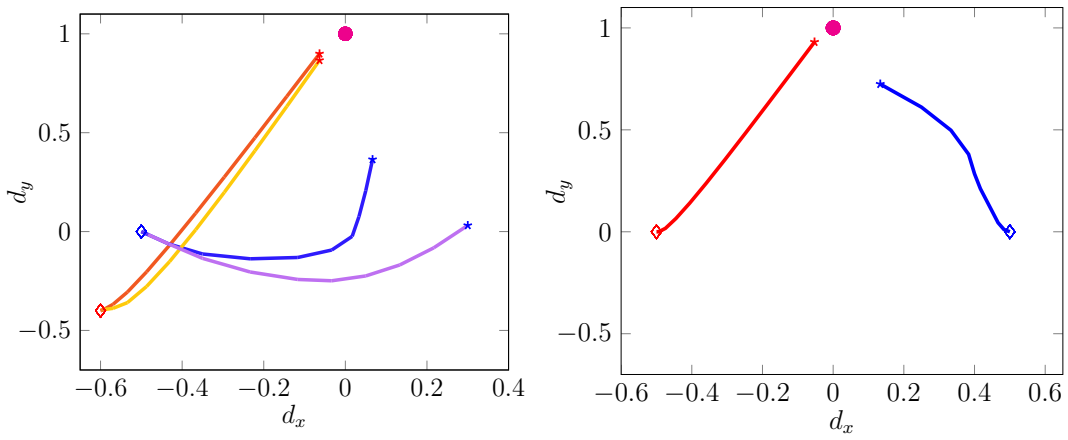


Figure 2.5. Trajectories where both players use their respective behavioral strategies. P1 keeps track of p, whereas P2 keeps track of $\hat{p}$.

2.7 Proofs and Additional Results

2.7.1 Proof of Theorem 2.1

The following proofs extend results and techniques from Cardaliaguet (2007) to zero-sum differential games with one-sided information and state constraints. To overview, we start by showing that the upper and lower values $V^{\pm}$ are Lipschitz continuous within the safe and unsafe state sets (Lemma 2.2) and convex with respect to p (Lemma 2.3). We then show that: (1) V^{-*} satisfies a subdynamic principle and is therefore a subsolution of a dual HJ equation and hence V^{-} is a dual supersolution of the corresponding primal HJ (Lemma 2.4, Lemma 2.4, Lemma 2.5); and (2) V^{+} also

satisfies a subdynamic principle (Lemma 2.6) and is therefore a dual subsolution of the primal HJ (Lemma 2.7). We can then use a comparison principle (see Cardaliaguet (2007)) to show that since V^- is a dual supersolution and V^+ is a dual subsolution of the primal HJ while both share the same terminal value, $V^- \geq V^+$. On the other hand, $V^- \leq V^+$ by definition and hence $V^- = V^+$.

We start with the following regularity result (see proof of Lemma 3.1 in Cardaliaguet (2007)):

Lemma 2.2. *(regularity of $V^\pm$). $V^\pm(t_0, x_0, p)$ are Lipschitz continuous for all $x_0 \in \mathcal{Q}(t_0)$. $V^\pm(t_0, x_0, p) = +\infty$ for all $x_0 \in \bar{\mathcal{Q}}(t_0)$.*

The following convexity result was originally developed for repeated games with incomplete information De Meyer (1996) and was later extended to differential games Cardaliaguet (2007). The same convexity result holds for imperfect-information dynamic games Brown *et al.* (2020a).

Lemma 2.3. *(convexity property of $V^\pm$). For any $(t, x) \in [0, T] \times \mathbb{R}^{d_x}$, $V^\pm$ are convex in p on $\Delta(I)$.*

Proof. Let $p^\lambda = (1-\lambda)p^0 + \lambda p^1$ for some p^0, $p^1 \in \Delta(I)$. Let $((\eta_i^0), \zeta^0)$ and $((\eta_i^1), \zeta^1)$ be the equilibrial strategies for $V(t, x, p^0)$ and $V(t, x, p^1)$, respectively. Introduce a set of "splitting" behavioral strategies (η_i^λ) for (t, x, p^λ) such that for any type i, $\eta_i^\lambda = \eta_i^0$ with probability $(1-\lambda)p_i^0/p_i^\lambda$ and $\eta_i^\lambda = \eta_i^1$ with probability $\lambda p_i^1/p_i^\lambda$. Then we have

$$\begin{aligned}
&\sup_\zeta \sum_i p_i^\lambda G_i(t, x, \eta_i^\lambda, \zeta) \\
=&\sup_\zeta \sum_i \left(p_i^\lambda \frac{(1-\lambda)p_i^0}{p_i^\lambda} G_i(t, x, \eta_i^0, \zeta) + p_i^\lambda \frac{\lambda p_i^1}{p_i^\lambda} G_i(t, x, \eta_i^1, \zeta) \right) \qquad (2.16) \\
\leq&(1-\lambda) \sup_\zeta \sum_i p_i^0 G_i(t, x, \eta_i^0, \zeta) + \lambda \sup_\zeta \sum_i p_i^1 G_i(t, x, \eta_i^1, \zeta).
\end{aligned}$$

Since the inequality in Eq. (2.16) holds for any "splitting" (η_i^λ), we have

$$V^{\pm}(t,x,p^\lambda) \leq (1-\lambda)V^{\pm}(t,x,p^0) + \lambda V^{\pm}(t,x,p^1) \tag{2.17}$$

for any $t \in [0,T]$ and $x \in \mathcal{Q}(t)$. For $x \in \bar{\mathcal{Q}}(t)$, the equality holds since $V^{\pm}(t,x,\cdot) = +\infty$. □

Remarks. (1) The proof says that by playing a "splitting" strategy, the value at (t,x,p^λ) should at least be as good as a linear interpolation between those at (t,x,p^0) and at (t,x,p^1). Hence the value is a convex hull in $\Delta(I)$ at any $(t,x) \in [0,T] \times \mathbb{R}^{d_x}$. (2) Assuming that the "splitting" strategy of Player 1 is known by Player 2, then the latter can perform Bayesian inference on Player 1's type based on his actions. For any type i, let u_i^0 and u_i^1 be two distinct actions to be taken at (t,x) following η_i^0 and η_i^1, respectively, and let p^λ be the current common belief. Then under observation of u_i^0 (resp. u_i^1), the common belief becomes p^0 (resp. p^1) with probability $(1-\lambda)$ (resp. λ). Since $p^\lambda = (1-\lambda)p^0 + \lambda p^1$, common belief is a martingale.

Next, we introduce a reformulation of V^{-*} to facilitate the derivation of its subdynamic principle. The proof of this reformulation is extended from Lemma 4.1 of Cardaliaguet (2007) to incorporate the state constraints.

Lemma 2.4. *(reformulation of V^{-*}). We have*

$$V^{-*}(t_0,x_0,\hat{p}) = \inf_{\zeta \in \mathcal{Z}_r(t_0)} \sup_{\eta \in \mathcal{H}_r(t_0)} \max_i \{\hat{p}_i - G_i(t_0,x_0,\eta,\zeta)\} \tag{2.18}$$

Proof. For later use, we first note that

$$\begin{aligned} V^{-}(t_0,x_0,p) &= \sup_{\zeta} \inf_{(\eta_i)} \sum_i p_i G_i(t_0,x_0,\eta,\zeta) \\ &= \sup_{\zeta} \sum_i p_i \inf_{\eta} G_i(t_0,x_0,\eta,\zeta). \end{aligned} \tag{2.19}$$

Let the right-hand side of Eq. (2.18) be $z = z(t_0, x_0, \hat{p})$. We can show that z is convex with respect to $\hat{p}$ using a technique similar to that of Lemma 2.3.

Then by the definition of z:

$$\begin{aligned} z^*(t_0, x_0, p) &= \sup_{\hat{p}} p^T \hat{p} - \inf_{\zeta} \sup_{\eta} \max_i \{\hat{p}_i - G_i(t_0, x_0, \eta, \zeta)\} \\ &= \sup_{\hat{p}} p^T \hat{p} - \inf_{\zeta} \max_i \left\{ \hat{p}_i - \inf_{\eta} G_i(t_0, x_0, \eta, \zeta) \right\} \\ &= \sup_{\zeta} \sup_{\hat{p}} \min_i \left\{ p^T \hat{p} - \hat{p}_i + \inf_{\eta} G_i(t_0, x_0, \eta, \zeta) \right\}. \end{aligned} \tag{2.20}$$

In this last expression, $\sup_{\hat{p}}$ is attained by setting $\hat{p}_i = \inf_\eta G_i(t_0, x_0, \eta, \zeta)$, in which case we have

$$z^*(t_0, x_0, p) = \sup_{\zeta} \sum_i p_i \inf_{\eta} G_i(t_0, x_0, \eta, \zeta) = V^-(t_0, x_0, p, z). \tag{2.21}$$

Since z is convex with respect to $\hat{p}$, we have $V^{-*} = z^{**} = z$. □

Next, to introduce the subdynamic principle of V^{-*}, we first introduce

$$U^{-*}(t_0, x_0, \hat{p}) := \inf_{\zeta \in \mathcal{Z}_r(t_0)} \sup_{\eta \in \mathcal{H}_r(t_0)} \max_i \left\{ \hat{p}_i - \mathbb{E}_{\eta,\zeta} \left[g_i(X_T^{t_0,x_0,\eta,\zeta}) \right] \right\} \tag{2.22}$$

as the conjugate lower value of the *unconstrained* version of the game, and $U^{\pm}$ as the corresponding upper and lower values. From Lemma 2.2 and Lemma 2.4, U^{-*} is Lipschitz continuous and convex in $\hat{p}$.

Lemma 2.5. *(subdynamic principle for V^{-*}). For any $(t_0, x_0, \hat{p}) \in [0, T) \times \mathbb{R}^{d_x} \times \mathbb{R}^I$ and any $t_1 \in (t_0, T]$, denote $x_1 = X_{t_1}^{t_0,x_0,\eta,\zeta}$ and $\mathcal{X}_1 = \mathcal{X}_{t_1}^{t_0,x_0,\eta,\zeta}$. We have*

$$V^{-*}(t_0, x_0, \hat{p}) \le \inf_{\zeta \in \mathcal{Z}(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} \min \left\{ \rho(\mathcal{X}_1) U^{-*} \left(t_1, x_1, \frac{\hat{p}}{\rho(\mathcal{X}_1)} \right), V^{-*}(t_1, x_1, \hat{p}) \right\} \tag{2.23}$$

Proof. Denote $V_1^{-*}(t_0, t_1, x_0, \hat{p}) := \inf_{\zeta \in \mathcal{Z}(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} V^{-*}\left(t_1, X_{t_1}^{t_0,x_0,\eta,\zeta}, \hat{p}\right)$. U_1^{-*} is similarly defined. We need the following preparations for the proof.

Player 1 plays a pure strategy in V^{-*}. We show below that best responses are always pure. In particular, Player 1 can play in pure strategies in V^{-*}, namely,

$$V^{-*}(t_0, x_0, \hat{p}) = \inf_{\zeta \in \mathcal{Z}_r(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} \max_i \{\hat{p}_i - G_i(t_0, x_0, \eta, \zeta)\} \tag{2.24}$$

for any $(t_0, x_0, \hat{p})$. First from Theorem 2.4 and using $\mathcal{H}(t) \subset \mathcal{H}_r(t)$, we have

$$\begin{aligned} V^{-*}(t_0, x_0, \hat{p}) &= \inf_{\zeta \in \mathcal{Z}_r(t_0)} \sup_{\eta \in \mathcal{H}_r(t_0)} \max_i \{\hat{p}_i - G_i(t_0, x_0, \eta, \zeta)\} \\ &\geq \inf_{\zeta \in \mathcal{Z}_r(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} \max_i \left\{\hat{p}_i - \mathbb{E}_\zeta \left[g_i(X_T^{t_0,x_0,\eta,\zeta}) \rho(\mathcal{X}_T^{t_0,x_0,\eta,\zeta})\right]\right\}. \end{aligned} \tag{2.25}$$

For the reverse inequality, we first note that for any $\eta \in \mathcal{H}_r(t_0)$ and $\omega_1 \in \Omega_\eta$, $\eta(\omega_1, \cdot) \in \mathcal{H}(t_0)$. With a fixed $\zeta \in \mathcal{Z}_r(t_0)$, and by using the convexity of $\max_i$ (i.e., $\max_i \mathbb{E}_\omega[f_i(\omega)] \leq \mathbb{E}_\omega[\max_i f_i(\omega)]$), we have

$$\begin{aligned} &\sup_{\eta \in \mathcal{H}_r(t_0)} \max_i \{\hat{p}_i - G_i(t_0, x_0, \eta, \zeta)\} \\ &\leq \sup_{\eta \in \mathcal{H}_r(t_0)} \int_{\Omega_\eta} \max_i \left\{\hat{p}_i - \mathbb{E}_\zeta \left[g_i(X_T^{t_0,x_0,\eta(\omega_1,\cdot),\zeta}) \rho(\mathcal{X}_T^{t_0,x_0,\eta(\omega_1,\cdot),\zeta})\right]\right\} d\mathbf{P}_\eta(\omega_1) \\ &\leq \sup_{\eta \in \mathcal{H}_r(t_0)} \sup_{\omega_1 \in \Omega_\eta} \max_i \left\{\hat{p}_i - \mathbb{E}_\zeta \left[g_i(X_T^{t_0,x_0,\eta(\omega_1,\cdot),\zeta}) \rho(\mathcal{X}_T^{t_0,x_0,\eta(\omega_1,\cdot),\zeta})\right]\right\} \\ &\leq \sup_{\eta \in \mathcal{H}(t_0)} \max_i \left\{\hat{p}_i - \mathbb{E}_\zeta \left[g_i(X_T^{t_0,x_0,\eta(\omega_1,\cdot),\zeta}) \rho(\mathcal{X}_T^{t_0,x_0,\eta(\omega_1,\cdot),\zeta})\right]\right\}. \end{aligned} \tag{2.26}$$

Since Eq. (2.26) holds for any ζ, together with Eq. (2.25), we have

$$V^{-*}(t, x, \hat{p}) = \inf_{\zeta \in \mathcal{Z}_r(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} \max_i \left\{\hat{p}_i - \mathbb{E}_\zeta \left[g_i(X_T^{t_0,x_0,\eta,\zeta}) \rho(\mathcal{X}_T^{t_0,x_0,\eta,\zeta})\right]\right\}. \tag{2.27}$$

Note that one can reach the same conclusion for U^{-*}.

ϵ-optimal strategy of Player 2. Let $\epsilon > 0$ and $\zeta^0 \in \mathcal{Z}(t_0)$ be some pure ϵ-optimal strategy for $V_1^{-*}(t_0, t_1, x, \hat{p})$. For any $x_1 \in \mathbb{R}^{d_x}$, we can find some ϵ-optimal strategy $\zeta^{x_1} \in \mathcal{Z}_r(t_1)$ for Player 2 in the game $V^{-*}(t_1, x_1, \hat{p})$. Let $B_\rho(x)$ be a ball around x with radius ρ, and let $\partial\mathcal{Q}(t)$ be the boundary of $\mathcal{Q}(t)$, i.e., for any $x \in \partial\mathcal{Q}(t)$ and $\rho > 0$, there exist $y \in B_\rho(x) \bigcap \mathcal{Q}(t)$ and $y' \in B_\rho(x) \bigcap \bar{\mathcal{Q}}(t)$.

For $x_1, y \in \mathcal{Q}(t_1) \setminus \partial\mathcal{Q}(t_1)$, from Lipschitz continuity of the map $y \to V^{-*}(t_1, y, \hat{p})$, ζ^{x_1} is also (2ϵ)-optimal for $V^{-*}(t_1, y, \hat{p})$ if $y \in B_\rho(x_1)$ for some radius $\rho > 0$. The same applies to $x_1, y \in \bar{\mathcal{Q}}(t_1)$ and $y \in B_\rho(x_1)$ since $V^{-*}(t_1, x, \hat{p}) = +\infty$ is constant for $x \in \bar{\mathcal{Q}}(t_1)$.

Since the dynamics f is bounded, we also know that the reachable states from (t_0, x_0) is bounded in some ball $B_R(0)$. Let us set $M = \|f\|_\infty$ and some small $\sigma > 0$ such that $M\sigma \leq \rho/2$. Then we choose $(x_l)_{l=1,...,l_0}$ such that $\bigcup_{l=1}^{l_0} B_{\rho/2}(x_l)$ contains $B_R(0)$. Let $(E_l)_{l=1,...,l_0}$ be a Borel partition of $B_R(0)$ such that, for any l, $E_l \subset B_{\rho/2}(x_l)$. We also require (x_l) to be chosen properly so that $E_l \subset \mathcal{Q}(t_1)$ or $E_l \subset \bar{\mathcal{Q}}(t_1)$.

We set

$$\zeta^l = \zeta^{x_l},\ \Omega^l = \Omega_{\zeta^{x_l}},\ \mathcal{F}^l = \mathcal{F}_{\zeta^{x_l}},\ \text{and } \mathbf{P}^l = \mathbf{P}_{\zeta^{x_l}} \tag{2.28}$$

for $l = 1, ..., l_0$. We choose some delay $\tau \in (0, \sigma]$ for all the strategies ζ^l. Note that if for some open-loop control $(\alpha, \delta) \in \mathcal{A}(t_0) \times \mathcal{D}(t_0)$ and for some l, we have $X_{t_1-\tau}^{t_0,x_0,\alpha,\delta} \in E_l$, then

$$|X_{t_1-\tau}^{t_0,x_0,\alpha,\delta} - X_{t_1}^{t_0,x_0,\alpha,\delta}| \leq \|f\|_\infty \tau \leq M\sigma \leq \rho/2, \tag{2.29}$$

so that $X_{t_1}^{t_0,x_0,\alpha,\delta}$ belongs to $B_\rho(x_l)$. Hence ζ^l is (2ϵ)-optimal for V^{-*} at $(t_1, X_{t_1}^{t_0,x_0,\alpha,\delta}, \hat{p})$.

Let us now define a new strategy $\zeta \in \mathcal{Z}_r(t_0)$ by setting

$$\Omega_\zeta = \prod_{l=1}^{l_0} \Omega^l,\ \mathcal{F}_\zeta = \mathcal{F}^1 \otimes ... \otimes \mathcal{F}^{l_0},\ \text{and } \mathbf{P}_\zeta = \mathbf{P}^1 \otimes ... \otimes \mathbf{P}^{l_0}. \tag{2.30}$$

For any $\omega = (\omega^1, ..., \omega^{l_0}) \in \Omega_\zeta$ and $\alpha \in \mathcal{A}(t_0)$, set

$$\zeta(\omega, \alpha) = \begin{cases} \zeta^0(\alpha)(\tau) & \text{if } \tau \in [t_0, t_1) \\ \zeta^l(\omega^l, \alpha)(\tau) & \text{if } \tau \in [t_1, T] \text{ and } X_{t_1-\tau}^{t_0,x_0,\alpha,\zeta^0} \in E_l. \end{cases} \tag{2.31}$$

For any pure strategy $\eta \in \mathcal{H}(t_0)$, we have

$$g_i(X_T^{t_0,x_0,\eta,\zeta}) = \sum_{l=1}^{l_0} g_i\left(X_T^{t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0}, \eta, \zeta^l}\right) \mathbf{1}_{O^l}, \tag{2.32}$$

where $O^l := \left\{ X_{t_1-\tau}^{t_0,x_0,\eta,\zeta^0} \in E_l \right\}$.

Property of $\rho(\cdot)$. Let $\rho_0 := \rho\left(\mathcal{X}_{t_1}^{t_0,x_0,\eta,\zeta^0}\right)$ and $\rho_1^l := \rho\left(\mathcal{X}_T^{t_1,x_1^l,\eta,\zeta^l}\right)$, where $x_1^l := X_{t_1}^{t_0,x_0,\eta,\zeta^0}$ if the state falls in E_l following pure strategies (η, ζ^0). Then we have

$$\begin{aligned} g_i\left(X_T^{t_0,x_0,\eta,\zeta}\right)\rho\left(\mathcal{X}_T^{t_0,x_0,\eta,\zeta}\right) &= g_i(X_T^{t_0,x_0,\eta,\zeta})\sum_l \max\{\rho_0,\rho_1^l\}\mathbf{1}_{O^l} \\ &= \sum_l g_i(X_T^{t_1,x_1^l,\eta,\zeta})\mathbf{1}_{O^l}\sum_l \max\{\rho_0,\rho_1^l\}\mathbf{1}_{O^l} \\ &= \sum_l g_i(X_T^{t_1,x_1^l,\eta,\zeta})\max\{\rho_0,\rho_1^l\}\mathbf{1}_{O^l}. \end{aligned} \tag{2.33}$$

For any set $(\omega^l) \in (\Omega^l)$, let $a^l := g_i(X_T^{t_1,x_1^l,\eta,\zeta(\omega^l,\cdot)}) \geq 0$. Also note that ρ_0 and ρ_1^l only take values from $\{1, +\infty\}$. One can show that the following always holds:

$$\sum_l a^l \max\{\rho_0,\rho_1^l\}\mathbf{1}_{O^l} = \max\left\{\sum_l a^l\rho_0\mathbf{1}_{O^l}, \sum_l a^l\rho_1^l\mathbf{1}_{O^l}\right\}. \tag{2.34}$$

Similarly we have

$$\begin{aligned} &\int_{\Omega^l}\sum_l\left[g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta,\zeta^l(\omega^l,\cdot)}\right)\max\{\rho_0,\rho_1^l\}\mathbf{1}_{O^l}\right]d\mathbf{P}^l(\omega^l) \\ &= \max\left\{\begin{array}{l} \int_{\Omega^l}\sum_l\left[g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta,\zeta^l(\omega^l,\cdot)}\right)\rho_0\mathbf{1}_{O^l}\right]d\mathbf{P}^l(\omega^l), \\ \\ \int_{\Omega^l}\sum_l\left[g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta,\zeta^l(\omega^l,\cdot)}\right)\rho_1^l\mathbf{1}_{O^l}\right]d\mathbf{P}^l(\omega^l) \end{array}\right\} \end{aligned} \tag{2.35}$$

Now we derive an upper bound of $\max_i \left\{\hat{p}_i - \mathbb{E}_\zeta\left[g_i(X_T^{t_0,x_0,\eta,\zeta})\rho(\mathcal{X}_T^{t_0,x_0,\eta,\zeta})\right]\right\}$:

$$
\begin{aligned}
&\max_i \left\{\hat{p}_i - \mathbb{E}_\zeta\left[g_i(X_T^{t_0,x_0,\eta,\zeta})\rho(\mathcal{X}_T^{t_0,x_0,\eta,\zeta})\right]\right\} \\
=&\max_i \left\{\hat{p}_i - \int_{\Omega^l}\sum_l \left[g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta,\zeta^l(\omega^l,\cdot)}\right)\max\{\rho_0,\rho_1^l\}\mathbf{1}_{O^l}\right]d\mathbf{P}^l(\omega^l)\right\} \\
\leq& \sup_{\eta'\in\mathcal{H}(t_1)}\max_i \left\{\hat{p}_i - \sum_l\left[\int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)\max\{\rho_0,\rho_1^l\}d\mathbf{P}^l(\omega^l)\mathbf{1}_{O^l}\right]\right\} \\
=& \sup_{\eta'\in\mathcal{H}(t_1)}\max_i \left\{\min\left\{\begin{array}{l}\hat{p}_i - \sum_l\left[\int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)\rho_0 d\mathbf{P}^l(\omega^l)\mathbf{1}_{O^l}\right], \\ \hat{p}_i - \sum_l\left[\int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)\rho_1^l d\mathbf{P}^l(\omega^l)\mathbf{1}_{O^l}\right]\end{array}\right\}\right\} \\
&\qquad\qquad\qquad\qquad\leq \\
&\min\left\{\begin{array}{l}\sup_{\eta'\in\mathcal{H}(t_1)}\max_i\left\{\hat{p}_i - \sum_l\left[\int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)\rho_0 d\mathbf{P}^l(\omega^l)\mathbf{1}_{O^l}\right]\right\}, \\ \sup_{\eta'\in\mathcal{H}(t_1)}\max_i\left\{\hat{p}_i - \sum_l\left[\int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)\rho_1^l d\mathbf{P}^l(\omega^l)\mathbf{1}_{O^l}\right]\right\}\end{array}\right\} \\
&\qquad\qquad\qquad\qquad\leq \\
&\min\left\{\begin{array}{l}\sum_l\left[\rho_0\sup_{\eta'\in\mathcal{H}(t_1)}\max_i\left\{\frac{\hat{p}_i}{\rho_0} - \int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)d\mathbf{P}^l(\omega^l)\right\}\mathbf{1}_{O^l}\right], \\ \sum_l\left[\sup_{\eta'\in\mathcal{H}(t_1)}\max_i\left\{\hat{p}_i - \int_{\Omega^l} g_i\left(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}\right)\rho_1^l d\mathbf{P}^l(\omega^l)\right\}\mathbf{1}_{O^l}\right]\end{array}\right\}.
\end{aligned}
\tag{2.36}
$$

Shorten the first and second terms in the above upper bound $\min\{\cdot,\cdot\}$ as A and

B, respectively. For B:

$$\begin{aligned}
&\sum_l \left[\sup_{\eta' \in \mathcal{H}(t_1)} \max_i \left\{ \hat{p}_i - \int_{\Omega^l} g_i \left(X_T^{t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0}, \eta', \zeta^l(\omega^l,\cdot)} \right) \rho_1^l d\mathbf{P}^l(\omega^l) \right\} \mathbf{1}_{O^l} \right] \\
\leq &\sum_l \left(V^{-*}(t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0}, \hat{p}) + 2\epsilon \right) \mathbf{1}_{O^l} \\
&\text{(because } \zeta^l \text{ is } (2\epsilon)\text{-optimal for } V^{-*} \text{ at } (t_1, x_1, \hat{p}) \text{ for any } x_1 \in E_l.) \qquad (2.37) \\
= &V^{-*}(t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0}, \hat{p}) + 2\epsilon \\
\leq &V_1^{-*}(t_0, t_1, x_0, \hat{p}) + 3\epsilon \\
&\text{(because } \zeta^0 \text{ is } \epsilon\text{-optimal for } V_1^{-*}(t_0, t_1, x_0, \hat{p}).)
\end{aligned}$$

For A, we consider the following scenarios: (1) If $\inf_{\eta' \in \mathcal{H}(t_1)} \sum_l \rho_1^l \mathbf{1}_{O^l} = +\infty$, i.e., there is always a chance for Player 2 to achieve constraint violation when the game starts at $(t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0})$, and $\rho_0 = 1$, i.e., (η, ζ^0) does not induce constraint violation, then

$$\begin{aligned}
A = &\sum_l \left[\sup_{\eta' \in \mathcal{H}(t_1)} \max_i \left\{ \hat{p}_i - \inf_{\eta' \in \mathcal{H}(t_1)} \int_{\Omega^l} g_i \left(X_T^{t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0}, \eta', \zeta^l(\omega^l,\cdot)} \right) d\mathbf{P}^l(\omega^l) \right\} \mathbf{1}_{O^l} \right] \\
> &\sum_l \left[\sup_{\eta' \in \mathcal{H}(t_1)} \max_i \left\{ \hat{p}_i - \inf_{\eta' \in \mathcal{H}(t_1)} \int_{\Omega^l} g_i \left(X_T^{t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0}, \eta', \zeta^l(\omega^l,\cdot)} \right) \rho_1^l d\mathbf{P}^l(\omega^l) \right\} \mathbf{1}_{O^l} \right] \\
= &B = -\infty.
\end{aligned} \qquad (2.38)$$

If $\inf_{\eta' \in \mathcal{H}(t_1)} \sum_l \rho_1^l \mathbf{1}_{O^l} = \rho_0 = 1$, then $A = B$. This is the scenario where the game reduces to its unconstrained version. Applying the same analysis from B to have $A \leq \rho_0(U_1^{-*}(t_0, t_1, x_0, \hat{p}/\rho_0) + 3\epsilon)$. If $\inf_{\eta' \in \mathcal{H}(t_1)} \sum_l \rho_1^l \mathbf{1}_{O^l} = \rho_0 = +\infty$, then $A = B = -\infty$.

If $\inf_{\eta' \in \mathcal{H}(t_1)} \sum_l \rho_1^l \mathbf{1}_{O^l} = 1$ and $\rho_0 = +\infty$, the game starting from $(t_1, X_{t_1}^{t_0,x_0,\eta,\zeta^0})$ will be played as an unconstrained one. $A < B$. Hence $A \leq \rho_0(U_1^{-*}(t_0, t_1, x_0, \hat{p}/\rho_0) + 3\epsilon) = -\infty$.

Combining these scenarios we have

$$\min\{A, B\} \leq \min\{\rho_0(U_1^{-*}(t_0, t_1, x_0, \hat{p}/\rho_0) + 3\epsilon), V_1^{-*}(t_0, t_1, x_0, \hat{p}) + 3\epsilon\}. \tag{2.39}$$

Since ϵ can be arbitrarily small, we have

$$\begin{aligned} V^{-*}(t_0, x_0, \hat{p}) &\leq \min\left\{\rho_0 U_1^{-*}(t_0, t_1, x_0, \hat{p}/\rho_0), V_1^{-*}(t_0, t_1, x_0, \hat{p})\right\} \\ &= \inf_{\zeta \in \mathcal{Z}(t_0)} \min\left\{ \sup_{\eta \in \mathcal{H}(t_0)} \rho_0 U^{-*}(t_1, X_{t_1}^{t_0, x_0, \eta, \zeta}, \hat{p}/\rho_0), \right. \\ &\qquad \left. \sup_{\eta \in \mathcal{H}(t_0)} V^{-*}(t_1, X_{t_1}^{t_0, x_0, \eta, \zeta}, \hat{p}) \right\} \end{aligned} \tag{2.40}$$

Lastly, if $x_0 \in \bar{\mathcal{Q}}(t_0)$, $\inf_\zeta \sup_\eta \rho_0 = -\infty$ by definition; otherwise, $\rho_0 = 1$ and $U^{-*} = V^{-*}$ at t_1. In both cases, the RHS of Eq. (2.40) becomes

$$\inf_{\zeta \in \mathcal{Z}(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} \min\left\{\rho_0 U^{-*}(t_0, X_{t_1}^{t_0, x_0, \eta, \zeta}, \hat{p}/\rho_0), V^{-*}(t_0, X_{t_1}^{t_0, x_0, \eta, \zeta}, \hat{p})\right\}. \tag{2.41}$$

□

Theorem 2.4. *(V^{-*} is a subsolution of HJ). For any $\hat{p} \in \mathbb{R}^I$, the map $(t, x) \to V^{-*}(t, x, \hat{p})$ is a viscosity subsolution of the dual Hamilton-Jacobi equation:*

$$\min\left\{\rho(x) U^{-*}(t, x, \hat{p}/\rho(x)) - w,\ w_t + H^*(x, Dw)\right\} = 0 \text{ in } [0, T] \times \mathbb{R}^{d_x}, \tag{2.42}$$

where H is defined by Eq. (2.4) and $H^*(x, \xi) = -H(x, -\xi)$. $\bar{\rho}(t_0, x_0) = 1$ if $x_0 \in \mathcal{Q}(t_0)$.

Proof. Let $\hat{p} \in \mathbb{R}^I$ be fixed, and let ϕ be a smooth test function such that

$$\phi(t, x) \geq V^{-*}(t, x, \hat{p}) \quad \forall (t, x) \in [0, T] \times \mathbb{R}^{d_x}, \tag{2.43}$$

with an equality at (t_0, x_0), where $t_0 \in [0, T)$. For any $v \in \mathcal{V}$, define a pure strategy $\zeta \in \mathcal{Z}(t_0)$ by setting

$$\zeta(\alpha)(t) = v \quad \forall \alpha \in \mathcal{A}(t_0),\ t \in [t_0, T]. \tag{2.44}$$

Since V^{-*} satisfies the subdynamic programming principle of Lemma 2.5, these exist $\epsilon > 0$, $h > 0$, and a pure strategy $\eta_h \in \mathcal{H}(t_0)$ such that

$$V^{-*}(t_0, x_0, \hat{p}) \leq \min \left\{ \rho \left(\mathcal{X}_{t_0+h}^{t_0,x_0,\eta_h,\zeta} \right) U^{-*} \left(t_0 + h, X_{t_0+h}^{t_0,x_0,\eta_h,\zeta}, \hat{p}/\rho \left(\mathcal{X}_{t_0+h}^{t_0,x_0,\eta_h,\zeta} \right) \right), \right.$$
$$\left. V^{-*} \left(t_0 + h, X_{t_0+h}^{t_0,x_0,\eta_h,\zeta}, \hat{p} \right) \right\} + \epsilon h, \tag{2.45}$$

or equivalently

$$\rho \left(\mathcal{X}_{t_0+h}^{t_0,x_0,\eta_h,\zeta} \right) U^{-*} \left(t_0 + h, X_{t_0+h}^{t_0,x_0,\eta_h,\zeta}, \hat{p}/\rho \left(\mathcal{X}_{t_0+h}^{t_0,x_0,\eta_h,\zeta} \right) \right) - V^{-*} \left(t_0, x_0, \hat{p} \right) + \epsilon h \geq 0, \tag{2.46}$$

and

$$V^{-*} \left(t_0 + h, X_{t_0+h}^{t_0,x_0,\eta_h,\zeta}, \hat{p} \right) - V^{-*}(t_0, x_0, \hat{p}) + \epsilon h \geq 0. \tag{2.47}$$

Set the open-loop control $\alpha_h(s) := \eta_h(v)(s)$ and the trajectory $x_h(s) = X_s^{t_0,x_0,\eta_h,\beta} = X_s^{t_0,x_0,\alpha_h,v}$. Then

$$x_h(t_0 + h) = x_0 + \int_{t_0}^{t_0+h} f(x_h(s), \alpha_h(s), v) ds = x_0 + \int_{t_0}^{t_0+h} f(x_0, \alpha_h(s), v) ds + h\epsilon(h), \tag{2.48}$$

where $\epsilon(h) \to 0$ as $h \to 0^+$. For Eq. (2.46), let $\epsilon \to 0^+$ and $h \to 0^+$, we have

$$\rho(x_0) U^{-*}(t_0, x_0, \hat{p}/\rho(x_0)) - \phi(t_0, x_0) \geq 0. \tag{2.49}$$

For Eq. (2.47), we have

$$\begin{aligned} 0 &\leq V^{-*} \left(t_0 + h, X_{t_0+h}^{t_0,x_0,\eta_h,\zeta}, \hat{p} \right) - V^{-*}(t_0, x_0, \hat{p}) + \epsilon h \\ &\leq \phi \left(t_0 + h, x_0 + \int_{t_0}^{t_0+h} f(x_0, \alpha_h(s), v) ds + h\epsilon(h), z \right) - \phi(t_0, x_0) + \epsilon h \\ &\leq h\phi_t(t_0, x_0) + \int_{t_0}^{t_0+h} D\phi(t_0, x_0)^T f(x_0, \alpha_h(s), v) ds + h\epsilon_1(h) + \epsilon h \\ &\leq h\phi_t(t_0, x_0) + h \sup_{u \in \mathcal{U}} D\phi(t_0, x_0)^T f(x_0, u, v) + h\epsilon_1(h) + \epsilon h, \end{aligned} \tag{2.50}$$

where $\epsilon_1(h) \to 0$ as $h \to 0^+$. Dividing the last inequality by h, letting $h \to 0^+$, $\epsilon \to 0^+$, and taking the infimum over $v \in \mathcal{V}$ to have

$$\phi_t(t_0, x_0) + \inf_{v \in \mathcal{V}} \sup_{u \in \mathcal{U}} D\phi(t_0, x_0)^T f(x_0, u, v) \geq 0. \tag{2.51}$$

Now notice that by definition

$$H^*(x, D\phi) = -H(x, -D\phi) = \inf_{v \in \mathcal{V}} \sup_{u \in \mathcal{U}} f(x, u, v)^T D\phi. \tag{2.52}$$

Hence

$$\phi_t(t_0, x_0) + H^*(x_0, D\phi(t_0, x_0)) \geq 0. \tag{2.53}$$

Hence

$$\min\left\{\rho(x)U^{-*}(t, x, \hat{p}/\rho(x)) - \phi,\ \phi_t + H^*(x, D\phi)\right\} \geq 0 \text{ in } [0, T] \times \mathbb{R}^{d_x}. \tag{2.54}$$

□

Using the same proof techniques we can derive the subdynamic principle for V^+ (Lemma 2.6) and prove that V^+ is a viscosity subsolution to the primal HJ (Lemma 2.7):

Lemma 2.6. *(subdynamic principle for V^+). We have for any $(t_0, x_0, p) \in [0, T) \times \mathbb{R}^{d_x} \times \Delta(I)$ and any $t_1 \in (t_0, T]$*

$$\begin{aligned} V^+(t_0, x_0, p) \leq \inf_{\eta \in \mathcal{H}(t_0)} \sup_{\zeta \in \mathcal{Z}(t_0)} \max\Big\{ & \rho(\mathcal{X}_{t_1}^{t_0, x_0, \eta, \zeta}) U^+\left(t_1, X_{t_1}^{t_0, x_0, \eta, \zeta}, p\right), \\ & V^+\left(t_1, X_{t_1}^{t_0, x_0, \eta, \zeta}, p\right)\Big\} \end{aligned} \tag{2.55}$$

Lemma 2.7. *(V^+ is a subsolution of HJ). For any $p \in \Delta(I)$, the map $(t, x) \to V^+(t, x, p)$ is a viscosity subsolution of the primal Hamilton-Jacobi equation:*

$$\min\left\{\rho(x)U^+(t, x, p) - w,\ w_t + H(x, Dw)\right\} = 0 \text{ in } [0, T] \times \mathbb{R}^{d_x}, \tag{2.56}$$

where H is defined by Eq. (2.4).

2.7.2 Proof of Theorem 2.2

Proof. The following proof follows that of Cardaliaguet (2009), with the additional treatment of the state constraint. Note that in numerical approximation, we use $K > 0$ to replace infinite values so that we can introduce bounded test functions. The HJ equations thus become

$$\begin{cases} w_t + H(x, Dw) = 0 & (t,x) \in \Omega \\ \min\{K - w, w_t + H(x, Dw)\} = 0 & (t,x) \in \bar{\Omega}, \end{cases}$$

where Ω (resp. $\bar{\Omega}$) contains all (t,x) such that $V(t,x) < K$ (resp. $V(t,x) = K$).

Consider w be any cluster point in the topology of uniform convergence on compact subsets of $[0,T] \times \mathbb{R}^{d_x} \times \Delta(I)$ of V_τ as $\tau \to 0^+$. w is convex with respect to p and satisfies:

$$w(T,x,p) = \sum_{i=1}^{I} p_i g_i(x), \tag{2.57}$$

for any $(T,x,p) \in \Omega \times \Delta(I)$ and $w(T,x,p) = K$ for any $(T,x,p) \in \bar{\Omega} \times \Delta(I)$. Let ϕ be a test function such that $w(\cdot,\cdot,p) - \phi$ has a strict local maximum at (t_0, x_0), and $w(t_0, x_0, p) = \phi(t_0, x_0)$. Then there are (t_k, x_k) converging to (t_0, x_0) such that $V_\tau(\cdot,\cdot,p) - \phi$ has a local maximum at (t_k, x_k).

First consider $(t_k, x_k) \in \Omega$. For any $x \in \mathbb{R}^{d_x}$,

$$V_\tau(t_{k+1}, x, p) - \phi(t_{k+1}, x) \le V_\tau(t_k, x_k, p) - \phi(t_k, x_k)$$

Then, rearranging (2.10) to have

$$\begin{aligned} 0 &= \mathrm{Vex}_p \left(\min_u \max_v V_\tau(t_{k+1}, x_k + \tau f(x_k, u, v), p) \right) - V_\tau(t_k, x_k, p) \\ &\le \min_u \max_v V_\tau(t_{k+1}, x_k + \tau f(x_k, u, v), p) - V_\tau(t_k, x_k, p) \\ &\le \min_u \max_v \phi(t_{k+1}, x_k + \tau f(x_k, u, v)) - \phi(t_k, x_k) \end{aligned}$$

Then, from standard arguments (see Cardaliaguet (2009) and references therein)

$$\frac{\partial \phi}{\partial t}(t_0, x_0) + \min_{u \in U} \max_{v \in V} f(x_0, u, v) \frac{\partial \phi}{\partial x}(t_0, x_0) \geq 0. \tag{2.58}$$

Now consider $(t_k, x_k) \in \bar{\Omega}$, in which case $\phi(t_k, x_k) = V_\tau(t_k, x_k, p) = K$. When $\tau \to 0^+$,

$$\min_u \max_v \phi(t_{k+1}, x_k + \tau f(x_k, u, v)) = K,$$

hence

$$\frac{\partial \phi}{\partial t}(t_0, x_0) + \min_{u \in U} \max_{v \in V} f(x_0, u, v) \frac{\partial \phi}{\partial x}(t_0, x_0) = 0.$$

Then we have

$$\min \{K - \phi,\ \phi_t + H(x, D\phi)\} = 0 \text{ in } \bar{\Omega}.$$

Hence, w is a dual subsolution of the HJI. We can follow the same technique to show that w is a supersolution in the dual sense, and therefore $w = V$. □

2.7.3 Proof of Proposition 2.1

Proof. Let $f^0 : [0,1]^{I-1} \to \mathbb{R}$ be a bounded and Lipschitz continuous function, $\mathcal{P} \subset [0,1]^{I-1}$ be a lattice, and f be a convex hull computed from the data $\{f(p), p\}_{p \in \mathcal{S}}$. Let the true convex hull of f^0 be $Vex(f^0)$: $Vex(f^0)(p) \leq f(p)$ for all $p \in [0,1]^{I-1}$, with equality reached at least for $p \in \mathcal{S}$.

Introduce a set $P^0 = \{p^{(i)} \in \mathcal{S}\}_{i=1}^I$ and a space $\mathcal{P}^0 = \{p \in [0,1]^{I-1} \mid \exists \lambda \in \Delta(I) \text{ s.t. } p = \sum_{i=1}^I \lambda[i] p^{(i)},\ p^{(i)} \in P^0\}$ so that $f(p) = \sum_{i=1}^I \lambda[i] f(p^{(i)})$ for all $p \in \mathcal{P}^0$, i.e., P^0 are vertices of a segment $\mathcal{P}^0$ of $[0,1]^{I-1}$ within which f is affine.

Let $U := \{u^{(i)}\}_{i=1}^N = \mathcal{P} \bigcap \mathcal{P}^0$ be the set of lattice nodes contained in $\mathcal{P}^0$. Since f is a convex hull of f^0, we have $f(u^{(i)}) \leq f^0(u^{(i)})$ for all $i = 1, ..., N$. U defines a segmentation $\mathcal{E}$ of $\mathcal{P}^0$: Each $e \in \mathcal{E}$ is associated with $U_e := \{u^{(e_i)}\}_{i=1}^I \subset U$ such that

$e = \{p \in [0,1]^{I-1} \mid \exists \lambda \in \Delta(I) \text{ s.t. } p = \sum_{i=1}^{I} \lambda[i] u^{(e_i)},\ u^{(e_i)} \in U_e\}$ and $u \notin e$ for any $u \in U \setminus U_e$.

For any $p \in e$, we have the following loose lower bound on $f^0(p)$:

$$f^0(p) \geq \min_i f^0(u^{e_i}) - \Delta L \geq \min_i f(u^{e_i}) - \Delta_e L, \tag{2.59}$$

where $\Delta := \max_{i,j} \|u^{e_i} - u^{e_j}\|_2$, and L is the Lipschitz constant of f^0. Δ is a constant for a given lattice $\mathcal{P}$.

Therefore within e, the convexification error is lower bounded by

$$\max_{\lambda \in \Delta(I)} \left\{ f(\sum_i \lambda[i] u^{(e_i)}) - \min_i f(u^{e_i}) + \Delta L \right\} = \max_i f(u^{e_i}) - \min_i f(u^{e_i}) + \Delta L \leq 2\Delta L. \tag{2.60}$$

Since this error is constant, and $f - 2\Delta L$ is a convex lower bound of $Vex(f^0)$, we have $\varepsilon_{vex} \leq 2\Delta L$. □

2.7.4 Bellman Backup of the Conjugate Value with the Presence of Instantaneous Loss

Here we extend the subdynamic principle for V^{-*} (Lemma 2.5) when instantaneous loss is present. Since we will use letter l to index possible states reached at t_1 from t_0, we denote the instantaneous loss for pure strategies (η, ζ) at time s as $L(\eta, \zeta, s)$ instead. For conciseness, let us consider $t_0 \in [0, T]$ and $(x_0, \hat{p}) \in \mathcal{Q}(t_0) \times \mathbb{R}^I$, i.e., states for which Player 1 can play to avoid state constraint violation. To recap, let η be a pure strategy of Player 1; let ζ be such that $\zeta = \zeta^0$ in $[t_0, t_1]$ where ζ^0 is pure and ϵ-optimal for $V_1^{-*}(t_0, t_1, x_0, \hat{p}) :=:= \inf_{\zeta \in \mathcal{Z}(t_0)} \sup_{\eta \in \mathcal{H}(t_0)} V^{-*}\left(t_1, X_{t_1}^{t_0, x_0, \eta, \zeta}, \hat{p} - \int_{t_0}^{t_1} L(\eta, \zeta, s) ds\right)$, and $\zeta = \zeta^l$ in $[t_1, T]$ where ζ^l is mixed and (2ϵ)-optimal for V^{-*} at $(t_1, x_1, \hat{p})$ for any $x_1 \in E_l$.

By definition and using the fact that Player 1 plays a pure strategy in the dual game, we have

$$V^{-*}(t_0,x_0,\hat{p}) = \inf_{\zeta}\sup_{\eta}\max_{i}\left\{\hat{p}_i - \int_{\omega}\left(g_i(X_T^{t_0,x_0,\eta,\zeta(\omega,\cdot)}) + \int_{t_0}^{T} L(\eta,\zeta(\omega,\cdot),s)ds\right)d\mathbf{P}(\omega)\right\}. \tag{2.61}$$

Here

$$\begin{aligned}
&\max_i\left\{\hat{p}_i - \int_{\omega}\left(g_i(X_T^{t_0,x_0,\eta,\zeta(\omega,\cdot)}) + \int_{t_0}^{T} L(\eta,\zeta(\omega,\cdot),s)ds\right)d\mathbf{P}(\omega)\right\}\\
=&\max_i\left\{\hat{p}_i - \sum_l\left(\int_{\omega^l}\left(g_i(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta,\zeta^l(\omega,\cdot)}) + \int_{t_0}^{t_1} l(\eta(s),\zeta^0(\omega,\cdot)(s))ds + \right.\right.\right.\\
&\left.\left.\left.\int_{t_1}^{T} L(\eta,\zeta^l(\omega,\cdot),s)ds\right)d\mathbf{P}^l(\omega^l)\right)\mathbf{1}_{O^l}\right\}\\
\le&\sum_l\sup_{\eta'}\max_i\left\{\hat{p}_i - \left(\int_{\omega^l} g_i(X_T^{t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\eta',\zeta^l(\omega^l,\cdot)}) + \int_{t_0}^{t_1} L(\eta,\zeta^0,s)ds + \right.\right.\\
&\left.\left.\int_{t_1}^{T} L(\eta',\zeta^l(\omega^l,\cdot),s)dsd\mathbf{P}^l(\omega^l)\right)\right\}\mathbf{1}_{O^l}\\
\le&\sum_l\left(V^{-*}\left(t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\hat{p} - \int_{t_0}^{t_1} L(\eta,\zeta^0,s)ds\right) + 2\epsilon\right)\mathbf{1}_{O^l}\\
&\text{(because } \zeta^l \text{ is } (2\epsilon)\text{-optimal for } V^{-*} \text{ at } (t_1,x_1,\hat{p}) \text{ for any } x_1\in E_l.)\\
=&V^{-*}\left(t_1,X_{t_1}^{t_0,x_0,\eta,\zeta^0},\hat{p} - \int_{t_0}^{t_1} L(\eta,\zeta^0,s)ds\right) + 2\epsilon\\
\le&V_1^{-*}(t_0,t_1,x_0,\hat{p}) + 3\epsilon\\
&\text{(because } \zeta^0 \text{ is } \epsilon\text{-optimal for } V_1^{-*}(t_0,t_1,x_0,\hat{p}).)
\end{aligned} \tag{2.62}$$

Since ϵ can be arbitrarily small, we have

$$V^{-*}(t_0,x_0,\hat{p}) \le \inf_{\zeta\in\mathcal{Z}(t_0)}\sup_{\eta\in\mathcal{H}(t_0)} V^{-*}(t_1,X_T^{t_0,x_0,\eta,\zeta},\hat{p} - \int_{t_0}^{t_1} L(\eta,\zeta,s)ds) \tag{2.63}$$

2.8 Conclusion and Future Work

We proved the existence of value for zero-sum differential games with state constraints and one-sided information and developed a backward induction scheme to approximate the value. Our method enables mechanistic synthesis of behavioral strategies and allows explanation of the resultant splitting of strategies and beliefs. Future work will investigate more efficient learning+search methods that take advantage of value convexity and alleviate error propagation.

Chapter 3

EQUILIBRIUM STRUCTURE EXPLOITATION FOR EFFICIENT GAME SOLVING

3.1 Introduction

The strength of game solvers has grown rapidly in the last decade, beating elite-level human players in Chess (Silver *et al.*, 2017a), Go (Silver *et al.*, 2017b), Poker (Brown and Sandholm, 2019; Brown *et al.*, 2020b), Diplomacy (FAIR† *et al.*, 2022), Stratego (Perolat *et al.*, 2022b), among others with increasing complexity. However, most of the existing solvers with proven convergence, either based on regret matching (Tammelin, 2014; Burch *et al.*, 2014; Moravčík *et al.*, 2017; Brown *et al.*, 2020b; Lanctot *et al.*, 2009) or gradient descent-ascent (McMahan, 2011; Perolat *et al.*, 2021; Sokota *et al.*, 2022; Cen *et al.*, 2021; Vieillard *et al.*, 2020), have computational complexities increasing with the size of the *finite* action set, and suffer from game-tree complexities growing exponentially with both the action size U and the tree depth K. Real-world games, however, often have continuous actions and happen in continuous time and state spaces, making them *differential* in nature. Applying existing solvers to differential games would require game-specific insight or extra computational overhead

This chapter is based on the work described in the paper "Solving Football by Exploiting Equilibrium Structure of 2p0s Differential Games with One-Sided Information" published in the **Proceedings of the 14^{th} International Conference on Learning Representations**.

for automated abstraction (Kroer and Sandholm, 2015; Hawkin *et al.*, 2011; Brown and Sandholm, 2015).

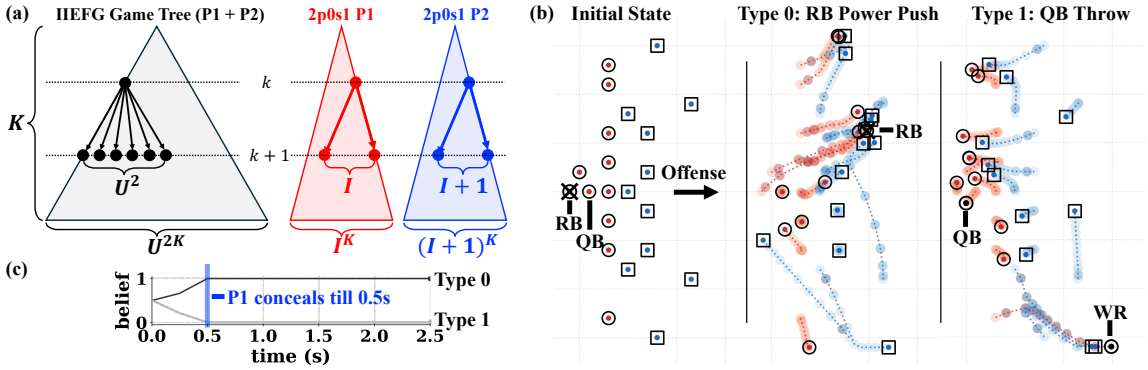


Figure 3.1. (a) IIEFG with U actions per player per infostate and K time steps has a game-tree complexity of U^{2K}. For 2p0s1 with I payoff types, deterministic dynamics, and Isaacs' condition, we show that the NE is I-atomic for P1 and $(I+1)$-atomic for P2, leading to a game-tree complexity of I^K for P1 in the primal game where P2 plays best responses and $(I+1)^K$ for P2 in the dual game where P1 plays best responses. (b) American Football with 22 players and continuous action spaces ($U = \infty$) with $K = 10$ time steps. P1 (red) attacks with two private game types ($I = 2$): Running back (RB) power-runs through the space created by blockers, and quarterback (QB) throws the ball to the leading wide receiver (WR). See animation. (c) At NE, P1 conceals type until 0.5 sec., similar to the reported 1.0 sec. Due to significant tree size reduction, the game can be solved in 30 minutes.

In this chapter, we address this scalability challenge for an K subset of 2p0s differential games where the informed player (P1) knows the payoff type while the uninformed player (P2) only holds a public belief $p_0 \in \Delta(I)$ over the finite set of I possible types. At the beginning of the game, nature draws a game type according to p_0 and informs P1 about the type. As the game progresses, the public belief p about the true game type is updated from p_0 based on the action sequence taken by P1 and its strategy via the Bayes' rule. P1's (resp. P2's) goal is to minimize (resp. maximize) the expected payoff over p_0. Due to the zero-sum nature, P1 may need to delay information release or manipulate P2's belief to take full advantage

of information asymmetry. While restricted, such games represent a wide range of attack-defense scenarios including football set-pieces where the attacker has private information about which play is to be executed, and missile defense where multiple potential targets are concerned. The setting of one-sided information, i.e., P1 knows everything about P2, is necessary for P2 to derive defense strategies in risk-sensitive games. We call this focused set of games "**2p0s1**".

We claim the following contributions:

- We prove two unique Nash equilibrium (NE) structures for 2p0s1: (1) The equilibrium behavioral strategy for P1 (resp. P2) is I-atomic (resp. $(I+1)$-atomic) on their continuous action space, and (2) the equilibrium strategies for P1 and P2 can be computed via separated primal and dual reformulations of the game. Together, these structures collapse the game-tree complexity to at most I^K for P1 and $(I+1)^K$ for P2. In comparison, solving the same game through the lens of IIEFG would have a game-tree complexity of U^{2K}, with a discretized action size of U (Fig. 3.1a).
- We demonstrate how this structural knowledge can significantly accelerate game solving: (1) For value and policy approximation settings where the ground-truth NE is available, we achieve qualitative improvements on solution accuracy and efficiency from SOTA normal- and behavioral-form solvers (CFR+, MMD, Deep-CFR, JPSPG, PPO, R-NaD). (2) We further demonstrate the practical value of the equilibrium structure of 2p0s1 by solving an American football setting where the attacking team needs to strategically conceal their true intention between "RB power push" and "QB throw". While this IIEFG has a complexity of 10^{440} even with a coarse mesh of 10 values per action dimension, the atomic structure

of its NE makes the game solvable in less than 30 minutes on a M1 Pro Macbook. See Fig. 3.1b,c.

3.2 Related Work

3.2.1 2p0s games with incomplete information.

Games where players have missing information only about the game types are called *incomplete-information*. These are a special case of imperfect-information games where nature only plays a chance move at the beginning Harsanyi (1967). The seminal work of Aumann *et al.* (1995) developed equilibrium strategies for a repeated game with one-sided incomplete information through the "Cav u" theorem, which reveals that belief-manipulating behavioral strategies optimize value through value convexification. Building on top of this, De Meyer (1996) introduced a dual game reformulation where the behavioral strategy of P2 becomes Markov. This technique helped Cardaliaguet (2007); Ghimire *et al.* (2024) to establish the value existence proof for 2p0s differential games with incomplete information. Unlike repeated games where belief manipulation occurs only in the first round of the game, differential games may have multiple critical time-state-belief points where belief manipulation is required to achieve equilibrium, depending on the specifications of system dynamics, payoffs, and state constraints (Ghimire *et al.*, 2024). Our paper builds on top of Cardaliaguet (2007); Ghimire *et al.* (2024), providing a thorough analysis of the convexification nature of behavioral strategies and, as a consequence, their atomic structure.

3.2.2 IIEFGs.

IIEFGs are partially-observable stochastic games (POSGs) with finite horizons. Significant efforts have been taken to approximate equilibrium of large IIEFGs with finite U (Koller and Megiddo, 1992; Billings *et al.*, 2003; Gilpin and Sandholm, 2006; Gilpin *et al.*, 2007; Sandholm, 2010; **?**; Brown *et al.*, 2020a; Perolat *et al.*, 2022b; Schmid *et al.*, 2023b). Regret matching algorithms (Zinkevich *et al.*, 2007; Lanctot *et al.*, 2009; Abernethy *et al.*, 2011; Brown *et al.*, 2019; Tammelin, 2014; Johanson *et al.*, 2012) have computational complexity $\mathcal{O}(U\varepsilon^{-2})$ to ε-Nash, while gradient-based solvers (McMahan, 2011; Perolat *et al.*, 2021; Sokota *et al.*, 2022) have $\mathcal{O}\left(\ln(U)\varepsilon^{-1}\ln\left(\varepsilon^{-1}\right)\right)$ to ε-QRE. All these complexities increase with U, and convergence guarantee only exists if the NE lies in the *interior* of the simplex $\Delta(U)$ (Perolat *et al.*, 2021). Critically, this latter assumption does not hold for 2p0s1, as we explain in Sec. 3.4. For games with continuous action spaces, existing pseudo-gradient based approaches (Martin and Sandholm, 2023, 2024) lack convergence guarantee, and perform poorly in our case studies in both normal- and extensive-form settings.

3.2.3 Multigrid for value approximation.

Since solving 2p0s1 essentially requires solving a Hamilton-Jacobi PDE, we briefly review multigrid methods for accelerating PDE solving(Trottenberg *et al.*, 2000). In a typical V-cycle solver (Braess and Hackbusch, 1983), a fine-mesh solve is first performed briefly, the residual is then restricted to a coarser mesh where a correction is solved and prolongated back to the fine mesh. Essentially, multigrid uses coarse solves to reduce the low-frequency approximation errors in the PDE solution at low costs,

leaving only high-frequency errors to be resolved through the fine mesh. Multigrid was successfully applied to solving Hamilton-Jacobi PDEs for optimal control and differential games (Han and Wan, 2013), although characterizing its computational complexity for such nonlinear PDEs is yet to be done. During value approximation, we extend multigrid to solving Hamilton-Jacobi PDEs underlying 2p0s1.

3.3 Problem Statement

We denote by $\Delta(I)$ the simplex in $\mathbb{R}^I$, $[I] := \{1, ..., I\}$, $a[i]$ the ith element of vector a, $\nabla_p V$ and $\partial_p V$ the respective gradient and subgradient of function V with respect to p. $\|\cdot\|_2$ is the l_2-norm and $\|v\|_A := \sqrt{v^T A v}$. Consider a time-invariant dynamical system that defines the evolution of the joint state $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ of P1 and P2 with control inputs $u \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$ and $v \in \mathcal{V} \subseteq \mathbb{R}^{d_v}$, respectively:

$$\dot{x}(t) = f(x(t), u, v). \tag{3.1}$$

The initial belief of players is set to nature's distribution about the game type. Denote by $\{\mathcal{H}_r^i\}^I$ the joint sets of I behavioral strategies of P1, and $\mathcal{Z}_r$ the set of behavioral strategies of P2. The subscript r indicates the random nature of behavioral strategies: At any infostate $(t, x, p) \in [0, T] \times \mathcal{X} \times \Delta(I)$ and with type i, P1 draws an action based on $\eta_i \in \mathcal{H}_r^i$, which is a probability measure over $\mathcal{U}$ parameterized by (t, x, p). P2's strategy $\zeta \in \mathcal{Z}_r$ is a probability measure over $\mathcal{V}$ independent of i. Denote by $X_{t_1}^{t_0,x_0,\eta_i,\zeta}$ the random state arrived at $t_1 \in (t_0, T]$ from (t_0, x_0) following (η_i, ζ) and the system dynamics in Eq. 3.1. With mild abuse of notation, let $(\eta_i(t), \zeta(t))$ denote the random controls at time t induced by (η_i, ζ). P1 accumulates a running cost $l_i(x(t), \eta_i(t), \zeta(t))$ during the game and receives a terminal cost $g_i(X_T^{t_0,x_0,\eta_i,\zeta})$. Together, a type-$i$ P1

minimizes:

$$J_i(t_0, x_0; \eta_i, \zeta) := \mathbb{E}_{\eta_i,\zeta}\left[g_i\left(X_T^{t_0,x_0,\eta_i,\zeta}\right) + \int_{t_0}^{T} l_i(x(s), \eta_i(s), \zeta(s))ds\right],$$

while P2 maximizes $J(t_0, x_0, p_0; \{\eta_i\}, \zeta) = \mathbb{E}_{i\sim p_0}[J_i]$. We say the game has a value V if and only if the upper value $V^+(t_0, x_0, p_0) = \inf_{\{\eta_i\}} \sup_\zeta J$ and the lower value $V^-(t_0, x_0, p_0) = \sup_\zeta \inf_{\{\eta_i\}} J$ are equal: $V = V^+ = V^-$. The game is proven to have a value under the following sufficient conditions (Cardaliaguet, 2007):

A1. $\mathcal{U} \subseteq \mathbb{R}^{d_u}$ and $\mathcal{V} \subseteq \mathbb{R}^{d_v}$ are compact and finite-dimensional.

A2. $f : \mathcal{X} \times \mathcal{U} \times \mathcal{V} \to \mathcal{X}$ is C^1 and has bounded value and first-order derivatives.

A3. $g_i : \mathcal{X} \to \mathbb{R}$ and $l_i : \mathcal{X} \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ are bounded and Lipschitz continuous.

A4. Isaacs' condition holds for the Hamiltonian $H : \mathcal{X} \times \mathbb{R}^{d_x} \times \Delta(I) \to \mathbb{R}$:

$$\begin{aligned} H(x, \xi, p) &:= \min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} f(x, u, v)^\top \xi + \mathbb{E}_{i\sim p}[l_i(x, u, v)] \\ &= \max_{v\in\mathcal{V}} \min_{u\in\mathcal{U}} f(x, u, v)^\top \xi + \mathbb{E}_{i\sim p}[l_i(x, u, v)]. \end{aligned} \tag{3.2}$$

Isaacs' condition allows any complete-information versions of this game to have *pure* Nash equilibria, including *nonrevealing* games where neither player knows the actual game type.

A5. Both players have full knowledge about f, $\{g_i\}_{i=1}^I$, $\{l_i\}_{i=1}^I$, p_0. Control inputs and states are fully observable. Players have perfect recall.

Our goal is to compute a Nash equilibrium (NE) $(\{\eta_i^\dagger\}, \zeta^\dagger)$ that attains V, given the game $G = \{\mathcal{X}, (\mathcal{U}, \mathcal{V}), \{g_i\}, \{l_i\}, f, T\}$.

3.4 A Primal-Dual Reformulation of the Game

In this section, we introduce discrete-time primal and dual reformulations of G, denoted by G_τ and G_τ^*, respectively, for which dynamic programming principles (DP)

exist. We show that P1's equilibrium behavioral strategy $\{\eta^\dagger_{i,\tau}\}$ in G_τ is I-atomic, i.e., $\eta^\dagger_{i,\tau}$ concentrates on at most I actions in $\mathcal{U}$, and P2's strategy $\zeta^\dagger_\tau$ in G^*_τ is $(I+1)$-atomic. Then we show that $(\{\eta^\dagger_{i,\tau}\}, \zeta^\dagger_\tau)$ approaches the Nash equilibrium of the differential game G as the time interval $\tau \to 0^+$.

3.4.1 The primal game G_τ.

G_τ is a discrete-time Stackelberg version of G where P2 plays a pure best response *after* P1 announces its next action. Let the value of G_τ be V_τ. V_τ satisfies the following DP for $(t, x, p) \in [0, T] \times \mathcal{X} \times \Delta(I)$:

$$V_\tau(t, x, p) = \min_{\{\eta_i\} \in \{\mathcal{H}_r\}^I} \mathbb{E}_{i \sim p, u \sim \eta_i} \Big[\max_{v \in \mathcal{V}} V_\tau(t + \tau, x'(u, v), p'(u)) + \tau l_i(x, u, v) \Big], \tag{3.3}$$

with a terminal boundary $V_\tau(T, x, p) = \sum_i p[i] g_i(x)$. For small enough τ, $x'(u, v) = x + \tau f(x, u, v)$. $p'(u)$ is the Bayes update of the public belief after P1 announces u: $p'(u)[i] = \eta_i(u; t, x, p) p[i] / \bar{\eta}(u; t, x, p)$, where $\bar{\eta}(u; t, x, p) = \sum_{i \in [I]} \eta_i(u; t, x, p) p[i]$ is the marginal over $\mathcal{U}$ across types. Note that P2's equilibrium behavioral strategy cannot be derived from Eq. 3.3.

3.4.2 The dual game G^*_τ.

For P2's strategy, we need a separate DP where P2 announces its next action and P1 best responses to it. We do so by first introducing the convex conjugate V^* of the value:

$$\begin{aligned} V^*(t_0, x_0, \hat{p}_0) &:= \max_p p^T \hat{p}_0 - V(t_0, x_0, p) = \max_p p^T \hat{p}_0 - \sup_{\zeta \in \mathcal{Z}_r} \inf_{\{\eta_i\} \in \{\mathcal{H}_r\}^I} \mathbb{E}_{\eta_i, \zeta, i} \Big[J_i \Big] \\ &= \max_p \inf_{\zeta \in \mathcal{Z}_r} \sup_{\{\eta_i\} \in \{\mathcal{H}_r\}^I} p^T \hat{p}_0 - \mathbb{E}_{\eta_i, \zeta, i} \Big[J_i \Big] = \inf_{\zeta \in \mathcal{Z}_r} \sup_{\eta \in \mathcal{H}} \max_{i \in \{1, \dots, I\}} \Big\{ \hat{p}_0[i] - \mathbb{E}_\zeta \Big[J_i \Big] \Big\}. \end{aligned} \tag{3.4}$$

Eq. 3.4 describes a dual game G^* with complete information, where the strategy space of P1 becomes $\mathcal{H} \times [I]$, i.e., the game type is now chosen by P1 rather than the nature. We prove in App. C.3 that P2's equilibrium in the dual game is also an equilibrium in the primal game if $\hat{p}_0 \in \partial_p V(t_0, x_0, p_0)$, and such $\hat{p}_0[i]$ represents the loss of type-i P1 should it play best responses to P2's equilibrium strategy. Therefore $\hat{p}_0[i] - \mathbb{E}_\zeta[J_i]$ measures P2's risk, and P2's equilibrium strategy minimizes the worst-case risk across all game types. We now introduce a discrete-time version of the dual game G^*_τ where P1 plays a pure best response after P2 announces their action at each time step. Let the value of G^*_τ be V^*_τ, the dual DP is:

$$V^*_\tau(t, x, \hat{p}) = \min_{\zeta, \hat{p}'(\cdot)} \mathbb{E}_{v\sim\zeta}\left[\max_{u\in\mathcal{U}} V^*_\tau(t+\tau, x'(u,v), \hat{p}'(v) - \tau l(x,u,v))\right], \tag{3.5}$$

with a terminal boundary $V^*(T, x, \hat{p}) = \max_{i\in[I]}\{\hat{p}[i] - g_i(x)\}$. Here $\hat{p}'(\cdot) : \mathcal{V} \to \mathbb{R}^I$ is constrained by $\mathbb{E}_{v\sim\zeta}[\hat{p}'(v)] = \hat{p}$ (similar to the martingale nature of p in G_τ), and $l(u,v)[i] = l_i(u,v)$.

3.4.3 Equilibrium strategies of G_τ and G^*_τ are atomic.

Our first theoretical result is Thm. 3.1, which states that P1's strategy that solves G_τ is I-atomic, and P2's strategy that solves G^*_τ is $(I+1)$-atomic:

Theorem 3.1. *The RHS of Eq. 3.3 can be reformulated as*

$$\begin{aligned}
&\min_{\{u^k\},\{\alpha_i^k\}} \max_{\{v^k\}} \quad \sum_{k=1}^{I} \lambda^k \Big(V(t+\tau, x+\tau f(x,u^k,v^k), p^k) + \tau \mathbb{E}_{i\sim p^k}[l_i(x,u^k,v^k)] \Big) \\
&s.t. \quad u^k \in \mathcal{U},\ v^k \in \mathcal{V}, \alpha_i^k \in [0,1],\ \sum_{k=1}^{I} \alpha_i^k = 1,\ \lambda^k = \sum_{i=1}^{I} \alpha_i^k p[i], \\
&p^k[i] = \frac{\alpha_i^k p[i]}{\lambda^k}, \quad \forall i,k \in [I],
\end{aligned} \tag{P$_1$}$$

i.e., $\eta^\dagger_{i,\tau}$ concentrates on actions $\{u^k\}_{k=1}^I$ for $i \in [I]$. The RHS of Eq. 3.5 can be reformulated as

$$\begin{aligned} &\min_{\{v^k\},\{\lambda^k\},\{\hat{p}^k\}} \max_{\{u^k\}} \sum_{k=1}^{I+1} \lambda^k \left(V^*(t+\tau, x+\tau f(x,u^k,v^k), \hat{p}^k - \tau l(x,u^k,v^k))\right) \\ &s.t. \quad u^k \in \mathcal{U},\ v^k \in \mathcal{V},\ \lambda^k \in [0,1],\ \sum_{k=1}^{I+1} \lambda^k \hat{p}^k = \hat{p},\ \sum_{k=1}^{I+1} \lambda^k = 1, \quad k \in [I+1]. \end{aligned} \tag{P_2}$$

Proof sketch. (1) Using Isaacs' condition, we show that P2's best response in Eq. 3.3 is implicitly governed by P1's action u, and u is in turn governed by the posterior belief p'. (2) With this insight, we can rewrite Eq. 3.3 as $V_\tau(t,x,p) = \min_\nu \int_{\Delta(I)} \tilde{V}_\tau(t,x,p')\nu(dp')$, where we control a pushforward density $\nu(dp')$ for the posterior belief to be p'. ν is subjected to $\int_{\Delta(I)} p'\nu(dp') = p$. For each p', $(u(p'), v(p'))$ is found by solving $\tilde{V}(t,x,p') = \min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} V_\tau(t+\tau, x', p') + \mathbb{E}_{i\sim p'} l_i$. Here $\tilde{V}(t,x,p')$ is P1's loss should both players play pure strategies at (t,x,p') within $[t, t+\tau)$, and its existence is guaranteed by Isaacs' condition. Since playing pure strategies does not change the public belief, we call $\tilde{V}_\tau$ the *non-revealing* value. (3) We can then show that $\min_\nu \int_{\Delta(I)} \tilde{V}_\tau(p')\nu(dp')$ is a convexification of $\tilde{V}_\tau$ in $\Delta(I)$. Since convexification requires at most I vertices in $\Delta(I)$, ν is at most I-atomic. Since ν determines $\eta^\dagger_{i,\tau}$, the latter is also I-atomic. (4) A similar argument can be made for G^*_τ, in which case the convexification is with respect to the dual variable $\hat{p}$. Since $\hat{p}$ is defined on $\mathbb{R}^I$ rather than constrained on $\Delta(I)$, $\zeta^\dagger$ is at most $(I+1)$-atomic. Proof in App. 3.7.1.

An example. To support the proof sketch, we use Fig. 3.2 to illustrate why an equilibrium behavioral strategy is atomic. Here $I = 2$, thus the value is defined on $\Delta(2)$ for fixed (t,x). To solve $\min_\nu \int_{\Delta(2)} \tilde{V}_\tau(p')\nu(dp')$, we scan the non-revealing value $\tilde{V}$ across $\Delta(2)$. One notices that if $\tilde{V}$ is not convex in p, it is always possible for P1 to achieve a lower loss by convexifying $\tilde{V}$ through a mixed strategy, leading to P_1.

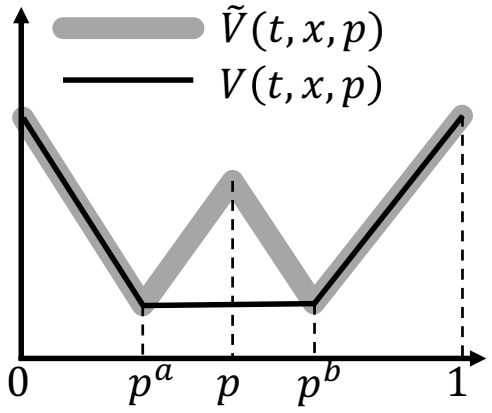


Figure 3.2. Value convexification causes NE to be atomic.

In the figure, P1 identifies $[\lambda^a, \lambda^b]^\top \in \Delta(2)$ and $\{p^a, p^b\}$ such that $\lambda^a p^a + \lambda^b p^b = p$. Solving the non-revealing pure NE (u^k, v^k) for each $k \in \{a, b\}$, P1's mixed strategy that convexifies $\tilde{V}$ is to play action u^k with probability $\alpha_i^k = p^k[i]\lambda^k / p[i]$ if P1 is type-i. By announcing this strategy, the public belief shifts to p^k via the Bayes' rule when P2 observes action u^k. As a result, P1 receives $V(p) = \lambda^a \tilde{V}(p^a) + \lambda^b \tilde{V}(p^b)$ on the convex hull of $\tilde{V}(p)$.

Remarks. (1) The atomic structure of the equilibrium strategies was first discovered for 2p0s repeated games with one-sided information (Aumann *et al.*, 1995; De Meyer, 1996). Our new contribution is in explaining its presence in differential games and in demonstrating its significant utilities in improving the effectiveness of multiagent reinforcement learning and model predictive control schemes. (2) P1's actions in 2p0s1 simultaneously advance system states and achieve signaling. The deterministic dynamics allows precise belief control (e.g., (p^a, p^b) in Fig. 3.2) for value convexification. For games with stochastic dynamics/observations, however, P1 will not be able to precisely control the belief, and the convex Bellman operators (P_1 and P_2) become lower bounds of the value. Finding tight upper-bounding operators that preserve atomic NEs is left for future work.

$(\{\eta^\dagger_{i,\tau}\}, \zeta^\dagger_\tau)$ **approaches NE of** G**.** Next we present Thm. 3.2, which proves that $(\{\eta^\dagger_{i,\tau}\}, \zeta^\dagger_\tau)$ computed from P_1 and P_2 approaches the equilibrium of G when τ is sufficiently small. The theorem uses P1's loss in G when they adapt $\{\eta^\dagger_{i,\tau}\}$ to G by playing a constant action within each time interval. This loss is denoted by $V_1(t, x, \hat{p}) := \max_{\zeta \in \mathcal{Z}} J(t, x, p; \{\eta^\dagger_{i,\tau}\}, \zeta)$. Similarly, P2's loss in G^* when they adapt $\zeta^\dagger_\tau$ to G is denoted by $V_1^*(t, x, p) := \max_{\{\eta_i\} \in \{\mathcal{H}^i\}^I} J^*(t, x, \hat{p}; \{\eta_i\}, \zeta^\dagger_\tau)$.

Theorem 3.2. *If A1-5 hold, then for any* $\epsilon > 0$*, there exists* $\tau^* > 0$*, such that for any* $\tau \in (0, \tau^*]$ *and the corresponding* $(\{\eta^\dagger_{i,\tau}\}, \zeta^\dagger_\tau)$*,* $V_1(t, x, p) - V(t, x, p) \in [0, \epsilon]$ *for any* $(t, x, p) \in [0, T] \times \mathcal{X} \times \Delta(I)$*, and* $V_1^*(t, x, \hat{p}) - V^*(t, x, \hat{p}) \in [0, \epsilon]$ *for any* $(t, x, \hat{p}) \in [0, T] \times \mathcal{X} \times \mathbb{R}^I$*.*

Proof sketch. We first note that $V \leq V_1$ because P1's NE in G_τ is not necessarily a NE in G, i.e., P1 can get a better value in G using $\{\eta^\dagger_i\}$ than $\{\eta^\dagger_{i,\tau}\}$. We also have, from Cardaliaguet (2009), the uniform convergence of $V_\tau \to V$ as $\tau \to 0^+$. Then, it remains to show that V_1 is "not far" from V_τ. To do so, we first show that the per-stage value gap between V_1 and V_τ is within $\mathcal{O}(\tau^2)$, which accumulates over time into $\mathcal{O}(\tau)$: $V_1 \leq V_\tau + \mathcal{O}(\tau)$. Then as $\tau \to 0^+$, combined with the uniform convergence of V_τ, we conclude that the gap $V_1 - V \in [0, \epsilon]$. This implies that the equilibrium $\{\eta^\dagger_{i,\tau}\}$ derived in G_τ is ϵ-optimal in the original game G as long as τ is small enough. Full proof in Sec. 3.7.2. The same techniques can be applied to G^*_τ.

With Thms. 3.1 and 3.2, and given that G is differential ($\tau \to 0^+$), we now proved that NEs of G are atomic, and can be approximated via Eqs. P_1 and P_2 with sufficiently small τ.

3.5 Incorporating the Atomic Structure into MARL and Control

We study the efficacy of atomic NEs when applied to (1) value approximation, (2) model-free MARL, and (3) control settings. For (1), we approximate value and strategies in the entire $[0, T] \times \mathcal{X} \times \Delta(I)$ through P_1 and P_2, and introduce multigrid to further reduce the compute. We compare with discrete-action solvers CFR+, MMD, CFR-BR, and continuous-action solver JPSPG. For (2) and (3), we solve NE strategies for a fixed initial (t_0, x_0, p_0). We compare with discrete-action solvers MMD, PPO, and R-NaD. We call the proposed continuous-action mixed-strategy solvers CAMS, CAMS-DRL, and CAMS-MPC for value approximation, MARL, and control, respectively.

3.5.1 Value approximation

3.5.2 CAMS for G_τ.

We discretize time as $\{k\tau\}_{k=0}^{K}$, $\tau = T/K$. Let $\mathcal{S} = \{(x, p)_i\}_{i \in [|\mathcal{S}|]}$ be a collocation set. We solve P_1 starting from $t = (K-1)\tau$ at all collocation points in $\mathcal{S}$. The resultant nonconvex-nonconcave minimax problems have size $(\mathcal{O}(I(I + d_u)), \mathcal{O}(Id_v))$ and are solved by DS-GDA (Zheng *et al.*, 2023), which guarantees sublinear convergence on nonconvex-nonconcave problems. To generalize value and strategies across $\mathcal{X} \times \Delta(I)$, a value network is trained on the minimax solutions and used to formulate the next round of minimax at $t - \tau$. G_τ^* is solved similarly.

3.5.3 Computational challenge.

From Thm. 3.2, large K (small τ) is required for strategies derived from P_1 and P_2 to be good approximations of the NE. Yet suppressing the value prediction error at $t = 0$ requires a computational complexity exponential in K. Specifically, let $\hat{V}_0(x, p) : \mathcal{X} \times \Delta(I) \to \mathbb{R}$ be the trained value networks at $t = 0$, we have the following result (see proof in App. B):

Theorem 3.3. *Given K, a minimax approximation error $\epsilon > 0$, a prediction error threshold $\delta > 0$, there exists $C \geq 1$, such that with a computational complexity of $\mathcal{O}(K^3 C^{2K} I^2 \epsilon^{-4} \delta^{-2})$, CAMS achieves*

$$\max_{(x,p)\in\mathcal{X}\times\Delta(I)} |\hat{V}_0(x, p) - V(0, x, p)| \leq \delta. \tag{3.6}$$

A similar result applies to the dual game. Zanette *et al.* (2019) discussed a linear value approximator that achieves $C = 1$. However, their method requires solving a linear program (LP) for every inference $\hat{V}_t(x, p)$ if (x, p) does not belong to the training set $\mathcal{S}$. In our context, incorporating their method would require auto-differentiating through the LP solver for each descent and ascent steps in minimax, which turned out to be too expensive. To this end, we introduce a multigrid scheme to reduce the cost for games with a large K.

3.5.4 Multigrid

Since strategies at time t are implicitly nonlinear functions of the value at $t + \tau$, the Hamilton-Jacobi PDEs underlying P_1 and P_2 are nonlinear. Our method extends the Full Approximation Scheme (FAS) for solving nonlinear PDEs (Trottenberg *et al.*, 2000; Henson *et al.*, 2003).

A two-grid FAS has four steps (see illustration in Fig. 3.3): (1) Restrict the fine-grid approximation and its residual; (2) solve the coarse-grid problem using the fine-grid residual; (3) compute the coarse-grid correction; (4) prolong the coarse-grid correction to fine-grid and add the correction to fine-grid approximation. For conciseness, we focus on the primal problem. Let $\hat{V}_t^l$ be the value network for time t on grid size (time interval) l. Let $\mathcal{R}^l$ be the restriction operator to a coarser grid with size $2l$: $\mathcal{R}^l(\hat{V}_t^l) = (\hat{V}_t^l + \hat{V}_{t+l}^l)/2$ is the value restriction from l to $2l$. Similarly, we define the prolongation operators $\mathcal{P}^{2l}$ as $\mathcal{P}^{2l}(\hat{V}_t^{2l}) = \hat{V}_t^{2l}$ if $t \in \mathcal{T}^{2l}$ or $\hat{V}_{t+l}^{2l}$ otherwise, where $\mathcal{T}^{2l} := \{n \cdot 2l : n \in \mathbb{N}_0, n < T/2l\}$. Let $\mathbb{O}^l(t, x, p; \hat{V})$ solve P_1 at (t, x, p) where $\hat{V}$ is the value at $t + l$, and outputs an approximation for $V(t, x, p)$. The dataset $\{(t, x^{(j)}, p^{(j)}, \mathbb{O}^l(t, x^{(j)}, p^{(j)}; \hat{V}_{t+l}^l))\}$ is used to train $\hat{V}_t^l(\cdot, \cdot)$. Let $r_t^l(x, p) = \hat{V}_t^l(x, p) - \mathbb{O}^l(t, x, p; \hat{V}_{t+l}^l)$ be the residual. To achieve $r_t^l(x, p) \approx 0$ for all $(t, x, p) \in \mathcal{T}^l \times \mathcal{X} \times \Delta(I)$, we restrict the fine grid approximations and residuals to the coarse grid and solve to determine the corrections. To do so, let $e_t^l(x, p)$ be the correction in grid l at (t, x, p). The coarse-grid problem is:

$$\underbrace{\mathcal{R}^l r_t^l}_{\text{residual}} = \underbrace{\mathbb{O}^{2l}(t, x, p; \mathcal{R}^l \hat{V}_{t+2l}^l + e_{t+2l}^{2l}) - \left(\mathcal{R}^l \hat{V}_t^l + e_t^{2l}(x, p)\right)}_{\text{coarse-grid eq. w/ corrections}} - \underbrace{\left(\mathbb{O}^{2l}(\mathcal{R}^l \hat{V}_{t+2l}^l) - \mathcal{R}^l \hat{V}_t^l\right)}_{\text{coarse-grid eq. w/o corrections}}, \tag{3.7}$$

where $e_t^{2l}(x, p)$ is computed backward from $T - 2l$ using $e_T^{2l} = 0$:

$$e_t^{2l}(x, p) = \mathbb{O}^{2l}(t, x, p; \mathcal{R}^l \hat{V}_{t+2l}^l + e_{t+2l}^{2l}) - \mathbb{O}^{2l}(\mathcal{R}^l \hat{V}_{t+2l}^l) - \mathcal{R}^l r_t^l. \tag{3.8}$$

This correction ensures consistency: If $\hat{V}_t^l = V(t, \cdot, \cdot)$ for all $t \in \mathcal{T}^l$, $e_t^{2l}(\cdot, \cdot) = 0$ for all $t \in \mathcal{T}^{2l}$. The coarse grid corrections are prolonged to the fine grid to update the fine-grid value approximation. Note that from Eq. 3.8, computing the coarse correction in our case requires two separate minimax calls with similar loss formulations. We further accelerate the multigrid solver by warm-starting these minimax problems

using the recorded minimax solution derived from the fine grid (during the residual computation).

This can be further extended to n-level multigrid by recursively reducing the coarse-grid size until the desired coarsest grid is reached. Alg. 2 presents the n-level multigrid algorithm.

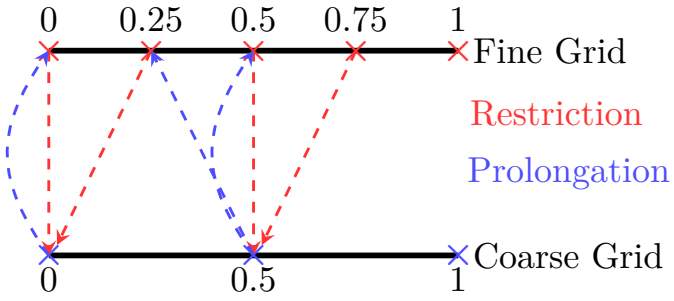


Figure 3.3. Illustration of 2-level multigrid method.

3.5.5 Model-free multiagent reinforcement learning

Recent studies (Rudolph *et al.*, 2025; Sokota *et al.*, 2022) showed that policy gradient methods for MARL, such as PPO and MMD, can effectively solve IIEFGs with properly tuned hyperparameters. We show that the atomic structure can be directly applied to this unified model-free framework with minimal code changes, while yielding significant solution improvement for 2p0s1. For CAMS-DRL, the policy network of P1 takes in the infostate (t, x, p) and outputs I logit vectors $\ell_k \in \mathbb{R}^I$ and I action prototypes $u^k \in \mathcal{U}$ for $k \in [I]$. Each logit vector ℓ_k is transformed via softmax to define the behavioral strategy of type-i P1, i.e., the probability $\eta_i(u^k; t, x, p)$ of choosing action u^k. The policy network for P2 takes in (t, x, p) and outputs a single logit $\ell \in \mathbb{R}^{I+1}$ and action prototypes μ^k for $k \in [I+1]$. We directly solve the NE of these policy models using PPO and MMD.

Algorithm 2 n-Level Multigrid for Value Approximation

Require: $k_{\max}, k_{\min}, \mathbb{O}$ (minimax solver), T (time horizon), N (number of data points), $\mathcal{R}$ (restriction operator), $\mathcal{P}$ (prolongation operator)

Initialize: $\mathcal{T}^l \leftarrow [0, l, 2l, \ldots, T-l],\ \forall\, l \in \{2^{-k_{\max}}, \ldots, 2^{-k_{\min}}\},\ L \leftarrow 2^{-k_{\min}}$

Initialize: Value networks $\hat{V}_t^l,\ \forall\, t \in \mathcal{T}^l,\ \forall\, l \in \{2^{-k_{\max}}, \ldots, 2^{-k_{\min}}\}$, policy set $\Pi \leftarrow \varnothing$

while resources not exhausted or until convergence **do**

  $R \leftarrow \varnothing,\ E^L \leftarrow \varnothing,\ \mathcal{S} \leftarrow \varnothing$

  Initialize coarsest-grid correction networks $\varepsilon_t^L, \forall\, t \in \mathcal{T}^L$

  $\mathcal{S}[t] \leftarrow$ sample N $(t, x, p),\ \forall\, t \in \mathcal{T}^{k_{\max}}$

  {down-cycle}

  for $k = k_{\max}$ **down to** $k_{\min} + 1$ **do**

    Compute target via $\mathbb{O}^k$ (init. with π_t if $\Pi[k] \neq \varnothing$), and store updated policies π_t in $\Pi[k],\ \forall\, t \in \mathcal{T}^k$

    Compute residuals $r^k[t],\ \forall\, t \in \mathcal{T}^k$

    if $k \neq k_{\max}$ **then**

      $r_t^k \leftarrow r_t^k + \mathcal{R} r_t^{k+1},\ \forall\, t \in \mathcal{T}^k$

    end if

    Store r_t^k in $R[k]$

  end for

  {coarse-solve backwards in time}

  for $t = T - L$ **to** 0 **do**

    $e_t^L \leftarrow \mathbb{O}^L(\mathcal{R}\hat{V}_{t+L}^l + \varepsilon_{t+L}^L) - \mathbb{O}^L(\mathcal{R}\hat{V}_{t+L}^l) - \mathcal{R}\, r_t^{k_{\min}+1}\ \{e_T^L = 0,\ \varepsilon_T^L = \varnothing\}$

    Store e_t^L in E^L

    Fit ε_L^t to e_L^t

  end for

  {up-cycle}

  for $k = k_{\min} + 1$ **to** $k_{\max}$ **do**

    $e_t^k \leftarrow \mathcal{P}(e_t^{k-1}),\ \forall\, t \in \mathcal{T}^k$

    Update $\hat{V}_t^k \leftarrow \hat{V}_t^k + e_t^k$

  end for

  {post smoothing (for all t's and l's)}

  $\texttt{target}, \pi_t \leftarrow \mathbb{O}^l(\hat{V}_{t+l}^l)$ (initialized with π_t)

  Fit $\hat{V}_t^l$ to $\texttt{target}$ and replace π_t in $\Pi[l]$

end while

3.5.6 Model predictive control

When dynamics f is known and differentiable, and an initial state (t_0, x_0, p_0) is given, we can formulate the primal (resp. dual) minimax objective as the sum over I^K (resp. $(I+1)^K$) game tree paths (Fig. 3.1a). CAMS-MPC builds the computational graph for the entire tree where the policy networks follow CAMS-DRL, and applies DS-GDA to this differentiable loss. This method is feasible thanks to the atomic structure of NEs and for small I. Since P1 in 2p0s1 games reveals their payoff mid-game, the game tree further collapses after information revelation. E.g., Hexner's game (see below) has a proven game-tree complexity of I, where P1 plays a fixed nonrevealing strategy before splitting to I type-dependent revealing strategies. Due to this collapse, modeling strategies using neural networks turns out to be more effective for the convergence to NEs than using infostate-wise parameterization.

3.6 Empirical Validation

Hexner's game. We introduce Hexner's game (Hexner, 1979; Ghimire *et al.*, 2024) to compare CAMS variants with baselines on solution quality and computational cost. This game has an analytical NE. The dynamics is $\dot{x}_j = A_j x_j + B_j u_j$ for $j = [2]$, where $x_j \in \mathcal{X}_j$, $u_j \in \mathcal{U}_j$, and A_j and B_j are known matrices. The target state of P1 is $z\theta$ where θ is drawn with distribution p_0 from Θ, $|\Theta| = I$, and $z \in \mathbb{R}^{d_x}$ is fixed and public. The expected payoff to P1 is:

$$J(\{\eta_i\}, \zeta) = \mathbb{E}_{i\sim p_0}\left[\int_0^T \|\eta_i(t)\|^2_{R_1} - \|\zeta(t)\|^2_{R_2} dt + \|x_1(T) - z\theta_i\|^2_{K_1} - \|x_2(T) - z\theta_i\|^2_{K_2}\right],$$

where $R_1, R_2, K_1, K_2 \succ 0$ are control- and state-penalty matrices. The goal of P1 is to get closer to the target $z\theta$ than P2. To take information advantage, P1 needs to decide when to home-in to and reveal the target.

Analytical NE:

There exists a critical time $t_r := t_r(T, \{A_j\}, \{B_j\}, \{R_j\}, \{K_j\})$. If $t_r \in (0, T)$, P1 moves towards $\mathbb{E}[\theta]$ as if it does not know the actual target until t_r when it fully reveals the target, i.e., value convexification happens at t_r. If $t_r \leq 0$, P1 homes towards the actual target at $t = 0$. P2's NE mirrors P1. See App. A.2.

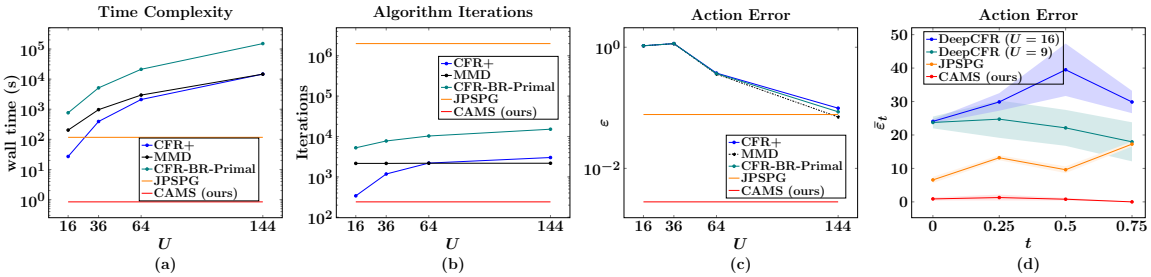


Figure 3.4. (a-c) Comparisons b/w CAMS, JPSPG, CFR+, MMD, CFR-BR-Primal on 1-stage Hexner's game. (d) Comparison b/w CAMS, JPSPG, and DeepCFR on 4-stage Hexner's w/ similar compute.

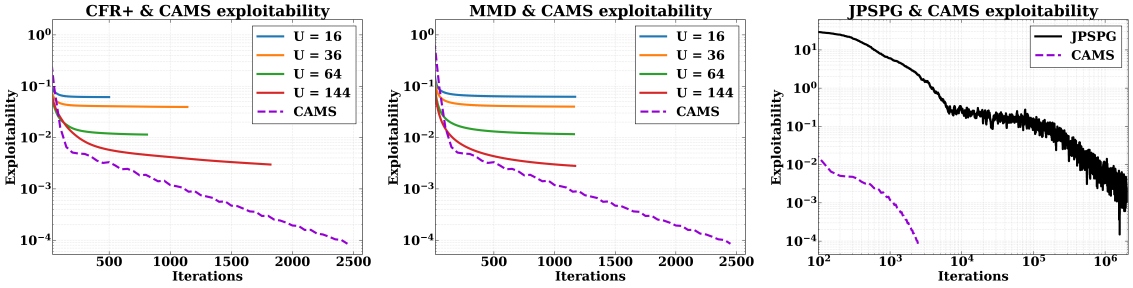


Figure 3.5. Exploitability comparison in the normal-form Hexner's game.

3.6.1 Effect of atomic NE on value approximation

Action-size-invariant convergence of CAMS. We use a normal-form Hexner's game with $\tau = T$ and a fixed initial state $x_0 \in \mathcal{X}$ to demonstrate that baseline algorithms suffer from increasing discrete action sizes while CAMS does not. The baselines are SOTA normal-form solvers including CFR+ (Tammelin, 2014), MMD (Sokota *et al.*, 2022), JPSPG (Martin and Sandholm, 2024), and a modified CFR-BR (Johanson *et al.*, 2012) (dubbed CFR-BR-Primal), where we focus on converging P1's strategy and only compute P2's best responses. Among these, only JPSPG naturally handles continuous action spaces. All baselines except JPSPG are standard implementations in OpenSpiel (Lanctot *et al.*, 2019). JPSPG implementation details are in App. C.5. The normal-form primal game has a trivial ground-truth strategy where P1 goes directly to its target. For visualization, we use $d_x = 4$ (position and velocity in 2D). For baselines except JPSPG, we use discrete action sets defined by 4 grid sizes so that $U = |\mathcal{U}_j| \in \{16, 36, 64, 144\}$. All algorithms terminate when a threshold of $NashConv(\pi_i) = \max_{\pi_i'} V_i(\pi_i') - V_i(\pi_i)$ is met. For conciseness, we only consider solving P1's strategy and thus use P1's NashConv. We set the threshold to 10^{-3} for baselines and use a more stringent threshold of 10^{-5} for CAMS. We then use DeepCFR and JPSPG as baselines for a 4-stage game where $T = 1$ and $\tau = 0.25$. DeepCFRs were run for 1000 CFR iterations (resp. 100) with 10 (resp. 5) traversals for $U = 9$ (resp. 16). The wall-time costs for game solving are 17 hours using CAMS (baseline), 24 hours for JPSPG, 29 hours ($U = 9$) and 34 hours ($U = 16$) using DeepCFR, all on an A100 GPU. More details on experiment settings can be found in App. C.4. Furthermore, JPSPG was run for $2 \cdot 10^8$ iterations, where each iteration consists

Table 3.1. Runtime comparison of CAMS with and without multigrid

# time steps	no multigrid	multigrid ↓
4	9.3 hrs	**2.3** hrs
10	27.6 hrs	**10.9** hrs
16	46.2 hrs	**17.8** hrs

of solving a game with a random initial state and type, and performing a strategy update.

Fig. 3.4 summarizes the comparisons with baselines. For the normal-form game, we compare both computational cost and the expected action error ε from the ground-truth action of P1: $\varepsilon(x_0) := \mathbb{E}_{i\sim p_0}\Big[\sum_{k=1}^{|\mathcal{U}|}\alpha_{ki}||u_k - u_i^*(x_0)||_2\Big]$, where $u_i^*(x_0)$ is the ground truth for type i at x_0. In the 4-stage game, we compare the expected action errors at each time-step: $\bar{\varepsilon}_t := \mathbb{E}_{x_t\sim\pi}\Big[\varepsilon(x_t)\Big]$, where π is the strategy learned by the respective learning method. For each strategy, we estimate $\{\bar{\varepsilon}_t\}_{t=1}^4$ by generating 100 trajectories with initial states uniformly sampled from $\mathcal{X}$. In terms of computational cost, all baselines (except JPSPG) have complexity and wall-time costs increasing with U, while CAMS is invariant to U. We also compare exploitability of CFR+, MMD, and JPSPG against CAMS for the normal-form Hexner's game in Fig. 3.5. For a fair comparison, best responses are computed in the continuous action spaces.

With similar or less compute, CAMS achieves significantly better strategies than DeepCFR and JPSPG in the 4-stage game. Sample trajectories for the 4-stage game are visualized in App. A.2.0.2. Fig. 3.6 compares the ground-truth vs. approximated NEs for 10-stage games with different initial states. While approximation errors exist, CAMS successfully learns the target-concealing behavior of P1. Averaging over 50 trajectories derived from CAMS, P1 conceals the target until $t_r = 0.60s \pm 0.06s$ (compared to the ground-truth $t_r = 0.5s$).

CAMS scalability with multigrid. We solve Hexner's games using CAMS with and without multigrid to demonstrate the scalability of our approach and the effect of multigrid. We report the runtime on one H100 GPU for 4-, 10-, and 16-stage games in Tab. 3.1. We run the 2-level multigrid on the 4- and 10-stage games, and 4-level multigrid on the 16-stage game (see Alg. 2 for pseudo code). We report resulting trajectories in App. 3.5.4.

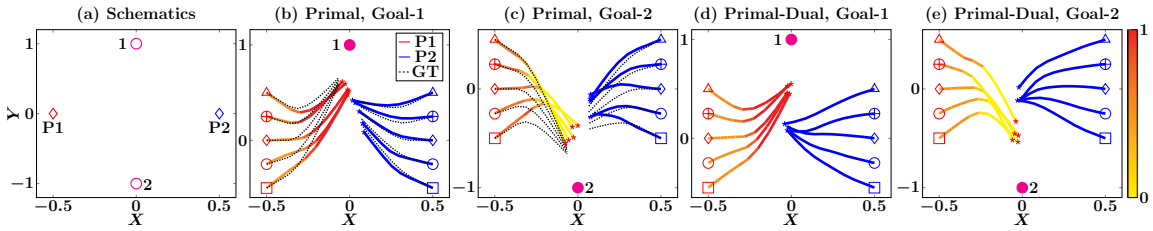


Figure 3.6. (a) Hexner's game schematics: one goal is selected out of two possible goals: Goal-1 and Goal-2, and communicated to P1. (b-e) Sample trajectories for the primal game (b-c) where P1 plays Nash and P2 plays best response, and primal-dual game (d-e) where both players play Nash. Dotted lines are ground-truth Nash. Color shades indicate evolution of public belief (Pr[Goal is 1]). Filled Magenta circle represents the true goal. Initial position pairs are marked with the same markers.

3.6.2 Effect of atomic NE on model-free MARL

For model-free MARL, we compare CAMS-DRL with MMD, PPO and R-NaD with discrete actions of size 100 formed by pairing each of the ten linearly spaced x-direction acceleration between $(-1, 1)$ and y-direction acceleration between $(-4, 4)$. We test the policy convergence in the normal-form game by comparing P1's learned policy every 8192 steps (corresponds to 1 iteration in Fig. 3.7(a)) against the ground truth policy. MMD, PPO, and R-NaD are implemented as in Rudolph *et al.* (2025). Results in Fig. 3.7 (a-b), which uses the same comparison metric as outlined in

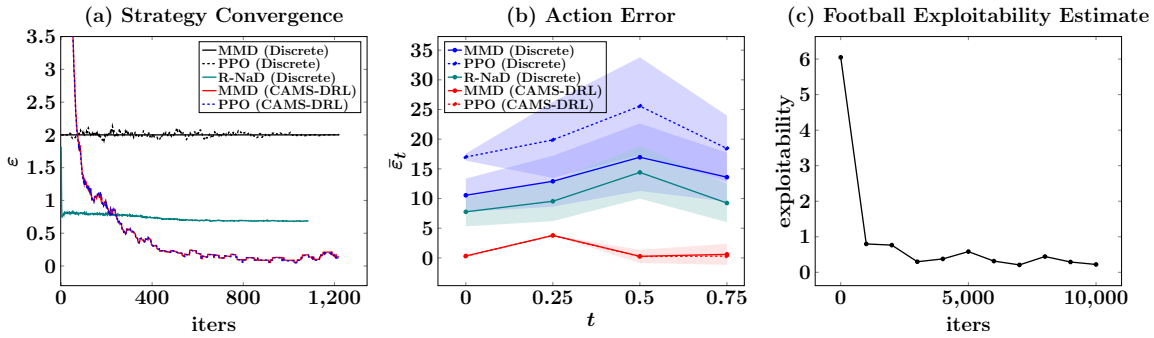


Figure 3.7. Comparisons b/w CAMS-DRL and standard PG methods on (a) 1-stage and (b) 4-stage games. (c) Exploitability estimate of the strategies in the Football game at different iterations.

Sec. 3.6.1, show that CAMS-DRL approximates NEs accurately, while PPO and MMD fail. R-NaD, on the other hand, shows better performance than MMD and PPO with little to no hyperparameter tuning (see App. C.6). This contradicts with observations in Rudolph *et al.* (2025). We conjecture that PG methods such as PPO and MMD have difficulty converging to non-interior solutions, which is the case in 2p0s1 due to the atomic structure of NEs. We performed the same comparison for a 4-stage game, with a discrete action size of 100 composed by pairing linearly spaced x- and y-direction accelerations, all between $(-12, 12)$. Results in Fig. 3.7b show significant improvement in solution accuracy when the atomic structure is exploited.

3.6.3 Effect of atomic NE on MPC

3.6.4 Hexner's game

With known differentiable dynamics and $I = 2$, a 10-stage Hexner's (primal) game has at most 2^{10} rollouts, making exact policy gradient computation feasible. Using

DS-GDA, the convergence to the ground-truth NE takes 5 minutes on a M1 Pro Macbook. See results in App. C.7.

3.6.5 Football game

The independence of the game-tree complexity from the action space allows us to solve games beyond what IIEFG solvers can afford. Here we model an 11-vs-11 American Football play as a short, two-team game in a 2D plane. The horizontal axis represents "downfield" progress and the vertical axis is sideline-to-sideline. Each player follows double-integrator kinematics and chooses acceleration at discrete time steps. Physical contact is captured by a smooth "merge" weight that grows as opponents approach, softly blending their velocities and attenuating their ability to apply new acceleration, so tackles emerge continuously rather than through hard impulses. The offense has two types ($I = 2$): an "RB power push" that prefers the RB to advance straight upfield, and a "QB throw" that rewards whichever offensive player gets furthest downfield. With loose bounds on velocity and acceleration, the soft tackle dynamics is control-affine, allowing Isaacs' condition to hold. Together with differentiable rollouts, the game settings satisfy A1-5, leading to atomic NE. With $K = 10$ and known differentiable dynamics, it is feasible to solve P1's (and then P2's) strategy by applying DS-GDA directly to the minimax objective constituting all I^K ($(I+1)^K$) rollouts. See App. C.8 for detailed game settings. The convergence takes 30 minutes on a M1 Pro Macbook. The resulting plays are summarized in Fig. 3.1b,c and animated in the github repo. Qualitatively, the results resemble real football tactics: the offense either tries to push through the defense (aka inside zone play), or goes out wide by faking a move (aka waggle play), while the defense, in response, either close-in on

the player with the ball possession or stay back to guard. Importantly, our results show that the offense conceal their play selection for 0.5 seconds, which is comparable to coaching analyses that estimate roughly a one-second window before the play becomes clear (Grabowski, 2020). Quantitatively, we estimate the exploitability of the strategies and plot its convergence in Fig. 3.7c. We also provide a comparison of resulting trajectories in Sec. 3.6.5.1 when each player switches to their respective BR policy to show that the final policies are minimally exploitable.

3.6.5.1 Making Sense of Exploitability

While exploitability provides a good indication of whether or not the learned policy is converging to the equilibrium, without a well known baseline outcome, it is often difficult to make sense of how drastic the true nash equilibrium is compared to the resulting policy with non-zero ϵ exploitability. Hence, here we compare the resulting trajectories when each player switch to their best-response policy to get a sense of how "exploitable" are the strategies, qualitatively. We plot the comparison in Fig. 3.8. Note that unlike in Fig. 3.1(b), in Fig. 3.8 P2 plays a dual game, which results in slight change in P1's policy.

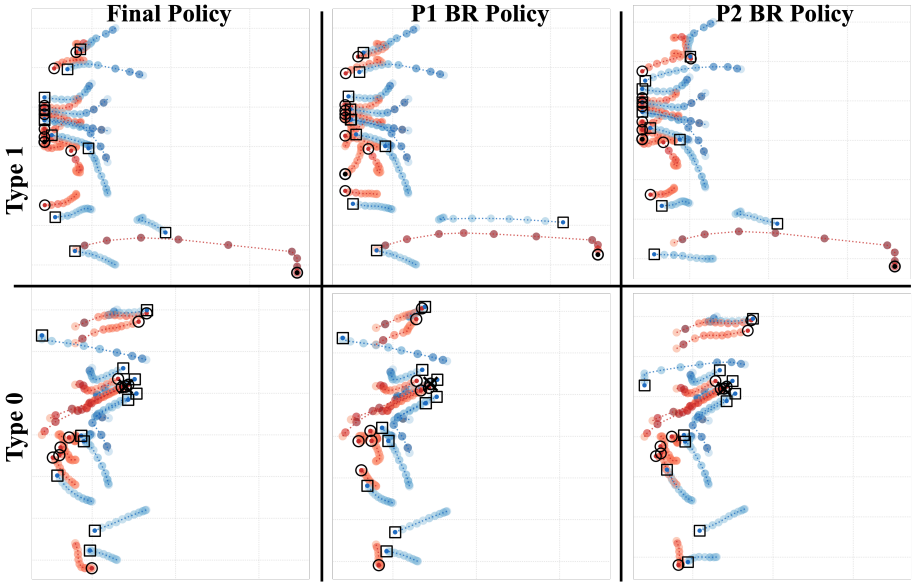


Figure 3.8. Comparison of resulting behavior from the final P1 and P2's policies vs their respective best-response (BR) policies. Both players' behavior do not shift qualitatively when switching to their BR policies highlighting that the resulting policy with non-zero exploitability is indeed close to the equilibrium.

3.7 Proofs and Additional Results

3.7.1 Proof of Theorem 3.1

Theorem 3.1 (A splitting reformulation of the primal and dual DPs). The RHS of Eq. 3.3 can be reformulated as

$$
\begin{aligned}
&\min_{\{u^k\},\{\alpha_{ki}\}} \max_{\{v^k\}} \quad \sum_{k=1}^{I} \lambda^k \Big(V(t+\tau, x^k, p^k) + \tau \mathbb{E}_{i\sim p^k}[l_i(x, u^k, v^k)] \Big) \\
&\text{s.t.} \quad u^k \in \mathcal{U}, \quad x^k = x + \tau f(x, u^k, v^k), \quad v^k \in \mathcal{V}, \\
&\qquad \alpha_{ki} \in [0,1], \ \sum_{k=1}^{I} \alpha_{ki} = 1, \ \lambda^k = \sum_{i=1}^{I} \alpha_{ki} p[i], \\
&\qquad p^k[i] = \frac{\alpha_{ki} p[i]}{\lambda^k}, \quad \forall i, k \in [I].
\end{aligned}
\tag{P$_1$}
$$

And the RHS of Eq. 3.5 can be reformulated as

$$\begin{aligned}
&\min_{\{v^k\},\{\lambda^k\},\{\hat{p}^k\}} \max_{\{u^k\}} \sum_{k=1}^{I+1} \lambda^k \left(V^*(t+\tau, x^k, \hat{p}^k - \tau l(x, u^k, v^k))\right) \\
&\text{s.t.} \quad u^k \in \mathcal{U}, \quad v^k \in \mathcal{V}, \quad x^k = x + \tau f(x, u^k, v^k), \qquad (\mathrm{P}_2) \\
&\lambda^k \in [0,1],\ \sum_{k=1}^{I+1} \lambda^k \hat{p}^k = \hat{p},\ \sum_{k=1}^{I+1} \lambda^k = 1,\ k \in [I+1].
\end{aligned}$$

Proof. Recall that the primal DP is:

$$\begin{aligned}
V_\tau(t_0, x_0, p) &= \min_{\{\eta_i\}} \mathbb{E}_{i\sim p,\ u\sim\eta^i} \left[\max_{v\in\mathcal{V}} \left\{V_\tau(t_0+\tau, x'(u,v), p'(u)) + \tau l_i(x,u,v)\right\}\right] \\
&= \min_{\{\eta_i\}} \int_{\mathcal{U}} \bar{\eta}(u) \max_{v\in\mathcal{V}} \left\{V_\tau(t_0+\tau, x'(u,v), p'(u)) + \tau \mathbb{E}_{i\sim p'(u)}[l_i(x,u,v)]\right\} du \\
&= \min_{\{\eta_i\}} \int_{\mathcal{U}} \bar{\eta}(u) a(u, p'(u)) du,
\end{aligned} \tag{3.9}$$

where the last equality uses Isaacs' condition to reduce v as an implicit function of u:

$$a(u, p'(u)) = \max_{v\in\mathcal{V}} V_\tau(t_0+\tau, x'(u,v), p'(u)) + \tau \mathbb{E}_{i\sim p'(u)}[l_i(x,u,v)]. \tag{3.10}$$

Now we introduce a pushforward measure ν on $\Delta(I)$ for any $E \subset \Delta(I)$: $\nu(E) = \int_{\{u:p'(u)\in E\}} \bar{\eta}(u)\, du$. Let $\eta_{p'}$ be the conditional measure on $\mathcal{U}$ for each p'. Then we have

$$\begin{aligned}
\min_{\{\eta_i\}} \int_{\mathcal{U}} \bar{\eta}(u) a(u, p'(u)) du &= \min_{\nu} \int_{\Delta(I)} \min_{\eta_{p'}} \Big[\int_{p'(u)=p'} a(u,p')\, \eta_{p'}(du)\Big] \nu(dp') \\
&= \min_{\nu} \int_{\Delta(I)} \min_{u\in\mathcal{U}} a(u,p') \nu(dp') \\
&= \min_{\nu} \int_{\Delta(I)} \tilde{a}(p') \nu(dp').
\end{aligned}$$

This leads to the following reformulation of V_τ:

$$\begin{aligned}
V_\tau(t_0, x_0, p) = &\min_{\nu} \int_{\Delta(I)} \tilde{a}(p') \nu(dp') \\
&\text{s.t.} \quad \mathbb{E}_\nu[p'] = p.
\end{aligned} \tag{3.11}$$

Now we show that the RHS of Eq. 3.11 computes the convexificiation of $\tilde{a}(p')$ at $p' = p$.

To do so, we first show V_τ is convex on $\Delta(I)$: Let two probability measures ν^1 and ν^2 be the solutions for $V_\tau(t_0, x_0, p^1)$ and $V_\tau(t_0, x_0, p^2)$, respectively. For any $\theta \in [0,1]$ and $p^\theta = \theta p^1 + (1-\theta)p^2$, the mixture $\nu^\theta := \theta\nu^1 + (1-\theta)\nu^2$ satisfies $\int_{\Delta(I)} p'\nu(dp') = \theta \int_{\Delta(I)} p'\nu^1(dp') + (1-\theta)\int_{\Delta(I)} p'\nu^2(dp') = p^\theta$ and is a feasible solution to Eq. 3.11 for p^θ. Therefore

$$V(p^\theta) \leq \int_{\Delta(I)} a(p')\nu^\theta(dp') = \theta V(p^1) + (1-\theta)V(p^2).$$

Then we show $V(p) \leq \tilde{a}(p)$ for any $p \in \Delta(I)$: The Dirac measure δ_p concentrated at p satisfies $\int_{\Delta(I)} p'\delta_p(dp') = p$. Thus δ_p is a feasible solution to Eq. 3.11 and by definition $V(p) \leq \int_{\Delta(I)} \tilde{a}(p')\delta_p(dp') = \tilde{a}(p)$.

Lastly, we show V is the largest convex minorant of $\tilde{a}$: Let h be any convex function on $\Delta(I)$ such that $h(p) \leq \tilde{a}(p)$ for all p. Given $p \in \Delta(I)$, for any probability measure ν that satisfies $\int_{\Delta(I)} p'\nu(dp') = p$, we have

$$h(p) = h\left(\int_{\Delta(I)} p'\nu(dp')\right) \leq \int_{\Delta(I)} h(p')\nu(dp') \leq \int_{\Delta(I)} \tilde{a}(p')\nu(dp').$$

Since this inequality holds for arbitrary ν, including the optimal ones that define $V(p)$ through Eq. 3.11, it follows that $h(p) \leq V(p)$. With these, $V(\cdot)$ in Eq. 3.11 is the convexification of $\tilde{a}(\cdot)$.

Since convexification in $\Delta(I)$ requires at most I vertices, ν^* that solves Eq. 3.11 is I-atomic. We will denote by $\{p^k\}_{k\in[I]}$ the set of "splitting" points that have non-zero probability masses according to ν^*, and let $\lambda^k := \nu^*(p^k)$. Using Isaacs' condition, $\arg\min_{u\in\mathcal{U}} a(u,p)$ is non-empty for any $p \in \Delta(I)$, and therefore each p^k is associated with (at least) one action in $\arg\min_{u\in\mathcal{U}} a(u,p^k)$. As a result, $\{\eta_i\}$ is also concentrated on a common set of I actions in $\mathcal{U}$. Specifically, denote this set by $\{u^k\}_{k\in[I]}$, we should

have $\alpha_{ki} := \eta_i(u^k) = \lambda^k p^k[i]/p[i]$. Thus we reach P_1. The same proof technique can be applied to the dual DP to derive P_2. □

3.7.2 Proof of Theorem 3.2

Theorem 3.2. For any $(t, x, p) \in [0, T] \times \mathcal{X} \times \Delta(I)$, if A1-5 hold, then for any $\epsilon > 0$, there exist $C > 0$, and $\tau \in (0, \tau^*]$, such that

$$0 \leq \max_{\zeta \in \mathcal{Z}} J(t, x, p; \{\eta^\dagger_{i,\tau}\}, \zeta) - V(t, x, p) \leq \epsilon \tag{3.12}$$

Similarly, for any $(t, x, \hat{p}) \in [0, T] \times \mathcal{X} \times \mathbb{R}^I$,

$$0 \leq \max_{\{\eta_i\} \in \{\mathcal{H}^i\}^I} J^*(t, x, \hat{p}; \{\eta_i\}, \zeta^\dagger_\tau) - V^*(t, x, \hat{p}) \leq \epsilon \tag{3.13}$$

3.7.3 Settings

We recall all necessary settings for the proof.

From Thm. 3.1, the Bellman backup for G_τ is P_1, which can be written as an operator T_τ:

$$T_\tau[V_\tau](t + \tau, x, p) := V_\tau(t, x, p) = \min_{\{\lambda^k\}, \{p^k\}} \quad \sum_{k=1}^{I} \lambda^k \tilde{V}_\tau(t, x, p^k), \tag{3.14}$$

where $\tilde{V}_\tau(t, x, p^k) = \min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} V_\tau(t + \tau, x + \tau f(x, u, v), p^k) + \tau \mathbb{E}_{i \sim p^k}[l_i(x, u, v)]$ is the non-revealing value at (t, x, p^k), and $\sum_{k=1}^I \lambda^k = 1,\ \sum_{k=1}^I \lambda^k p^k = p,\ p^k \in \Delta(I),\ \forall k \in [I]$.

From Cardaliaguet (2007), the value V of the original game G is the unique viscosity solution to the following Hamilton-Jacobi equation

$$\nabla_t w + H(t, x, \nabla_x w) = 0, \tag{3.15}$$

with terminal boundary $w(T,x,p) = \sum_{i=1}^{I} p[i]g_i(x)$, and is convex in p.

3.7.4 Preliminary lemmas

The following lemmas will be used in the main proof.

Lemma 3.1 (Uniform Convergence (Thm 3.1 of Cardaliaguet (2009))**.** *Assuming (A1-A5) hold, the map V_τ converges uniformly to the value function V on compact subsects of $[0,T] \times \mathcal{X} \times \Delta(I)$, in the following sense:*

$$\lim_{\tau \to 0^+,\, t_k \to t,\, x' \to x,\, p' \to p} V_\tau(t_k, x', p') = V(t,x,p) \quad \forall (t,x,p) \in [0,T] \times \mathcal{X} \times \Delta(I).$$

Proof. Proved by Cardaliaguet (2009) for both V_τ and V_τ^*. Souquière (2010) generalized this result to games with running costs and lack of information on both sides. □

Lemma 3.2. *Let $\bar{\mathcal{V}}(t_0) = \{v : [t_0, T] \to \mathcal{V}\}$ denote a set of open-loop controls for P2. Then, given f, a constant control $u \in \mathcal{U}$ on $[t_k, t_{k+1}]$, a time-varying control $v \in \bar{\mathcal{V}}(t)$, there exists a w such that*

$$w \in Cof(x, u, \mathcal{V}) \tag{3.16}$$

where Co denotes the convex hull.

Proof. Let

$$w := \frac{1}{\tau} \int_{t_k}^{t_{k+1}} f(x, u, v(s))\, ds \tag{3.17}$$

We show that w as defined above lies in the convex hull of the set $f(x,u,\mathcal{V})$.

Let $S := f(x,u,\mathcal{V}) = \{f(x,u,\bar{v}) : \bar{v} \in \mathcal{V}\}$. For any time s, since $v(s) \in \mathcal{V}$, we have:

$$f(x,u,v(s)) \in S \quad \text{for all } s \in [t_k, t_{k+1}] \tag{3.18}$$

Case 1: when $v(\cdot)$ is piecewise-constant. Assume first that $v(\cdot)$ is piecewise constant on the partition $[t_k, t_{k+1}]$:

$$[t_k, t_{k+1}] = \bigcup_{j=1}^{m} N_j, \quad |N_j| = \Delta t_j, \quad \sum_{j=1}^{m} \Delta t_j = \tau,$$

and $v(s) = v_j$ on N_j. Then

$$\int_{t_k}^{t_{k+1}} f(x, u, v(s))ds = \sum_{j=1}^{m} \Delta t_j f(x, u, v_j).$$

Divide by τ to get

$$w = \frac{1}{\tau} \sum_{j=1}^{m} \Delta t_j f(x, u, v_j) = \sum_{j=1}^{m} q_j f(x, u, v_j), \qquad \text{where } q_j = \frac{\Delta t_j}{\tau}.$$

$q_j \geq 0$ and $\sum_{j=1}^{m} q_j = 1$, so w is a finite weighted average of points in S. Hence,

$$w \in \mathrm{Co}(S) = \mathrm{Co} f(x, u, \mathcal{V}) \tag{3.19}$$

Case 2: general measurable case. For a general measurable $v(\cdot)$, we can approximate it by piecewise-constant $v_n(\cdot)$ such that

$$f(x, u, v_n(\cdot) \to f(x, u, v(\cdot) \quad \text{ in } L^1([t_k, t_{k+1}]$$

This is true because $f(x, u, \cdot)$ is bounded and measurable through $v(\cdot)$.

Define

$$w_n := \frac{1}{\tau} \int_{t_k}^{t_{k+1}} f(x, u, v_n(s)\ ds.$$

From Case 1, each $w_n \in \mathrm{Co}(S)$.

Moreover,

$$\begin{aligned} |w_n - w| &= \left| \frac{1}{\tau} \int_{t_k}^{t_{k+1}} (f(x, u, v_n(s)) - f(x, u, v(s))\ ds \right| \\ &\leq \frac{1}{\tau} \int_{t_k}^{t_{k+1}} |f(x, u, v_n(s)) - f(x, u, v(s))|\ ds \to 0 \end{aligned}$$

So $w_n \to w$

Finally, S is compact because $\mathcal{V}$ is compact and f is continuous. Hence, $\mathrm{Co}(S)$ is compact and therefore closed. Thus, the limit of $w_n \in \mathrm{Co}(S)$ also belongs to $\mathrm{Co}(S)$:

$$w \in \mathrm{Co}(S) = \mathrm{Co} f(x, u, \mathcal{V}) \tag{3.20}$$

□

Lemma 3.3 (Local Quadratic Error)**.** *For any $k \in \{0, \ldots, L-1\}$, any x, a fixed control $u \in \mathcal{U}$, any control $v \in \bar{\mathcal{V}}(t_k)$, there exists some $w \in Cof(x, u, \mathcal{V})$, such that*

$$|X_{t_{k+1}}^{t_k, x, u, v} - (x + \tau w)| \leq C_1 \tau^2 \tag{3.21}$$

Proof. We have, from the ODE,

$$X(t_{k+1}) = X_{t_{k+1}}^{t_k, x, u, v} = x + \int_{t_k}^{t_{k+1}} f(X(s), u, v(s))\, ds. \tag{3.22}$$

From the definition (3.17) of w,

$$x + \tau w = x + \int_{t_k}^{t_{k+1}} f(x, u, v(s)) ds. \tag{3.23}$$

Subtract (3.23) from (3.22)

$$\begin{aligned} X(t_{k+1}) - (x + \tau w) &= \int_{t_k}^{t_{k+1}} \Big(f(X(s), u, v(s)) - f(x, u, v(s)) \Big)\, ds \\ \Big| X(t_{k+1}) - (x + \tau w) \Big| &= \Big| \int_{t_k}^{t_{k+1}} \Big(f(X(s), u, v(s)) - f(x, u, v(s)) \Big)\, ds \Big| \\ &\leq \int_{t_k}^{t_{k+1}} \Big| f(X(s), u, v(s)) - f(x, u, v(s)) \Big|\, ds \end{aligned}$$

From Lipschitz continuity of f, $|f(X(s), u, v(s) - f(x, u, v(s))| \leq L_f |X(s) - x|$,

$$\Big| X(t_{k+1}) - (x + \tau w) \Big| \leq \int_{t_k}^{t_{k+1}} L_f |X(s) - x|\, ds.$$

From boundedness of f: Let $||f||_\infty \leq F$. Then for $s \in [t_k, t_{k+1}]$

$$\Big|X(s) - x\Big| = \Big|\int_{t_k}^{s} f(X(r), u, v(r)\, dr\Big| \leq \int_{t_k}^{s} |f(\cdot)|\, dr \leq \int_{t_k}^{s} F dr = F(s - t_k),$$

which leads to

$$\Big|X(t_{k+1}) - (x + \tau w)\Big| \leq \int_{t_k}^{t_{k+1}} L_f \cdot F \cdot (s - t_k)\, ds = L_f \cdot F \int_0^{\tau} r\, dr = \frac{L_f \cdot F}{2}\tau^2$$

which yields equation (3.21) with $C_1 = \frac{L_f \cdot F}{2}$. □

Lemma 3.4 (Grid Error)**.** *There exists some constant $C > 0$ independent of τ such that, for any $k \in \{0, \cdots, L\}$, $x \in \mathbb{R}^{d_x}$, and $p \in \Delta(I)$, we have*

$$\max_{\zeta \in \mathcal{Z}} J(t_k, x, p; \{\eta^{\dagger}_{i,\tau}\}, \zeta) \leq V_\tau(t_k, x, p) + C(T - t_k)\tau \tag{3.24}$$

Proof. Let $\tau = T/L$, $\{\eta^{\dagger}_{i,\tau}\}$ be the random strategy associated with the feedback strategy (u^k, α_{ki}) in the game G_τ, and $\zeta \in \mathcal{Z}$. Then, for any $k \in \{0, \ldots, L\}$, $x \in \mathbb{R}^{d_x}$, and $p \in \Delta(I)$, claim that:

$$\max_{\zeta \in \mathcal{Z}} J(t_k, x, p; \{\eta^{\dagger}_{i,\tau}\}, \zeta) \leq V_\tau(t_k, x, p) + C(L - k)\tau^2 \tag{3.25}$$

We use backward induction to prove the equation (3.25). At $k = L$, the inequality holds trivially as there is no time left in the game and the error term goes to 0. Then, assume that (3.25) holds for $k + 1$:

$$\max_{\zeta \in \mathcal{Z}(t_{k+1})} J(t_{k+1}, x, p; \{\eta^{\dagger}_{i,\tau}\}, \zeta) \leq V_\tau(t_{k+1}, x, p) + C(L - (k+1))\tau^2 \tag{3.26}$$

Now, to prove for any k, let us analyze what happens on $[t_k, t_{k+1})$. From the feedback given by (P_1), for a fixed (x, p) we have:

$$X_T^{t_k, x, \{\eta^{\dagger}_{i,\tau}\}, \zeta} = X_T^{t_{k+1}, x^j, \{\eta^{j\dagger}_{i,\tau}\}, \zeta^j} \quad \text{with probability } \lambda^j p^j[i]/p[i],$$

where

- $x^j = X_{t_{k+1}}^{t_k,x,u^j,\zeta(u^j)}$,
- $\{\eta_{i,\tau}^{j\dagger}\}$ is the continuation of the random strategy from time t_{k+1} associated with the same feedback but starting at (t_{k+1}, x^j, p^j),
- $\zeta^j \in \mathcal{Z}(t_{k+1})$ is the continuation of the non-anticipative ζ after u^j is chosen.

Then, using Lemma 3.3, there is some $w^j \in \mathrm{Co} f(x, u^j, \mathcal{V})$ such that:

$$|x^j - (x + \tau w^j)| \leq C_1 \tau^2. \tag{3.27}$$

From the definition of J,

$$\begin{aligned}
&J(t_k, x, \{\eta_{i,\tau}^\dagger\}, \zeta, p) \\
&= \sum_{i=1}^{I} p[i] \mathbb{E}_{\{\eta_{i,\tau}^\dagger\}} \Big[g_i\Big(X_T^{t_k,x,\{\eta_{i,\tau}^\dagger\},\zeta}\Big) + \int_{t_k}^{T} l_i\Big(X(s), \{\eta_{i,\tau}^\dagger\}(s), \zeta(s)\Big)\, ds \Big] \\
&= \sum_{i=1}^{I} \sum_{j=1}^{I+1} p[i]\lambda^j p^j[i]/p[i] \mathbb{E}_{\{\eta_{i,\tau}^{j\dagger}\}} \Big[g_i\Big(X_T^{t_{k+1},x,\{\eta_{i,\tau}^{j\dagger}\},\zeta^j}\Big) + \int_{t_k}^{t_{k+1}} l_i(x(s), u^j, \zeta(u^j))\, ds \\
&\qquad\qquad + \int_{t_{k+1}}^{T} l_i\Big(X(s), \{\eta_{i,\tau}^{j\dagger}\}(s), \zeta^j(s)\Big)\, ds \Big] \\
&\leq \sum_{j=1}^{I+1} \lambda^j \max_{\zeta' \in \mathcal{Z}(t_{k+1})} \sum_{i=1}^{I} p^j[i] \mathbb{E}_{\{\eta_{i,\tau}^{j\dagger}\}} \Big[g_i\Big(X_T^{t_{k+1},x,\{\eta_{i,\tau}^{j\dagger}\},\zeta'}\Big) + \int_{t_k}^{t_{k+1}} l_i(x(s), u^j, \zeta(u^j))\, ds \\
&\qquad\qquad + \int_{t_{k+1}}^{T} l_i\Big(X(s), \{\eta_{i,\tau}^{j\dagger}\}(s), \zeta'(s)\Big)\, ds \Big]
\end{aligned}$$

Combining the above result with (3.26), we get:

$$J(t_k, x, \{\eta_{i,\tau}^\dagger\}, \zeta, p) \leq \sum_{j=1}^{I+1} \lambda^j \Big[V_\tau(t_{k+1}, x^j, p^j) + C(L - (k+1))\tau^2 \Big].$$

Finally, applying (3.27) and assuming C_2-Lipschitz continuity of V_τ, we arrive at:

$$\begin{aligned}
J(t_k, x, \{\eta^\dagger_{i,\tau}\}, \zeta, p) &\leq \sum_{j=1}^{I+1} \lambda^j \Big[V_\tau(t_{k+1}, x + \tau w^j, p^j) + C_2 |x^j - (x + \tau w^j)| + \\
&\qquad\qquad C(L - (k+1))\tau^2 \Big] \\
&\leq \sum_{j=1}^{I+1} \lambda^j \Big[V_\tau(t_{k+1}, x + \tau w^j, p^j) + C_1 C_2 \tau^2 + C(L - (k+1))\tau^2 \Big] \\
&\leq \sum_{j=1}^{I+1} \lambda^j \max_{w \in \mathrm{Co} f(x, u^j, \mathcal{V})} \Big[V_\tau(t_{k+1}, x + \tau w, p^j) + C_1 C_2 \tau^2 + C(L - (k+1))\tau^2 \Big] \\
&\leq V_\tau(t_k, x, p) + (C_1 C_2 + 2)\tau^2 + C(L - (k+1))\tau^2.
\end{aligned}$$

The term $2\tau^2$ arises from two approximation steps: first τ^2 from selecting an approximately optimal convex decomposition in the definition of $V_\tau(t_k, x, p)$, and the second τ^2 from choosing u^j to be only τ^2-optimal for $\min_u \max_v(\cdot)$ so that they are Borel measurable with respect to (x, p).

Then, choosing $C = (C_1 C_2 + 2)$ satisfies (3.25).

Since we have, $L = T/\tau$, and $t_k = k\tau$. We get

$$\max_{\zeta \in \mathcal{Z}} J(t_k, x, p; \{\eta^\dagger_{i,\tau}\}, \zeta) \leq V_\tau(t_k, x, p) + C(T - t_k)\tau$$

Note that compared with the proof provided in Cardaliaguet (2009) Theorem 4.1, our proof incorporates instantaneous costs and clarified the stage-wise $\mathcal{O}(\tau)$ error in (3.26).

□

Proof of Theorem 3.2

Proof. Our proof focuses on the primal game. The same technique can be applied to the dual game. First, it is easy to see $V(t,x,p) \leq \max_{\zeta\in\mathcal{Z}} J(t,x,p;\{\eta^{\dagger}_{i,\tau}\},\zeta)$ because $\{\eta^{\dagger}_{i,\tau}\}$ is not necessarily NE in G.

Next, we need to show

$$0 \leq \max_{\zeta\in\mathcal{Z}} J(t,x,p;\{\eta_{i,\tau}\}^{\dagger},\zeta) - V(t,x,p) \leq \epsilon \tag{3.28}$$

i.e., applying $\{\eta^{\dagger}_{i,\tau}\}$ (which solves G_τ) to G will yield a value not far from the true value of G.

Let us fix some t', x', and p', and some k such that $t' \in [t_{k-1}, t_k)$. Then,

$$\begin{aligned}\max_{\zeta\in\mathcal{Z}_r(t')} J(t',x',p';\{\eta^{\dagger}_{i,\tau}\},\zeta) &= \max_{\zeta\in\mathcal{Z}(t')} J(t',x',p';\{\eta^{\dagger}_{i,\tau}\},\zeta) \\ &\leq \max_{|x-x'|\leq F\tau} \max_{\zeta\in\mathcal{Z}(t_k)} J(t_k,x,p';\{\eta^{x\dagger}_{i,\tau}\},\zeta) + \tau||l||_\infty\end{aligned}$$

where $\{\eta^{x\dagger}_{i,\tau}\}$ is the strategy at (t_k,x,p'), F is a bound on f, and $||l||_\infty$ term is due to bounded instantaneous cost. Hence, from Lemma 3.4

$$\max_{\zeta\in(t')Z} J(t',x',p',\{\eta_{i,\tau}\}^{\dagger},\zeta) \leq \max_{|x-x'|\leq F\tau} V_\tau(t_k,x,p') + \mathcal{O}(\tau)$$

Here, $\mathcal{O}(\tau)$ absorbs $C(T-t_k)\tau + \tau||l||_\infty$. Together with $\tau = T/L \to 0$ and Lemma 3.1, $\max_{|x-x'|\leq F\tau} V_\tau(t_k,x,p')$ converges to $V(t',x',p')$ and $\mathcal{O}(\tau) \to 0$

Therefore, for any $\epsilon > 0$ there exists $\tau^* > 0$ such that for all $\tau \in (0,\tau^*]$ inequality (3.28) holds. □

3.7.5 Sample Trajectories

Here we present sample trajectories for three different initial states for each P1 type. The policies learned by CAMS results in trajectories that are significantly close to the ground truth than the other two baseline algorithms.

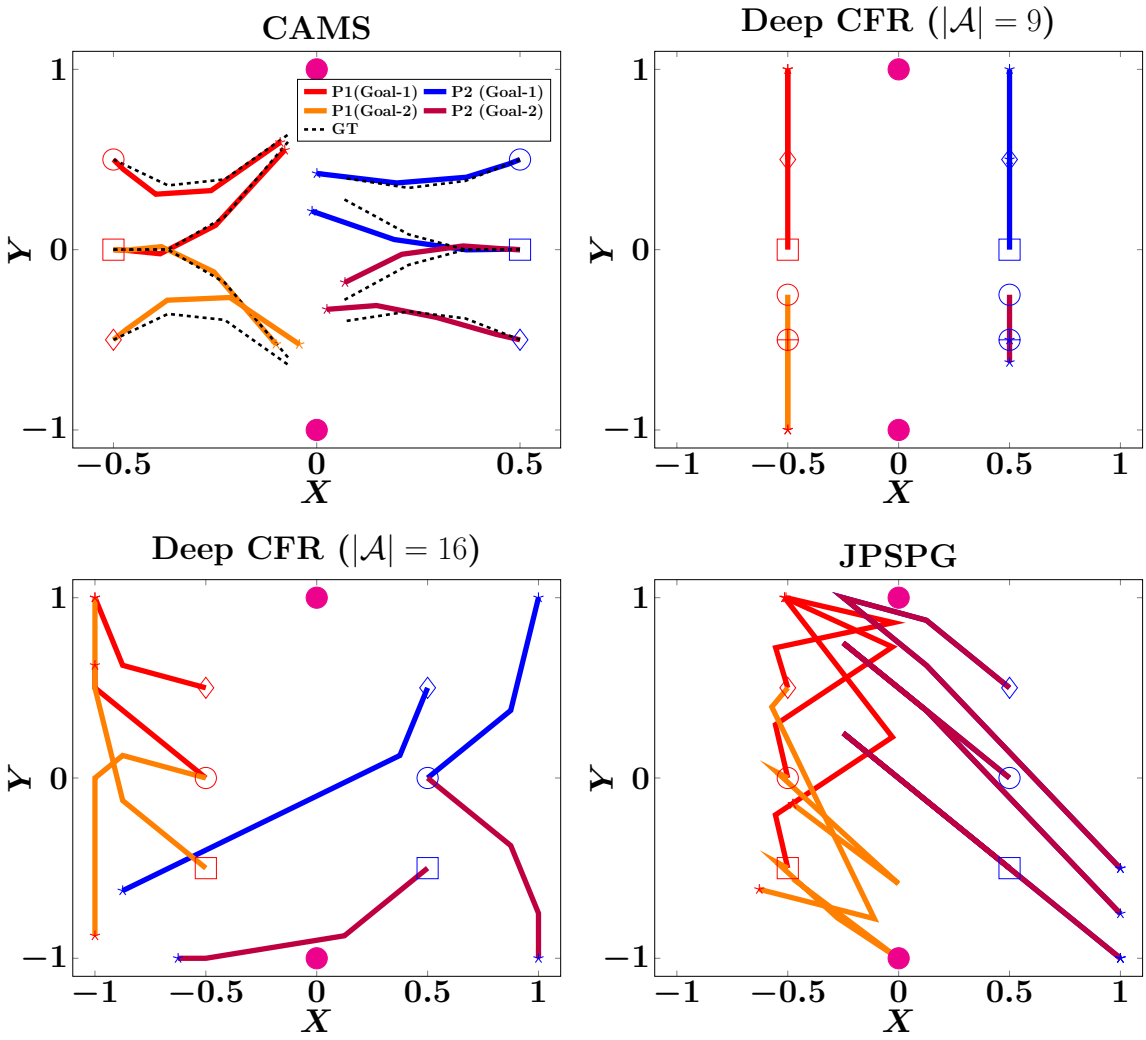


Figure 3.9. Trajectories generated using CAMS (primal game), DeepCFR, and JPSPG. The initial position pairs are marked with same marker and the final with star. The trajectories from CAMS are close to the ground-truth while those from DeepCFR and JPSPG are not.

3.8 Conclusion

Unlike general IIEFGs where NE strategies are distributions over the entire action space, we showed that 2p0s1 games enjoy an atomic NE structure when P1 can precisely control the public belief, leading to an exponentially reduced game-tree complexity. We demonstrated the utilities of this NE structure in solving games with continuous action spaces in model-free and model-based modes, in terms of computational cost and solution quality. Our methods enable fast approximation of deceptive and counter-deception *team* strategies, with potential applications to sports, missile/drone defense, and risk-sensitive robotics applications, tailored for specific team dynamics, action spaces, and task specifications.

Chapter 4

ENABLING ONLINE SOLVING

4.1 INTRODUCTION

In this chapter, we shift the focus to a special class of 2p0s1 games that have linear dynamics and quadratic costs. Previous chapters dealt with a general 2p0s1 games with non-linear dynamics and payoffs and revealed the atomic structure of the NE leading to its tractable approximation. Implementing such approximation schemes for real-time sub-game solving, however, has not been achieved, yet is desired for applications where sim-to-real gaps exist and robust control is required. This chapter improves the computational efficiency of sub-game solving for P1 in 2p0s1 with linear dynamics and quadratic losses (hereby represented as LQ). Specifically, we show that P1's NE computation can be formulated as a bi-level optimization problem where the outer level optimizes the "signaling" strategy, i.e., when and how to reveal information through control, and the inner level is a game-tree linear-quadratic regulator (LQR) that solves for the optimal closed-loop control. This bi-level problem is solved via an adjoint-enabled backpropagation scheme: A "backward" LQR pass is followed by a "forward" gradient descent pass for improving the signaling.

This chapter is based on the work "Fast Strategy Solving for the Informed Player in Two-Player Zero-Sum Linear-Quadratic Differential Games with One-sided Information", under review during the time of the dissertation submission.

We then investigate the effectiveness of the game solver that combines the two structural insights to enable sub-second strategy computation in 2p0s1 LQ games.

The main contributions are:

1. The formulation of P1's NE strategy learning as a bi-level optimization problem where a belief-induced LQ game is nested in signaling optimization.
2. An adjoint-enabled backpropagation algorithm that differentiates the Riccati recursion to optimize signaling policy, resulting in ≈10Hz sub-game solving for a homing game with an 8D state space, 2D action spaces, and a time horizon of $K = 10$.
3. Two case studies of the homing game: we use Hexner's game with an analytical NE to verify the correctness of our algorithm, and a "drone landing under attack" scenario to demonstrate the value of fast sub-game solving under P2's random velocity disturbances.

4.2 PROBLEM FORMULATION

4.2.1 Game Setup and Information

We consider 2p0s1 LQ on a finite time horizon $[0, T]$, and approximate its NE through time discretization with a time step $\tau > 0$. Let $K \coloneqq T/\tau$ and $t_k \coloneqq k\tau$ for $k \in [K]$. At $t = 0$, a payoff type θ is drawn from $\Theta = \{\theta_i\}_{i\in[I]}$ with a public prior $p_0 \in \Delta(\Theta)$. P1 observes θ; P2 only knows p_0. In continuous time, the joint state $x(t) \in \mathbb{R}^n$ evolves according to linear dynamics

$$\dot{x}(t) = A_c x(t) + B_{1,c} u(t) + B_{2,c} v(t),$$

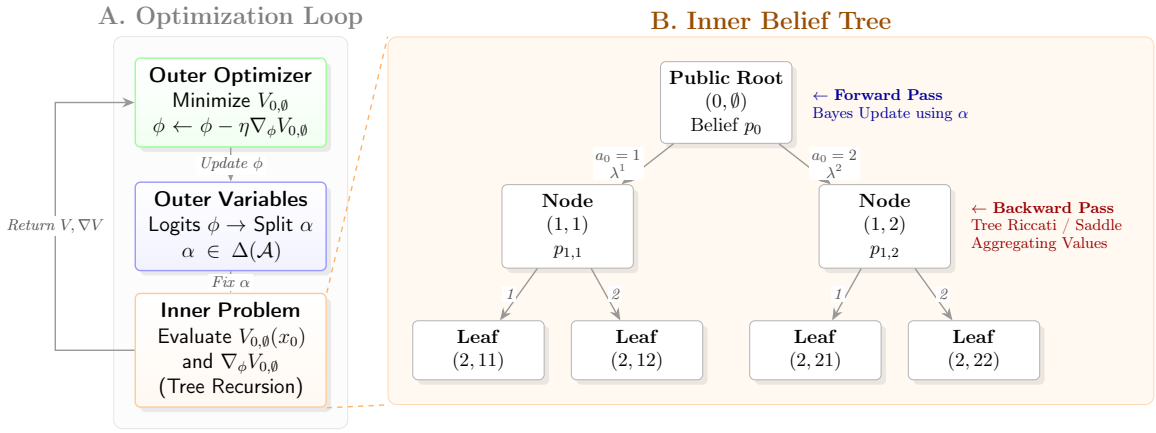


Figure 4.1. Bilevel structure used in the solver. (A) The outer level optimizes belief-splitting coefficients α (parameterized by logits ϕ). (B) Given α, the inner level constructs the public belief tree (forward Bayes propagation), computes belief-averaged costs, and evaluates the induced tree of LQ saddle subgames via a tree-structured Riccati recursion, returning the root value and its gradient.

where $u(t) \in \mathbb{R}^{m_1}$ is P1's control and $v(t) \in \mathbb{R}^{m_2}$ is P2's control. The corresponding discrete-time dynamics is:

$$x_{k+1} = Ax_k + B_1 u_k + B_2 v_k, \tag{4.1}$$

with $A = I + \tau A_c$, $B_i = \tau B_{i,c} + A_c B_{i,c} \frac{\tau^2}{2}$ for $i \in [2]$. The public belief about θ is denoted by $p(t)$ in continuous time and p_k in discrete time, and is updated by Bayes' rule. Throughout the game both players observe the state, action, and belief trajectories.

4.2.2 Strategy and Payoff Structures

We denote by $\{\mathcal{H}_r^i\}^I$ the joint sets of I behavioral strategies of P1[2]. A type-θ_i behavioral strategy $\eta_i \in \mathcal{H}_r^i$ maps the current public information (e.g., (t, x, p) in

[2]We define "behavioral strategy" as the time-delayed non-anticipative strategies standard for differential games Elliott and Kalton (1972).

continuous time or (k, x_k, p_k) in discrete time) to a probability measure over $\mathbb{R}^{m_1}$. Similarly, P2's behavioral strategy $\zeta \in \mathcal{Z}_r$ produces a probability measure over $\mathbb{R}^{m_2}$. The subscript r highlights the randomness of behavioral strategies.

Let X_T be the random terminal state induced by (η_i, ζ) under the continuous-time dynamics. Type-θ_i P1 minimizes

$$J_{\theta_i}(t_0, x_0) := \mathbb{E}_{\eta_i,\zeta}\left[g_{\theta_i}(X_T) + \int_{t_0}^{T} \ell_{\theta_i}(\eta_i(s), \zeta(s))\, ds\right], \tag{4.2}$$

where the running cost is quadratic and strictly convex-concave in controls

$$\ell_{\theta_i}(u, v) = \frac{1}{2}u^\top R_i u \; - \; \frac{1}{2}v^\top S_i v, \quad R_i \succ 0,\; S_i \succ 0, \tag{4.3}$$

and the terminal cost is convex in P1's state and concave in P2's:

$$g_{\theta_i}(x) = \frac{1}{2}x^\top Q_i x + q_i^\top x + c_i, \;\; Q_i = Q_i^\top,\; q_i \in \mathbb{R}^n,\; c_i \in \mathbb{R}. \tag{4.4}$$

P2 maximizes the expected payoff $J(t_0, x_0, p_0) = \mathbb{E}_{\theta\sim p_0}[J_\theta(t_0, x_0)]$. The upper and lower values are defined as $V^+(t_0, x_0, p_0) = \inf_{\{\eta_i\}} \sup_\zeta J$ and $V^-(t_0, x_0, p_0) = \sup_\zeta \inf_{\{\eta_i\}} J$, respectively. The game has a value V if $V = V^+(t_0, x_0, p_0) = V^-(t_0, x_0, p_0)$. Existence of V in 2p0s1 is proved in Cardaliaguet (2009), notably under the Isaacs' condition which holds for 2p0s1 LQ . Ghimire *et al.* (2026) develops a discrete-time primal-dual reformulation of the game to approximate NE strategies. Specifically, P1 solves a Stackelberg primal game where P2 best responses to P1 via the following Bellman backup:

$$V_\tau(k, x_k, p_k) = \min_{\{\eta_i\}\in\{\mathcal{H}_r\}^I} \mathbb{E}_{\theta_i\sim p_k, u_k\sim\eta_i}\Big[\max_{v_k} V_\tau(k+1, x_{k+1}, p_{k+1}) + \tau\ell_{\theta_i}(u_k, v_k)\Big], \tag{4.5}$$

with boundary $V_\tau(K, x, p) = \sum_{\theta_i} p_{\theta_i} g_{\theta_i}(x)$. Here p_{k+1} is obtained after observing action u: $p_{k+1}(i) \propto p_k(i)\,\mathbb{P}(u \mid \theta = \theta_i)$ under P1's strategy set $\{\eta_i\}$.

The rest of the paper focuses on efficiently solving (4.5) at all game tree nodes. Specifically, we approximate $\{\eta_i\}$ by separating (i) a signaling component that governs the evolution of the public belief (i.e., how p_k is split into posteriors p_{k+1}), and (ii) a control component that specifies the corresponding optimal state-feedback actions given the public belief. The next subsection makes this separation explicit via an I-branch public belief-tree representation.

4.2.3 Reparameterization of $\{\eta_i\}$

Our goal is to decompose $\{\eta_i\}$ into variables for signaling and those for control. We start with a critical property of 2p0s1: It is proved that P1's NE strategies in (4.5) are *atomic* Ghimire *et al.* (2026): η_i for $i \in [I]$ randomizes over at most I common action prototypes at each tree node, and thus the primal game can be considered a tree-structured minimax problem, with each node consisting of I branches. See visualization in Fig. 4.1B for $I = 2$ and $K = 2$.

Let $\mathcal{A} \coloneqq \{1, 2, \ldots, I\}$ be the index set for action prototypes, and $\omega \in \mathcal{A}^k$ denote P1's length-k action sequence. Each tree node can be uniquely identified by (k, ω). At each node (k, ω) and for each private type θ_i, P1 selects the next action index $a \in \mathcal{A}$ according to a distribution specified by η_i:

$$\begin{aligned} &\alpha^a_{k,\omega,i} \;=\; \mathbb{P}(a_k = a \mid \theta = \theta_i,\ \text{public node } (k,\omega)), \\ &\alpha^a_{k,\omega,i} \geq 0,\ \sum_{a \in \mathcal{A}} \alpha^a_{k,\omega,i} = 1. \end{aligned} \tag{4.6}$$

We call the collection $\alpha = \{\alpha^a_{k,\omega,i}\}$ the signaling policy as it governs how the public belief evolves: Given a public belief $p \in \Delta(\Theta)$ at node (k, ω), the probability of

observing prototype a is

$$\lambda^a_{k,\omega}(p) \coloneqq \sum_{i=1}^{I} p_i \, \alpha^a_{k,\omega,i}. \tag{4.7}$$

After observing a, Bayes' rule yields the posterior belief at child node $(k{+}1, \omega a)$:

$$p_{k+1,\omega a}(i) = \frac{\alpha^a_{k,\omega,i} \, p_{k,\omega}(i)}{\lambda^a_{k,\omega}(p_{k,\omega})} \qquad (\lambda^a_{k,\omega}(p_{k,\omega}) > 0). \tag{4.8}$$

As a result, belief p across the game tree follows a martingale determined by α: $\sum_{a \in \mathcal{A}} \lambda^a_{k,\omega}(p_{k,\omega}) \, p_{k+1,\omega a} = p_{k,\omega}$.

Now we describe the LQ game for fixed α. Denote the value at (k, ω) as $V_{k,\omega}(x)$. At leaves, we have:

$$V_{K,\omega}(x) = \sum_{i=1}^{I} p_{K,\omega}(i) \, g_i(x). \tag{4.9}$$

Consider an edge (k, ω, a) from node (k, ω) to child $(k+1, \omega a)$. Conditioned on observing a, the posterior belief is $p_{k+1,\omega a}$ by (4.8). Taking expectation of the type-dependent running cost (4.3) under this posterior yields a standard quadratic stage cost with belief-averaged matrices

$$\bar{R}^a_{k,\omega} \;\coloneqq\; \sum_{i=1}^{I} p_{k+1,\omega a}(i) \, R_i, \quad \bar{S}^a_{k,\omega} \;\coloneqq\; \sum_{i=1}^{I} p_{k+1,\omega a}(i) \, S_i. \tag{4.10}$$

Let $\lambda^a_{k,\omega} := \lambda^a_{k,\omega}(p_{k,\omega})$ for brevity. The edge value conditioning on a solves an LQ subgame:

$$V^a_{k,\omega}(x) \;\coloneqq\; \min_{u \in \mathbb{R}^{m_1}} \max_{v \in \mathbb{R}^{m_2}} \Big\{ \tau\Big(\tfrac{1}{2} u^\top \bar{R}^a_{k,\omega} \, u \;-\; \tfrac{1}{2} v^\top \bar{S}^a_{k,\omega} \, v \Big) + V_{k+1,\omega a}\big(Ax + B_1 u + B_2 v\big) \Big\}. \tag{4.11}$$

Since Isaacs' condition holds for 2p0s1 LQ, each edge subproblem (4.11) has a unique solution. This will be formalized in Theorem 4.1. The node value is an expectation over branches:

$$V_{k,\omega}(x) = \sum_{a \in \mathcal{A}} \lambda^a_{k,\omega} \, V^a_{k,\omega}(x), \quad k = 0, \ldots, K-1. \tag{4.12}$$

To summarize, the equilibrium computation of (4.5) across the game tree can be viewed as a bilevel program: the outer level chooses the signaling α, while the inner level solves the induced game (4.9)–(4.12) to define an evaluation map $\alpha \mapsto V_{0,\emptyset}(x_0, p_0)$. With initial state x_0 and prior p_0, the outer signaling problem is

$$\min_{\alpha} \; V_{0,\emptyset}(x_0, p_0) \quad \text{s.t.} \quad \alpha_{k,\omega,i} \in \Delta(\mathcal{A}) \;\; \forall k, \omega, i. \tag{4.13}$$

In Section 4.3 we show how to (i) evaluate $V_{0,\emptyset}(x_0, p_0)$ efficiently for a fixed α via a tree-structured Riccati recursion, and (ii) differentiate through this evaluation to optimize α.

4.3 DIFFERENTIABLE RICCATI SOLVER

We now describe a differentiable tree-structured Riccati recursion composed of (1) *edge steps* that solve the local LQ games (4.11) to obtain feedback laws and edge values $V^a_{k,\omega}$ given the values at child nodes $V_{k+1,\omega a}$, and (2) *node steps* (4.12) that average edge values using branch probabilities $\lambda^a_{k,\omega}$ to obtain the *quadratic* value at parent node.

4.3.1 Quadratic Value Representation on Nodes

A key enabler of the Riccati recursion is Theorem 4.1, which states that node values are quadratic and convex in state:

$$V_{k,\omega}(x) = \frac{1}{2} x^\top P_{k,\omega} x + r_{k,\omega}^\top x + c_{k,\omega}, \tag{4.14}$$

where coefficients $(P_{k,\omega}, r_{k,\omega}, c_{k,\omega})$ depend on $p_{k,\omega}$ (hence on α) and can be computed by backward induction.

Theorem 4.1 (Quadratic closure on the public tree)**.** *For fixed α, $V_{k,\omega}(\cdot)$ at any (k,ω) is quadratic convex in state.*

Proof. We prove by induction. At a leaf node (K,ω) (4.9), since g_i is quadratic and strictly convex in state, $V_{K,\omega}$ is also quadratic and convex. Using (4.4), leaf coefficients are

$$\begin{aligned} P_{K,\omega} &= \sum_{i=1}^{I} p_{K,\omega}(i)\, Q_i, \qquad r_{K,\omega} = \sum_{i=1}^{I} p_{K,\omega}(i)\, q_i, \\ c_{K,\omega} &= \sum_{i=1}^{I} p_{K,\omega}(i)\, c_i. \end{aligned} \tag{4.15}$$

For induction, fix an edge (k,ω,a) with child node $(k{+}1,\omega a)$. Let the child value be

$$V_{k+1,\omega a}(x) = \frac{1}{2} x^\top P^+ x + r^{+\top} x + c^+, \quad P^+ \coloneqq P_{k+1,\omega a},$$
$$r^+ \coloneqq r_{k+1,\omega a}, \quad c^+ \coloneqq c_{k+1,\omega a}.$$

On this edge, the posterior belief-averaged running matrices are $\bar{R} \coloneqq \bar{R}^a_{k,\omega}$ and $\bar{S} \coloneqq \bar{S}^a_{k,\omega}$ from (4.10). The edge value (4.11) can therefore be written as

$$V^a_{k,\omega}(x) = \min_u \max_v \left\{ \tfrac{1}{2} u^\top (\tau \bar{R})\, u - \tfrac{1}{2} v^\top (\tau \bar{S})\, v + \tfrac{1}{2} x^{+\top} P^+ x^+ + r^{+\top} x^+ + c^+ \right\}, \tag{4.16}$$

where $x^+ = Ax + B_1 u + B_2 v$.

Define the stacked input $w = \begin{bmatrix} u \\ v \end{bmatrix}$ and the input matrix $B = [B_1 \;\; B_2]$. The saddle stage Hessian is

$$\bar{R}_w \;\coloneqq\; \begin{bmatrix} \tau \bar{R} & 0 \\ 0 & -\tau \bar{S} \end{bmatrix}. \tag{4.17}$$

Then the objective inside (4.16) becomes a quadratic function of (w,x):

$$\Phi^a_{k,\omega}(w;x) = \frac{1}{2} w^\top \bar{R}_w w + \frac{1}{2} (Ax + Bw)^\top P^+ (Ax + Bw) + r^{+\top} (Ax + Bw) + c^+. \tag{4.18}$$

Expanding (4.18) and collecting terms in w yields

$$\Phi^a_{k,\omega}(w;x) = \frac{1}{2} w^\top \underbrace{\left(\bar{R}_w + B^\top P^+ B\right)}_{H^a_{k,\omega}} w + \; w^\top B^\top (P^+ A x + r^+) \\ + \; \frac{1}{2} x^\top A^\top P^+ A x \; + \; (A^\top r^+)^\top x \; + \; c^+. \tag{4.19}$$

We first verify that the minimax problem is well-posed and admits a *unique* saddle. Let $\|\cdot\|$ denote the Euclidean norm for vectors and the induced (spectral) operator norm for matrices, and since $P^+ = P^{+\top}$, $\|P^+\| = \max_i |\lambda_i(P^+)|$, where $\lambda(M)$ is the eigenvalue of M. From (4.19),

$$H^a_{k,\omega} = \begin{bmatrix} H_{uu} & H_{uv} \\ H_{uv}^\top & H_{vv} \end{bmatrix},$$

with $H_{uu} = \tau \bar{R} + B_1^\top P^+ B_1, \quad H_{vv} = -\tau \bar{S} + B_2^\top P^+ B_2$. Since $\bar{R}, \bar{S}$ are posterior convex combinations of $\{R_i \succ 0\}$ and $\{S_i \succ 0\}$, $\lambda_{\min}(\bar{R}) \geq \underline{r} := \min_i \lambda_{\min}(R_i) > 0$ and $\lambda_{\min}(\bar{S}) \geq \underline{s} := \min_i \lambda_{\min}(S_i) > 0$. Thus, for any $\delta u, \delta v$,

$$\delta u^\top H_{uu} \delta u \geq \big(\tau \underline{r} - \|P^+\| \, \|B_1\|^2\big) \|\delta u\|^2, \tag{4.20}$$

$$\delta v^\top H_{vv} \delta v \leq \big(- \tau \underline{s} + \|P^+\| \, \|B_2\|^2\big) \|\delta v\|^2, \tag{4.21}$$

with $\delta u^\top \bar{R} \delta u \geq \underline{r} \|\delta u\|^2$, $y^\top P^+ y \geq -\|P^+\| \|y\|^2$, and $\|B_j \delta\| \leq \|B_j\| \, \|\delta\|$. Let $\bar{P} := \sup_{k,\omega} \|P_{k,\omega}\| < \infty$ and assume a standard discretization so that $\|B_j(\tau)\| \leq \beta_j \tau$ for all $\tau \in (0, \tau_0]$ and some $\beta_j, \tau_0 > 0$. Then $H_{uu} \succ 0$ and $H_{vv} \prec 0$ for any $\tau \in (0, \tau^*)$, where

$$\tau^* := \min \left\{ \frac{\underline{r}}{\bar{P}\beta_1^2}, \; \frac{\underline{s}}{\bar{P}\beta_2^2}, \; \tau_0 \right\}.$$

Hence $\Phi^a_{k,\omega}(\cdot; x)$ is strictly convex in u and strictly concave in v, and the edge game (4.16) admits a unique saddle point.

Therefore, we have

$$\nabla_w \Phi^a_{k,\omega}(w;x) = H^a_{k,\omega} \, w + B^\top (P^+ A x + r^+) = 0.$$

The unique saddle point is then the following:

$$\begin{aligned} w^{a*}_{k,\omega}(x) &= \mathcal{K}^a_{k,\omega} x + \kappa^a_{k,\omega}, \;\; \mathcal{K}^a_{k,\omega} = -\big(H^a_{k,\omega}\big)^{-1} B^\top P^+ A, \\ \kappa^a_{k,\omega} &= -\big(H^a_{k,\omega}\big)^{-1} B^\top r^+. \end{aligned} \tag{4.22}$$

Partitioning $w = \begin{bmatrix} u \\ v \end{bmatrix}$ gives the edge-specific affine feedback laws

$$u^{a*}_{k,\omega}(x) = \mathcal{K}^a_{u,k,\omega} x + \kappa^a_{u,k,\omega}, \tag{4.23}$$

$$v^{a*}_{k,\omega}(x) = \mathcal{K}^a_{v,k,\omega} x + \kappa^a_{v,k,\omega}, \tag{4.24}$$

where $(\mathcal{K}_u, \kappa_u)$ and $(\mathcal{K}_v, \kappa_v)$ are the blocks of $(\mathcal{K}, \kappa)$. Let $b(x) \coloneqq B^\top(P^+Ax + r^+)$ be the affine term in (4.19). Then (4.19) can be written as

$$\begin{aligned} \Phi^a_{k,\omega}(w; x) = \frac{1}{2}(w + H^{-1}b(x))^\top H (w + H^{-1}b(x)) \\ - \frac{1}{2} b(x)^\top H^{-1} b(x) + \frac{1}{2} x^\top A^\top P^+ A x + (A^\top r^+)^\top x + c^+, \end{aligned}$$

where $H \equiv H^a_{k,\omega}$. At the saddle point

$$w = w^*(x) = -H^{-1} b(x),$$

the squared term vanishes and we obtain

$$V^a_{k,\omega}(x) = \frac{1}{2} x^\top A^\top P^+ A x \;+\; (A^\top r^+)^\top x \;+\; c^+ - \;\frac{1}{2}\, b(x)^\top \big(H^a_{k,\omega}\big)^{-1} b(x). \tag{4.25}$$

Expanding the last term with $b(x) = B^\top(P^+Ax + r^+)$ and collecting quadratic, linear, and constant parts in x gives the quadratic coefficients of the edge value:

$$P^a_{k,\omega} = A^\top P^+ A - A^\top P^+ B \big(H^a_{k,\omega}\big)^{-1} B^\top P^+ A, \tag{4.26}$$

$$r^a_{k,\omega} = A^\top r^+ - A^\top P^+ B \big(H^a_{k,\omega}\big)^{-1} B^\top r^+, \tag{4.27}$$

$$c^a_{k,\omega} = c^+ - \frac{1}{2} r^{+\top} B \big(H^a_{k,\omega}\big)^{-1} B^\top r^+. \tag{4.28}$$

Returning to the node recursion (4.12), since each $V^a_{k,\omega}(x)$ is quadratic in x, the node coefficients are simply the λ-weighted averages:

$$\begin{aligned} P_{k,\omega} &= \sum_{a\in\mathcal{A}} \lambda^a_{k,\omega}\, P^a_{k,\omega}, \qquad r_{k,\omega} = \sum_{a\in\mathcal{A}} \lambda^a_{k,\omega}\, r^a_{k,\omega}, \\ c_{k,\omega} &= \sum_{a\in\mathcal{A}} \lambda^a_{k,\omega}\, c^a_{k,\omega}. \end{aligned} \tag{4.29}$$

□

Remark 4.2. In our case studies, the dynamics are player-separable: $x = [x^{(1)}, x^{(2)}]$, $A = \texttt{diag}(A_1, A_2)$, $B_1 = [\tilde{B}_1; 0]$, $B_2 = [0; \tilde{B}_2]$, and the terminal quadratic has the signed block form $Q_i = \texttt{diag}(Q_i^{(1)}, -Q_i^{(2)})$ with $Q_i^{(1)}, Q_i^{(2)} \succeq 0$. The tree recursion preserves $P_{k,\omega} = \texttt{diag}(P^{(1)}_{k,\omega}, -P^{(2)}_{k,\omega})$ with $P^{(1)}_{k,\omega}, P^{(2)}_{k,\omega} \succeq 0$, hence $H_{uv} = B_1^\top P^+ B_2 = 0$ and $H_{uu} = \tau\bar{R} + \tilde{B}_1^\top P^{(1)} \tilde{B}_1 \succ 0$, $H_{vv} = -\tau\bar{S} - \tilde{B}_2^\top P^{(2)} \tilde{B}_2 \prec 0$ for all $\tau > 0$. Therefore every edge subproblem is strictly convex–concave and has a unique saddle without requiring $\tau < \tau^*$.

4.3.2 The Solver

For fixed α, the tree Riccati recursion has two steps: The forward (Bayes) pass propagates beliefs $\{p_{k,\omega}\}$ and branch weights $\{\lambda^a_{k,\omega}\}$ using (4.7)-(4.8), and compute averaged edge matrices $\{\bar{R}^a_{k,\omega}, \bar{S}^a_{k,\omega}\}$ via (4.10). Then backward (Riccati) pass initializes leaf quadratics by (4.15); then for $k = K-1, \ldots, 0$ computes edge updates (4.22)-(4.28) and aggregate to nodes via (4.29).

To enforce simplex constraints in (4.6), we parameterize α via logits ϕ and a softmax:

$$\alpha^a_{k,\omega,i} = \frac{\exp(\phi^a_{k,\omega,i})}{\sum_{b\in\mathcal{A}} \exp(\phi^b_{k,\omega,i})}. \tag{4.30}$$

Alternatively, α can also be modeled by a neural network $F : \mathbb{R}^n \times \Delta(\Theta) \to \alpha$. Given logits ϕ, we compute $\alpha(\phi)$ and run the tree Riccati recursion to obtain the root quadratic $(P_{0,\emptyset}, r_{0,\emptyset}, c_{0,\emptyset})$. The scalar objective is then

$$\mathcal{L}(\phi) := V_{0,\emptyset}(x_0, p_0) = \frac{1}{2} x_0^\top P_{0,\emptyset} x_0 + r_{0,\emptyset}^\top x_0 + c_{0,\emptyset}. \tag{4.31}$$

All operations are differentiable, and thus gradients $\nabla_\phi \mathcal{L}$ can be obtained via automatic differentiation and used in first- or second-order optimizers. We can further accelerate the optimization by leveraging adjoint-enabled gradient descent Amos *et al.* (2018). The solver is summarized in Algorithm 3.

Algorithm 3 2p0s1 LQ Solver

Input: Game data $\{R_i, S_i, Q_i, q_i, c_i\}_{i=1}^I$, (A, B_1, B_2), horizon K, time step τ, initial state x_0, prior p_0.
Initialize logits $\{\phi^a_{k,\omega,i}\}$, choose optimizer and step size.
for outer iterations $t = 0, 1, 2, \ldots$ **do**
  Compute $\alpha(\phi)$ via softmax (4.30).
  (Forward) Build belief tree $\{p_{k,\omega}, \lambda^a_{k,\omega}\}$ (4.7)–(4.8).
  (Forward) Compute $\bar{R}^a_{k,\omega}, \bar{S}^a_{k,\omega}$ (4.10).
  (Backward) Initialize $(P_{K,\omega}, r_{K,\omega}, c_{K,\omega})$ at leaf (4.15).
  for $k = K-1, \ldots, 0$ **do**
    for each node $\omega \in \mathcal{A}^k$ and prototype $a \in \mathcal{A}$ **do**
      Compute edge feedback $(K^a_{k,\omega}, \kappa^a_{k,\omega})$ via (4.22).
      Compute $(P^a_{k,\omega}, r^a_{k,\omega}, c^a_{k,\omega})$ via (4.26)–(4.28).
    end for
    Aggregate to node $(P_{k,\omega}, r_{k,\omega}, c_{k,\omega})$ via (4.29).
  end for
  Compute loss $\mathcal{L}(\phi) = V_{0,\emptyset}(x_0)$ via (4.31).
  Backpropagate to obtain $\nabla_\phi \mathcal{L}$ from root to leaf.
  Optimizer step on ϕ.
end for
Output: Optimized α and edge feedback $\{K^a_{k,\omega}, \kappa^a_{k,\omega}\}$ to compose $\{\eta_i\}$.

4.4 Case Studies

We implement the proposed method to solve two 2p0s1 LQs. Both games use 2D double integrator dynamics and have a time horizon of $K = 10$ with $\tau = 0.1$. Game-specific parameters are discussed as follows.

4.4.1 Case 1: Hexner's Game

We use Hexner's game Hexner (1979) with an analytical NE to verify the correctness of the learned policies. The game has a type-dependent terminal cost and type-independent running costs. P1 has two potential targets at $(0, \theta)$ with $\theta = \{-1, 1\}$. The terminal cost function is:

$$g_{\theta_i}(x) = (x^{(1)} - z\theta_i)^\top \tilde{Q}_i (x^{(1)} - z\theta_i) - (x^{(2)} - z\theta_i)^\top \tilde{Q}_i (x^{(2)} - z\theta_i), \tag{4.32}$$

where $x^{(1)}$ and $x^{(2)}$ represent the states of P1 and P2, respectively. $z = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$. $\tilde{Q}_i = \tilde{Q} = \texttt{diag}(1, 1, 0, 0)$ is the terminal cost matrix. Rewriting (4.32) in the form of (4.4), we get,

$$\begin{aligned} Q_i &= Q = 2 \cdot \texttt{diag}(\tilde{Q}, -\tilde{Q}), \\ q_i &= 2 \begin{bmatrix} -\theta_i (\tilde{Q}^\top z) & \theta_i (\tilde{Q}^\top z) \end{bmatrix}^\top, \\ c_i &= \theta_i^2 (z^\top \tilde{Q} z - z^\top \tilde{Q} z) = 0 \end{aligned}$$

The control penalty matrices for P1 and P2 are $R = \texttt{diag}(0.05, 0.025)$ and $S = \texttt{diag}(0.05, 0.1)$, respectively.

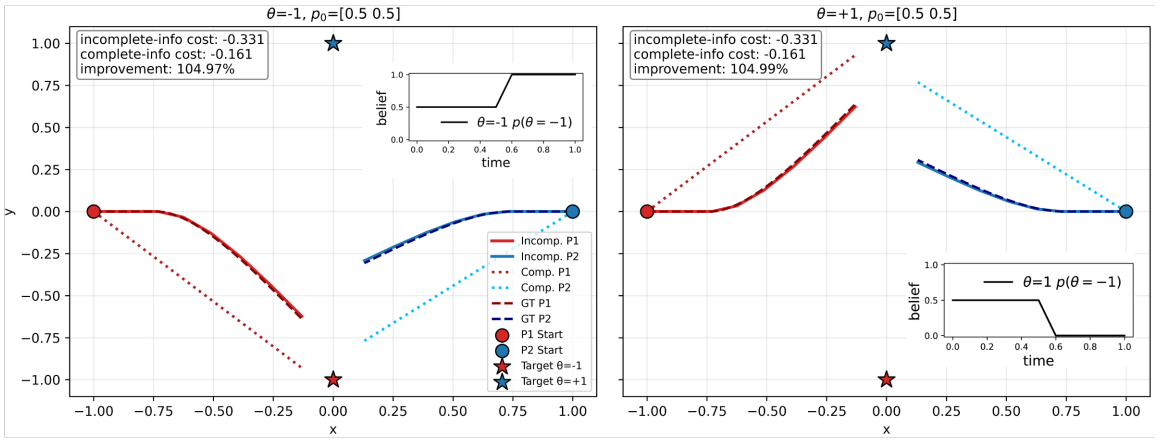


Figure 4.2. Comparisons between ground-truth NEs ("GT") of Hexner's game and learned policies for $\theta = -1$ (left) and $\theta = 1$ (right) under incomplete information. Dotted lines are NEs for the corresponding *complete-information* games. P1 is able to improve significantly by exploiting its information advantage.

The analytical solution for Hexner's game is studied in Ghimire *et al.* (2026); Hexner (1979): P1 plays non-revealing (i.e., its strategy is same across types) until a game-dependent critical time t_r when they completely reveal the target. Ghimire *et al.* (2026) reported a cold-solve time of 5 minutes on a Macbook M1 Pro for P1's NE along with P2's best responses. In comparison, our algorithm achieves the same in under 1 minute. Results are visualized in Fig. 4.2, where the learned policies match the analytical ground-truth, recovering the critical time of $t_r \approx 0.5$. We also compare the NEs against those for the corresponding complete-information games (where P2 knows P1's targets) along with the value improvement percentages to show the necessity of information control in strategic planning.

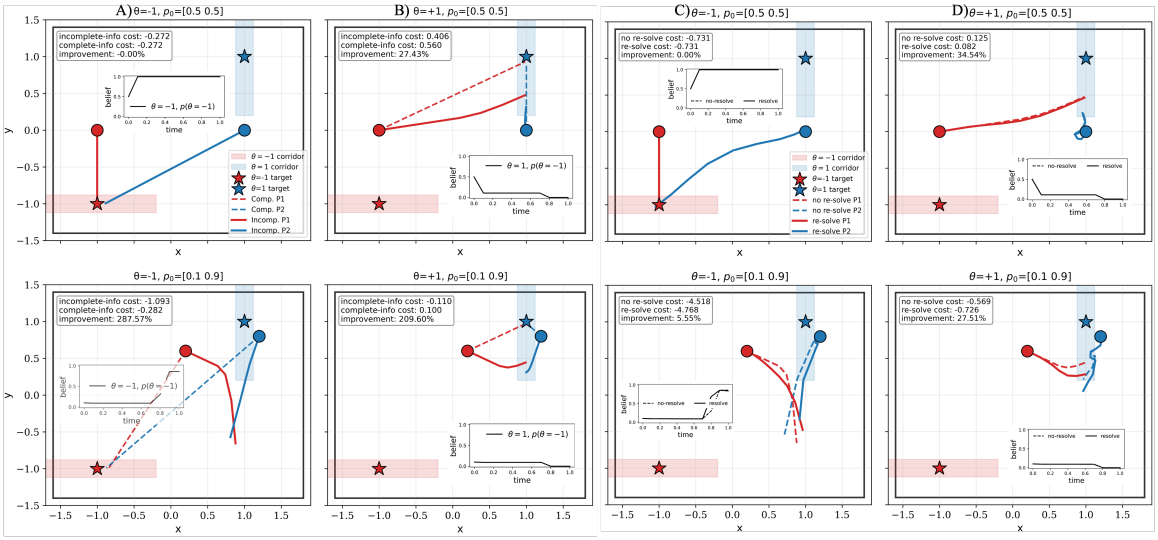


Figure 4.3. Sample trajectories for the drone landing under attack case. A-B: Comparison between complete vs incomplete information games. P1 is better off in expectation when playing the incomplete-info game. Improvements for specific realizations are labeled in the plots. B-C: Comparison between re-solved vs offline policy when P2's dynamics is stochastic. On average, resolving helps P1.

4.4.2 Case 2: Drone Landing Under Attack

In this game, a drone (P1) has two possible landing zones $(-1,-1)$ and $(1,1)$ in a horizontal and a vertical orientation, and is evading from an adversary (P2). Final penalty matrices are type-dependent and shared by both players: $\tilde{Q}_1 = \texttt{diag}(1, 20, 0, 0)$ and $\tilde{Q}_2 = \texttt{diag}(20, 1, 0, 0)$. Running penalty matrices are type-independent: $R = \texttt{diag}(0.05, 0.025)$, and $S = \texttt{diag}(0.02, 0.04)$. Dynamics, running and terminal cost functions follow Hexner's game, with $z = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$

4.4.2.1 Optimal policy under nominal conditions

First we present the result when the deployment and offline training share same condition. We highlight that, unlike in the Hexner's game, the optimal policy in this case is state- and belief-dependent. We provide few cases to showcase the resulting learned policy in Fig. 4.3A-B.

4.4.2.2 Optimal policy under stochastic disturbance

Next, we introduce noise in P2's state during deployment, and argue that by re-solving the sub-game at each time-step, P1 can play robustly against a noisy P2. The noise is added to P2's velocity from an i.i.d. Normal distribution $\mathcal{N}(0, \Sigma^2)$. To better reflect the underlying problem, P1 considers stochastic LQ during online play, which results in an accumulation of additional cost $\mathbf{tr}(P_k\Sigma)$ due to noise at every time step. However, note that the feedback gain remains unchanged in stochastic LQ with zero-mean noise. We use $\Sigma = \texttt{diag}(0, 0, 0.25, 0.25)$ in our results.

We compare the cost of the game realized by re-solving the sub-games to that realized without re-solving. We report this difference in addition to the mean resolve time (milliseconds) per time step reported from 1600 runs in Tab. 4.1. Sample trajectories are shown in Fig. 4.3C-D.

Table 4.1. Δcost of P1 and mean subgame solving time

Δ cost ($\downarrow$)	95% Confidence Interval	resolve time (ms)
-0.1051 ± 0.3656	$[-0.1232,\ -0.0877]$	71.51 ± 59.57

4.5 Conclusion

We provided an efficient gradient-based algorithm for solving the informed player's strategy in two-player zero-sum linear-quadratic differential games with one-sided information by leveraging its equilibrium structure. Future work will explore methods to improve algorithmic efficiency for the uninformed player, which solves a dual game that cannot exploit the LQ structure as much as the primal, yet may leverage the optimal signaling from the latter.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This dissertation studied two-player zero-sum differential games with one-sided information (2p0s1) as a mathematically focused yet practically meaningful class of safety-critical interactions. In the hierarchy between complete-information differential games and fully general imperfect-information dynamic games, 2p0s1 occupies an important middle ground. Complete-information models ignore hidden intent, while fully general partially observable stochastic formulations are often too broad to admit structure that can be rigorously exploited. By contrast, 2p0s1 already contains the central difficulty that makes many real-world safety-critical problems hard: one agent acts with private information, the other must defend under uncertainty, and both players must do so in continuous time, in continuous state and action spaces, and often under safety constraints. In many attack-defense settings, this one-sided information pattern is not merely a convenient simplification; it is the minimal setting in which the defender must explicitly reason about hidden intent. In that sense, solving 2p0s1 is a principled first step toward broader safety-critical game solving.

Within this framework, the dissertation makes three main contributions. First, Chapter 2 extends the theory of 2p0s1 games to the presence of state constraints by proving existence of value and deriving primal-dual subdynamic principles, together with a backward induction scheme for value approximation and behavioral strategy synthesis. This part clarifies how equilibrium strategies manipulate belief and how

state constraints change not only trajectories, but also the timing and advantage of information release. Second, Chapter 3 focuses on algorithm design exploiting the unique atomic equilibrium structure inherent to 2p0s1 games: under deterministic dynamics, both the informed and uninformed players need randomize only over finitely many actions at each decision point. This collapses the effective game-tree complexity relative to standard imperfect-information extensive-form formulations and enables a family of scalable implementations, including value approximation, model-free multiagent reinforcement learning, and model predictive control. Third, Chapter 4 specializes the problem to linear-quadratic systems and develops a differentiable tree-Riccati methodology that formulates the informed player's equilibrium computation as a bilevel optimization problem over signaling and control, enabling near real-time re-solving during deployment.

Taken together, these contributions bridge rigorous foundations and scalable computation of Nash equilibrium strategies in 2p0s1. Chapter 2 establishes the theoretical foundation. Chapter 3 shows that the equilibrium structure is not only an interesting property, but can also be algorithmically useful, making it possible to solve continuous-action games that are out of reach for generic IIEFG solvers, including the 22-player football setting. Chapter 4 pushes this line further toward online use by enabling sub-second subgame solving in the linear-quadratic case and by demonstrating robust re-solving in the drone-landing-under-attack scenario under deployment noise. The unifying message is that equilibrium structure should not be viewed as a by-product of analysis. Rather, it should be viewed as computational leverage.

5.2 Future Directions and Broader Outlook

Several directions follow naturally from this work. The most immediate is real-time implementation on hardware. The linear-quadratic solver already points toward this possibility: because online re-solving is on the order of tens of milliseconds per subgame, a natural next step is closed-loop deployment on a physical platform such as a drone. Such a step would move the framework from proof-of-concept simulation toward an end-to-end safety-critical system. At the same time, it would surface the practical issues that are abstracted away in simulation, including estimator error, latency, actuator saturation, packet loss, and mismatch between assumed and realized dynamics. A particularly interesting question is whether belief-aware equilibrium policies remain robust when the signaling channel itself is corrupted by perception and control error.

A second direction is extension to stochastic dynamics and observations. This is mathematically important because, once stochasticity is introduced, the informed player can no longer steer the public belief as precisely through state evolution. Indeed, Chapter 3 already notes that in such cases the convex Bellman operators used in the deterministic setting become lower bounds, leaving open the problem of finding tight upper-bounding operators that preserve useful equilibrium structure. This difficulty is consistent with the broader literature. Cardaliaguet and Rainer (2009) proved value existence for stochastic differential games with asymmetric information and characterized the value through a second-order Hamilton-Jacobi equation, while Cardaliaguet and Rainer (2013) later revisited the underlying notion of strategy through pathwise strategies. A natural next step, therefore, is to develop value and strategy approximation schemes for stochastic 2p0s1 that recover as much as

possible of the deterministic structure exploited in this dissertation—ideally, an approximate primal-dual decomposition together with a quantitative notion of near-atomic equilibrium under noise.

A third direction is extension to partial observability. The present dissertation assumes that states and actions are publicly observed, so the main hidden variable is the payoff type. In partially observable settings, however, the public information state would need to jointly represent latent physical state, hidden type, and potentially past prescriptions or signaling rules. As discussed in the related-work chapters, common-information approaches by Nayyar et al. (2013) and the zero-sum extension of Kartik and Nayyar (2021) suggest one possible route, but it remains unclear how to retain both value existence and scalable strategy synthesis in continuous-time safety-critical games. Extending 2p0s1 in this direction would substantially broaden its applicability, especially to surveillance, pursuit-evasion, and human-robot interaction settings in which sensing is limited, delayed, or corrupted.

At a broader level, the context in which this line of work sits is also changing. Over the past decade, AI capability was often measured by super-human performance in board and card games such as Go, chess, shogi, and poker. Those milestones were foundational, but they are no longer the only yardstick for strategic intelligence. Increasingly, evaluation is shifting toward agents that can navigate the digital world. Benchmarks such as WebArena and OSWorld assess long-horizon interaction with websites, software, and full computer environments rather than fixed-rule games, and current state-of-the-art agents remain far below human performance on these tasks. Looking ahead, a reasonable expectation is that, as digital agents become more reliable, the focus will continue to move toward embodied agents operating in the physical world, where success depends not only on reasoning and planning, but also

on safe execution under uncertainty. Recent embodied benchmarks, and especially hybrid settings that combine web-scale reasoning with physical action, already point in this direction. From this perspective, the study of safety-critical differential games with asymmetric information is well aligned with where agent evaluation is heading — toward systems that must reason strategically, act continuously, and remain safe while interacting with other agents in the real world.

Abernethy, J., P. L. Bartlett and E. Hazan, "Blackwell approachability and no-regret learning are equivalent", in "Proceedings of the 24th Annual Conference on Learning Theory", pp. 27–46 (JMLR Workshop and Conference Proceedings, 2011).

Amos, B., I. Jimenez, J. Sacks, B. Boots and J. Z. Kolter, "Differentiable mpc for end-to-end planning and control", Advances in neural information processing systems **31** (2018).

Amos, B., L. Xu and J. Z. Kolter, "Input convex neural networks", in "International Conference on Machine Learning", pp. 146–155 (PMLR, 2017).

Aumann, R. J., M. Maschler and R. E. Stearns, *Repeated games with incomplete information* (MIT press, 1995).

Bansal, S. and C. J. Tomlin, "Deepreach: A deep learning approach to high-dimensional reachability", in "2021 IEEE International Conference on Robotics and Automation (ICRA)", pp. 1817–1824 (IEEE, 2021).

Billings, D., N. Burch, A. Davidson, R. Holte, J. Schaeffer, T. Schauenberg and D. Szafron, "Approximating game-theoretic optimal strategies for full-scale poker", in "IJCAI", vol. 3, p. 661 (2003).

Braess, D. and W. Hackbusch, "A new convergence proof for the multigrid method including the v-cycle", SIAM journal on numerical analysis **20**, 5, 967–975 (1983).

Brown, N., A. Bakhtin, A. Lerer and Q. Gong, "Combining deep reinforcement learning and search for imperfect-information games", Advances in Neural Information Processing Systems **33**, 17057–17069 (2020a).

Brown, N., A. Bakhtin, A. Lerer and Q. Gong, "Combining deep reinforcement learning and search for imperfect-information games", Advances in Neural Information Processing Systems **33**, 17057–17069 (2020b).

Brown, N., A. Lerer, S. Gross and T. Sandholm, "Deep counterfactual regret minimization", in "International conference on machine learning", pp. 793–802 (PMLR, 2019).

Brown, N. and T. Sandholm, "Superhuman ai for heads-up no-limit poker: Libratus beats top professionals", Science **359**, 6374, 418–424 (2018).

Brown, N. and T. Sandholm, "Superhuman ai for multiplayer poker", Science **365**, 6456, 885–890 (2019).

Brown, N. and T. W. Sandholm, "Simultaneous abstraction and equilibrium finding in games", (2015).

Burch, N., M. Johanson and M. Bowling, "Solving imperfect information games using decomposition", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 28 (2014).

Cardaliaguet, P., "Differential games with asymmetric information", SIAM journal on Control and Optimization **46**, 3, 816–838 (2007).

Cardaliaguet, P., "Numerical approximation and optimal strategies for differential games with lack of information on one side", Advances in Dynamic Games and Their Applications: Analytical and Numerical Developments pp. 1–18 (2009).

Cardaliaguet, P. and C. Rainer, "Stochastic differential games with asymmetric information", Applied Mathematics and Optimization **59**, 1, 1–36 (2009).

Cardaliaguet, P. and C. Rainer, "Pathwise strategies for stochastic differential games with an erratum to "stochastic differential games with asymmetric information"", Applied Mathematics & Optimization **68**, 1, 75–84 (2013).

Cen, S., Y. Wei and Y. Chi, "Fast policy extragradient methods for competitive games with entropy regularization", Advances in Neural Information Processing Systems **34**, 27952–27964 (2021).

De Meyer, B., "Repeated games, duality and the central limit theorem", Mathematics of Operations Research **21**, 1, 237–251 (1996).

Durkota, K., V. Lisỳ, C. Kiekintveld, K. Horák, B. Bošanskỳ and T. Pevnỳ, "Optimal strategies for detecting data exfiltration by internal and external attackers", in "International Conference on Decision and Game Theory for Security", pp. 171–192 (Springer, 2017).

Elliott, R. J. and N. J. Kalton, *The existence of value in differential games*, vol. 126 (American Mathematical Soc., 1972).

FAIR†, M. F. A. R. D. T., A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu *et al.*, "Human-level play in the game of diplomacy by combining language models with strategic reasoning", Science **378**, 6624, 1067–1074 (2022).

Garcia, E., D. W. Casbeer and M. Pachter, "Design and analysis of state-feedback optimal strategies for the differential game of active defense", IEEE Transactions on Automatic Control **64**, 2, 553–568 (2018).

Garcia, E., D. W. Casbeer and M. Pachter, "The complete differential game of active target defense", Journal of Optimization Theory and Applications **191**, 2, 675–699 (2021).

Ghimire, M., L. Zhang, Z. Xu and Y. Ren, "State-constrained zero-sum differential games with one-sided information", in "Proceedings of the 41st International Conference on Machine Learning", edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett and F. Berkenkamp, vol. 235 of *Proceedings of Machine Learning Research*, pp. 15512–15539 (PMLR, 2024), URL https://proceedings.mlr.press/v235/ghimire24a.html.

Ghimire, M., L. Zhang, Z. Xu and Y. Ren, "Solving football by exploiting equilibrium structure of 2p0s differential games with one-sided information", in "The Fourteenth International Conference on Learning Representations", (2026), URL https://openreview.net/forum?id=vRwuBOxbsJ.

Gilpin, A., S. Hoda, J. Pena and T. Sandholm, "Gradient-based algorithms for finding nash equilibria in extensive form games", in "Internet and Network Economics: Third International Workshop, WINE 2007, San Diego, CA, USA, December 12-14, 2007. Proceedings 3", pp. 57–69 (Springer, 2007).

Gilpin, A. and T. Sandholm, "Finding equilibria in large sequential games of imperfect information", in "Proceedings of the 7th ACM conference on Electronic commerce", pp. 160–169 (2006).

Grabowski, K., "Wake forest's slow mesh rpo for explosive plays", URL https://coachandcoordinator.com/2020/10/wake-forests-slow-mesh-rpo-for-explosive-plays/ (2020).

Han, D. and J. W. Wan, "Multigrid methods for second order hamilton–jacobi–bellman and hamilton–jacobi–bellman–isaacs equations", SIAM Journal on Scientific Computing **35**, 5, S323–S344 (2013).

Harsanyi, J. C., "Games with incomplete information played by "bayesian" players, i–iii part i. the basic model", Management science **14**, 3, 159–182 (1967).

Hawkin, J., R. Holte and D. Szafron, "Automated action abstraction of imperfect information extensive-form games", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 25, pp. 681–687 (2011).

Henson, V. *et al.*, "Multigrid methods nonlinear problems: an overview", Computational imaging **5016**, 36–48 (2003).

Hexner, G., "A differential game of incomplete information", Journal of Optimization Theory and Applications **28**, 213–232 (1979).

Horák, K. and B. Bošanskỳ, "A point-based approximate algorithm for one-sided partially observable pursuit-evasion games", in "International conference on decision and game theory for security", pp. 435–454 (Springer, 2016).

Hu, H., Z. Zhang, K. Nakamura, A. Bajcsy and J. F. Fisac, "Learning-aware safety for interactive autonomy", arXiv preprint arXiv:2309.01267 (2023).

Jacot, A., F. Gabriel and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks", Advances in neural information processing systems **31** (2018).

Johanson, M., N. Bard, N. Burch and M. Bowling, "Finding optimal abstract strategies in extensive-form games", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 26, pp. 1371–1379 (2012).

Kartik, D. and A. Nayyar, "Upper and lower values in zero-sum stochastic games with asymmetric information", Dynamic Games and Applications **11**, 363–388 (2021).

Koller, D. and N. Megiddo, "The complexity of two-person zero-sum games in extensive form", Games and economic behavior **4**, 4, 528–552 (1992).

Koller, D. and A. Pfeffer, "Generating and solving imperfect information games", in "IJCAI", pp. 1185–1193 (Citeseer, 1995).

Kroer, C. and T. Sandholm, "Discretization of continuous action spaces in extensive-form games", in "Proceedings of the 2015 international conference on autonomous agents and multiagent systems", pp. 47–56 (2015).

Lanctot, M., E. Lockhart, J.-B. Lespiau, V. Zambaldi, S. Upadhyay, J. Pérolat, S. Srinivasan, F. Timbers, K. Tuyls, S. Omidshafiei, D. Hennes, D. Morrill, P. Muller, T. Ewalds, R. Faulkner, J. Kramár, B. D. Vylder, B. Saeta, J. Bradbury, D. Ding, S. Borgeaud, M. Lai, J. Schrittwieser, T. Anthony, E. Hughes, I. Danihelka and J. Ryan-Davis, "OpenSpiel: A framework for reinforcement learning in games", CoRR **abs/1908.09453**, URL http://arxiv.org/abs/1908.09453 (2019).

Lanctot, M., K. Waugh, M. Zinkevich and M. Bowling, "Monte carlo sampling for regret minimization in extensive games", Advances in neural information processing systems **22** (2009).

Lee, D., *Safety-Guaranteed Autonomy under Uncertainty* (University of California, Berkeley, 2022).

Li, Y., G. Carboni, F. Gonzalez, D. Campolo and E. Burdet, "Differential game theory for versatile physical human–robot interaction", Nature Machine Intelligence **1**, 1, 36–43 (2019).

Liang, L., F. Deng, Z. Peng, X. Li and W. Zha, "A differential game for cooperative target defense", Automatica **102**, 58–71 (2019).

Martin, C. and T. Sandholm, "Finding mixed-strategy equilibria of continuous-action games without gradients using randomized policy networks", in "Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence", pp. 2844–2852 (2023).

Martin, C. and T. Sandholm, "Joint-perturbation simultaneous pseudo-gradient", arXiv preprint arXiv:2408.09306 (2024).

Mc Carthy, S. M., A. Sinha, M. Tambe and P. Manadhata, "Data exfiltration detection and prevention: Virtually distributed pomdps for practically safer networks", in "International Conference on Decision and Game Theory for Security", pp. 39–61 (Springer, 2016).

McMahan, B., "Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization", in "Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics", edited by G. Gordon, D. Dunson and M. Dudík, vol. 15 of *Proceedings of Machine Learning Research*, pp. 525–533 (PMLR, Fort Lauderdale, FL, USA, 2011), URL https://proceedings.mlr.press/v15/mcmahan11b.html.

Moravčík, M., M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson and M. Bowling, "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker", Science **356**, 6337, 508–513 (2017).

Nayyar, A., A. Gupta, C. Langbort and T. Başar, "Common information based markov perfect equilibria for stochastic games with asymmetric information: Finite games", IEEE Transactions on Automatic Control **59**, 3, 555–570 (2013).

Perolat, J., B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony *et al.*, "Mastering the game of stratego with model-free multiagent reinforcement learning", Science **378**, 6623, 990–996 (2022a).

Perolat, J., B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony *et al.*, "Mastering the game of stratego with model-free multiagent reinforcement learning", Science **378**, 6623, 990–996 (2022b).

Perolat, J., R. Munos, J.-B. Lespiau, S. Omidshafiei, M. Rowland, P. Ortega, N. Burch, T. Anthony, D. Balduzzi, B. De Vylder *et al.*, "From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization", in "International Conference on Machine Learning", pp. 8525–8535 (PMLR, 2021).

Rudolph, M., N. Lichtle, S. Mohammadpour, A. Bayen, J. Z. Kolter, A. Zhang, G. Farina, E. Vinitsky and S. Sokota, "Reevaluating policy gradient methods for imperfect-information games", arXiv preprint arXiv:2502.08938 (2025).

Sandholm, T., "The state of solving large incomplete-information games, and application to poker", Ai Magazine **31**, 4, 13–32 (2010).

Schmid, M., M. Moravčík, N. Burch, R. Kadlec, J. Davidson, K. Waugh, N. Bard, F. Timbers, M. Lanctot, G. Z. Holland *et al.*, "Student of games: A unified learning algorithm for both perfect and imperfect information games", Science Advances **9**, 46, eadg3256 (2023a).

Schmid, M., M. Moravčík, N. Burch, R. Kadlec, J. Davidson, K. Waugh, N. Bard, F. Timbers, M. Lanctot, G. Z. Holland *et al.*, "Student of games: A unified learning algorithm for both perfect and imperfect information games", Science Advances **9**, 46, eadg3256 (2023b).

Shin, Y., J. Darbon and G. E. Karniadakis, "On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type pdes", arXiv preprint arXiv:2004.01806 (2020).

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm", arXiv preprint arXiv:1712.01815 (2017a).

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge", nature **550**, 7676, 354–359 (2017b).

Sokota, S., R. D'Orazio, J. Z. Kolter, N. Loizou, M. Lanctot, I. Mitliagkas, N. Brown and C. Kroer, "A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games", arXiv preprint arXiv:2206.05825 (2022).

Souquière, A., "Approximation and representation of the value for some differential games with asymmetric information", International Journal of Game Theory **39**, 699–722 (2010).

Souquière, A., "Approximation and representation of the value for some differential games with asymmetric information", International Journal of Game Theory **39**, 4, 699–722 (2010).

Tammelin, O., "Solving large imperfect information games using cfr+", arXiv preprint arXiv:1407.5042 (2014).

Trottenberg, U., C. W. Oosterlee and A. Schuller, *Multigrid* (Elsevier, 2000).

Van Dijk, M., A. Juels, A. Oprea and R. L. Rivest, "Flipit: The game of "stealthy takeover"", Journal of Cryptology **26**, 4, 655–713 (2013).

Vieillard, N., T. Kozuno, B. Scherrer, O. Pietquin, R. Munos and M. Geist, "Leverage the average: an analysis of kl regularization in reinforcement learning", Advances in Neural Information Processing Systems **33**, 12163–12174 (2020).

Zanette, A., A. Lazaric, M. J. Kochenderfer and E. Brunskill, "Limiting extrapolation in linear approximate value iteration", Advances in Neural Information Processing Systems **32** (2019).

Zhang, L., M. Ghimire, W. Zhang, Z. Xu and Y. Ren, "Value approximation for two-player general-sum differential games with state constraints", arXiv preprint arXiv:2311.16520 (2023).

Zheng, T., L. Zhu, A. M.-C. So, J. Blanchet and J. Li, "Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization", Advances in Neural Information Processing Systems **36**, 54075–54110 (2023).

Zinkevich, M., M. Johanson, M. Bowling and C. Piccione, "Regret minimization in games with incomplete information", Advances in neural information processing systems **20** (2007).

APPENDIX A

EXAMPLES OF ZERO-SUM GAMES WITH ONE-SIDED INFORMATION

Here we discuss two games in detail, namely, the zero-sum beer-quiche game which is extensive-form with sequential actions, and Hexner's game (Hexner, 1979) which is differential and with simultaneous actions. Both games have one-sided information. We show that the characterization of value proposed by Cardaliaguet (2007) leads to the equilibrium behavioral strategies for both games.

A.1 Zero-sum beer-quiche game

We present a zero-sum variant of the classic beer-quiche game[3], which is an incomplete-information game with a perfect Bayesian equilibrium.

Game settings. In this sequential game, Player 1 first chooses to take either quiche (Q) or beer (B), and based on his choice, Player 2 chooses to either defer (d) or bully (b). Player 1 has a probability of p_T to be tough (T) and $p_W = 1 - p_T$ to be weak (W). The exact type is unknown to Player 2 but $p = [p_T, p_W]^T$ is common knowledge. The payoffs to be maximized by Player 1 follow Table A.1. For example, if Player 1 is tough and chooses to eat quiche (Q) while Player 2 chooses to bully (b), then Player 1 receives a payoff of 1.

Table A.1. Payoff table for a zero-sum beer-quiche game

Tough

	b	d
B	2	1
Q	1	0

Weak

	b	d
B	-2	0
Q	-1	2

Perfect Bayesian equilibrium. The standard approach finds the behavioral strategies of both players for a particular p. Consider the extensive form of the game as shown in Fig. A.1. Dotted lines represent info sets that Player 2 cannot distinguish. Here, Player 1 has $p_T = \frac{1}{3}$ to be Tough. The behavioral strategies for each player are derived as follows:

Let Q, B, q, b represent probabilities of Player 1 choosing quiche given he is tough, beer given he is tough, quiche given he is weak, and beer given he is weak, respectively. Assume x and y be the probability of Player 2 bullying Player 1 who chooses quiche

[3]For more information, please see: https://gametheory101.com/courses/game-theory-101/the-beer-quiche-game/

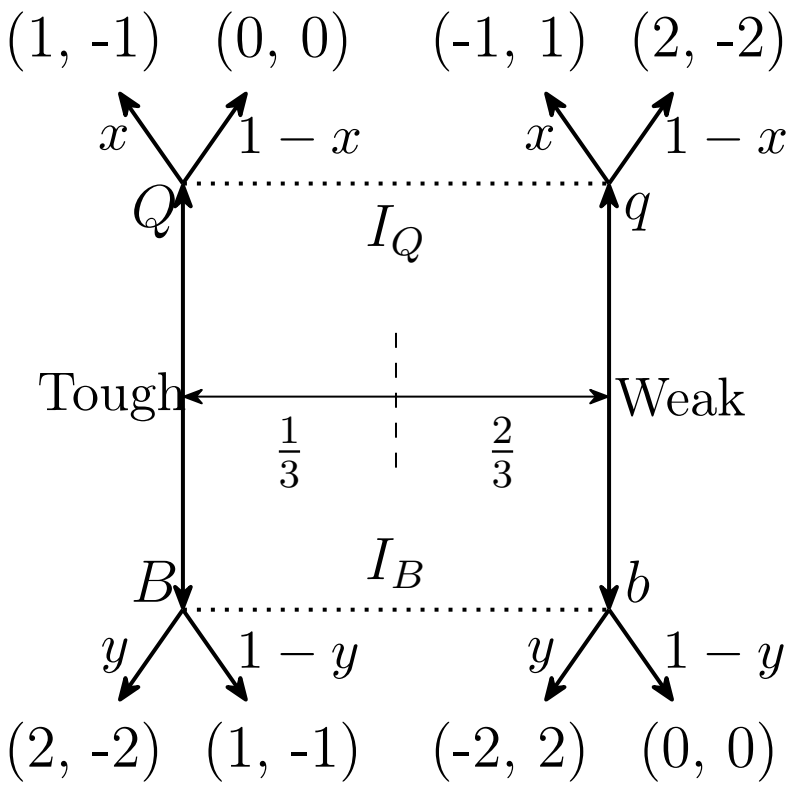


Figure A.1. Zero-Sum Variant of the Beer-Quiche Game

and beer, respectively. First, we find the belief for Player 2 on the info-set $\mathcal{I}_{\mathcal{Q}}$:

$$\text{if } (Q,q) \neq (0,0), \quad \mu_2(T|\mathcal{I}_{\mathcal{Q}}) = \frac{\frac{1}{3}Q}{\frac{1}{3}(Q) + \frac{2}{3}(q)} = \frac{Q}{Q+2q} \quad \text{and}$$

$$\mu_2(W|\mathcal{I}_{\mathcal{Q}}) = \frac{\frac{2}{3}q}{\frac{1}{3}Q + \frac{2}{3}q} = \frac{2q}{Q+2q}$$

$$\text{if } (B,b) \neq (0,0), \quad \mu_2(T|\mathcal{I}_{\mathcal{B}}) = \frac{B}{B+2b}, \quad \text{and} \quad \mu_2(W|\mathcal{I}_{\mathcal{B}}) = \frac{2b}{B+2b}$$

Then, the expected payoffs for Player 2 on $\mathcal{I}_{\mathcal{Q}}$ are:

$$E_2(\text{bully}|\mathcal{I}_{\mathcal{Q}}) = \frac{Q}{Q+2q}(-1) + \frac{2q}{Q+2q}(1) = -\frac{Q-2q}{Q+2q}$$

$$E_2(\text{defer}|\mathcal{I}_{\mathcal{Q}}) = \frac{Q}{Q+2q}(0) + \frac{2q}{Q+2q}(-2) = -\frac{4q}{Q+2q}$$

The expected payoff can be expressed as:

$$\begin{aligned} E_2(\mathcal{I}_{\mathcal{Q}}) &= -\frac{Q-2q}{Q+2q}x - (1-x)\frac{4q}{Q+2q} \\ &= \frac{-(Q-6q)x - 4q}{Q+2q} \end{aligned}$$

The value of x that maximizes $E_2(\mathcal{I}_{\mathcal{Q}})$ is:

$$x = \begin{cases} \text{any} & \text{if} \quad (Q,q) = (0,0) \\ 1 & \text{if} \quad Q < 6q \\ \text{any} & \text{if} \quad Q = 6q \\ 0 & \text{if} \quad Q > 6q \end{cases}$$

Following the same reasoning on info-set $\mathcal{I}_{\mathcal{B}}$, we find the value of y that maximizes $E_2(\mathcal{I}_{\mathcal{B}})$ as:

$$y = \begin{cases} \text{any} & \text{if} \quad (B,b) = (0,0) \\ 0 & \text{if} \quad 4b < B \\ \text{any} & \text{if} \quad 4b = B \\ 1 & \text{if} \quad 4b > B \end{cases}$$

The expected payoffs to Player 1 for his strategies are:

$$E_1(Q) = x, \quad E_1(B) = y + 1, \quad E_1(q) = 2 - 3x, \quad E_1(b) = -2y$$

Assume $B \geq 4b$.

Then,

$$\begin{aligned} & 1 - Q = B \geq 4b = 4(1-q) \\ \implies & 1 - Q \geq 4 - 4q \\ \implies & 4q \geq Q + 3 \\ \implies & 6q > Q \end{aligned}$$

Hence, $x = 1$. Thus, $E_1(Q) = 1 < E_1(B) = y + 1$.

As a result, $B = 1$, and $Q = 0$.

Assuming $B > 4b$, following the process as above, we reach to a contradiction. Therefore,

$$B = 4b \implies b = \frac{B}{4} = \frac{1}{4}$$

Then, For $b = 1/4$ to be feasible,

$$y = \frac{1}{2}$$

Therefore, we find an equilibrium with

$$x = 1,\ y = \frac{1}{2},\ q = \frac{3}{4},\ b = \frac{1}{4},\ \text{and } B = 1, Q = 0$$

To summarize, Player 2 always bullies the person who eats quiche and bullies the person drinking beer half the time. The tough guy always drinks beer while the weak guy drinks beer a quarter of the time and eats quiche three-quarters of the time. One can easily check that the $B < 4b$ case also leads to a contradiction, resulting in a unique equilibrium for the game.

Solution using primal and dual backward induction. Now we solve the game through backward induction of its primal and dual values (denoted by V and C respectively). Here we introduce discrete-time $t = 0, 1, 2$: Players 1 and 2 make their respective decisions at $t = 0$ and $t = 1$, and the game ends at $t = 2$. We describe the states of the game as the decisions being made up to the corresponding time, e.g., $x = (B, b)$ at $t = 2$ means that Player 1 has chosen beer and Player 2 to bully.

Primal game: At the terminal time step ($t = 2$), based on the payoff table, we have

$$V(2, x, p) = \begin{cases} 4p_T - 2 & \text{if } x = (B, b) \\ p_T & \text{if } x = (B, d) \\ 2p_T - 1 & \text{if } x = (Q, b) \\ 2 - 2p_T & \text{if } x = (Q, d) \end{cases}. \tag{A.1}$$

At the intermediate time step ($t = 1$), we have

$$V(1, x, p) = \min_{v \in \{b,d\}} V(2, (x, v), p). \tag{A.2}$$

We can find the best responses of Player 2 for both actions of Player 1. This leads to

$$V(1, x, p) = \begin{cases} p_T & \text{if } x = B,\ 3p_T - 2 \geq 0 \quad (v^* = d) \\ 4p_T - 2 & \text{if } x = B,\ 3p_T - 2 < 0 \quad (v^* = b) \\ 2 - 2p_T & \text{if } x = Q,\ 4p_T - 3 \geq 0 \quad (v^* = d) \\ 2p_T - 1 & \text{if } x = Q,\ 4p_T - 3 < 0 \quad (v^* = b) \end{cases}. \tag{A.3}$$

Note that since Player 1 does not take an action in this time step, we do not need to take a concave hull of $V(1, x, \cdot)$. At the beginning of the game ($t = 0$), we have

$$V(0, \emptyset, p) = \text{Cav}\left(\max_{u \in \{B,Q\}} V(1, u, p)\right). \tag{A.4}$$

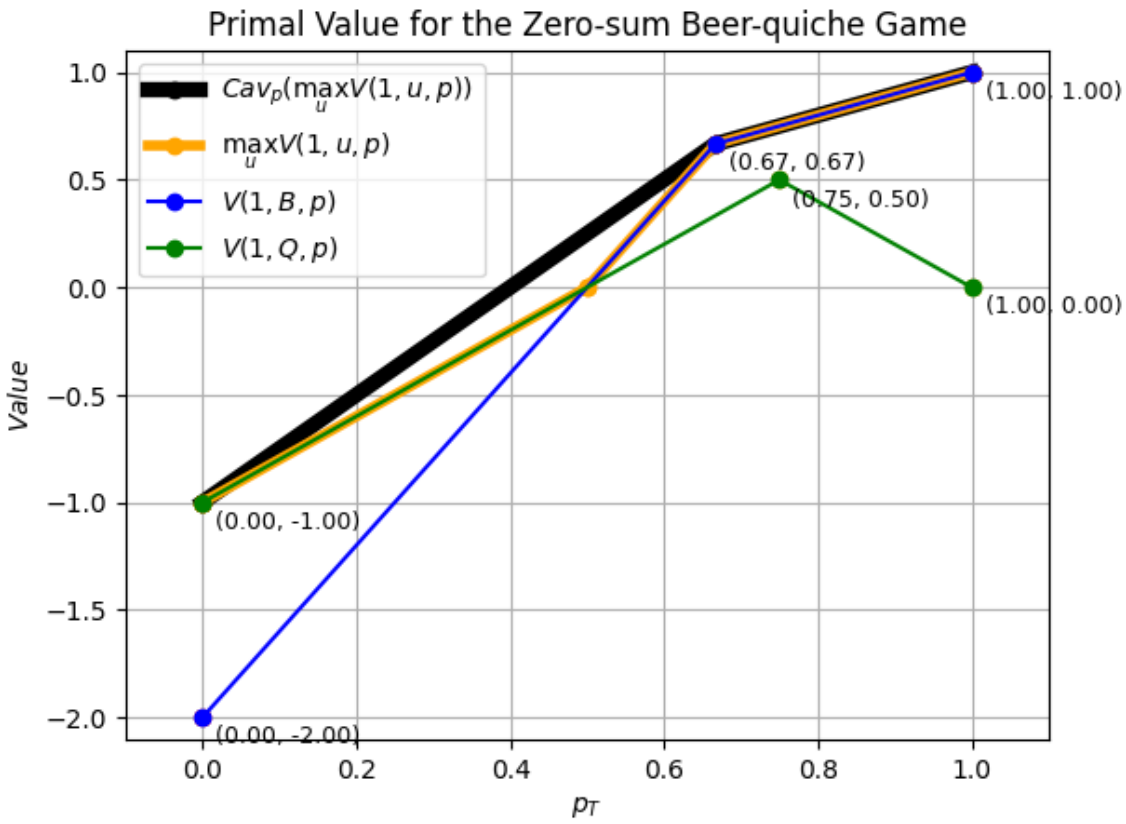


Figure A.2. Primal value $V(0, \emptyset, p_T)$ at $t = 0$.

By taking the concave hull (see Fig. A.2), we get

$$V(0,\emptyset,p) = \begin{cases} 5p_T/2 - 1 & \text{if } p_T < 2/3 \\ p_T & \text{if } p_T \geq 2/3 \end{cases}. \tag{A.5}$$

Note that from Fig. A.2, when $p_T \in [0, 2/3)$, $V(0,\emptyset,p) = \lambda \max_u V(1,u,p^1) + (1-\lambda)\max_u V(1,u,p^2)$, where $p^1 = [0,1]^T$, $p^2 = [2/3, 1/3]^T$, and $\lambda p^1 + (1-\lambda)p^2 = p$. When $p_T = 1/3$, $\lambda = 1/2$, Player 1's strategy is thus

$$\begin{aligned} &\Pr(u = Q|T) = \frac{\lambda p^1[1]}{p[1]} = 0, && \Pr(u = Q|W) = \frac{\lambda p^1[2]}{p[2]} = 3/4, \\ &\Pr(u = B|T) = \frac{(1-\lambda)p^2[1]}{p[1]} = 1, && \Pr(u = B|W) = \frac{(1-\lambda)p^2[2]}{p[2]} = 1/4. \end{aligned} \tag{A.6}$$

This result is consistent with the perfect Bayesian equilibrium.

Dual game: For Player 2, we first derive the dual variable $\hat{p} \in \partial_p V(0,\emptyset,p)$ for $p = [1/3, 2/3]^T$. By definition, $\hat{p}^T p$ defines the concave hull of $V(0,\emptyset,p)$, and therefore

we have

$$\begin{aligned}[1/3, 2/3]\hat{p} &= V(0, \emptyset, p) = -1/6 \\ [0, 1]\hat{p} &= V(0, \emptyset, [0, 1]) = -1.\end{aligned} \tag{A.7}$$

This leads to $\hat{p} = [3/2, -1]^T$.

At the terminal time, we have

$$\begin{aligned}C(2, x, \hat{p}) &= \min\{\hat{p}[1] - g_T(x), \hat{p}[2] - g_W(x)\} \\ &= \begin{cases} \min\{\hat{p}[1] - 2, \hat{p}[2] + 2\} & \text{if } x = (B, b) \\ \min\{\hat{p}[1] - 1, \hat{p}[2]\} & \text{if } x = (B, d) \\ \min\{\hat{p}[1] - 1, \hat{p}[2] + 1\} & \text{if } x = (Q, b) \\ \min\{\hat{p}[1], \hat{p}[2] - 2\} & \text{if } x = (Q, d) \end{cases}\end{aligned} \tag{A.8}$$

At $t = 1$, we have

$$C(1, u, \hat{p}) = \mathrm{Cav}_{\hat{p}} \left(\max_v C(2, (u, v), \hat{p}) \right). \tag{A.9}$$

When $u = B$, the conjugate value is a concave hull of a piece-wise linear function:

$$\begin{aligned}C(1, B, \hat{p}) &= \mathrm{Cav}_{\hat{p}} \left(\begin{cases} \hat{p}[1] - 1 & \text{if } \hat{p}[2] \geq \hat{p}[1] - 1 & (v^* = d) \\ \hat{p}[2] & \text{if } \hat{p}[2] \in [\hat{p}[1] - 2, \hat{p}[1] - 1) & (v^* = b) \\ \hat{p}[1] - 2 & \text{if } \hat{p}[2] \in [\hat{p}[1] - 4, \hat{p}[1] - 2) & (v^* = d) \\ \hat{p}[2] + 2 & \text{if } \hat{p}[2] < \hat{p}[1] - 4 & (v^* = b) \end{cases} \right) \\ &= \begin{cases} \hat{p}[1] - 1 & \text{if } \hat{p}[2] \geq \hat{p}[1] - 1 & (v^* = d) \\ 2/3\hat{p}[1] + 1/3\hat{p}[2] - 2/3 & \text{if } \hat{p}[2] \in [\hat{p}[1] - 4, \hat{p}[1] - 1) & (\text{mixed strategy}) \\ \hat{p}[2] + 2 & \text{if } \hat{p}[2] < \hat{p}[1] - 4 & (v* = b) \end{cases}\end{aligned} \tag{A.10}$$

Fig. A.3 visualizes $C(1, B, \hat{p})$. For $\hat{p} = [3/2, -1]^T$ which satisfies $\hat{p}[2] \in [\hat{p}[1] - 4, \hat{p}[1] - 1)$, Player 2 follows a mixed strategy determined based on $\{\lambda_1, \lambda_2, \lambda_3\} \in \Delta(3)$ and $\hat{p}^j \in \mathbb{R}^2$ for $j = 1, 2, 3$ such that

(i) At least one of $\hat{p}^j$ for $j = 1, 2, 3$ should satisfy $\hat{p}[2] = \hat{p}[1] - 1$ (denoted as line 1) and another $\hat{p}[2] = \hat{p}[1] - 4$ (denoted as line 2). The last could be on either line 1 or 2. These conditions are necessary for $C(1, B, \hat{p})$ to be a concave hull:

$$C(1, B, \hat{p}) = \sum_{j=1}^{3} \lambda_j \max_v C(2, (B, v), \hat{p}^j). \tag{A.11}$$

(ii) Without loss of generality, we will set $\hat{p}^1$ on line 1 and both $\hat{p}^2$ and $\hat{p}^3$ on line 2; $\sum_{j=1}^{3} \lambda_j \hat{p}^j = \hat{p}$.

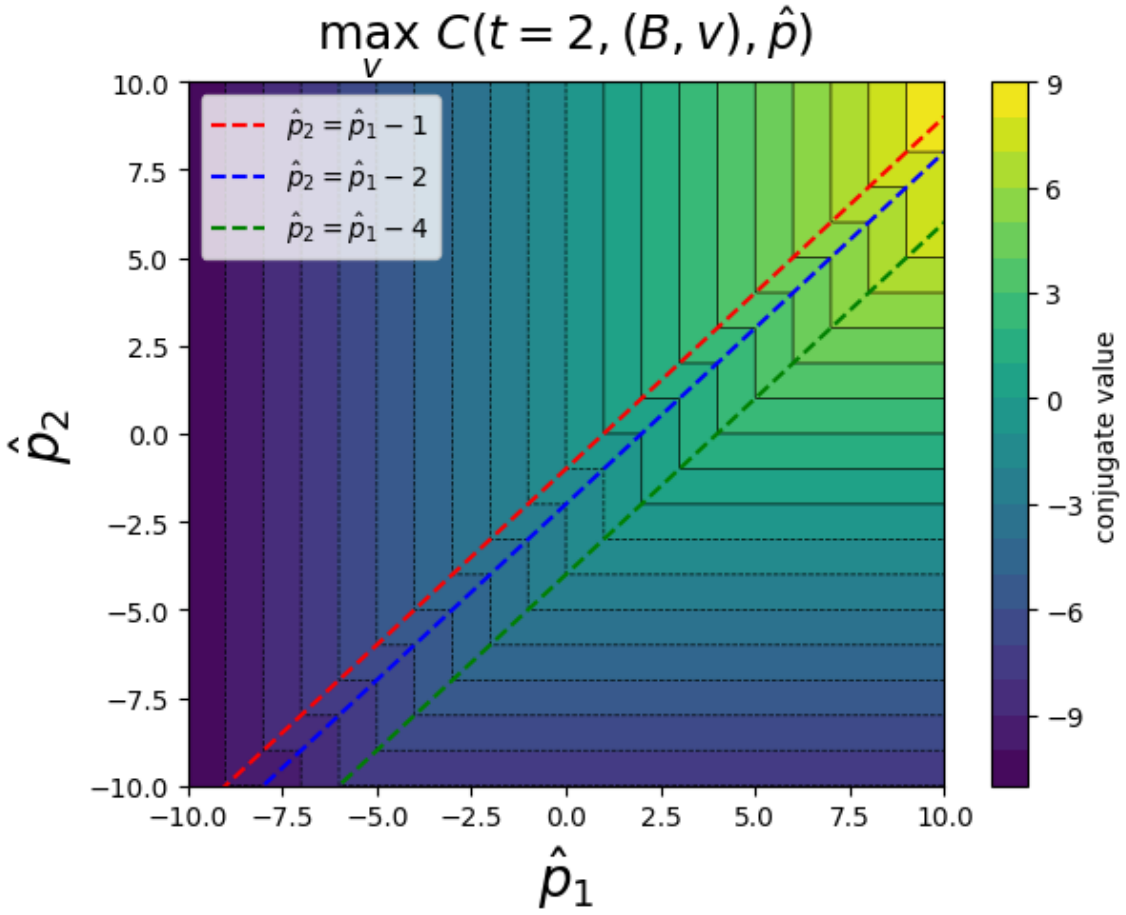


Figure A.3. Conjugate value $\max_v\ C(2, B, \hat{p})$ at $t = 2$.

These conditions leads to $\lambda_1 = 1/2$ and $\lambda_2 + \lambda_3 = 1/2$. Therefore Player 2 chooses to defer and bully with equal chance when Player 1 takes beer.

When $u = Q$, we similarly have

$$C(1, Q, \hat{p}) = \begin{cases} \hat{p}[1] & \text{if } \hat{p}[2] \geq \hat{p}[1] + 2 & (v^* = d) \\ \dots & \text{if } \hat{p}[2] \in [\hat{p}[1] - 2, \hat{p}[1] + 2) & (\text{mixed strategy}) \\ \hat{p}[2] + 1 & \text{if } \hat{p}[2] < \hat{p}[1] - 2 & (v* = b) \end{cases} \tag{A.12}$$

We omitted the derivation of the concave hull when $\hat{p}[2] \in [\hat{p}[1] - 2, \hat{p}[1] + 2)$ because for $\hat{p} = [3/2, -1]^T$, $C(1, Q, \hat{p}) = \hat{p}[2] + 1 = 0$ while $v^* = b$, i.e. if Player 1 takes quiche, Player 2 chooses to bully with certainty.

The value and its conjugate provide behavioral strategies for Player 1 (informed) and Player 2 (non-informed), respectively, for arbitrary initial belief p. Moreover, the convex nature of the value reveals subsets of p where Player 1 should use a mixed strategy that manipulates the belief in order to improve its value. Similarly, the

convex nature of the conjugate value reveals subsets of dual variables $\hat{p}$ where Player 2 should use a mixed strategy to mitigate risks due to its uncertainty about Player 1.

A.2 Hexner's game

A.2.0.1 Game Settings

The players move in an arena bounded between $[-1, 1]$ in all directions. All games in this work follow 2D/3D point dynamics as follows: $\dot{x}_j = Ax_j + Bu_j$, where x_j is a vector of position and velocity and u_j is the action for player j. Note that we use u and v in the optimization problems P_1 and P_2 to represent player 1 and player 2's actions respectively. The type independent effort loss for each player j is defined as $l_j(u_j) = u_j^\top R_j u_j$, where $R_1 = \texttt{diag}(0.05, 0.025)$ and $R_2 = \texttt{diag}(0.05, 0.1)$. For the higher dimensional case, $R_1 = \texttt{diag}(0.05, 0.05, 0.025)$ and $R_2 = \texttt{diag}(0.05, 0.05, 0.1)$. Note that, in the incomplete information case, P1 is able to get better payoff by hiding the target because P2 incurs higher effort cost, and hence cannot accelerate as fast as P1.

A.2.0.2 Analytical Solution

For the unconstrained simplified football game discussed in Sec. 2.6, the strategies depend on the trajectory of the $d_1 - d_2$. In Fig.A.4, we plot the trajectory and determine the critical time from Eq.(A.17). For the choices of parameters, we determine $t_r = 0.5$s. We set $R_A = \texttt{diag}(0.05, 0.025)$, and $R_D = \texttt{diag}(0.05, 0.1)$.

Concretely, the analytical solution to the equilibrium policies is the solution to respective player's HJB:

$$u_j = -R_j^{-1} B_j^\top K_j x_j + R_j^{-1} B_j^\top K_j \Phi_j z \tilde{\theta}_j,$$

based on the reformulation outlined below in which players' action $\tilde{\theta}_j \in \mathbb{R}$ become 1D and are decoupled from the state: where Φ_j is a state-transition matrix that solves $\dot{\Phi}_j = A_j \Phi_j$, with $\Phi_j(T)$ being an identity matrix, and K_j is a solution to a continuous-time differential Ricatti equation:

$$\dot{K}_j = -A_j^\top K_j - K_j A_j + K_j^\top B_j R_j^{-1} B_j^\top K_j, \tag{A.13}$$

Finally, by defining

$$d_j = z^\top \Phi_j^\top K_j B_j R_j^{-1} B_j^\top K_j^\top \Phi_i z$$

and the critical time

$$t_r = \arg\min_t \int_0^t (d_1(s) - d_2(s)) ds$$

and

$$\tilde{\theta}_j(t) = \begin{cases} 0, & t \in [0, t_r] \\ \theta, & t \in (t_r, T] \end{cases}.$$

As explained in Sec.3.6, P1 chooses $\theta_1 = 0$ until the critical time t_r and P2 follows.

Note that in order to compute the ground truth when time is discretized with some τ, we need the discrete counterpart of equation A.13, namely the discrete-time Ricatti difference equation and compute the matrices K recursively.

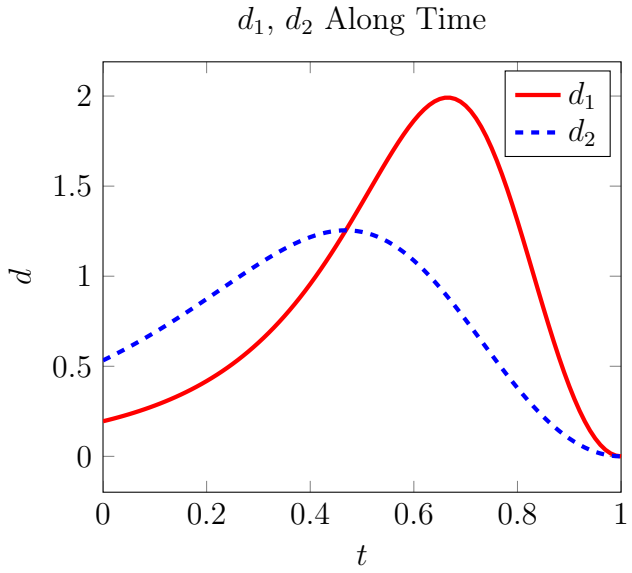


Figure A.4. Plot of d_1 and d_2 along time. The critical time t_r occurs at $t \approx 0.4$. The attacker will conceal its type until t_r and reveal it after t_r.

A.2.0.3 Solution Using Subdynamic Principle

Now we discuss the solution using Cardaliaguet's method based on the reformulation by Hexner. To recap, the payoff to be minimized by Player 1 is

$$J(t, \tilde{\theta}_1, \tilde{\theta}_2) = \mathbb{E}_\theta \left[\int_{\tau=t}^{T} (\tilde{\theta}_1(\tau) - \theta)^2 d_1(\tau) - (\tilde{\theta}_2(\tau) - \theta)^2 d_2(\tau) d\tau \right], \tag{A.14}$$

where d_1, d_2, p_θ are common knowledge; θ is only known to Player 1; $\tilde{\theta}_1$ (resp. $\tilde{\theta}_2$) is Player 1's (resp. Player 2's) strategy. We consider two player types $\theta \in \{-1, 1\}$ and

therefore $p_\theta \in \Delta(2)$. Since the reformulation contains no system state, the strategies are functions of only time. Hexner's solution is as follows:

$$\tilde{\theta}_1(s) = \tilde{\theta}_2(s) = 0 \quad \forall s \in [0, t_r] \tag{A.15}$$

$$\tilde{\theta}_1(s) = \tilde{\theta}_2(s) = \theta \quad \forall s \in (t_r, T], \tag{A.16}$$

where

$$t_r = \arg\min_t \int_0^t (d_1(s) - d_2(s))ds, \tag{A.17}$$

and (d_1, d_2) are defined in Eq. (2.2).

We will need the following preparation before introducing Cardaliaguet's solution. First, introduce time stamps $[T_k]_{k=1}^{2r}$ as roots of $d_1 - d_2$, with $T_0 = 0$ and $T_{2r+1} = t_r$. Without loss of generality, we assume that:

$$d_1 - d_2 < 0 \quad \forall t \in (T_{2k}, T_{2k+1}) \; \forall k = 0, ..., r, \tag{A.18}$$

$$d_1 - d_2 \geq 0 \quad \forall t \in [T_{2k-1}, T_{2k}] \; \forall k = 1, ..., r. \tag{A.19}$$

We also introduce $D_k := \int_{T_k}^{T_{k+1}} (d_1 - d_2)ds$ and

$$\tilde{D}_k = \begin{cases} \tilde{D}_{k+1} + D_k & \text{if } \tilde{D}_{k+1} + D_k < 0 \\ 0 & \text{otherwise} \end{cases}, \tag{A.20}$$

with $\tilde{D}_{2r+1} = 0$.

Lemma A.1. *(Properties of D_k and $\tilde{D}_k$) The following properties will be useful:*

1. $\int_k^{2q+1} (d_1 - d_2)ds = \sum_k^{2q} D_k < 0$, $\forall k = 0, ..., 2q$;
2. $\int_{2q+1}^{k} (d_1 - d_2)ds = \sum_{2q+1}^{k-1} D_k > 0$, $\forall k = 2q+2, ..., 2r+1$;
3. $\tilde{D}_{2q+2} + D_{2q+1} > 0$;
4. $\tilde{D}_k < 0, \; \forall k < 2q+1$.

Proof. Properties 1 and 2 are results directly from the definition of D_k.

For property 3, if $\tilde{D}_{2q+2} + D_{2q+1} \leq 0$, then $\tilde{D}_{2q+2} = \tilde{D}_{2q+3} + D_{2q+2} \leq -D_{2q+1}$, then $\tilde{D}_{2q+3} \leq -(D_{2q+2} + D_{2q+1}) < 0$ (property 2). This leads to $\tilde{D}_{2q+k} \leq -\sum_{i=1}^{k-1} D_{2q+i} < 0$ for $k = 1, ..., 2r - 2q$. Thus $\tilde{D}_{2r} < 0$. Contradiction.

For property 4, first we have $\tilde{D}_{2q+1} = 0$ (property 3). Since $D_{2q} < 0$ (property 1), $\tilde{D}_{2q} = D_{2q} < 0$. □

Primal game. We start with $V(T, p) = 0$ where $p := p_\theta[1]$ is the probability of $\theta = -1$. Let the Hamiltonian be

$$\begin{aligned} H(p) &= \min_{\tilde{\theta}_1} \max_{\tilde{\theta}_2} \mathbb{E}_\theta \left[(\tilde{\theta}_1 - \theta)^2 d_1 - (\tilde{\theta}_2 - \theta)^2 d_2 \right] \\ &= 4p(1-p)(d_1 - d_2). \end{aligned}$$

We note that the optimal actions to achieve the above Hamiltonian are $\tilde{\theta}_1 = \tilde{\theta}_2 = 1-2p$. From Bellman backup, we can get

$$V(T_k, p) = 4p(1-p)\tilde{D}_k.$$

Therefore, at T_{2q+1}, we have

$$\begin{aligned} V(T_{2q+1}, p) &= Vex_p \left(V(T_{2q+2}, p) + 4p(1-p)D_{2q+1}\right) \\ &= Vex_p \left(4p(1-p)(\tilde{D}_{2q+2} + D_{2q+1})\right). \end{aligned}$$

Notice that $\tilde{D}_{2q+2} + D_{2q+1} > 0$ (property 3) and $\tilde{D}_k < 0$ for all $k < 2q+1$ (property 4), T_{2q+1} is the first time such that the right-hand side term inside the convexification operator, i.e., $4p(1-p)(\tilde{D}_{2q+2} + D_{2q+1})$, becomes concave. Therefore, splitting of belief happens at T_{2q+1} with $p^1 = 0$ and $p^2 = 1$. Player 1 plays $\tilde{\theta}_1 = -1$ (resp. $\tilde{\theta}_1 = 1$) with probability 1 if $\theta = -1$ (resp. $\theta = 1$), i.e., Player 1 reveals its type. This result is consistent with Hexner's.

Dual game. We can show that the conjugate value follows

$$C(t, \hat{p}) = \begin{cases} \max_{i\in\{1,2\}} \hat{p}[i] & \forall t \geq T_{2q+1} \\ \hat{p}[2] - \tilde{D}_t \left(1 - \frac{\hat{p}[1]-\hat{p}[2]}{4\tilde{D}_t}\right)^2 & \forall t < T_{2q+1},\ 4\tilde{D}_t \leq \hat{p}[1] - \hat{p}[2] \leq -4\tilde{D}_t \\ \hat{p}[1] & \forall t < T_{2q+1},\ \hat{p}[1] - \hat{p}[2] \geq 4\tilde{D}_t \\ \hat{p}[2] & \forall t < T_{2q+1},\ \hat{p}[1] - \hat{p}[2] < 4\tilde{D}_t \end{cases}$$

Here $\hat{p} \in \nabla_{p_\theta} V(0, p_\theta)$ and $V(0, p_\theta) = 4p[1]p[2]\tilde{D}_0$. For any particular $p_* \in \Delta(2)$, from the definition of subgradient, we have $\hat{p}[1]p_*[1] + \hat{p}[2]p_*[2] = 4p_*[1]p_*[2]\tilde{D}_0$ and $\hat{p}[1] - \hat{p}[2] = 4(p_*[2] - p_*[1])\tilde{D}_0$. Solving these to get $\hat{p} = [4p_*[2]^2\tilde{D}_0, 4p_*[1]^2\tilde{D}_0]^T$. Therefore $\hat{p}[1] - \hat{p}[2] = 4\tilde{D}_0(1 - 2p_*[1]) \in [4\tilde{D}_0, -4\tilde{D}_0]$, and

$$C(0, \hat{p}) = \hat{p}[2] - \tilde{D}_0 \left(1 - \frac{\hat{p}[1] - \hat{p}[2]}{4\tilde{D}_0}\right)^2.$$

Notice that $C(t, \hat{p})$ is convex to $\hat{p}$ since $\tilde{D}_0 < 0$ (property 4) for all $t \in [0, T]$. Therefore, there is no splitting of $\hat{p}$ during the dual game, i.e., $\tilde{\theta}_2 = 1 - 2p$. This result is also consistent with Hexner's.

APPENDIX B

PREDICTION ERROR OF VALUE APPROXIMATION

Here we show that CAMS shares the same exponential error propagation as in standard approximate value iteration (AVI). The only difference from AVI is that the measurement error in CAMS comes from numerical approximation of the minimax problems rather than randomness in state transition and rewards. To start, let the true value be $V(t,x,p)$. Following Zanette *et al.* (2019), the prediction error $\epsilon_t^{bias} := \max_{x,p} |\hat{V}_t(x,p) - V(t,x,p)|$ is affected by (1) the prediction error $\epsilon_{t+\tau}^{bias}$ propagated back from $t+\tau$, (2) the minimax error ϵ_t^{minmax} caused by limited iterations in solving the minimax problem at each collocation point: $\epsilon_t^{minmax} = \max_{(x,p)\in\mathcal{S}_t} |\tilde{V}(t,x,p) - V(t,x,p)|$, and (3) the approximation error due to the fact that $V(t,\cdot,\cdot)$ may not lie in the model hypothesis space $\mathcal{V}_t$ of $\hat{V}_t$: $\epsilon_t^{app} = \max_{x,p} \min_{\hat{V}_t\in\mathcal{V}_t} |\hat{V}_t(x,p) - V(t,x,p)|$.

Approximation error. For simplicity, we will abuse the notation by using x in place of (x,p) and omit time dependence of variables when possible. In practice we consider $\hat{V}_t$ as neural networks that share the architecture and the hypothesis space. Note that $\hat{V}_T(\cdot) = V(T,\cdot)$ is analytically defined by the boundary condition and thus $\epsilon_T^{app} = \epsilon_T^{bias} = 0$. To enable the analysis on neural networks, we adopt the assumption that $\hat{V}$ is infinitely wide and that the resultant neural tangent kernel (NTK) is positive definite. Therefore from NTK analysis (Jacot *et al.*, 2018), $\hat{V}$ can be considered a kernel machine equipped with a kernel function $r(x^{(i)}, x^{(j)}) := \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$ defined by a feature map $\phi : \mathcal{X} \to \mathbb{R}^{d_\phi}$. Given training data $\mathcal{S} = \{(x^{(i)}, V^{(i)})\}$, let $r(x)[i] := r(x^{(i)}, x)$, $R_{ij} := r(x^{(i)}, x^{(j)})$, $V_\mathcal{S} := [V^{(1)}, ..., V^{(N)}]^\top$, $\Phi_\mathcal{S} := [\phi(x^{(1)}), ..., \phi(x^{(N)})]$, and $w_\mathcal{S} := \Phi_\mathcal{S}(\Phi_\mathcal{S}^\top \Phi_\mathcal{S})^{-1} V_\mathcal{S}$ be model parameters learned from $\mathcal{S}$, then

$$\hat{V}(x) = r(x)^\top R^{-1} V_\mathcal{S} = \langle \phi(x), w_\mathcal{S} \rangle \tag{B.1}$$

is a linear model in the feature space. Let $\theta^{\phi(x)} := r(x)^\top R^{-1}$ and $C := \max_x \|\theta^{\phi(x)}\|_1$. Further, let $\mathcal{S}^\dagger := \arg\min_\mathcal{S} |\langle \phi(x), w_\mathcal{S} \rangle - V(x)|$ and $w^\dagger := w_{\mathcal{S}^\dagger}$, i.e., $w^\dagger$ represents the best hypothetical model given sample size N. Since N is finite, the data-dependent hypothesis space induces an approximation error $\epsilon_t^{app} := \max_x |\langle \phi(x), w^\dagger \rangle - V(x)|$. From standard RKHS analysis, we have $\epsilon_t^{app} \propto N^{-\frac{1}{2}}$.

Error propagation. Recall that we approximately solve P_1 at each collocation point. Let $z := \{\lambda, p, u, v\}$ be the collection of variables and $\tilde{z}$ be the approximated saddle point resulting from DS-GDA. Let $\tilde{V}(t,x,\tilde{z})$ be the approximate value at (t,x) and let $V(t,x,z^*)$ be the value at the true saddle point z^*. Lemma B.1 bounds the error of $\tilde{V}(t,x,\tilde{z})$:

Lemma B.1. $\max_x |\tilde{V}(t,x,\tilde{z}) - V(t,x,z^*)| \le \epsilon_{t+\tau}^{bias} + \epsilon_t^{minmax}$.

Proof. Note that $\sum_{k=1}^{I} \lambda^k = 1$. Then

$$\begin{aligned}\max_x |\tilde{V}(t,x,\tilde{z}) - V(t,x,z^*)| &\leq \max_x |\tilde{V}(t,x,\tilde{z}) - \tilde{V}(t,x,z^*)| \\ &\quad + \max_x |\tilde{V}(t,x,z^*) - V(t,x,z^*)| \\ &\leq \epsilon_t^{minmax} + \max_x |\sum_{k=1}^{I} \lambda^k (\tilde{V}(t+\tau,x',p^k) \\ &\quad - V(t+\tau,x',p^k)| \\ &\leq \epsilon_t^{minmax} + \epsilon_{t+\tau}^{bias}.\end{aligned} \tag{B.2}$$

□

Now we can combine this measurement error with the inherent approximation error ϵ_t^{app} to reach the following bound on the prediction error ϵ_t^{bias}:

Lemma B.2. $\max_x |\hat{V}_t(x) - V(t,x)| \leq C_t(\epsilon_t^{minmax} + \epsilon_{t+\tau}^{bias} + \epsilon_t^{app}) + \epsilon_t^{app}$.

Proof.

$$\begin{aligned}&\max_x |\hat{V}_t(x) - V(t,x)| \leq \max_x |\hat{V}_t(x) - \langle \phi(x), w^\dagger \rangle| + \max_x |\langle \phi(x), w^\dagger \rangle - V(t,x)| \\ &\leq \max_x |\langle \theta^{\phi(x)}, \tilde{V}(t,x) - V(t,x) \rangle| + \max_x |\langle \theta^{\phi(x)}, V(t,x) - \phi(x)^\top w^\dagger \rangle| + \epsilon_t^{app} \\ &\leq C(\epsilon_t^{minmax} + \epsilon_{t+\tau}^{bias} + \epsilon_t^{app}) + \epsilon_t^{app}.\end{aligned} \tag{B.3}$$

□

Lem. B.3 characterizes the propagation of error:

Lemma B.3. *Let $\epsilon_t^{app} \leq \epsilon^{app}$, $\epsilon_t^{minmax} \leq \epsilon^{minmax}$, and $C_t \leq C$ for all $t \in [T]$. If $\epsilon_T^{app} = 0$, then $\epsilon_0^{bias} \leq TC^T(\epsilon^{app} + C(\epsilon^{minmax} + \epsilon^{app}))$.*

Proof. Using Lem. B.2 and by induction, we have

$$\epsilon_0^{bias} \leq (\epsilon^{app} + C(\epsilon^{minmax} + \epsilon^{app}))\frac{1 - C^T}{1 - C} \leq TC^T(\epsilon^{app} + C(\epsilon^{minmax} + \epsilon^{app})). \tag{B.4}$$

□

We can now characterize the computational complexity of the baseline algorithm through Thm. B.1, by taking into account the number of DS-GDA iterations and the per-iteration complexity:

Theorem B.1. *For a fixed T and some error threshold $\delta > 0$, with a computational complexity of $\mathcal{O}(T^3 C^{2T} I^2 \epsilon^{-4} \delta^{-2})$, CAMS achieves*

$$\max_{(x,p) \in \mathcal{X} \times \Delta(I)} |\hat{V}_0(x,p) - V(0,x,p)| \leq \delta. \tag{B.5}$$

Proof. From Lem. B.3 and using the fact that $\epsilon^{app} \propto N^{-1/2}$, achieving a prediction error of δ at $t = 0$ requires $N = \mathcal{O}(C^{2T}T^2\delta^{-2})$. CAMS solves TN minimax problems, each requires a worst-case $\mathcal{O}(\epsilon^{-4})$ iterations, and each iteration requires computing gradients of dimension $\mathcal{O}(I^2)$, considering the dimensionalities of action spaces as constants. This leads to a total complexity of $\mathcal{O}(T^3C^{2T}I^2\epsilon^{-4}\delta^{-2})$. □

APPENDIX C

IMPLEMENTATION DETAILS

C.1 Data Sampling

Unconstrained Game. For the unconstrained game, we sample positions (d_x, d_y) and velocities $(\dot{d}_x, \dot{d}_y)$ for both players. As the arena is bounded between $[-1, 1]$ in both x and y directions, we sample the positions of the two players in $[-1, 1]$. However, when it comes to velocities, we experimentally determine the range from the LQR problem as the following: $\dot{d}_{x_1} \in [-6, 6]$, $\dot{d}_{y_1} \in [-6, 6]$, $\dot{d}_{x_2} \in [-4, 4]$, and $\dot{d}_{y_2} \in [-4, 4]$. We then normalize the velocities between $[-1, 1]$ and compute the values as described in algorithm 1. The resulting normalized joint states $(\mathcal{X})$ and values (V) are stored for training the value network. At each time step we sample 10000 states and set $|\mathcal{P}| = 100$. This brings the total training data at each time step to 1M for the unconstrained case.

Constrained Primal Game. For the constrained game, we sample the positions between $[-1, 1]$ and all velocities between the ranges discussed above. As in the unconstrained case, these are normalized to $[-1, 1]$ before computing the values and storing the training data. With the same $\mathcal{P}$, we sample 5000 states from the feasible set $\mathcal{Q}(t)$, resulting in 500,000 training data at each time step. Solving constrained game requires evaluating $\min_u \max_v V(t, x + \tau f(x, u, v), p)$ over all possible pairs of x'(i.e. $x + \tau f(x, u, v)$), which is memory intensive. Based on the available resources we set the total number of initial states to be sampled to 5000. To speed up the calculation, and capture a wide range of data, we divide the state space into 50 uniform intervals, and distribute the computation to 56 CPU cores, with 515,271 MB of total memory. Each minimax computation is independent and hence can be evaluted in parallel.

Constrained Dual Game. In the dual game, the uninformed player keeps track of the process $\hat{p} \in \mathbb{R}^I$. As a result, the dual value is a 10-D function, which increases the complexity of the computation due to the need for convexification of the value along I dimensions (here, $I = 2$). We follow the same procedure as in the primal game and collect 100,000 samples for training. The range of $\hat{p}$ was determined to be $[-14, 14]$ based on the primal value at the initial time as discussed in Sec. 2.5. Furthermore, due to the additional input dimension in the dual value network, the dual value approximation suffers from relatively higher error compared to the primal value. Ultimately this affects the strategy of the uninformed player (Player 2). We compare the resulting strategy of Player 2 from the dual value with that of the ground truth strategy in the unconstrained game.

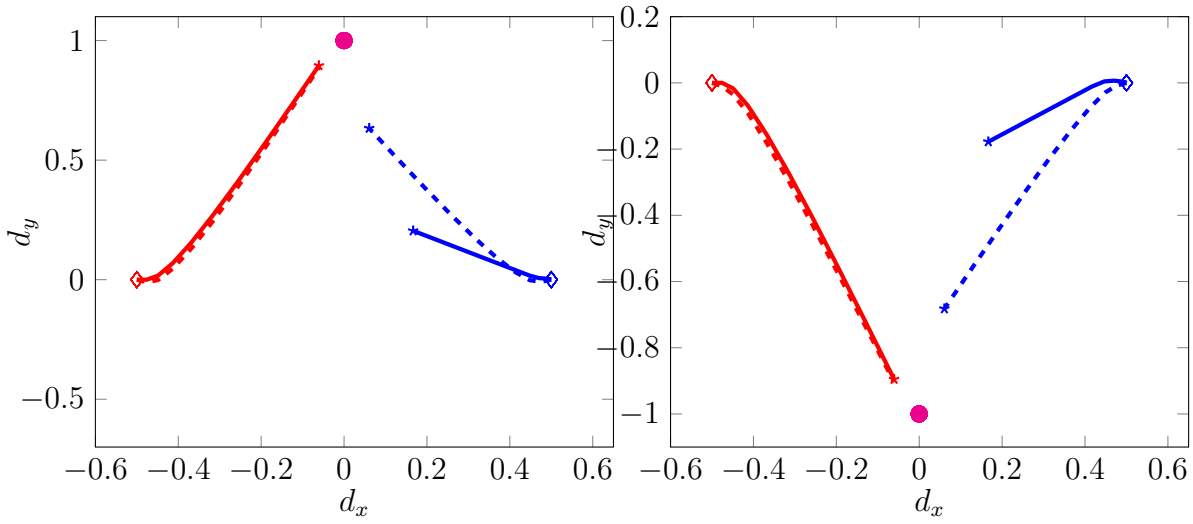


Figure C.1. Comparison between the P2's ground truth strategy and the strategy synthesized from the dual value. P1's trajectory is shown red and P2's in blue. Solid trajectories correspond to that obtained when P2 plays its equilibrium strategy. Dotted trajectories represent the ground truth solution.

C.2 Backward Reachable Tube

The computation of the Backward Reachable Tube (BRT) allows us to classify the state space into feasible and infeasible regions at different times from Player 1's perspective.

Computation of BRT. For the simplified football game, the state constraint is defined as $c(x) := \|(d_{x_1}, d_{y_1}) - (d_{x_2}, d_{y_2})\|_2 - r$, and $\mathcal{C} = \{x : c(x) \leq 0\}$. The Hamilton-Jacobi-Isaacs Variational Inequality (HJI VI) is denoted by L and satisfies the boundary condition D Bansal and Tomlin (2021):

$$\begin{aligned} L(\tilde{V}, t, x) &= \min\{\nabla_t \tilde{V}(t,x) + H(t,x), c(x) - \tilde{V}(t,x)\} = 0, \\ D(\tilde{V}, x) &= \tilde{V}(T,x) - c(x) = 0, \end{aligned} \tag{C.1}$$

where H is the Hamiltonian:

$$H(t,x) = \max_u \min_v \langle \nabla_x \tilde{V}(t,x), f(x,u,v) \rangle. \tag{C.2}$$

We use Physics-Informed Neural Network (PINN) to learn the value function $\tilde{V}(t,x)$, the sub-zero level set of which represents the BRT:

$$\bar{\mathcal{Q}}(t) = \{x \in \mathbb{R}^{d_x} : \tilde{V}(t,x) \leq 0\}. \tag{C.3}$$

We denote PINN dataset $\mathcal{D} = \left\{\left(t^{(k)}, x^{(k)}\right)\right\}_{k=1}^{K}$ containing uniformly sampled data points in $[0, T] \times \mathbb{R}^{d_x}$ and define the loss function as:

$$\min_{\tilde{V}} \quad \mathcal{L}\left(\tilde{V}\right) = \sum_{k=1}^{K} \left\| L(\tilde{V}^{(k)}, t^{(k)}, x^{(k)}) \right\|_1 + C_1 \left\| D(\tilde{V}^{(k)}, x^{(k)}) \right\|_1, \tag{C.4}$$

where $\tilde{V}^{(k)}$ is an abbreviation for $\tilde{V}\left(t^{(k)}, x^{(k)}\right)$ and C_1 is the hyperparameter that balances the loss term $\|L\|_1$ and $\|D\|_1$.

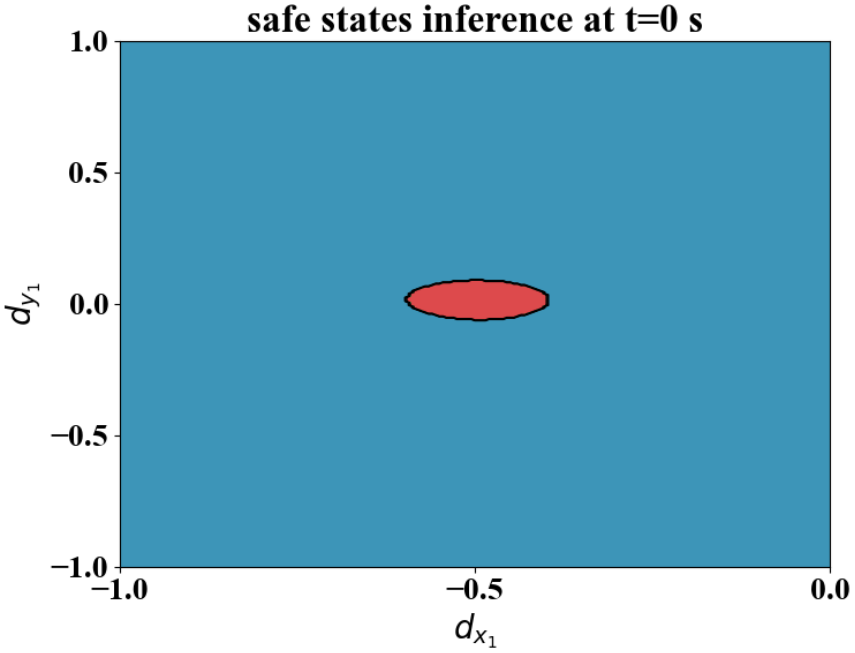


Figure C.2. A visualization of safe/unsafe initial position for the attacker when the defender is fixed at $(-0.05, 0)$. The initial velocities for both players are zero in the beginning. The red (blue) region represents unsafe (safe) states.

Training. We uniformly sample 60k input states $x \in [-1, 1]$ (specifically, 10k states $x \in \mathcal{C}$) and use curriculum learning proposed in Bansal and Tomlin (2021) to improve the training convergence. The rest of the dynamics parameters are chosen as: $T = 1$, $r = 0.05$, $u_x \in [-6, 6]$, $u_y \in [-12, 12]$, $v_x \in [-6, 6]$, $v_y \in [-4, 4]$, velocities are sampled as mentioned in Sec. C.1 and are normalized between $[-1, 1]$. The PINN utilizes a fully-connected network with 3 hidden layers, each comprising 512 neurons with `sin` activation function. The network adopts the Adam optimizer with a fixed learning rate of 2×10^{-5}. We first pretrain the network over 10k iterations to satisfy

the boundary condition D and then refine the network through 100k gradient descent steps, with states sampled from an expanding time window starting from the terminal. Fig. C.2 shows the visualization of BRT in a 2D frame given $t = 0$ and fixed states except (d_{x_1}, d_{y_1}).

C.3 Connection between Primal and Dual Games

Here we use the infinitely-repeated game setting to explain the connection between the primal and dual games and the interpretation of the dual variable $\hat{p}$. Please see Theorem 2.2 in De Meyer (1996) and the extension to differential games in Cardaliaguet (2007).

C.3.1 Game setting.

In an infinitely-repeated normal-form game with one-sided information, there is a set of I possible payoff tables $\{A_i\}_{i=1}^I$. A table is initially drawn from p_0 and shown to P1, while P2 only knows p_0. At stage $t \in [T]$, the players draw actions from their strategies η_i and ζ, respectively, under the current belief p_t. The actions are public, which trigger belief updates, yet the actual payoffs are hidden (or otherwise P2 may infer the true payoff easily). At the terminal T, an average payoff is received by P1 from P2. This game is proven to have a value (Aumann *et al.*, 1995) for finite and infinite T. Here we only consider the latter case where the value is defined as

$$V(p_0) = \max_{\{\eta_i\}_{i=1}^I} \min_{\zeta} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{i\sim p_0}[\eta_i^\top A_i \zeta]. \tag{C.5}$$

Notice that this is a special case of the differential games we are interested in, where the incomplete information is on the running loss, the action spaces are discrete, and the dynamics is identity.

C.3.2 Primal-dual properties.

Let the primal game be $G(p)$ for $p \in \Delta(I)$, the dual game be $G^*(\hat{p})$ for $\hat{p} \in \mathbb{R}^I$, and let $\{\eta_i\}_{i=1}^I$ be the set of strategies for P1 and ζ the strategy for P2. $\eta_i \in \Delta(d_u)$ and $\zeta \in \Delta(d_v)$. We note that P1's strategy $\{\eta_i\}_{i=1}^I$ can also be together represented in terms of $\pi := \{\pi_{ij}\}^{I,d_u}$ such that $\sum_j^{d_u} \pi_{ij} = p[i]$ and $\eta_i[j] = \pi_{ij}/p[i]$, i.e., nature's distribution is the marginal of π and P1's strategy the conditional of π. Let $G^i_{\eta\zeta}$ be

the payoff to P1 of type i for strategy profile (η, ζ). From De Meyer (1996) we have the following results connecting $G(p)$ and $G^*(\hat{p})$:

1. If π is Nash for P1 in $G(p)$ and $\hat{p} \in \partial V(p)$, then $\{\eta_i\}_{i=1}^I$ is also Nash for P1 in $G^*(\hat{p})$.
2. If π is Nash for P1 in $G^*(\hat{p})$ and p is induced by π, then $p \in \partial V^*(\hat{p})$ and π is Nash for P1 in $G(p)$.
3. If ζ is Nash for P2 in $G^*(\hat{p})$ and $p \in \partial V^*(\hat{p})$, then ζ is also Nash for P2 in $G(p)$.
4. If ζ is Nash for $G(p)$, and let $\hat{p}^i := \max_{\eta \in \Delta(d_u)} G^i_{\eta\zeta}$ and $\hat{p} := [\hat{p}^1, ..., \hat{p}^I]^T$, then $p \in \partial V^*(\hat{p})$ and ζ is also Nash for P2 in $G^*(\hat{p})$.

From the last two properties we have: If ζ is Nash for $G(p)$ and $G^*(\hat{p})$, then $\hat{p} = \max_{\eta \in \Delta(d_u)} G^i_{\eta\zeta}$, i.e., $\hat{p}[i]$ is the payoff of type i if P1 plays a best response for that type to P2's Nash.

C.4 OpenSpiel Implementations and Hyperparameters

We use OpenSpiel (Lanctot *et al.*, 2019), a collection of various environments and algorithms for solving single and multi-agent games. We select OpenSpiel due to its ease of access and availability of wide range of algorithms. The first step is to write the game environment with simultaneous moves for the stage-game and the multi-stage games (with 4 decision nodes). Note that to learn the policy, the algorithms in OpenSpiel require conversion from simultaneous to sequential game, which can be done with a built-in method.

In the single-stage game, P1 has two information states representing its type, and P2 has only one information state (i.e., the starting position of the game which is fixed). In the case of the 4-stage game, the information state (or infostate) is a vector consisting of the P1's type (2-D: [0, 1] for type-1, [1, 0] for type-2), states of the players (8-D) and actions of the players at each time step ($4 \times 2 \times U$). The 2-D "type" vector for P2 is populated with 0 as it has no access to P1's type. For example, the infostate at the final decision node for a type-1 P1 could be $[0, 1, x^{(8)}, \mathbb{1}^{(U)}_{u_0}, \mathbb{1}^{(U)}_{d_0}, \cdots, \mathbb{1}^{(U)}_{d_2}, \mathbf{0}^{(U)}, \mathbf{0}^{(U)}]$, and $[0, 0, x^{(8)}, \mathbb{1}^{(U)}_{u_0}, \mathbb{1}^{(U)}_{d_0}, \cdots, \mathbb{1}^{(U)}_{d_2}, \mathbf{0}^{(U)}, \mathbf{0}^{(U)}]$ for P2, where u_k, d_k represent the index of the actions at k^{th} decision node, $k = 0, 1, 2, 3$

The hyperparameters for DeepCFR is listed in table C.1

C.5 Joint-Perturbation Simultaneous Pseudo-Gradient (JPSPG)

The core idea in the JPSPG algorithm is the use of pseudo-gradient instead of computing the actual gradient of the utility to update players' strategies. By

Table C.1. Hyperparameters for DeepCFR Training

Policy Network Layers	(256, 256)
Advantage Network Layers	(256, 256)
Number of Iterations	1000 (100, for $U = 16$)
Number of Traversals	5 (10, for $U = 16$)
Learning Rate	1e-3
Advantage Network Batch Size	1024
Policy Network Batch Size	10000 (5000 for $U = 16$)
Memory Capacity	1e7 (1e5 for $U = 16$)
Advantage Network Train Steps	1000
Policy Network Train Steps	5000
Re-initialize Advantage Networks	True

perturbing the parameters of a utility function (which consists of the strategy), an unbiased estimator of the gradient of a smoothed version of the original utility function is obtained. Computing pseudo-gradient can often be cheaper than computing exact gradient, and at the same time suitable in scenarios where the utility (or objective) functions are "black-box" or unknown. Building on top of pseudo-gradient, Martin and Sandholm (2024) proposed a method that estimates the pseudo-gradient with respect to all players' strategies simultaneously. The implication of this is that instead of multiple calls to estimate the pseudo-gradient, we can estimate the pseudo-gradient in a single evaluation. More formally, let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ be a vector-valued function. Then, its smoothed version is defined as:

$$\mathbf{f}_\sigma(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim\mu}\mathbf{f}(\mathbf{x} + \sigma\mathbf{z}), \tag{C.6}$$

where μ is a d-dimensional standard normal distribution, $\sigma \neq 0 \in \mathbb{R}$ is a scalar. Then, extending the pseudo-gradient of a scalar-valued function to a vector-valued function, we have the following pseudo-Jacobian:

$$\nabla\mathbf{f}_\sigma(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim\mu}\frac{1}{\sigma}\mathbf{f}(\mathbf{x} + \sigma\mathbf{z}) \otimes \mathbf{z}, \tag{C.7}$$

where $\otimes$ is the tensor product.

Typically, in a game, the utility function returns utility for each player given their strategy. Let $\mathbf{u} : \mathbb{R}^{n\times d} \to \mathbb{R}^n$ be the utility function in a game with n players, where each player has a d-dimensional strategy. Then, the simultaneous gradient of $\mathbf{u}$ would be a function $\mathbf{v} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. That is, row i of $\mathbf{v}(\mathbf{u})$ is the gradient of the utility

of the player i with respect to its strategy, $\mathbf{v}_i = \nabla_i \mathbf{u}_i$. As a result, we can rewrite $\mathbf{v}$ concisely as: $\mathbf{v} = \texttt{diag}(\nabla \mathbf{u})$, where ∇ is the Jacobian. With these we have the following:

$$\begin{aligned}\mathbf{v}_\sigma(\mathbf{x}) &= \texttt{diag}(\nabla \mathbf{u}_\sigma(\mathbf{x})) \\ &= \texttt{diag}\left(\mathbb{E}_{\mathbf{z}\sim\mu}\frac{1}{\sigma}\mathbf{u}_\sigma(\mathbf{x}+\sigma\mathbf{z})\otimes\mathbf{z}\right) \\ &= \mathbb{E}_{\mathbf{z}\sim\mu}\frac{1}{\sigma}\texttt{diag}\left(\mathbf{u}_\sigma(\mathbf{x}+\sigma\mathbf{z})\otimes\mathbf{z}\right) \\ &= \mathbb{E}_{\mathbf{z}\sim\mu}\frac{1}{\sigma}\mathbf{u}_\sigma(\mathbf{x}+\sigma\mathbf{z})\odot\mathbf{z},\end{aligned} \tag{C.8}$$

where $\odot$ is element-wise product and a result of the fact that $\texttt{diag}(\mathbf{a}\otimes\mathbf{b}) = \mathbf{a}\odot\mathbf{b}$. Hence, by evaluating Eq. C.8 once, we get the pseudo-gradient associated with all players, making the evaluation constant as opposed to linear in number of players.

Once the pseudo-gradients are evaluated, the players update their strategy in the direction of the pseudo-gradient, assuming each player is interested in maximing their respective utility.

C.5.1 JPSPG Implementation.

In games with discrete-action spaces, where strategy is the probability distribution over the actions, JPSPG can be directly applied to get mixed strategy. However, for continuous-action games, a standard implementation would result in pure strategy solution than mixed. In order to compute a mixed strategy, we can turn into neural network as a strategy with an added randomness that can be learned as described in Martin and Sandholm (2023, 2024). We similarly define two strategy networks for each player, the outputs of which are scaled based on the respective action bounds with the help of hyperbolic tangent (`tanh`) activation on the final layer. The input to the strategy networks (a single hidden layered neural network with 64 neurons and output neuron of action-space dimension) are the state of the player and a random variable whose mean and variance are trainable parameters. We follow the architecture as outlined by Martin and Sandholm (2024) in their implementation of continuous-action Goofspiel. We would like to thank the authors for providing an example implementation of JPSPG on a normal-form game.

In the normal-form Hexner's game, P1's state $\mathbf{x}_1 = \{x_1, y_1, \texttt{type}\}$, and P2's state $\mathbf{x}_2 = \{x_2, y_2\}$. x_i, and y_i denote the x-y coordinates of the player i. In 4-stage case, we also include x-y velocities in the state and append the history of actions chosen by both P1 and P2 into the input to the strategy network. As an example, P1's input at the very last decision step a vector $[x_1, y_1, v_{x_1}, v_{y_1}, x_2, y_2, v_{x_2}, v_{y_2}, \texttt{type}, u_{1_x}, u_{1_y}, d_{1_x}, d_{1_y}, u_{2_x}, u_{2_y}, d_{2_x}, d_{2_y}, u_{3_x}, u_{3_y}, d_{3_x}, d_{3_y}] \in \mathbb{R}^{21}$, where u_j and d_j represent actions of P1 and P2, respectively, at j^{th} decision

point. P2's input, on the other hand, is the same without the `type` information making it a vector in $\mathbb{R}^{20}$.

C.6 Hyperparameter Sweep for PG Baselines

Here we report a sweep of hyperparameters across different learning rates and entropy coefficient for the PG MMD and PPO algorithms. Specifically, we run the algorithms with learning rates of $\{2.5e-5, 2.5e-4, 2.5e-2, 2.5e-1\}$, and entropy coefficient of $\{0.01, 0.05, 0.1, 0.2\}$. We also run RNaD with all four learning rates. The sweep is reported in Fig. C.3.

C.7 Solving Hexner's game via MPC

Here we solve a 2D Hexner's primal game with $I = 2$, $K = 10$, and other settings following App. A.2. With these settings and using the equilibrium in App. A.2.0.2, the true type revelation time is $t_r = 0.5$ second. We directly solve the minimax problem by autodiffing the gradient of the sum of payoffs from the 2^{10} paths of the game tree. At each infostate along each path, P1's strategy is modeled by a neural network that takes in (t, x, p) and outputs I action prototypes and an I-by-I logit matrix that encodes the type-dependent probabilities of taking each of the action prototypes, and P2's best response is modeled by a separate neural network that takes in (t, x, p) and outputs a single action. With a DS-GDA solver, the search successfully converges to the GT equilibrium. Fig. C.4 illustrates the NE trajectories for one particular initial state and the corresponding belief dynamics.

C.8 A Differentiable 11-vs-11 American Football Game

We model a single running/pass play as a 2p0s1 game between the offense (P1) and defense (P2) teams. Each player is a point mass with double-integrator dynamics on a 2D plane. Time is discretised with macro step $\Delta t = \tau$ and $K = T/\tau$ steps, and each macro step is resolved by n_{sub} semi-implicit Euler substeps for stability.

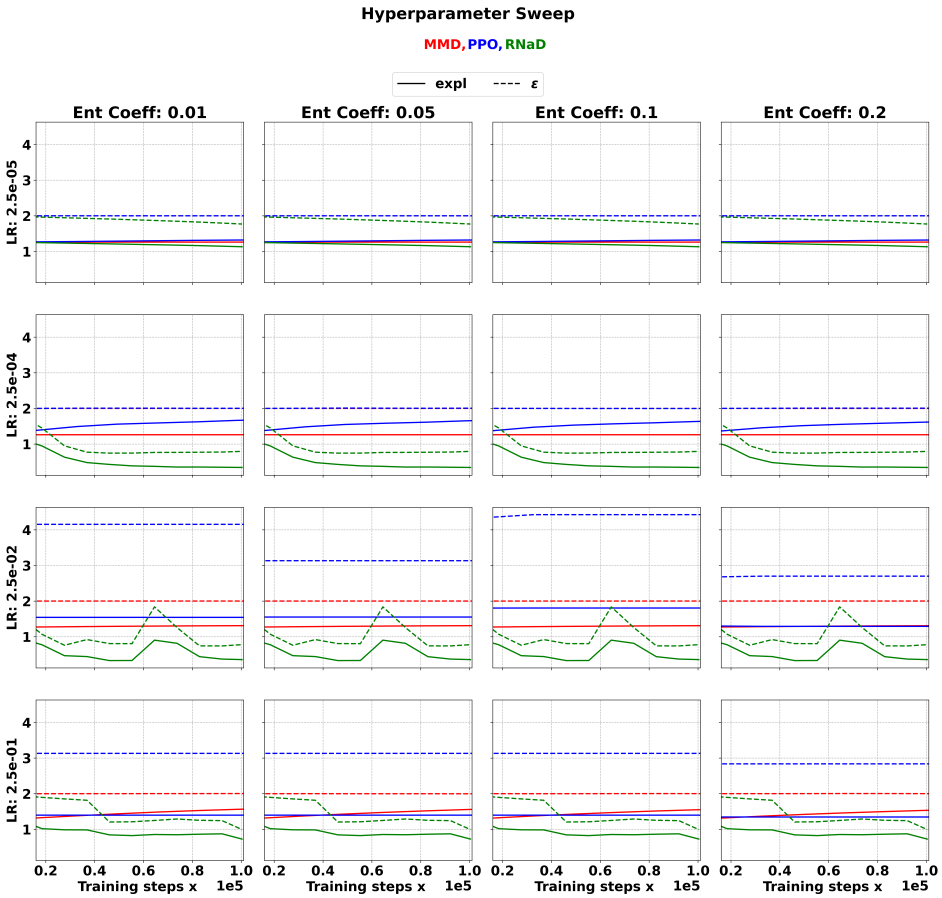


Figure C.3. Hyperparameter sweep for the baseline PG algorithms. RNaD, being a non-standard PG algorithm, doesn't use entropy coefficient the way MMD and PPO do; hence, we copy the same plot across different entropy coefficient value for reference.

C.8.1 State, controls, and bounds.

Let $N = 11$ be players per team. offense positions and velocities are $X^{(1)}, V^{(1)} \in \mathbb{R}^{N\times 2}$; defense $X^{(2)}, V^{(2)} \in \mathbb{R}^{N\times 2}$. We pack them into a state vector $x = [X^{(1)}, V^{(1)}, X^{(2)}, V^{(2)}] \in \mathbb{R}^{8N}$. At each step, the teams apply accelerations $U_1, U_2 \in \mathbb{R}^{N\times 2}$ (stacked later as $u = [u_1; u_2] \in \mathbb{R}^{4N}$). Kinematic saturations enforce a playable box of half-width BOX_POS and box-limited speeds and accelerations BOX_VEL, BOX_ACC by componentwise clamping after each substep.

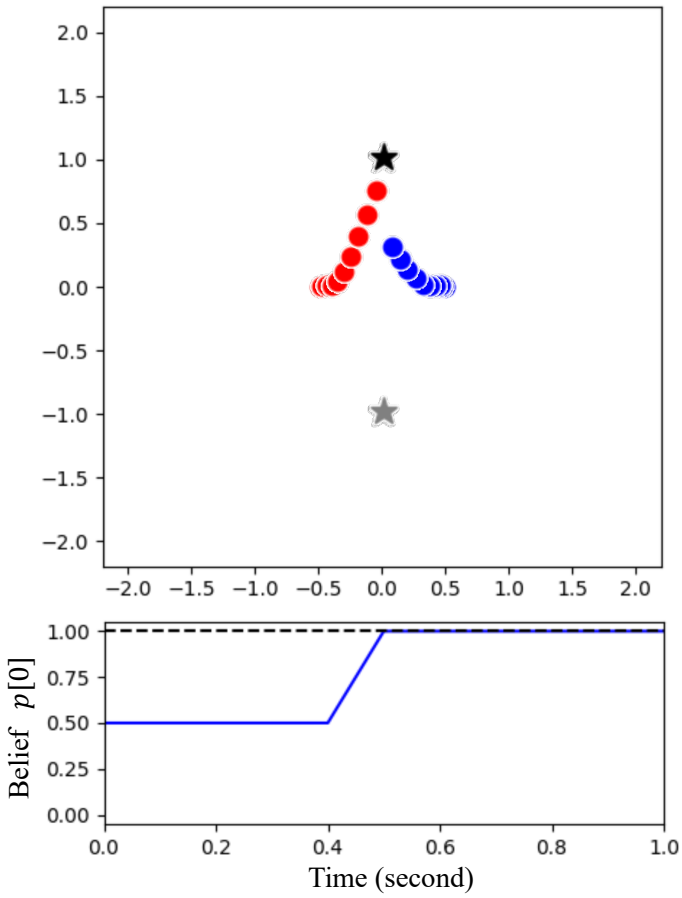


Figure C.4. 2D Hexner's game solved by MPC.

C.8.2 Differentiable tackle dynamics (smooth contact and merge).

During a substep with duration $\delta t = \tau / n_{\text{sub}}$, we first compute a soft, pairwise "stickiness" weight between an attacker i and a defender j:

$$w_{ij} \;=\; \sigma\Big(k_{\text{tackle}}\big(r_{\text{thr}}^2 - \|X_i^{(1)} - X_j^{(2)}\|^2\big)\Big)\,,$$

where $\sigma(z) = 1/(1+e^{-z})$, k_{tackle} sets steepness and $r_{\text{thr}}^2 = \text{MERGE_RADIUS}^2$. These weights form $W \in [0,1]^{N\times N}$. We then compute velocity "sharing" and contact accelerations via convex averaging across opponents:

$$\widehat{V}_i^{(1)} = \frac{V_i^{(1)} + \sum_j w_{ij} V_j^{(2)}}{1 + \sum_j w_{ij}}, \qquad \widehat{V}_j^{(2)} = \frac{V_j^{(2)} + \sum_i w_{ij} V_i^{(1)}}{1 + \sum_i w_{ij}},$$

$$A_{\text{c},i}^{(1)} = \frac{\sum_j w_{ij}\, A_{\text{c},j}^{(2)}}{1 + \sum_j w_{ij}}, \qquad A_{\text{c},j}^{(2)} = \frac{\sum_i w_{ij}\, A_{\text{c},i}^{(1)}}{1 + \sum_i w_{ij}},$$

with $A_{\mathrm{c}}^{(\cdot)}$ initialised at zero so the first pass merely defines a contact baseline. This produces smooth, differentiable coupling without hard impulses.

To blend *control* and *contact* accelerations we form state-dependent merge probabilities

$$p_{\mathrm{m},i}^{(1)} = 1 - \exp\Big(-\sum_j w_{ij}\Big), \qquad p_{\mathrm{m},j}^{(2)} = 1 - \exp\Big(-\sum_i w_{ij}\Big),$$

and set

$$A_i^{(1)} = \big(1 - p_{\mathrm{m},i}^{(1)}\big)\, U_{1,i} + p_{\mathrm{m},i}^{(1)}\, A_{\mathrm{c},i}^{(1)}, \qquad A_j^{(2)} = \big(1 - p_{\mathrm{m},j}^{(2)}\big)\, U_{2,j} + p_{\mathrm{m},j}^{(2)}\, A_{\mathrm{c},j}^{(2)}.$$

We then perform a semi-implicit Euler update with the shared velocities $\widehat{V}$:

$$\begin{aligned}
V^{(1)} &\leftarrow \mathrm{clip}\Big(\widehat{V}^{(1)} + A^{(1)}\,\delta t,\ \pm \mathrm{BOX_VEL}\Big)\,,\\
V^{(2)} &\leftarrow \mathrm{clip}\Big(\widehat{V}^{(2)} + A^{(2)}\,\delta t,\ \pm \mathrm{BOX_VEL}\Big)\,,\\
X^{(1)} &\leftarrow \mathrm{clip}\big(X^{(1)} + V^{(1)}\,\delta t,\ \pm \mathrm{BOX_POS}\big)\,,\\
X^{(2)} &\leftarrow \mathrm{clip}\big(X^{(2)} + V^{(2)}\,\delta t,\ \pm \mathrm{BOX_POS}\big)\,.
\end{aligned}$$

C.8.3 Control-affine analysis

Fix a macro time k and a substep, and treat the current state $(X^{(1)}, V^{(1)}, X^{(2)}, V^{(2)})$ as given. The weights W, the merge probabilities $p_{\mathrm{m}}^{(\cdot)}$, the shared velocities $\widehat{V}^{(\cdot)}$, and the contact terms $A_{\mathrm{c}}^{(\cdot)}$ are *functions of the state only* at that substep. Consequently,

$$A^{(1)} \;=\; \underbrace{\big(1 - p_{\mathrm{m}}^{(1)}\big)}_{\text{state-only}} \odot U_1 \;+\; \underbrace{p_{\mathrm{m}}^{(1)} \odot A_{\mathrm{c}}^{(1)}}_{\text{state-only}}, \qquad A^{(2)} \;=\; \big(1 - p_{\mathrm{m}}^{(2)}\big) \odot U_2 \;+\; p_{\mathrm{m}}^{(2)} \odot A_{\mathrm{c}}^{(2)}.$$

The semi-implicit update is affine in $(A^{(1)}, A^{(2)})$, hence affine in (U_1, U_2):

$$x_{k+1} \;=\; f(x_k) \;+\; B_1(x_k)\, u_1 \;+\; B_2(x_k)\, u_2,$$

where the "input matrices" B_1, B_2 are diagonal masks with entries $(1 - p_{\mathrm{m}}^{(\cdot)})\,\delta t$ in the velocity rows and $(1 - p_{\mathrm{m}}^{(\cdot)})\,\delta t^2$ in the corresponding position rows, all depending only on x_k. Thus the map is *control-affine* for any fixed state, and globally *piecewise* control-affine due to the velocity/position clamping at the box limits; the latter introduces non-smooth but almost-everywhere differentiable saturations.

C.8.4 Tackle probability and running cost.

We summarise the likelihood of a tackle against the ball-carrier (RB) via a differentiable probabilistic OR across all defenders. Let "rb" index the RB on offense,

then with the same W,

$$p_{\text{tackle}} \;=\; 1 - \prod_{j=1}^{N} \big(1 - w_{\text{rb},j}\big).$$

The running loss at a macro step is

$$\ell_{\text{run}} \;=\; \frac{0.1}{2}\,\tau\,\Big(\text{vec}(U_1)^\top R_1\,\text{vec}(U_1) \;-\; \text{vec}(U_2)^\top R_2\,\text{vec}(U_2)\Big) \;+\; \lambda_{\text{tackle}}\,p_{\text{tackle}},$$

with $R_1 = R_2 = I_{4N}$ in our defaults, a small control weight to encourage purposeful motion, and λ_{tackle} is the penalty weight for RB being tackled.

C.8.5 Terminal payoffs: power-push (RB) vs. QB throw

The hidden type $i^\star \in \{0, 1\}$ selects the objective. For the power-push run ($i^\star = 0$), let $(x_{\text{rb}}, y_{\text{rb}})$ denote the RB coordinates and $\alpha_{\text{in}} = -0.8$. The terminal loss is

$$L_{\text{term}}^{\text{run}} \;=\; -\Big(x_{\text{rb}} + \alpha_{\text{in}}|y_{\text{rb}}|\Big),$$

which rewards downfield progress while softly encouraging an inside lane. For the QB throw ($i^\star = 1$), we reward the deepest downfield offensive player, regardless of role:

$$L_{\text{term}}^{\text{throw}} \;=\; -\max_{i\in\{1,\dots,N\}} X_{i,x}^{(1)}.$$

The implemented terminal function is

$$L_{\text{term}} \;=\; \begin{cases} L_{\text{term}}^{\text{run}}, & i^\star = 0, \\ L_{\text{term}}^{\text{throw}}, & i^\star = 1. \end{cases}$$

The overall zero-sum loss is the sum of running losses over $k = 0, \dots, K-1$ plus L_{term}.

C.8.6 Initial lineup.

For $N = 11$ we instantiate a realistic I-formation offense against a 4–3 base defense in a normalized field window. Coordinates use x as downfield (increasing towards the defense) and y as lateral. offense aligns its line on the line of scrimmage at $x = \text{LINEUP_OFF_X}$ with O-line y coordinates $\{-0.80, -0.40, 0.00, 0.40, 0.80\}$ labelled LT, LG, C, RG, RT, a tight end at $y = 1.10$ (right), wide receivers at $y = \pm 1.45$ at the same x, a quarterback at $x = \text{LINEUP_OFF_X} - 0.20$, a fullback at $x = \text{LINEUP_OFF_X} - 0.30$, $y = 0.20$, and the running back at $x = \text{LINEUP_OFF_X} - 0.40$, $y = 0.00$. defense places a four-man line at $x = \text{LINEUP_DEF_X}$ with $y \in \{-0.60, -0.20, 0.20, 0.60\}$, three linebackers at $x = \text{LINEUP_DEF_X} - 0.15$, $y \in \{-0.80, 0.00, 0.80\}$, cornerbacks slightly pressed at $x = \text{LINEUP_DEF_X} + 0.05$, $y = \pm 1.45$, and two safeties deep at $x = \text{LINEUP_DEF_X} - 0.45$, $y \in \{-0.90, 0.90\}$.

APPENDIX D

PUBLICATIONS ASSOCIATED WITH THE DISSERTATION

The chapters in this dissertation have been published or are currently under review, as outlined below. The author contributed as the first-author in all the resulting publications, and has the permissions from the co-authors to use the articles.

1. Chapter 2 is published in the *Proceedings of* 41^{st} *International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024.

   > Ghimire, M., L. Zhang, Z. Xu and Y. Ren, "State-Constrained Zero-Sum Differential Games with One-Sided Information." "Proceedings of the 41st International Conference on Machine Learning", in "Proceedings of Machine Learning Research 235:15512-15539", (2024), URL https://proceedings.mlr.press/v235/ghimire24a.html.

2. Chapter 3 is published in the 14^{th} *International Conference on Learning Representations*.

   > Ghimire, M., L. Zhang, Z. Xu and Y. Ren, "Solving football by exploiting equilibrium structure of 2p0s differential games with one-sided information", in "The Fourteenth International Conference on Learning Representations", (2026), URL https://openreview.net/forum?id=vRwuBOxbsJ.

3. Chapter 4 has been submitted for review to IEEE Control Systems Letters (L-CSS).