

HETEROGENEITY AND STRATEGIC SOPHISTICATION IN MULTI-AGENT REINFORCEMENT LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Yüksel Arslantaş
August 2024

Heterogeneity and Strategic Sophistication in Multi-agent Reinforcement Learning

By Yüksel Arslantaş

August 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Muhammed Ömer Sayın(Advisor)

Orhan Arıkan

İsmail Uyanık

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

HETEROGENEITY AND STRATEGIC SOPHISTICATION IN MULTI-AGENT REINFORCEMENT LEARNING

Yüksel Arslantaş

M.S. in Electrical and Electronics Engineering

Advisor: Muhammed Ömer Sayın

August 2024

Decision-making powered by artificial intelligence (AI) is becoming increasingly prevalent in socio-technical systems such as finance, smart transportation, security, and robotics. Therefore, there is a critical need for developing the theoretical foundation of how multiple AI decision-makers interact with each other and with humans in order to ensure their reliable use in these systems. Since multiple AI decision-makers make decisions autonomously without central coordination, heterogeneity of their algorithms is inevitable. We establish a theoretical framework for the impact of heterogeneity on multi-agent sequential decision-making under uncertainty. First, we examine the potential heterogeneity of independent learning algorithms assuming that opponents play according to some stationary strategy. To this end, we present a broad family of algorithms that encompass widely-studied dynamics such as fictitious play and Q-learning. While existing convergence results only consider homogeneous cases, where each agent uses the same algorithm; we show that they can still converge to equilibrium if they follow any two different members of this algorithm family. This strengthens the predictive power of game-theoretic equilibrium analysis for heterogeneous systems. We then analyze how a strategically sophisticated agent can manipulate independent learning algorithms, revealing the vulnerability of such independent reinforcement learning algorithms. Finally, we demonstrate the practical implications of our findings by implementing our results in stochastic security games, highlighting its potential for real-life applications, and explore the impact of strategic AI in human-AI interactions in cyberphysical systems.

Keywords: Multi-agent reinforcement learning, game theory, stochastic control.

ÖZET

ÇOKLU ETMEN PEKİŞTİRMELİ ÖĞRENMEDE HETEROJENLİK VE STRATEJİK KARMAŞIKLIK

Yüksel Arslantaş

Elektrik ve Elektronik Mühendisliği, Yüksek Lisans

Tez Danışmanı: Muhammed Ömer Sayın

Ağustos 2024

Yapay zekâ destekli karar verme, finans, akıllı ulaşım, güvenlik ve robotik gibi sosyoteknik sistemlerde giderek yaygınlaşmaktadır. Bu nedenle, bu sistemlerde güvenilir kullanımı sağlamak için birden fazla yapay zekâ karar vericisinin birbirleriyle ve insanlar ile nasıl etkileşimde bulunduğu dair teorik temel oluşturulması kritik bir ihtiyaçtır. Birden fazla yapay zekâ karar vericisi merkezi koordinasyon olmadan özerk olarak kararlar aldığından, algoritmalarının heterojenliği kaçınılmazdır. Belirsizlik altında çoklu etmen sıralı karar vermede heterojenliğin etkisi üzerine teorik bir çerçeve oluşturuyoruz. İlk olarak, rakiplerin sabit bir stratejiye göre oynadığını varsayarak bağımsız öğrenme algoritmalarının potansiyel heterojenliğini inceliyoruz. Bu amaçla, hayali oyun ve Q-öğrenme gibi geniş çapta incelenen dinamikleri kapsayan geniş bir algoritma ailesi sunuyoruz. Mevcut yakınsama sonuçları yalnızca her aracın aynı algoritmayı kullandığı homojen durumları ele alırken; bu algoritma ailesinin herhangi iki farklı üyesini takip ettiklerinde yine de dengeye yakınsayabileceklerini gösteriyoruz. Bu, oyun teorik denge analizinin heterojen sistemler için tahmin gücünü güçlendirmektedir. Daha sonra, stratejik olarak gelişmiş bir etmenin bağımsız öğrenme algoritmalarını nasıl manipüle edebileceğini analiz ederek, bu tür bağımsız pekiştirmeli öğrenme algoritmalarının savunmasızlığını ortaya koyuyoruz. Son olarak, bulgularımızın pratik sonuçlarını stokastik güvenlik oyunlarına uygulayarak gerçek dünya uygulamaları için potansiyelini vurguluyor ve siber fiziksel sistemlerde insan-yapay zekâ etkileşimlerinde stratejik yapay zekânın etkisini araştırıyoruz.

Anahtar sözcükler: Çoklu etmen pekiştirmeli öğrenme, oyun kuramı, stokastik kontrol.

Acknowledgement

I would like to express my sincere gratitude to my professors at Bilkent University for their invaluable support and guidance throughout my academic journey. I am particularly indebted to my thesis advisor, Asst. Prof. Muhammed Ömer Sayın, for his unwavering belief in my abilities and his guidance in navigating the complexities of research. His insightful feedback and encouragement were instrumental in shaping this thesis.

I would also like to thank my friends and colleagues, Ege, Ahmed Said, and Yiğit, for their companionship, support, and stimulating discussions throughout the research process.

Most importantly, I am eternally grateful to my mother and sister for their unwavering love, support, and encouragement. Their belief in me has been my guiding light throughout this journey.

This thesis is supported by The Scientific and Technological Research Council of Türkiye (TUBITAK) BİDEB 2232-B International Fellowship for Early Stage Researchers under Grant Number 121C124.

Contents

1	Introduction	1
2	Heterogeneous Learning Dynamics in Competitive Environments	7
2.1	Related Works	9
2.2	Game Formulation	11
2.3	A Family of Learning Dynamics	14
2.3.1	Algorithm Family 1 for Matrix Games	14
2.3.2	Algorithm Family 2 for Stochastic Games	16
2.4	Convergence Results	19
2.4.1	Convergent Heterogeneous Learning for ZSGs	20
2.4.2	Convergent Heterogeneous Learning for ZSSGs	29
2.5	Illustrative Examples	39
3	Strategizing against Independent Learning Dynamics	42

3.1	Related Works	44
3.2	Strategic Actors against Q-learners	45
3.3	Solving the Stochastic Game	50
3.4	Illustrative Examples	56
4	Strategic Security: Assessing Reinforcement Learning Vulnerabilities	62
4.1	Linear Influence Network Models	64
4.1.1	Attack and Defense on the Network	65
4.2	Stochastic Games and Independent Q-learning	66
4.2.1	Q-learning	68
4.3	Strategic Attacker with Complete Model Knowledge	68
4.4	Illustrative Examples	74
5	The Role of Strategic Agent in Human-AI Interactions	78
5.1	Strategic AI and Human Interaction	80
5.1.1	Human Agent	81
5.1.2	Strategic AI Agent	82
5.1.3	Approximation of the MDP	83
5.1.4	Model Prediction	85
5.2	Illustrative Examples	87

6 Conclusion	92
A Preliminary Information for Stochastic Approximation	106
A.1 Stochastic Differential Inclusion Approximation	106
A.2 Asynchronous Updates	108
B Proof of Lemma 1	110
C Proof of Lemma 2	112
D Proof of Lemma 3	114
E Proof of Lemma 4	115

List of Figures

2.1	Agent i 's learning dynamics in the repeated play of a matrix game when the agent assumes that the opponent plays according to a stationary strategy. We use dashed arrows for opponent actions since agents may not have access to them.	14
2.2	Agent i 's learning dynamics in SGs when the agent assumes that the opponent plays according to a Markov stationary strategy. We use dashed arrows for opponent actions since agents may not have access to them. The learning dynamics in SGs differ from the ones for repeated matrix games, illustrated in Figure 2.1 mainly due to the iterative MDP solver. Color coded arrows between stage game dynamics and the iterative MDP solver represent the coupling between them.	17
2.3	The evaluation of estimated value functions in a ZSSG with two states and two actions, where $S = \{s^1, s^2\}$, $\gamma = 0.3$ and $\tau = 0.002$. Each sub-figure corresponds to Scenarios 1, 2, and 3, respectively. "Full", "Temporal" and "None" refer to whether the agent has full, temporal (with $\theta^i = 0.5$) or no access to the opponent's actions. The different colors and line styles represent different agents and states. The shaded areas represent the standard deviation of the value function estimates. The dotted lines represent the bounds on the value function estimates according to Theorem 2.	41

3.1	Illustration of the interaction between N-types and A-types across the repeated play of the normal-form game $\mathcal{G}$. A-types compete against each other while considering N-types' q-function estimates as an underlying state of an SG. A-types can use dynamic programming to solve the SG and use trackers to track N-types' q-function estimates by leveraging the complete model knowledge.	45
3.2	The evolution of the empirical averages of the action profiles for Q-learner vs Q-learner in the prisoner's dilemma.	56
3.3	The evolution of the empirical averages of the action profiles for strategic actor vs Q-learner in the prisoner's dilemma.	57
3.4	The evolution of the empirical averages of the action profiles for Q-learner vs Q-learner in the two-agent four-action zero-sum game.	58
3.5	The evolution of the empirical averages of the action profiles for strategic actor vs Q-learner in the two-agent four-action zero-sum game.	59
4.1	A linear influence network illustrating compromised and secured security assets. Green servers represent uncompromised assets, red servers represent compromised assets, and the blue server represents a defended asset. Shaded arrows depict interdependencies between assets: solid arrows indicate usable attack paths, and dashed arrows indicate unusable paths. The solid black arrow represents the attacker's direct attack.	65
4.2	The illustration demonstrates how the attacker and the defender take actions during the stochastic game $\mathcal{M}$. The dashed lines indicate observations from the previous stage.	72
4.3	Average discounted rewards of the attacker against different exploration levels.	75

4.4	Average discounted rewards of the attacker with different learning models.	76
5.1	An illustration of the EWA learning dynamics based on the parameters τ , α , κ , and δ . Belief-based learning dynamics, represented with a solid line, correspond to the case where $\delta = 1$. Payoff-based learning dynamics (i.e., reinforcement learning) are depicted as a shaded area for $\delta = 0$	80
5.2	The illustration demonstrates how the strategic AI agent and the human agent take actions during the repeated play of $\mathcal{G}$. The dashed lines indicate observations from the previous stage.	84
5.3	The evolution of the empirical averages of the action profiles for the human agent vs human agent and strategic AI agent vs human agent in coordination game for perfect and imperfect knowledge.	86
5.4	The evolution of the empirical averages of the action profiles for the human agent vs human agent and strategic AI agent vs human agent in anti-coordination game for perfect and imperfect knowledge.	87
5.5	The evolution of the empirical averages of the action profiles for the human agent vs human agent and strategic AI agent vs human agent in zero-sum game for perfect and imperfect knowledge.	88

List of Tables

3.1	Single A-type vs Multiple N-types	60
3.2	Two Competing A-types vs Single N-type	61
5.1	Coordination Game	89
5.2	Anti-coordination Game	90
5.3	Zero-sum Game	90

Chapter 1

Introduction

Multi-agent reinforcement learning (MARL) has emerged as a powerful paradigm for training intelligent agents to collaborate or compete in complex environments. Its applications span various domains, including economics, where it can model strategic interactions between market players [1, 2, 3, 4, 5], robotics, where it can enable teams of robots to perform coordinated tasks [6, 7, 8], and smart cities, where it can optimize traffic flow and resource allocation [9, 10, 11]. However, a significant portion of the current MARL research focuses on homogeneous settings, where agents share the same learning dynamics. This assumption often falls short in real-world scenarios where agents are inherently heterogeneous. For example, the human-human interactions often exhibit significant heterogeneity such that agents may place varying emphasis on past versus recent experiences [12]. This has led to increased interest in heterogeneous learning within economics. Studies like [13] explore the impact of such heterogeneity, arising from diverse initial beliefs, update rule variations, and other factors, on the local stability of macroeconomic models. Similarly, the authors in [14, 15] investigate how different information sets used for forecasting affect learning speed towards equilibrium. The field of engineering also embraces heterogeneous learning, particularly in cooperative multi-agent systems. Examples include distributed coordination with communication delays [16], consensus optimization with varying computational

capabilities [17], and motion planning for robot swarms with hardware differences, perception variations, and environmental factors [18, 19, 20, 21, 22]. In this work, we explore the heterogeneity in MARL, specifically focusing on the following question:

How variations in learning rates, exploration strategies, model knowledge, and strategic sophistication can impact agent behavior and overall system performance?

Consider a scenario of two agents engaged in an (competitive or collaborative) interaction. Each agent aims to maximize its own utility function while the joint actions ultimately determine the individual payoffs received. In such environments, identifying the optimal action for each agent through introspective reasoning becomes a non-trivial optimization problem due to its inherent dependence on the opponent’s chosen strategy. While adopting a worst-case equilibrium strategy may appear cautious, it can be overly conservative against adversaries with limited cognitive and computational capabilities. Alternatively, agents can leverage interaction and the resulting environmental feedback to learn and adapt their behavior [23]. However, the specific learning dynamics employed by each agent are contingent upon their individual interests and available resources. These factors encompass the agent’s tolerance for exploration, the information structure of the environment, and the level of model knowledge possessed. Any differences in these aspects will inevitably lead to the emergence of heterogeneous learning dynamics.

We first focus on analyzing heterogeneous learning dynamics in competitive environments, specifically two-agent zero-sum games. We consider both repeated play of the normal form game and stochastic games. While stochastic games can be viewed as a sequence of stage games played at specific states, unlike zero-sum games, the payoffs in these stage games consider not just immediate rewards but also how the agents anticipate playing in the future. Current actions influence not only the immediate rewards but also the probability of transitioning to the

next state and the resulting future rewards. This dependence on anticipated future actions means the stage games within stochastic games can deviate from a strict zero-sum structure. Therefore, our research also focuses on near zero-sum games. We propose new families of algorithms for these types of two-agent (near) zero-sum games. These frameworks provide a unified approach to address how exploration strategies, access to opponent actions, model knowledge, and learning rates can influence agent behavior. We focus on best-response-type learning dynamics where the agents base their actions on their beliefs about how the opponent plays, assuming the opponent follows a fixed strategy. The level of access to opponent actions plays a crucial role in shaping these beliefs. For instance, with full access, agents observe all opponent actions, leading to belief-based dynamics such as fictitious play and its variants [24, 25]. Conversely, with no access, agents rely solely on past rewards, resulting in payoff-based dynamics such as individual Q-learning [26]. Finally, temporal access represents scenarios where agents observe opponent actions with some probability, creating a hybrid approach. The proposed algorithm families offer several desirable properties for multi-agent learning [27]:

- *Uncoupled Learning Dynamics:* The dynamics function independently of the specific game being played. They do not require knowledge of the opponent’s objective, actions, or the underlying game model.
- *Rationality:* Under standard assumptions about learning parameters, these dynamics ensure agents achieve (near) optimal performance against opponents using fixed strategies.
- *Convergence for Diverse Learning Dynamics:* Even within our proposed family of algorithms, agents can employ different dynamics that can achieve convergence towards (near) equilibrium in zero-sum games (both repeated and stochastic) as long as the discount factor is sufficiently small. This property holds true for both homogeneous and heterogeneous learning scenarios.

We leverage the concept of continuous-time best-response dynamics to approximate the discrete updates of these algorithms. This allows us to analyze their convergence properties by formulating a Lyapunov function based on stochastic approximation methods, e.g., see [28, 29]. Notably, our proposed Lyapunov function is a unifying framework that can handle scenarios with heterogeneous learning dynamics, even in games that are not strictly zero-sum (near zero-sum games). This capability distinguishes our approach from previous work.

Having analyzed the convergence properties under heterogeneous learning rates, model knowledge, and exploration strategies, we then shift our focus to the impact of strategic sophistication on agent behavior. Here, we consider a normal-form game played repeatedly. To illustrate the worst-case scenario, we assume a strategically sophisticated agent with complete knowledge of the game, the opponent’s learning dynamics, and the entire game history. In contrast, the other agent, referred to as the naive agent, employs the widely used independent Q-learning algorithm [5, 30, 31, 32, 33]. To analyze this scenario, we model the game as a stochastic control problem for the strategic agent. In this framework, the state space incorporates the Q-values of the naive agent. Through this modeling approach, we demonstrate the existence of an optimal strategy for the strategic agent, allowing it to maximize its discounted sum of rewards. However, calculating the optimal policy for the strategic agent presents a challenge due to the continuum state space. To address this, we approximate the underlying Markov Decision Process (MDP) using quantization as a proof of the concept (see [34] for details). Here, we focus on quantization-based approximation, but other methods such as state aggregation [35, 36] or approximate linear programming [37, 38] could also be explored. Through a combination of analytical and numerical analysis, we evaluate the performance of the strategic agent.

One specific application we consider is the phenomenon of *tacit collusion* in economics, where agents may learn to cooperate implicitly without explicit communication [2, 5, 39, 40]. This is particularly relevant to scenarios involving learning algorithms used in market competition and price setting, where preventing tacit collusion becomes crucial. Our work demonstrates how a sophisticated agent, equipped with the knowledge of the naive agent’s learning dynamics, can

exploit such vulnerabilities in reinforcement learning algorithms, particularly independent Q-learning.

Our investigation extends beyond the two-agent case. We analyze scenarios with multiple naive agents and even settings with two competing strategic agents. Similar to the two-agent case, we perform both analytical and numerical evaluations of their performance.

An important theoretical contribution of our work is proving the Lipschitz continuity of the value function in the MDP derived from the naive agents' Q-values. This property ensures the value function is well-behaved and susceptible to approximation techniques based on function approximation. Leveraging this result, we scale up our problem by focusing on stochastic games [41], which offer a more general framework than repeated play in normal-form games. By exploiting the well-behaved nature of the value function, we introduce the use of neural networks to approximate it. This method, known as neuro-dynamic programming [42], allows us to tackle problems significantly more complex than repeated play in normal-form games. To illustrate the power of our approach in handling complex scenarios, we consider stochastic security games [43]. These games model situations where a strategic attacker and a defender employing reinforcement learning, interact within a linear influence network [43, 44, 45, 46]. Through experiments, we demonstrate the effectiveness of the strategic agent in exploiting vulnerabilities in the defender's learning dynamics.

Finally, we explore the impact of strategic AI in human-AI interactions within cyberphysical systems, where humans and AI agents collaborate or compete [47, 48]. We model the human agent using the experimentally validated Experience-Weighted Attraction (EWA) model, known for its effectiveness in capturing human behavior in repeated games [49, 50, 51]. Similar to our prior approach, we formulate the problem as a stochastic control problem, where the state space incorporates the human agent's attraction vector. Once again, quantization-based approximation is employed. However, real-world applications rarely involve complete knowledge of the opponent, especially when dealing with

humans whose behavior can be more nuanced. To address this challenge, we consider scenarios where the human agent adopts one of a finite number of learning dynamics, and the strategic agent aims to predict the correct type through in-game observations. Our findings demonstrate the vulnerability of the EWA model to exploitation by a strategic AI, particularly in coordination, anti-coordination, and zero-sum games. These results highlight the importance of considering strategic AI not only as a potential adversary but also as a collaborative partner when objectives are aligned.

Chapter 2

Heterogeneous Learning Dynamics in Competitive Environments

*“Variety’s the very spice of life,
that makes it so savory.”*

William Cowper

Recently, provable convergence of multi-agent reinforcement learning (MARL) in stochastic games has gained significant attention due to their success in various artificial intelligence applications, e.g., see [52, 53, 54, 55, 56, 57, 58, 59, 60]. This line of work focuses on achieving convergence with radically uncoupled and rational learning dynamics. However, the existing convergence results for zero-sum games primarily deal with *homogeneous* settings where all agents adopt the same learning dynamic. This homogeneity assumption limits the applicability of these findings to real-world scenarios with multi-agent interactions. In practice, non-cooperative agents may not have the incentive or capability to follow identical learning dynamics, especially in competitive environments. Addressing provable convergence in *heterogeneous* settings, where agents employ diverse learning dynamics or have varying information access, presents a critical challenge and a

necessary paradigm shift for MARL.

One key challenge lies in establishing a unifying framework that can handle the heterogeneous learning dynamics. The difficulty of this challenge is amplified in stochastic games (SGs) compared to the repeated play of the games. We can view SGs as a sequence of state-specific *stage games* played whenever a particular state is reached. However, the payoffs of these stage games are not determined by the immediate rewards received. They also depend on how the agents anticipate playing in future stages, as current actions influence not only the immediate rewards but also the probability of transitioning to the next state and the resulting *continuation payoff*. Recently two-timescale learning frameworks have addressed the issue of possible non-stationarity in stage-game payoffs from continuation payoffs [54, 55, 56]. However, in real-world scenarios, the stage-game dynamics may not necessarily sum to zero. The deviation of the stage-games from the zero-sum structure gets boosted under heterogeneities due to varying ways agents estimate the continuation payoffs. Consequently, there is a critical need for developing robust convergence guarantees that hold even for near zero-sum games with heterogeneous learning dynamics.

We introduce new families of algorithms designed for two-agent zero-sum games, encompassing both repeated and stochastic scenarios. The algorithm families offer a unified framework capable of handling heterogeneous learning dynamics in terms of exploration, access to opponent actions, model knowledge and step sizes. We focus on best-response-type learning dynamics, where agents respond greedily to the belief they formed about how the opponent plays under the assumption that the opponent plays according to some stationary strategy. The algorithms can adapt to different levels of information access regarding the opponent’s actions. In the case of *full-access*, the agent directly observes the opponent’s actions (*belief-based dynamics*). Conversely, with *no-access*, the agent rely on past rewards to infer the opponent’s strategy (*payoff-based dynamics*). In *temporal-access*, the agent occasionally observes the opponent’s action with some probability independent of the history of the game.

This chapter is organized as follows: We discuss the relevant literature and position our contributions in Section 2.1. We provide preliminary information about SGs in Section 2.2. We introduce two families of best-response-type learning dynamics for repeated matrix games and SGs in Section 2.3. We characterize the convergence properties of any mixture of dynamics from these families for near zero-sum matrix games and SGs in Section 2.4. Section 2.5 includes numerical examples.

2.1 Related Works

Heterogeneous learning, where individuals adopt diverse strategies in response to the same information, is a common phenomenon observed in human-human interactions during repeated game play due to how they weight past and recent experiences when making decisions [12]. This heterogeneity has also attracted the attention in the economics literature. For example, the studies in macroeconomics demonstrate that differing prior beliefs and inertia updates can lead to different outcomes compared to models assuming homogeneous behavior [15]. Furthermore, the authors of [14, 15] highlight the impact of heterogeneous information sets on the speed at which learning process converge to equilibrium. While these works study the heterogeneities, the seminal work of [49] offers a unifying framework called Experience-Weighted Attraction (EWA) for capturing human learning dynamics through four key parameters. Experimental results have shown that EWA to be highly effective in interpreting human learning behavior, including the modeling of hybrid learning frameworks [49, 51]. Although the work of [61] explore heterogeneities among certain EWA parameters, the findings in this area remains primarily experimental. Our work differs from these lines of works by providing a unifying framework to address heterogeneities for best-response-type dynamics with theoretical guarantees.

Heterogeneous learning has been extensively studied in the context of cooperative multi-agent systems. Examples include distributed coordination under communication delays [16], consensus optimization with varying computational

capabilities [17], and motion planning for robot swarms with diverse hardware, sensors, and operating environments [18, 19, 20, 21, 22]. However, these cooperative settings typically involve algorithms that are inherently coupled and often assume a level of global coordination, which differs from the independent and rational behavior characteristic of our model.

In MARL, heterogeneity is also explored in cooperative settings where teams are composed of different type of agents based their roles in the system. In [62], the authors show that a sophisticated agent leading the team can achieve optimal results for the tabular setting. However, most other studies in MARL primarily focus on function approximation to address large-scale problems instead of providing theoretical guarantees [63, 64, 65]. For example, in [63, 64], the authors leverage the heterogeneity among team members to improve the overall performance and address the challenges with sample efficiency and exploration by assigning distinct roles to agents. Heterogeneity studies in MARL are also motivated by real-life applications. For example, the authors in [66] study heterogeneous learning in traffic light control systems since traffic junctions may not be homogeneous in practical applications. Similarly, the authors in [67] investigate heterogeneity among agents in scenarios with unknown environment mappings.

Heterogeneous learning in non-cooperative settings has recently received attention, primarily within the context of non-competitive environments. For example, in [68] the author explore decentralized learning for Markov potential games where agents employ different learning rates within a two-timescale framework. Furthermore, studies have focused on decentralized learning for stochastic games with specific structural properties allowing for heterogeneous exploration phases among agents [69]. Another line of research investigates learning in convex games, considering agents with varying access to information [70].

Heterogeneous learning in competitive settings has primarily focused on scenarios where agents employ different learning rates [53, 71, 72, 73]. In [71] the authors explore the decentralized learning in two-agent zero-sum games and multi-agent identical interest normal-form games with different learning rates. In [72] the

authors investigate the gradient-based learning for continuous games with heterogeneous learning rates. Moreover, the authors in [53] study the independent policy gradient algorithms for two-agent zero-sum stochastic games with different timescales. Subsequently, in [73] the authors demonstrate that the novel fictitious play variant for stochastic games introduced by [74] converges to equilibrium in two-agent zero-sum stochastic games under the conditions of comparable learning rates and a sufficiently small discount factor. Our work departs from these studies by adopting a more comprehensive approach to heterogeneous learning in competitive settings. We analyze the convergence properties of a broad spectrum of learning dynamics under heterogeneity in terms of learning rates, exploration preferences, information availability and knowledge of the game structure.

2.2 Game Formulation

Two-agent zero-sum SGs (ZSSGs) are generalizations of Markov Decision Processes (MDPs) to two-agent competitive environments. Let agents be indexed by $i = 1, 2$ and agents i and j denote, respectively, the typical agent and the typical opponent. Then, we can characterize a ZSSG by a tuple $\mathcal{M} = \langle S, (A^i, r^i)_{i=1}^2, p, \gamma \rangle$. Here, S denotes the finite set of states and A^i denotes agent i 's finite set of actions at each state.¹ The reward $r^i : S \times A^i \times A^j \rightarrow \mathbb{R}$ represents agent i 's stage payoff function. Due to its zero-sum nature, we have $r^1(s, a^1, a^2) + r^2(s, a^2, a^1) = 0$ for all (s, a^i, a^j) . Similarly, $p(s_+ | s, a)$ for each (s, a, s_+) denotes the transition probability from s to s_+ when agents play the action profile $a = (a^i, a^j)$. Lastly, $\gamma \in [0, 1)$ is the discount factor.

Agents take their actions to maximize the expected sum of discounted rewards they receive over infinite horizon. To this end, they can follow Markov stationary strategies that determine the probability of any action played depending only on the current state. For example, denote agent i 's Markov stationary strategy and the strategy space, respectively, by $\pi^i : S \rightarrow \Delta^i$ and Π^i , where Δ^i is the probability simplex over A^i . Then, at every visit to s , agent i plays action $a^i \sim$

¹The generalization to state-variant action sets is rather straightforward.

$\pi^i(s)$ independent of the past plays and the opponent play.

If the opponent j follows Markov stationary strategy π^j , then agent i faces an MDP, for which there exists a Markov stationary solution. Without loss of generality, agent i 's goal is to find $\pi^i \in \Pi^i$ maximizing

$$U^i(\pi^i, \pi^j) := \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k) \right], \quad (2.1)$$

where (s_k, a_k) is the pair of state and action profile at stage $k = 0, 1, \dots$ and the expectation is taken with respect to the randomness on states visited and actions played when agents play according to the strategy profile (π^i, π^j) . However, this objective is not a well-defined optimization problem due to its dependence on how the opponent plays. The following is a definition of Markov stationary equilibrium to predict the outcome of such non-cooperative interactions and Shapley proved its existence for ZSSGs in his inaugural paper [41].

Definition 1 (Markov Stationary Equilibrium) *A strategy profile $(\pi_*^1, \pi_*^2) \in \Pi^1 \times \Pi^2$ is near Markov stationary equilibrium of $\mathcal{M}$ with approximation error $\epsilon \geq 0$ provided that $U^i(\pi_*^i, \pi_*^j) \geq U^i(\pi^i, \pi_*^j) - \epsilon$ for all $\pi^i \in \Pi^i$ and $i = 1, 2$. Call (π_*^1, π_*^2) Markov stationary equilibrium if $\epsilon = 0$.*

Note that for $|S| = 1$ and $\gamma = 0$, ZSSGs reduce to zero-sum matrix games (ZSGs) characterized by the tuple $\mathcal{G} = \langle A^i, R^i \rangle_{i=1}^2$, where $R^i \in \mathbb{R}^{|A^i| \times |A^j|}$ such that $(\pi^i)^T R^i \pi^j = \mathbb{E}_{(a^i, a^j) \sim (\pi^i, \pi^j)} [r^i(a^i, a^j)]$ for all (π^i, π^j) . View any action $a^i \in A^i$ as a pure strategy in which the associated action gets played with probability 1, i.e., $A^i \subset \Delta^i$, and define $\text{BR}_\epsilon^i : \mathbb{R}^{|A^i|} \rightarrow 2^{\Delta^i}$ by

$$\text{BR}_\epsilon^i(q^i) := \{a^i \in A^i : (a^i)^T q^i \geq (\tilde{a}^i)^T q^i - \epsilon \forall \tilde{a}^i \in A^i\}. \quad (2.2)$$

Let $\text{BR}^i(q^i) := \text{BR}_\epsilon^i(q^i)$ for $\epsilon = 0$. Then, we can define Nash equilibrium in $\mathcal{G}$ as follows.

Definition 2 (Nash Equilibrium (NE)) *A strategy profile $(\pi_*^1, \pi_*^2) \in \Delta^1 \times \Delta^2$*

is ϵ -Nash equilibrium of $\mathcal{G}$ provided that $\pi_*^i \in \text{BR}_\epsilon(R^i \pi_*^j)$ for $i = 1, 2$ and $j \neq i$. Call (π_*^1, π_*^2) Nash equilibrium if $\epsilon = 0$.

Let $\text{BR}_\tau^i : \mathbb{R}^{|A^i|} \rightarrow \text{int}(\Delta^i)$ denote the smoothed best response under entropy regularization, defined by²

$$\text{BR}_\tau^i(q^i) := \underset{\mu^i \in \Delta^i}{\text{argmax}} \{ (\mu^i)^T q^i + \tau^i H^i(\mu^i) \} \quad (2.3)$$

for all $q^i \in \mathbb{R}^{|A^i|}$, where $\tau^i > 0$ is a temperature parameter determining the impact of the perturbation and the perturbation

$$H^i(\mu^i) := - \sum_{a^i} \mu^i(a^i) \log(\mu^i(a^i)) \in [0, \log |A^i|]. \quad (2.4)$$

In the smoothed best response, any action gets played with some positive probability bounded from below by

$$\text{BR}_\tau^i(q^i)(a^i) = \frac{\exp(q^i(a^i)/\tau^i)}{\sum_{\tilde{a}^i \in A^i} \exp(q^i(\tilde{a}^i)/\tau^i)} \quad (2.5a)$$

$$\geq \exp(-2\|q^i\|_\infty/\tau^i)/|A^i| > 0 \quad \forall a^i. \quad (2.5b)$$

The following is a refinement of NE, also known as Nash distribution, to incorporate such bounded rationality [75].

Definition 3 (Quantal Response Equilibrium (QRE)) *A strategy profile $(\pi_*^1, \pi_*^2) \in \Delta^1 \times \Delta^2$ is quantal response equilibrium of $\mathcal{G}$ provided that $\pi_*^i = \text{BR}_\tau^i(R^i \pi_*^j)$ for $i = 1, 2$ and $j \neq i$.*

Note that $\text{BR}_\tau^i(R^i \pi^j)$ has single value different from the set-valued function $\text{BR}_\epsilon^i(R^i \pi^j)$. Furthermore, we have $\text{BR}_\tau^i(R^i \pi^j) \in \text{BR}_\epsilon^i(R^i \pi^j)$ for $\epsilon \geq \tau^i H^i$.

²Given a set A , let $\text{int}(A)$ denote its interior.

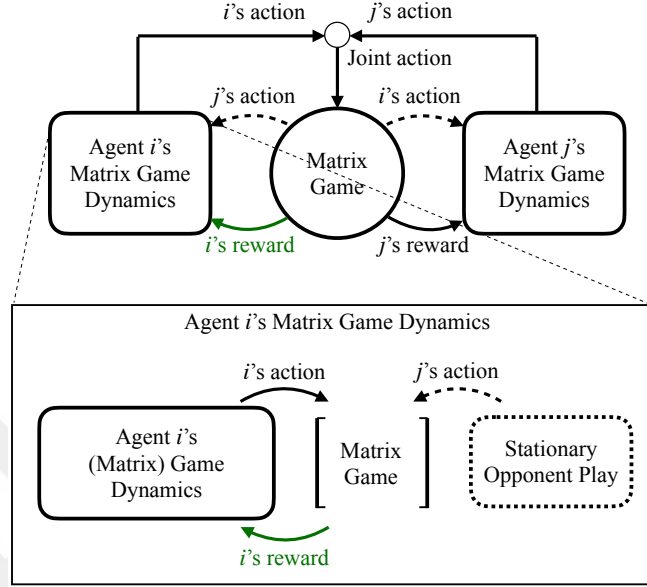


Figure 2.1: Agent i 's learning dynamics in the repeated play of a matrix game when the agent assumes that the opponent plays according to a stationary strategy. We use dashed arrows for opponent actions since agents may not have access to them.

2.3 A Family of Learning Dynamics

This section describes two families of learning dynamics for the repeated play of matrix games and SGs.

2.3.1 Algorithm Family 1 for Matrix Games

Consider the repeated play of a matrix game $\mathcal{G} = \langle A^i, R^i \rangle_{i=1}^2$, as illustrated in Figure 2.1. Each agent i assumes that the opponent j plays according to some underlying mixed strategy $\pi^j \in \Delta^j$. Then, the values of the local actions are given by $q^i = R^i \pi^j \in \mathbb{R}^{|A^i|}$, or equivalently $q^i(\cdot) = \mathbb{E}_{a^j \sim \pi^j} [r^i(\cdot, a^j)] \in \mathbb{R}^{|A^i|}$, with a slight abuse of notation since A^i is a finite set. We call q^i by *local Q -function*. Agent i can estimate q^i based on the observations he/she makes throughout the repeated play of the underlying game. The observations can include the opponent's actions and the rewards received. Let q_k^i denote the estimate at stage k . Then, agent i

Algorithm 1 Heterogeneous Learning in ZSGs

Require: $(\theta^i, \tau^i, (\alpha_k^i)_{k \geq 0})$

```

1: initialize:  $q_0^i$  arbitrarily
2: for each stage  $k = 0, 1, \dots$  do
3:   if  $a_k^j$  can be observed and  $R^i$  is known then
4:     play  $a_k^i \in \text{BR}^i(q_k^i)$  or  $a_k^i \sim \text{BR}_{\tau^i}^i(q_k^i)$   $\triangleright$  depending on  $\tau^i$ 
5:     observe  $a_k^j$   $\triangleright$  possible with probability  $\theta^i$ 
6:     set  $\hat{q}_k^i(\cdot) \equiv R^i a_k^j \in \mathbb{R}^{|A^i|}$ 
7:     set  $\bar{\alpha}_k^i(\tilde{a}^i) = \alpha_k^i$  for all  $\tilde{a}^i \in A^i$ 
8:   else
9:     play  $a_k^i \sim \text{BR}_{\tau^i}^i(q_k^i)$ 
10:    observe  $r_k^i = r^i(a_k^i, a_k^j) \in \mathbb{R}$ 
11:    set  $\hat{q}_k^i(\tilde{a}^i) = r_k^i$  for all  $\tilde{a}^i \in A^i$ 
12:    set  $\bar{\alpha}_k^i(\tilde{a}^i) = \mathbb{I}_{\{\tilde{a}^i = a_k^i\}} \min \left\{ 1, \frac{\alpha_k^i}{\text{BR}_{\tau^i}^i(q_k^i)(\tilde{a}^i)} \right\}$ 
13:   end if
14:   update  $q_{k+1}^i(\cdot) = q_k^i(\cdot) + \bar{\alpha}_k^i(\cdot)(\hat{q}_k^i(\cdot) - q_k^i(\cdot))$ 
15: end for

```

responds greedily to the estimate q_k^i .

As discussed above, the agents can have different means and interests, which can lead to heterogeneous learning dynamics. As a unifying framework, we present Algorithm Family 1, in which algorithms are parameterized by the probability of access to opponent actions $\theta^i \in [0, 1]$, the temperature parameter $\tau^i \geq 0$ and the step size $(\alpha_k^i)_{k \geq 0}$. The cases $\theta^i = 1$ and $\theta^i = 0$, respectively, correspond to the dynamics where the agent has full access (i.e., belief-based) or no access (i.e., payoff-based) to the opponent's actions while the case $\theta^i \in (0, 1)$ implies that the agent observes the opponent's action with probability θ^i at each stage independent of the history of the game. On the other hand, the case $\tau^i = 0$ implies that the agent takes actions according to the best response $\text{BR}^i(q_k^i)$, i.e., $a_k^i \in \text{BR}^i(q_k^i)$, whereas the case $\tau^i > 0$ implies the agent responds according to the smoothed best response $\text{BR}_{\tau^i}^i(q_k^i)$, i.e., $a_k^i \sim \text{BR}_{\tau^i}^i(q_k^i)$. The latter can be preferred for exploration if R^i is not known or for Hannan consistency, e.g., see [23], though it can also lead to occasional mistakes.

Well-studied fictitious play [24], smooth fictitious play [25], and individual Q-learning [26] are members of Algorithm Family 1. For example, in fictitious play,

agent i forms a belief, say π_k^j , about the opponent's mixed strategy based on the opponent's actions observed. If the belief gets updated according to

$$\pi_{k+1}^j = \pi_k^j + \alpha_k^i(a_k^j - \pi_k^j) \quad (2.6)$$

with some step size α_k^i , then $q_k^i = R^i \pi_k^j$ evolves as in step 14 in Algorithm Family 1. On the other hand, in the individual Q-learning dynamics, agent i normalizes the step size α_k^i with $\text{BR}_\tau^i(q_k^i)(a_k^i)$ to ensure that $q_k^i(a^i)$ for each a^i gets updated at the same rate in the expectation and the threshold on the normalized step size ensures that the iterates remain bounded. However, the convergence guarantees for (smooth) fictitious play and individual Q-learning are for homogeneous cases where every agent follows the same dynamic with identical step sizes, e.g., see [76, 77, 26].

2.3.2 Algorithm Family 2 for Stochastic Games

Consider learning in SGs. Following the trend in [54, 55, 56, 58, 59, 60], we can extend Algorithm Family 1 to SG settings through the *stage-game framework*. The stage-game framework was introduced by Shapley in [41], and also used in Minimax-Q [78] and Nash-Q [79] algorithms. Particularly, given $\mathcal{M} = \langle S, (A^i, r^i)_{i=1}^2, p, \gamma \rangle$, we can view stage-wise interactions among agents as they are playing certain matrix games (called *auxiliary stage games*) specific to each state whenever the associated state gets visited. Similar to the repeated play scheme, agents can follow an algorithm from Algorithm Family 1 for the current auxiliary stage game as if the opponent plays stationary strategy specific to the current state. In other words, agents can play the underlying auxiliary stage game as if the opponent plays according to Markov stationary strategy, as illustrated in Figure 2.2.

However, the payoffs of the auxiliary stage games depend both on the current reward received and the rewards to be received in the future stages, called *continuation payoff*. Given that opponent j follows Markov stationary strategy, say

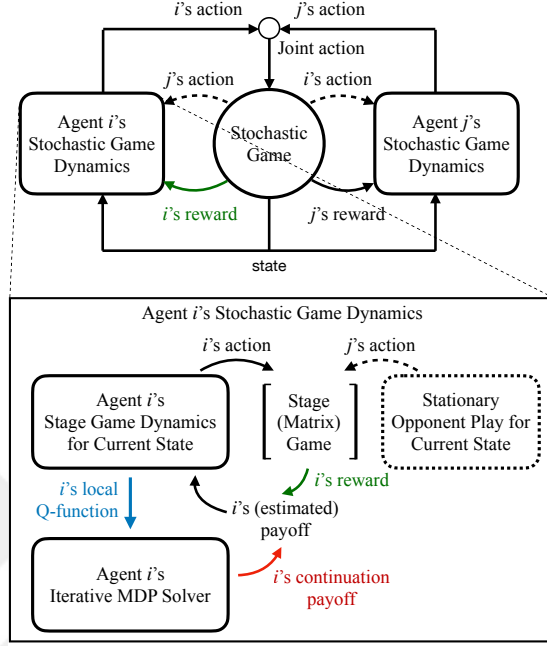


Figure 2.2: Agent i 's learning dynamics in SGs when the agent assumes that the opponent plays according to a Markov stationary strategy. We use dashed arrows for opponent actions since agents may not have access to them. The learning dynamics in SGs differ from the ones for repeated matrix games, illustrated in Figure 2.1 mainly due to the iterative MDP solver. Color coded arrows between stage game dynamics and the iterative MDP solver represent the coupling between them.

$\pi^j : S \rightarrow \Delta^j$, the principle of optimality yields that agent i 's payoffs in the auxiliary stage games, called *global Q-function* and denoted by $Q^i : S \times A^i \times A^j \rightarrow \mathbb{R}$, are given by

$$Q^i(s, a) = r^i(s, a) + \gamma \sum_{s_+ \in S} p(s_+ | s, a) v^i(s_+) \quad (2.7)$$

for all (s, a) , and the value function $v^i : S \rightarrow \mathbb{R}$ is defined by

$$v^i(s) = \max_{a^i \in A^i} \mathbb{E}_{a^j \sim \pi^j(s)} [Q^i(s, a^i, a^j)] \quad (2.8)$$

for all s . Under the stationary-opponent-play assumption, the agents can view that they play an underlying auxiliary stage game characterized by $\mathcal{G}(s) = \langle A^i, Q^i(s, \cdot) \rangle_{i=1}^2$ whenever s gets visited. Correspondingly, their local actions in

$\mathcal{G}(s)$ lead to the local Q-function

$$q^i(s, \cdot) = \mathbb{E}_{a^j \sim \pi^j(s)}[Q^i(s, \cdot, a^j)] \in \mathbb{R}^{A^i}.$$

They can estimate their value functions and Q-functions, by solving (2.7) and (2.8) iteratively, as illustrated in Figure 2.2.

Algorithm 2 Heterogeneous Learning for ZSSGs

Require: $(\theta^i, \tau^i, (\alpha_t^i)_{t \geq 0}, (\beta_t^i)_{t \geq 0})$

- 1: **initialize:** q_0^i and v_0^i arbitrarily, and the counter $t_0^s = 0, \forall s$
- 2: **for** each stage $k = 0, 1, \dots$ **do**
- 3: observe s_k
- 4: **if** $k=0$ **then**
- 5: jump step 21
- 6: **end if**
- 7: *Stage Game Dynamics for the Previous Stage Game:*
- 8: recall $s = s_{k-1}$ and $t = t_{k-1}^s$
- 9: **if** $a_t^j(s)$ got observed and $(r^i, p(\cdot|\cdot))$ is known **then**
- 10: recall $a^j = a_t^j(s)$ and set $\hat{q}_t^i(s, \cdot) \equiv r^i(s, \cdot, a^j) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s, \cdot, a^j) v_{t_{k-1}^{\tilde{s}}}^i(\tilde{s})$
- 11: set $\bar{\alpha}_t^i(s, \tilde{a}^i) = \alpha_t^i$ for all $\tilde{a}^i \in A^i$
- 12: **else**
- 13: recall $r_{k-1}^i \in \mathbb{R}$
- 14: set $\hat{q}_t^i(s, \tilde{a}^i) = r_{k-1}^i + \gamma v_{t_{k-1}^{\tilde{a}^i}}^i(s_k)$ for all $\tilde{a}^i \in A^i$
- 15: recall $a^i = a_t^i(s)$ and set $\bar{\alpha}_t^i(s, \tilde{a}^i) = \mathbb{I}_{\{\tilde{a}^i = a^i\}} \min \left\{ 1, \frac{\alpha_t^i}{\text{BR}_{\tau^i}(q_t^i(s, \cdot))(\tilde{a}^i)} \right\}$
- 16: **end if**
- 17: update $q_{t+1}^i(s, \cdot) \equiv q_t^i(s, \cdot) + \bar{\alpha}_t^i(s, \cdot)(\hat{q}_t^i(s, \cdot) - q_t^i(s, \cdot))$
- 18: *Iterative MDP Solver for the Previous Stage Game:*
- 19: set $\mu_t^i(s) = a_t^i(s) \in \text{BR}^i(q_t^i(s, \cdot))$ or $\mu_t^i(s) = \text{BR}_{\tau^i}(q_t^i(s, \cdot))$
- 20: update $v_{t+1}^i(s) = v_t^i(s) + \beta_t^i \left(\mu_t^i(s)^T q_t^i(s, \cdot) - v_t^i(s) \right)$
- 21: increment the counter $t_k^{s'} = t_{k-1}^{s'} + \mathbb{I}_{\{s'=s\}}$ for all $s' \in S$
- 22: *Play:*
- 23: recall $\tilde{s} = s_k$ and $\tilde{t} = t_k^{\tilde{s}}$
- 24: **if** a_k^j can be observed and $(r^i(\cdot), p(\cdot|\cdot))$ is known **then**
- 25: play $a_t^i(\tilde{s}) \in \text{BR}^i(q_t^i(\tilde{s}, \cdot))$ or $a_t^i(\tilde{s}) \sim \text{BR}_{\tau^i}(q_t^i(\tilde{s}, \cdot))$ $\triangleright$ depending on τ^i
- 26: observe a_k^j $\triangleright$ possible with probability θ^i
- 27: **else**
- 28: play $a_t^i(\tilde{s}) \sim \text{BR}_{\tau^i}(q_t^i(\tilde{s}, \cdot))$
- 29: observe $r_k^i = r^i(s_k, a_k^i, a_k^j)$
- 30: **end if**
- 31: **end for**

To address the heterogeneities that can arise due to the agents' different means and interests in a unifying framework, we present Algorithm Family 2, in which algorithms are parameterized by the probability of access to opponent actions $\theta^i \in [0, 1]$, the temperature parameter $\tau^i \geq 0$, and the step size $(\alpha_t^i)_{t \geq 0}$ as in Algorithm Family 1, and additionally the step size $(\beta_t^i)_{t \geq 0}$. The step size β_t^i is used in the iterative MDP solver.

Remark 1. *In Algorithm Family 1, the updates of local Q function estimates take place at the end. On the other hand, in Algorithm Family 2, the updates take place after the next state observed for a unified description addressing the one-stage look ahead for payoff-based learning. In other words, the update for the previous stage game takes place before the agents play in the current stage game.*

Remark 2. *The two-timescale model-based fictitious play dynamics presented by [55], and the decentralized Q -learning dynamics presented by [56] can be viewed as members of Algorithm Family 2. However, the family does not include the two-timescale model-free fictitious play dynamics from [55]. The family instead treats model-free cases as if they are payoff-based. Agents by following this approach do not need to rely on the opponent to play every action with some positive probability to explore the underlying global Q -function properly. This can play an important role in exploration when the opponent prefers the best response rather than smoothed best response.*

2.4 Convergence Results

This section addresses the convergence properties of any mixture of learning dynamics from Algorithm Families 1 and 2, respectively, for ZSGs and ZSSGs.

2.4.1 Convergent Heterogeneous Learning for ZSGs

In Algorithm Family 1, agents can use different step sizes satisfying the following standard assumption:

Assumption 1. *The local step sizes $(\alpha_k^i \in (0, 1))_{k=0}^\infty$ decay monotonically $\alpha_k^i \rightarrow 0$ as $k \rightarrow \infty$ and $\sum_{k \geq 0} \alpha_k^i = \infty$ for each $i = 1, 2$. Furthermore, $\sum_{k \geq 0} (\alpha_k^i)^2 < \infty$ if agent i plays according to the smoothed best response $\text{BR}_\tau^i(\cdot)$.*

In the convergence analysis, we apply *stochastic differential inclusion approximation* methods from [28] as the best response may not be unique. The preliminary information about these methods are provided in Appendix A. For a unified representation, we let τ^i (or τ^j) be zero if agent i (or agent j) plays according to the best response, and define

$$\text{val}^i(R^i) := \max_{\mu^i \in \Delta^i} \min_{\mu^j \in \Delta^j} \{(\mu^i)^T R^i \mu^j\} \quad (2.9a)$$

$$\underline{\text{val}}^i(R^i) := \max_{\tilde{\mu}^i \in \Delta^i} \min_{\tilde{\mu}^j \in \Delta^j} \left\{ (\tilde{\mu}^i)^T R^i \tilde{\mu}^j + \tau^i H^i(\tilde{\mu}^i) - \tau^j H^j(\tilde{\mu}^j) \right\}, \quad (2.9b)$$

corresponding, respectively, to the NE and QRE equilibrium values of the zero-sum game with payoff matrix R^i .³ Even under perturbations, $\underline{\text{val}}^i(\cdot)$ has the contraction property with respect to the maximum norm, denoted by $\|\cdot\|_{\max}$, as shown in the following lemma, whose proof is moved to Appendix B.

Lemma 1. *For $\text{val}^i : \mathbb{R}^{|A^i| \times |A^j|} \rightarrow \mathbb{R}$ and $\underline{\text{val}}^i : \mathbb{R}^{|A^i| \times |A^j|} \rightarrow \mathbb{R}$, as described in (2.9), we have*

$$\begin{aligned} \underline{r} &\leq |\text{val}^1(R^1) + \text{val}^2(R^2)| \leq \bar{r} \\ \underline{r} &\leq |\underline{\text{val}}^1(R^1) + \underline{\text{val}}^2(R^2)| \leq \bar{r}, \end{aligned}$$

where $\underline{r}$ and $\bar{r} \in \mathbb{R}$ are, resp., the minimum and maximum entries of the matrix

³The uniqueness of the equilibrium value follows from the Minimax Theorem.

$R^1 + (R^2)^T$, and

$$-\tau^j \log |A^j| \leq \underline{\text{val}}^i(R^i) - \text{val}^i(R^i) \leq \tau^i \log |A^i|$$

for any $R^i \in \mathbb{R}^{|A^i| \times |A^j|}$.

The following theorem characterizes the convergence properties of heterogeneous learning dynamics in near ZSGs.

Theorem 1. *Consider the repeated play of $\mathcal{G} = \langle A^i, R^i \rangle_{i=1}^2$. Suppose that agents follow a learning dynamic from Algorithm Family 1 and the local step sizes satisfy Assumption 1. Furthermore, the local step sizes are comparable such that $\lim_{k \rightarrow \infty} \alpha_k^i / \alpha_k^j \in (0, \infty)$ and, without loss of generality, we assume that $\alpha_k^1 / \alpha_k^2 \rightarrow d \in (0, 1]$ as $k \rightarrow \infty$. Let $\pi_k^i \in \Delta^i$ denote a weighted empirical average of the agent i 's past actions until stage k . Let $a_k^i \in A^i$ denote the action played by agent i at stage k . Then, π_k^i evolves according to*

$$\pi_{k+1}^i = \pi_k^i + \alpha_k^j (a_k^i - \pi_k^i) \quad (2.10)$$

with arbitrary π_0^i . We have $\|q_k^i - R^i \pi_k^j\|_2 \rightarrow 0$ as $k \rightarrow \infty$ and

$$\limsup_{k \rightarrow \infty} (\delta_k^1 + d \cdot \delta_k^2) \leq c \quad (2.11)$$

almost surely, where

$$\delta_k^i := (\mu_k^i)^T q_k^i + \tau^i H^i(\mu_k^i) - \tau^j H^j(\pi_k^j) - \underline{\text{val}}^i(R^i), \quad (2.12)$$

$$c := \lambda \|R^1 + (R^2)^T\|_{\max} - (\underline{\text{val}}^1(R^1) + \underline{\text{val}}^2(R^2)) \quad (2.13)$$

for some $\lambda > 1$, and $\mu_k^j = a_k^j \in \text{BR}^j(q_k^j)$ (or $\mu_k^j = \text{BR}_\tau^j(q_k^j)$) if agent j plays according to the best response (or smoothed best response). We also have $\delta_k^i \geq \|q_k^i - R^i \pi_k^j\|_2$.

Proof. The local Q-function estimates remain bounded in Algorithm Family 1 since $\bar{\alpha}_k^i(a^i) \in [0, 1]$ for all $a^i \in A^i$. For example, $\limsup_{k \rightarrow \infty} \|q_k^i\|_\infty \leq \|R^i\|_{\max}$.

By (2.5a), the boundedness of the estimates and decaying step size imply that the normalization at step 12 becomes redundant eventually. Hence, for any $\epsilon > 0$, there exists a stage K_ϵ such that $\bar{\alpha}_k^i(a^i) = \mathbb{I}_{\{a^i=a_k^i\}} \alpha_k^i / \text{BR}_\tau^i(q_k^i)(a^i)$ and $q_k^i \in \Xi^i$, where Ξ^i is a compact subset of $\mathbb{R}^{|A^i|}$, for all $k \geq K_\epsilon$. The evolution of the local Q-function estimates q_k^i and the empirical averages of actions π_k^j can be written as

$$\begin{bmatrix} q_{k+1}^i \\ \pi_{k+1}^j \end{bmatrix} = \begin{bmatrix} q_k^i \\ \pi_k^j \end{bmatrix} + \alpha_k^i \left(\begin{bmatrix} R^i \mu_k^j \\ \mu_k^j \end{bmatrix} + \begin{bmatrix} \underline{w}_k^i \\ \bar{w}_k^j \end{bmatrix} - \begin{bmatrix} q_k^i \\ \pi_k^j \end{bmatrix} \right) \quad (2.14)$$

for all $k \geq K_\epsilon$. In the update of π_k^j , the noise term $\bar{w}_k^j = a_k^j - \mu_k^j$. In the update of q_k^i , the noise term $\underline{w}_k^i = R^i a_k^j - R^i \mu_k^j$ if agent i observes a_k^j and knows R^i . If agent i does not observe a_k^j or know R^i , then we have

$$\underline{w}_k^i(a^i) = \frac{\mathbb{I}_{\{a^i=a_k^i\}}}{\mu_k^i(a^i)} (r^i(a^i, a_k^j) - q_k^i(a^i)) - (\mathbb{E}_{a^j \sim \mu_k^j} [r^i(a^i, a^j)] - q_k^i(a^i)). \quad (2.15)$$

Let $x_k = [q_k^i; \pi_{i=1}^j]^2$. If agent j plays according to the best response, let $F^{ij}(x_k) \subset \mathbb{R}^{|A^i|+|A^j|}$ be a set-valued function defined by⁴

$$F^{ij}(x_k) = \left\{ \begin{bmatrix} R^i \mu^j \\ \mu^j \end{bmatrix} - \begin{bmatrix} q_k^i \\ \pi_k^j \end{bmatrix} : \mu^j \in \text{Conv}(\text{BR}^j(q_k^j)) \right\}. \quad (2.16)$$

If agent j plays according to the smoothed best response, let $F^{ij}(x_k) \in \mathbb{R}^{|A^i|+|A^j|}$ be a single-valued function defined by

$$F^{ij}(x_k) = \begin{bmatrix} R^i \text{BR}_\tau^j(q_k^j) \\ \text{BR}_\tau^j(q_k^j) \end{bmatrix} - \begin{bmatrix} q_k^i \\ \pi_k^j \end{bmatrix}. \quad (2.17)$$

Note that the noise-free innovation term in (2.14) satisfies $\begin{bmatrix} R^i \mu_k^j \\ \mu_k^j \end{bmatrix} - \begin{bmatrix} q_k^i \\ \pi_k^j \end{bmatrix} \in$

⁴Let $\text{Conv}(X)$ denote the convex hull of the set X .

$F^{ij}(x_k)$. Therefore, the evolution of the local Q-function estimates q_k^i and empirical averages of actions π_k^j satisfies

$$\begin{bmatrix} q_{k+1}^i \\ \pi_{k+1}^j \end{bmatrix} - \begin{bmatrix} q_k^i \\ \pi_k^j \end{bmatrix} - \alpha_k^i \begin{bmatrix} \underline{w}_k^i \\ \overline{w}_k^j \end{bmatrix} \in \alpha_k^i \cdot F^{ij}(x_k). \quad (2.18)$$

We can show that $x_k \in X = \prod_{i=1}^2 \Xi^i \times \Delta^j \subset \mathbb{R}^{2m}$, where $m = |A^1| + |A^2|$, satisfies

$$x_{k+1} - x_k - \alpha_k(w_k + e_k) \in \alpha_k \cdot F(x_k), \quad (2.19)$$

for all $k \geq K_\epsilon$, where the step size $\alpha_k := \alpha_k^1$, the noise term

$$w_k = \left[\underline{w}_k^1, \overline{w}_k^2, \frac{\alpha_k^2}{\alpha_k^1} \underline{w}_k^2, \frac{\alpha_k^2}{\alpha_k^1} \overline{w}_k^1 \right],$$

the set-valued map $F(x_k) \subset \mathbb{R}^{2m}$ is given by

$$F(x_k) = \left\{ \begin{bmatrix} f^{12} \\ \frac{1}{d} \cdot f^{21} \end{bmatrix} : f^{ij} \in F^{ij}(x_k) \text{ for } i = 1, 2 \right\}, \quad (2.20)$$

and the error term is defined by

$$e_k = \begin{bmatrix} \mathbf{0} \\ \left(\frac{\alpha_k^2}{\alpha_k^1} - \frac{1}{d} \right) \left(\begin{bmatrix} R^2 \mu_k^1 \\ \mu_k^1 \end{bmatrix} - \begin{bmatrix} q_k^2 \\ \pi_k^1 \end{bmatrix} \right) \end{bmatrix}, \quad (2.21)$$

which is asymptotically negligible as $\alpha_k^2/\alpha_k^1 \rightarrow 1/d \in [1, \infty)$.

For the cases where agent j plays according to the best response, the set-valued function $F^{ij}(\cdot)$, as described in (2.16), is a Marchaud map (described in Appendix A.1) because the best response function $\text{BR}^j(\cdot)$ is a Marchaud map, e.g., see [28, Section 2.2]. For the cases where agent j plays according to the smoothed best response, the single-valued function $F^{ij}(\cdot)$, as described in (2.17), is Lipschitz continuous since the smoothed best response function $\text{BR}_\tau^j(\cdot)$ is $1/\tau$ -Lipschitz continuous, e.g., see [80, Proposition 4]. Since $F^{ij}(\cdot)$ is a Marchaud map in each case, by (2.20), we can show that $F(\cdot)$ is a Marchaud map.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space and $\{\mathcal{F}_k = \sigma(q_l^i, \pi_l^i \text{ for } i = 1, 2, \text{ and } l \leq k)\}_{k \geq 0}$ be a filtration of $\mathcal{F}$. Note that $\underline{w}_{k-1}^i$ and $\overline{w}_{k-1}^i$ for each $i = 1, 2$, and therefore, w_{k-1} are $\mathcal{F}_k$ -measurable for each $k > 0$. Furthermore, we have $\mathbb{E}[\overline{w}_k^i | \mathcal{F}_k] = \mathbf{0}$ and $\mathbb{E}[\|\overline{w}_k^i\|^2 | \mathcal{F}_k]$ is bounded from above by some finite number uniformly for all $k \geq K_\epsilon$ since

- $\overline{w}_k^i = \mathbf{0}$ if agent i plays according to the best response,
- $\overline{w}_k^i = a_k^i - \text{BR}_\tau^i(q_k^i)$, where $a_k^i \sim \text{BR}_\tau^i(q_k^i)$ with finite support, if agent i plays according to the smoothed best response.

On the other hand, if agent i does not know R^i , then $\underline{w}_k^i$ is as described in (2.15), and therefore, we can show that $\mathbb{E}[\underline{w}_k^i | \mathcal{F}_k] = \mathbf{0}$ and $\mathbb{E}[\|\underline{w}_k^i\|^2 | \mathcal{F}_k]$ is bounded from above by some finite number uniformly for all $k \geq K_\epsilon$. If agent i knows R^i , then $\underline{w}_k^i = R^i a_k^j - R^i \mu_k^j$ with probability θ^i and $\underline{w}_k^i$ is as described in (2.15) with probability $(1 - \theta^i)$ independent of $\mathcal{F}_k$. Therefore, we also have $\mathbb{E}[\underline{w}_k^i | \mathcal{F}_k] = \mathbf{0}$ and $\mathbb{E}[\|\underline{w}_k^i\|^2 | \mathcal{F}_k]$ is bounded from above by some finite number uniformly for all $k \geq K_\epsilon$.

Correspondingly, $\mathbb{E}[w_k^i | \mathcal{F}_k] = \mathbf{0}$ and $\mathbb{E}[\|w_k^i\|^2 | \mathcal{F}_k] < W$ for some finite W for all $k \geq K_\epsilon$ since w_k is a linear function of $\underline{w}_k^i$'s and $\overline{w}_k^i$'s while the deterministic α_k^2 / α_k^1 is uniformly bounded away from zero for all $k \geq K_\epsilon$. Furthermore, Assumption 1 holds for the step size α_k . Therefore, the iterate x_k converges to a compact connected internally chain transitive set of the differential inclusion $\dot{x} \in F(x)$.

Since $F(\cdot)$ is a Marchaud map, there always exists a solution to $\dot{x} \in F(x)$ with initial point $x_o \in X$ that is an absolutely continuous mapping $x : \mathbb{R} \rightarrow X$ such that $x(0) = x_o$, and $\frac{dx(t)}{dt} \in F(x(t))$ for almost every $t \in \mathbb{R}$ [28]. Define $q^i : \mathbb{R} \rightarrow \Xi^i$ and $\pi^j : \mathbb{R} \rightarrow \Delta^j$ such that $x(t) = [q^i(t); \pi^j(t)]_{i=1}^2$. Then, for almost every $t \in \mathbb{R}$, $q^i(t)$ and $\pi^j(t)$ satisfy

$$\frac{dq^i(t)}{dt} = d^i(R^i \mu^j(t) - q^i(t)) \quad (2.22a)$$

$$\frac{d\pi^j(t)}{dt} = d^j(\mu^j(t) - \pi^j(t)) \quad (2.22b)$$

for each $i = 1, 2$ and $j \neq i$, where $\mu^j(t) \in \text{BR}^j(q^j(t))$ (and $\mu^j(t) = \text{BR}_{\tau^j}(q^j(t))$) if agent j plays according to the best response (and the smoothed best response), and $d^1 = 1$ and $d^2 = 1/d \in [1, \infty)$.

Next, we define the continuous function $V : X \rightarrow [0, \infty)$ by⁵

$$V(q^1, \pi^2, q^2, \pi^1) = (L^1(q^1, \pi^2) + d \cdot L^2(q^2, \pi^1) - c)_+ + \sum_{i=1}^2 \|q^i - R^i \pi^j\|_2, \quad (2.23)$$

where c is as described in (2.13) and $L^i : \Xi^i \times \Delta^j \rightarrow \mathbb{R}$ is defined by

$$L^i(q^i, \pi^j) := \max_{\tilde{\mu}^i \in \Delta^i} \{(\tilde{\mu}^i)^T q^i + \tau^i H^i(\tilde{\mu}^i) - \tau^j H^j(\pi^j)\} + \|q^i - R^i \pi^j\|_2 - \underline{\text{val}}^i(R^i). \quad (2.24)$$

Lemma 1 yields that $c \geq (\lambda - 1)\|R^1 + (R^2)^T\|_{\max} \geq 0$ and $c = 0$ if and only if $\|R^1 + (R^2)^T\|_{\max} = 0$ as $\lambda > 1$.

Let $\bar{\mu}^i \in \text{argmax}_{\bar{\mu}^i} \{(\bar{\mu}^i)^T R^i \pi^j + \tau^i H^i(\bar{\mu}^i)\}$.⁶ Then, we have

$$\begin{aligned} L^i(q^i, \pi^j) &\geq (\bar{\mu}^i)^T q^i + \tau^i H^i(\bar{\mu}^i) - \tau^j H^j(\pi^j) + \|q^i - R^i \pi^j\|_2 - \underline{\text{val}}^i(R^i) \\ &= \underbrace{(\bar{\mu}^i)^T R^i \pi^j + \tau^i H^i(\bar{\mu}^i) - \tau^j H^j(\pi^j) - \underline{\text{val}}^i(R^i)}_{\geq 0} \\ &\quad + \underbrace{(\bar{\mu}^i)^T (q^i - R^i \pi^j) + \|q^i - R^i \pi^j\|_2}_{\geq 0} \end{aligned} \quad (2.25)$$

where the first term on the right-hand side is non-negative by the definitions of $\underline{\text{val}}^i(R^i)$ and $\bar{\mu}^i$, and the second term is non-negative as $\bar{\mu}^i \in \Delta^i$ is a probability distribution and $\|x\|_2 \geq \|x\|_{\infty}$ for any vector x . Hence, $L^i(q^i, \pi^j) \geq 0$ is a *non-negative* function.

Remark 3. *If the underlying game is zero-sum, then we have $c = 0$ and $(L^1 + dL^2)_+ = L^1 + dL^2$. Furthermore, if we also had $q^i = R^i \pi^j$ and $d = 1$, then*

⁵Let $(x)_+ := (x + |x|)/2$.

⁶We have set inclusion rather than equality to address the case $\tau^i = 0$.

the function (2.23) reduces to

$$\sum_{i=1}^2 \max_{\tilde{\mu}^i \in \Delta^i} \{ (\tilde{\mu}^i)^T R^i \pi^j + \tau^i H^i(\tilde{\mu}^i) - \tau^j H^j(\pi^j) \}, \quad (2.26)$$

as formulated in [76] (and [81]) if both agents play according to the best response with $\tau^1 = \tau^2 = 0$ (and the smoothed best response with $\tau^1, \tau^2 > 0$).

The candidate function (2.23) also resembles the Lyapunov functions formulated in [56] and [73]. The function differs from the former by addressing asymmetric step sizes and convergence to quantal response equilibrium when agents play according to the smoothed best response. The function differs from the latter by addressing the cases with smoothed best response and payoff-based learning where q^i may not be equal to $R^i \pi^j$. In that respect, we can view (2.23) as a unifying Lyapunov function addressing all these cases.

In the following, we show the validity of (2.23) as a Lyapunov function for (2.22). Given a solution $x(t) = [q^i(t); \pi^j(t)]_{i=1}^2$ to (2.22), the time derivatives of $\|q^i - R^i \pi^j\|$ and $L^i(q^i(t), \pi^j(t))$ are, resp. given by

$$\frac{d\|q^i - R^i \pi^j\|}{dt} = \frac{(q^i - R^i \pi^j)^T (\dot{q}^i - R^i \dot{\pi}^j)}{\|q^i - R^i \pi^j\|} \quad (2.27a)$$

$$\frac{dL^i}{dt} = (\mu^i)^T \dot{q}^i - \tau^j \nabla H^j(\pi^j) \dot{\pi}^j + \frac{d\|q^i - R^i \pi^j\|}{dt} \quad (2.27b)$$

almost everywhere, based on [76] (and [81]) if agent i plays according to the best response (and the smoothed best response). After some algebra, we obtain $\frac{d}{dt} \|q^i - R^i \pi^j\| = -d^i \|q^i - R^i \pi^j\|$ and

$$\frac{dL^i}{dt} \leq d^i [(\mu^i)^T R^i \mu^j - \underline{\text{val}}^i + \tau^i H^i(\mu^i) - \tau^j H^j(\mu^j)] - d^i \cdot L^i(q^i, \pi^j), \quad (2.28)$$

where the inequality follows from $H^j(\mu^j) - H^j(\pi^j) - \nabla H^j(\pi^j)(\mu^j - \pi^j) \leq 0$ due to the concavity of entropy function. Recall that $d^1 = 1$ and $d^2 = 1/d$. Therefore,

we have

$$\begin{aligned} \frac{d}{dt}(L^1 + d \cdot L^2) &\leq (\mu^1)^T(R^1 + (R^2)^T)\mu^2 - (\underline{\text{val}}^1(R^1) + \underline{\text{val}}^2(R^2)) \\ &\quad - L^1 - d \cdot L^2 + \underbrace{(d-1)L^2}_{\leq 0}, \end{aligned}$$

where the last term on the right-hand side is non-positive since $d \in (0, 1]$ and $L^2(\cdot)$ is a non-negative function. By (2.13), we obtain

$$\frac{d}{dt}(L^1 + d \cdot L^2) \leq -(L^1 + d \cdot L^2 - c), \quad (2.29)$$

with strict inequality in non-zero-sum games, i.e., when $\|R^1 + (R^2)^T\|_{\max} \neq 0$. Hence, the set $\Lambda = \{(q^1, \pi^2, q^2, \pi^1) \in X : L^1(q^1, \pi^2) + dL^2(q^2, \pi^1) \leq c \text{ and } q^i = R^i \pi^j, \text{ for } i = 1, 2\}$ is positively invariant for almost every solution of (2.22) and $V(\cdot)$ is a Lyapunov function for Λ . Note that $V(\Lambda) = 0$ is a singleton. Therefore, we have

$$\lim_{k \rightarrow \infty} V(q_k^1, \pi_k^2, q_k^2, \pi_k^1) = 0 \quad (2.30)$$

almost surely. Then, the definition of $V(\cdot)$ implies that there exists an asymptotically negligible sequence $\zeta_k \rightarrow 0$ such that

$$L^1(q_k^1, \pi_k^2) + d \cdot L^2(q_k^2, \pi_k^1) \leq c + \zeta_k \quad \forall k, \quad (2.31)$$

and $\|q_k^i - R^i \pi_k^j\|_2 \rightarrow 0$ almost surely for each $i = 1, 2$. By (2.24) and the non-negativity of $L^i(\cdot)$, we have

$$\delta_k^1 + d \cdot \delta_k^2 \leq c + \zeta_k - \|q_k^1 - R^1 \pi_k^2\|_2 - d \cdot \|q_k^2 - R^2 \pi_k^1\|_2,$$

and $\delta_k^i \geq \|q_k^i - R^i \pi_k^j\|_2$, where δ_k^i is as described in (2.12). The proof is completed as $\|q_k^i - R^i \pi_k^j\|_2 \rightarrow 0$. $\square$

The following corollary to Theorem 1 characterizes the near equilibrium convergence of the dynamics in near ZSGs. To this end, we define

$$u^i(\pi^i, \pi^j) := (\pi^i)^T R^i \pi^j + \tau^i H^i(\pi^i) - \tau^j H^j(\pi^j) \quad (2.32)$$

for all $(\pi^i, \pi^j) \in \Delta^i \times \Delta^j$ and $i = 1, 2$.

Corollary 1. *Consider the repeated play of $\mathcal{G} = \langle A^i, R^i \rangle_{i=1}^2$. Suppose that agents follow a learning dynamic from Algorithm Family 1 and the local step sizes satisfy Assumption 1. Furthermore, the local step sizes are comparable such that $\lim_{k \rightarrow \infty} \alpha_k^i / \alpha_k^j \in (0, \infty)$ and, without loss of generality, we assume that $\alpha_k^1 / \alpha_k^2 \rightarrow d \in (0, 1]$ as $k \rightarrow \infty$. Let $\pi_k^i \in \Delta^i$ denote a weighted empirical average of the agent i 's past actions, evolving according to (2.10) for all $k \geq 0$ with arbitrary π_0^i . Then, we have*

$$\begin{aligned} \limsup_{k \rightarrow \infty} (u^i(\pi_k^i, \pi_k^j) - \underline{\text{val}}^i(R^i)) &\leq \frac{2}{d^i} \|\bar{R}\|_{\max} \\ \liminf_{k \rightarrow \infty} (u^i(\pi_k^i, \pi_k^j) - \underline{\text{val}}^i(R^i)) &\geq -2 \left(1 + \frac{1}{d^j}\right) \|\bar{R}\|_{\max} \end{aligned}$$

almost surely for each $i = 1, 2$ and $j \neq i$, where $\bar{R} := R^1 + (R^2)^T$ and $(d^1, d^2) = (1, d)$. Hence, (π_k^1, π_k^2) converges almost surely to NE or QRE (depending on, resp., whether $\tau^i = 0$ or $\tau^i > 0$) in ZSGs, i.e., when $\|\bar{R}\|_{\max} = 0$, for any $d \in (0, 1]$.

Proof. Since $\delta_k^i \geq \|q_k^i - R^i \pi_k^j\|_2$ and $\|q_k^i - R^i \pi_k^j\|_2 \rightarrow 0$, Theorem 1 yields that

$$u^i(\pi_k^i, \pi_k^j) - \underline{\text{val}}^i(R^i) \leq \frac{c}{d^i} + \zeta_k \quad (2.33)$$

for all $k \geq 0$ and $i = 1, 2$, for some $\zeta_k \rightarrow 0$ as $k \rightarrow \infty$ almost surely. We have the upper bound as Lemma 1 yields that $c \leq (\lambda + 1) \|\bar{R}\|_{\max}$.

To compute a lower bound, we add and subtract $(\pi_k^i)^T (R^j)^T \pi_k^j$ to the left-hand side of (2.33). Then, we can write the inequality as

$$-(u^j(\pi_k^j, \pi_k^i) - \underline{\text{val}}^j(R^j)) \leq \frac{c}{d^j} + \zeta_k - (\pi_k^i)^T \bar{R} \pi_k^j + \sum_{t=1}^2 \underline{\text{val}}^t(R^t). \quad (2.34)$$

Note that $(\pi_k^i)^T \bar{R} \pi_k^j \geq -\|\bar{R}\|_{\max}$ as π_k^i 's are probability distributions. Then,

Lemma 1 yields that

$$(u^j(\pi_k^j, \pi_k^i) - \underline{\text{val}}^j(R^j)) \geq -\frac{c}{d^i} - 2\|\bar{R}\|_{\max} - \zeta_k. \quad (2.35)$$

By changing the agent index from j to i , we obtain the lower bound. $\square$

The following corollary to Theorem 1 shows the rationality of the dynamics from Algorithm Family 1.

Corollary 2. *Consider the repeated play of $\mathcal{G} = \langle A^i, R^i \rangle_{i=1}^2$. Suppose that agent i follows a learning dynamic from Algorithm Family 1 while the opponent j plays according to mixed strategy $\pi^j \in \Delta^j$. The local step size $(\alpha_k^i)_{k \geq 0}$ satisfy Assumption 1. Let $\pi_k^i \in \Delta^i$ denote a weighted empirical average of the agent i 's past actions, evolving according to (2.10) with arbitrary initialization. Then, we have $u^i(\pi_k^i, \pi^j) \rightarrow \max_{\pi^i \in \Delta^i} \{u^i(\pi^i, \pi^j)\}$ almost surely as $k \rightarrow \infty$.*

Proof. We can consider that agent i is playing a ZSG against an opponent (other than agent j) following the same dynamic with a singleton action space and view agent j as the Nature. Then, the proof follows from Corollary 1. $\square$

2.4.2 Convergent Heterogeneous Learning for ZSSGs

Algorithm Family 2 involves two local step sizes α_k^i and β_k^i . The following is the counterpart of Assumption 1 for β_k^i .

Assumption 2. *The local step sizes $(\beta_k^i \in (0, 1))_{k=0}^\infty$ decay monotonically $\beta_k^i \rightarrow 0$ as $k \rightarrow \infty$ and $\sum_{k \geq 0} \beta_k^i = \infty$ for each $i = 1, 2$.*

Furthermore, we assume that α_k^i and β_k^i decay in different time scales at certain rates as in [56, Assumption 1-ii].

Assumption 3. *Given any $M \in (0, 1)$, there exists a non-decreasing polynomial*

function $C(\cdot)$ (which may depend on M) such that we have

$$\max \{ \ell \in \mathbb{Z}_+ | \ell \leq c \text{ and } \beta_\ell / \alpha_c > \lambda \} \leq Mc \quad (2.36)$$

for all $c \geq C(\lambda^{-1})$ for any $\lambda \in (0, 1)$ provided that $\{ \ell \in \mathbb{Z}_+ | \ell \leq c \text{ and } \beta_\ell / \alpha_c > \lambda \} \neq \emptyset$.

In SGs, agents play stage games associated with the state visited. Whether each state gets visited infinitely often plays an important role in ensuring that agents play stage games associated with each state sufficiently many times to reach equilibrium. To this end, consider the directed graph G whose vertices correspond to the states of the underlying SG and there is an edge from s to s' provided that

- *Case i)* - Both agent play according to the best response: $p(s' | s, a) > 0$ for all $a \in A$,
- *Case ii)* - Only agent i plays according to the best response: $p(s' | s, a^i, a^j) > 0$ for all $a^i \in A^i$ and at least for one $a^j \in A^j$,
- *Case iii)* - Both agent play according to the smoothed best response: $p(s' | s, a) > 0$ at least for one action profile $a \in A$.

We make the following assumption for the graph G .

Assumption 4. *The graph G is strongly connected (each vertex is reachable from every other vertex).*

Let $Q_*^i(s, a)$ and $v_*^i(s)$ denote the unique Q-function and value function associated with Markov stationary equilibrium of the underlying ZSSG. Then, they satisfy

$$Q_*^i(s, a) = r^i(s, a) + \gamma \sum_{s_+ \in S} p(s_+ | s, a) v_*^i(s_+) \quad (2.37)$$

and $v_*^i(s) = \text{val}^i(Q_*^i(s, \cdot))$. We also have $v_*^1(s) + v_*^2(s) = 0$ for all s in ZSSGs.

The following theorem addresses the convergence of heterogeneous learning dynamics for ZSSGs.

Theorem 2. *Consider a ZSSG characterized by $\mathcal{M} = \langle S, (A^i, r^i)_{i=1}^2, p, \gamma \rangle$. Suppose that agents follow a learning dynamic from Algorithm Family 2 and the local step sizes α_k^i and β_k^i 's satisfy, respectively, Assumptions 1, 2 and 3. Furthermore,*

- *the local step sizes for the local Q -function update are comparable such that $\lim_{k \rightarrow \infty} \alpha_k^i / \alpha_k^j \in (0, \infty)$ and, without loss of generality, we assume that $\alpha_k^1 / \alpha_k^2 \rightarrow d \in (0, 1]$ as $k \rightarrow \infty$,*
- *the local step sizes for the value function updates may or may not be comparable.*

Let $\pi_t^i(s) \in \Delta^i$ denote a weighted empirical average of the agent i 's actions at state s until the t th visit. Let $a_t(s) \in A^i$ denote the action played by agent i at the t th visit to state s . Then, $\pi_t(s)$ evolves according to

$$\pi_{t+1}^i(s) = \pi_t^i(s) + \alpha_t^j(a_t^i(s) - \pi_t^i(s)) \quad (2.38)$$

with arbitrary $\pi_0^i(s)$. Let $Q_t^i(s, a) \in \mathbb{R}$ denote the global Q -function associated with the value function estimates and defined by

$$Q_t^i(s, a) := r^i(s, a) + \gamma \sum_{s_+ \in S} p(s_+ | s, a) v_t^i(s_+). \quad (2.39)$$

If Assumption 4 also holds and the discount factor $\gamma \in [0, d/2)$, then we have $\|q_t^i(s, \cdot) - Q_t^i(s, \cdot)\pi_t^j(s)\|_2 \rightarrow 0$ as $t \rightarrow \infty$ and

$$\limsup_{t \rightarrow \infty} |v_t^i(s) - v_*^i(s)| \leq \frac{2d + 2\gamma - 3\gamma d}{(1 - \gamma)(d - 2\gamma)} \sum_{l=1}^2 \tau^l \log |A^l|$$

almost surely for each $i = 1, 2$ and $s \in S$.

Proof. The local Q -function estimates again remain bounded in Algorithm Family 2 since the step sizes in steps 16 and 18 take values in $[0, 1]$. Due to

(2.5a), Assumption 4 and the boundedness of the estimates yield that any state gets visited infinitely often, i.e., $t_k^s \rightarrow \infty$ almost surely for each s as $k \rightarrow \infty$. Furthermore, the decaying step sizes imply that the normalization at step 14 becomes redundant eventually for each s . Hence, for any $\epsilon > 0$, there exists a (path-dependent) stage K_ϵ such that the normalization is redundant, and the iterates $q_{t_k^s}^i(s, \cdot) \in \Xi_q^i$ and $v_{t_k^s}^i(s) \in \Xi_v^i$, where Ξ_q^i and Ξ_v^i are compact subsets of, respectively, $\mathbb{R}^{|A^i|}$ and $\mathbb{R}$, for all $k \geq K_\epsilon$.

Define $\bar{v}_k^i(s) = v_{t_k^s}^i(s)$ for all s . Following the proof in [56], we fix state s and let k_t denote the stage at which state s gets visited for the t th times. The changes between the t th and the $(t+1)$ st visits at $K_\epsilon \leq k_t < k_{t+1}$ can be written as

$$\begin{bmatrix} q_{t+1}^i(s, \cdot) \\ \pi_{t+1}^j(s) \\ \bar{v}_{k_{t+1}}^i(\cdot) \end{bmatrix} = (1 - \alpha_t^i) \begin{bmatrix} q_t^i(s, \cdot) \\ \pi_t^j(s) \\ \bar{v}_{k_t}^i(\cdot) \end{bmatrix} + \alpha_t^i \left(\begin{bmatrix} Q_t^i(s, \cdot) \mu_t^j(s, \cdot) \\ \mu_t^j(s, \cdot) \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \underline{\omega}_t^i(s, \cdot) \\ \bar{\omega}_t^j(s, \cdot) \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \varepsilon_t^i(s, \cdot) \end{bmatrix} \right),$$

where the response $\mu_t^j(s, \cdot)$ is either the best or smoothed best response to $q_t^j(s, \cdot)$ depending on agent j 's preference. In the update of $\pi_t^j(s)$, the noise term $\bar{\omega}_t^j(s, \cdot) = a_t^j(s) - \mu_t^j(s, \cdot)$. In the update of $q_t^i(s, \cdot)$, the noise term $\underline{\omega}_t^i(s, \cdot) = Q_t^i(s, \cdot) a_t^j(s) - Q_t^i(s, \cdot) \mu_t^j(s, \cdot)$ if agent i observes the opponent action and knows the model. If the agent does not observe the opponent action or know the model, then we have

$$\begin{aligned} \underline{\omega}_t^i(s, a^i) = & - (\mathbb{E}_{a^j \sim \mu_t^j(s, \cdot)} [Q_t^i(s, a^i, a^j)] - q_t^i(s, a^i)) \\ & + \frac{\mathbb{I}_{\{a^i = a_t^i(s)\}}}{\mu_t^i(s, a^i)} (r^i(s, a^i, a_t^j(s)) + \gamma v_{k_t}^i(s_{k_{t+1}}) - q_t^i(s, a^i)). \end{aligned} \quad (2.40)$$

Different from (2.15), here $\underline{\omega}_t^i(s, \cdot)$ depends also on $s_{k_{t+1}} \sim p(\cdot \mid s_{k_t}, a_t^1(s), a_t^2(s))$. Lastly, the error term $\varepsilon_t^i(s, \cdot) : S \rightarrow \mathbb{R}$ is defined by

$$\varepsilon_t^i(s, s') = \frac{\bar{v}_{k_{t+1}}^i(s') - \bar{v}_{k_t}^i(s')}{\alpha_t^i}. \quad (2.41)$$

The numerator depends on the random k_{t+1} and the denominator α_t^i goes to zero as $t \rightarrow \infty$.

Note that Assumptions 3 and 4 are equivalent to [56, Assumptions 1-ii and 2-i], respectively. Therefore, we can invoke [56, Lemma 2] to conclude that $\{\varepsilon_t^i\}_{t \geq 0}$ is asymptotically negligible almost surely.

Let $y_t = [q_t^i(s, \cdot); \pi_t^j(s); \bar{v}_{k_t}^i(\cdot)]_{i=1}^2 \in Y$, which depends on state s implicitly and $Y = \prod_{i=1}^2 \Xi_q^i \times \Delta^j \times \Xi_v^i$ is a compact subset of $\mathbb{R}^{2(m+|S|)}$. The parameter y_t satisfies

$$y_{t+1} - y_t - \alpha_t(\omega_t + \bar{e}_t) \in \alpha_t \cdot \bar{F}(y_t). \quad (2.42)$$

where

$$\omega_t = \left[\underline{\omega}_t^1(s, \cdot); \bar{\omega}_t^2(s, \cdot); \mathbf{0}; \frac{\alpha_t^2}{\alpha_t^1} \underline{\omega}_t^2(s, \cdot); \frac{\alpha_t^2}{\alpha_t^1} \bar{\omega}_t^1(s, \cdot); \mathbf{0} \right], \quad (2.43)$$

the error term

$$\bar{e}_t = \begin{bmatrix} \mathbf{0} \\ \varepsilon_t^1(s, \cdot) \\ \left(\frac{\alpha_t^2}{\alpha_t^1} - \frac{1}{d} \right) \left(\begin{bmatrix} Q_t^2(s, \cdot) \mu_t^1(s) \\ \mu_t^1(s) \end{bmatrix} - \begin{bmatrix} q_t^2(s, \cdot) \\ \pi_t^1(s) \end{bmatrix} \right) \\ \frac{\alpha_t^2}{\alpha_t^1} \varepsilon_t^2(s, \cdot) \end{bmatrix}, \quad (2.44)$$

and $\bar{F}(\cdot)$ is defined by

$$\bar{F}(y) = \left\{ \begin{bmatrix} f^{12} \\ \mathbf{0} \\ \frac{1}{d} \cdot f^{21} \\ \mathbf{0} \end{bmatrix} : f^{ij} \in \bar{F}^{ij}(y) \text{ for } i = 1, 2 \right\} \quad (2.45)$$

and $\bar{F}^{ij}(y)$ is as described in (2.16) (or (2.17)) with R^i is replaced with $Q^i(\cdot) = r^i(s, \cdot) + \gamma \sum_{s_+} p(s_+|s, \cdot) v^i(s_+)$ if agent j plays according to the best (or the smoothed best) response. Since F^{ij} is a Marchaud map, $\bar{F}^{ij}$ and $\bar{F}$ are Marchaud maps.

As in the proof of Theorem 1, the noise terms satisfy $\mathbb{E}[\bar{\omega}_t^i; \underline{\omega}_t^i | \mathcal{F}_{k_t}] = \mathbf{0}$ and $\mathbb{E}[\|\bar{\omega}_t^i; \underline{\omega}_t^i\|^2 | \mathcal{F}_{k_t}] < W$ for some finite W for all t such that $k_t \geq K_\epsilon$, where $\mathcal{F}_{k_t} := \sigma(q_l^i(s, \cdot), \pi_l^i(s), v_{k_l}^i(\cdot))$ for $i = 1, 2$, and $l \leq t$. Since $\{\varepsilon_t^i\}_{t \geq 0}$ is asymptotically negligible almost surely, $\lim_{k \rightarrow \infty} \alpha_k^1 / \alpha_k^2 = d \in (0, 1]$ yield that $\bar{e}_t$ is also

asymptotically negligible almost surely. Therefore, the iterate y_t converges to a compact connected internally chain transitive set of the differential inclusion $\dot{y} \in \bar{F}(y)$.

Let $y : \mathbb{R} \rightarrow Y$ be the solution to $\dot{y} \in \bar{F}(y)$ with initial point $y_o \in Y$ such that $y(0) = y_o$ and $\frac{dy(t)}{dt} \in \bar{F}(y(t))$ for almost every $t \in \mathbb{R}$. Define $q^i : \mathbb{R} \rightarrow \Xi_q^i$, $\pi^j : \mathbb{R} \rightarrow \Delta^j$, and $v^i(s) : \mathbb{R} \rightarrow \Xi_v^i$ such that $y(t) = [q^i(t); \pi^j(t); [\bar{v}^i(\tilde{s})(t)]_{\tilde{s} \in S}]_{i=1}^2$. Then, for almost every $t \in \mathbb{R}$, we have

$$\frac{dq^i(t)}{dt} = d^i(Q^i(t)\mu^j(t) - q^i(t)) \quad (2.46a)$$

$$\frac{d\pi^j(t)}{dt} = d^j(\mu^j(t) - \pi^j(t)) \quad (2.46b)$$

$$\frac{d\bar{v}^i(\tilde{s})(t)}{dt} = 0 \quad \forall \tilde{s} \in S \quad (2.46c)$$

for each $i = 1, 2$ and $j \neq i$, where $\mu^j(t) \in \text{BR}^j(q^j(t))$ (and $\mu^j(t) = \text{BR}_\tau^j(q^j(t))$) if agent j plays according to the best response (and the smoothed best response), and $d^1 = 1$ and $d^2 = 1/d \in [1, \infty)$. The definition of Q^i and (2.46c) yield that there exists $\bar{Q}^i = Q^i(t)$ for all t . Therefore, (2.46) reduces to (2.22) and we can use the Lyapunov function (2.23), formulated in the proof of Theorem 1. To this end, we define

$$\bar{\delta}_t^i(s) := \mu_t^i(s, \cdot)^T q_t^i(s, \cdot) + \tau^i H^i(\mu_t^i(s, \cdot)) - \tau^j H^j(\pi_t^j(s)) - \underline{\text{val}}^i(Q_t^i(s, \cdot)), \quad (2.47)$$

$$\bar{c}_t(s) := \lambda \|\bar{Q}_t(s, \cdot)\|_{\max} - \sum_{i=1}^2 \underline{\text{val}}^i(Q_t^i(s, \cdot)) \quad (2.48)$$

for some $\lambda > 1$, where $\bar{Q}_t(s, \cdot) \equiv Q_t^1(s, \cdot) + Q_t^2(s, \cdot)^T$ for each s . Then, the Lyapunov function yields that

$$\bar{\delta}_t^i(s) \geq \|q_t^i(s, \cdot) - Q_t^i(s, \cdot)\pi_t^j(s)\|_2 \rightarrow 0 \quad (2.49a)$$

$$\bar{\delta}_t^1(s) + d \cdot \bar{\delta}_t^2(s) \leq \bar{c}_t(s) + \text{a.n.e.}, \quad (2.49b)$$

for some asymptotically negligible error (a.n.e.) decaying to zero almost surely.

We can also write (2.49b) as

$$\sum_{i=1}^2 \bar{\delta}_t^i(s) \leq \bar{c}_t(s) + (1-d)\bar{\delta}_t^2(s) + \text{a.n.e.} \quad (2.50)$$

On the right-hand side, we have $\bar{\delta}_t^2(s)$. By (2.49a), Lemma 1 yields that we can bound $\bar{\delta}_t^2(s)$ from above by

$$\begin{aligned} \bar{\delta}_t^2(s) &\leq \sum_{i=1}^2 \bar{\delta}_t^i(s) - \|q_t^1(s, \cdot) - Q_t^1(s, \cdot)\pi_t^2(s)\|_2 \\ &\quad + \underline{\text{val}}^1(Q_t^1(s, \cdot)) + \underline{\text{val}}^2(Q_t^2(s, \cdot)) - \min_a \bar{Q}_t(s, a). \end{aligned} \quad (2.51)$$

Therefore, by (2.48), (2.49a), (2.50) and (2.51), we obtain

$$\begin{aligned} \sum_{i=1}^2 \bar{\delta}_t^i(s) &\leq \frac{\lambda}{d} \|\bar{Q}_t(s, \cdot)\|_{\max} - \frac{1-d}{d} \min_a \bar{Q}_t(s, a) \\ &\quad - \underline{\text{val}}^1(Q_t^1(s, \cdot)) - \underline{\text{val}}^2(Q_t^2(s, \cdot)) + \text{a.n.e.} \end{aligned} \quad (2.52)$$

By (2.4), (2.47) and (2.52), we have

$$\sum_{i=1}^2 \mu_t^i(s, \cdot)^T q_t^i(s, \cdot) \leq \frac{\lambda}{d} \|\bar{Q}_t(s, \cdot)\|_{\max} - \frac{1-d}{d} \min_a \bar{Q}_t(s, a) + \xi + \text{a.n.e.}, \quad (2.53)$$

where $\xi := \sum_{i=1}^2 \tau^i \log |A^i|$. We highlight the role of $\mu_t^i(s, \cdot)^T q_t^i(s, \cdot)$ in step 18 in Algorithm Family 2. Here, the second term $\min_a \bar{Q}_t(s, a)$ on the right-hand side poses a challenge to establish the contraction property. By (2.39), the fact that $r^1(s, a) + r^2(s, a) = 0$ for all (s, a) implies that

$$\bar{Q}_t(s, a) = \gamma \sum_{s_+ \in S} p(s_+ \mid s, a) \sum_{i=1}^2 v_t^i(s_+) \quad (2.54)$$

for all (s, a) . Correspondingly, we focus on the value function update to address $\min_a \bar{Q}_t(s, a)$.

Let t denote t_{k-1}^s with implicit dependence on state and stage for notational

simplicity, and define $\tilde{v}_{k-1}^i(s) := v_t^i(s) - v_*^i(s)$ for all k . Then, $\tilde{v}_k^i(s)$ remains unchanged if s does not get visited and evolves according to

$$\tilde{v}_k^i(s) = (1 - \bar{\beta}_{k-1}^i(s))\tilde{v}_{k-1}^i(s) + \bar{\beta}_{k-1}^i(s) (\mu_t^i(s)^T q_t^i(s, \cdot) - v_*^i(s)),$$

where the step size $\bar{\beta}_k^i(s) := \mathbb{I}_{\{s=s_k\}}\beta_t^i \in [0, 1]$. Note that

$$\begin{aligned} \mu_t^i(s)^T q_t^i(s, \cdot) - v_*^i(s) &\geq \underline{\text{val}}^i(Q_t^i(s, \cdot)) - \text{val}^i(Q_t^i(s, \cdot)) + \text{val}^i(Q_t^i(s, \cdot)) - \text{val}^i(Q_*^i(s, \cdot)) \\ &\quad + \bar{\delta}_t^i(s) - \tau^i H^i(\mu_t^i(s, \cdot)) + \tau^j H^j(\pi_t^j(s)) \end{aligned} \quad (2.55)$$

which follows from (2.47) and $v_*^i(s) = \text{val}^i(Q_*^i(s, \cdot))$. Based on (2.4), (2.37), (2.39) and (2.49a), Lemma 1 yields that

$$\mu_t^i(s)^T q_t^i(s, \cdot) - v_*^i(s) \geq \gamma \min_{s_+ \in S} \{\tilde{v}_k^i(s_+)\} - \underline{e}_k^i(s) \quad (2.56)$$

for some $\lim_{k \rightarrow \infty} \underline{e}_k^i(s) \xi$ almost surely. Hence, $\tilde{v}_k^i(s)$'s satisfy

$$\tilde{v}_k^i(s) \geq (1 - \bar{\beta}_{k-1}^i(s))\tilde{v}_{k-1}^i(s) + \bar{\beta}_{k-1}^i(s) \left(\gamma \min_{s_+ \in S} \{\tilde{v}_k^i(s_+)\} - \underline{e}_k^i(s) \right). \quad (2.57)$$

Since $t_k^s \rightarrow \infty$ as $k \rightarrow \infty$ almost surely and β_k^i satisfies Assumption 2, we can invoke Lemma 6 (from Appendix A) for (2.57) to conclude that

$$\liminf_{k \rightarrow \infty} \tilde{v}_k^i(s) \geq -\frac{\xi}{1 - \gamma} \quad \forall s \quad (2.58)$$

almost surely. Since $v_*^1(s) + v_*^2(s) = 0$ for all s , we have $\tilde{v}_k^1(s) + \tilde{v}_k^2(s) = v_{t_k^s}^1(s) + v_{t_k^s}^2(s)$. By (2.54) and (2.58), we obtain

$$\liminf_{t \rightarrow \infty} \min_a \bar{Q}_t(s, a) \geq -\frac{2\gamma\xi}{1 - \gamma}. \quad (2.59)$$

Then, by (2.53) and (2.59), we have

$$\sum_{i=1}^2 (\mu_t^i(s, \cdot)^T q_t^i(s, \cdot) - v_*^i(s)) \leq \frac{\lambda}{d} \|\bar{Q}_t(s, \cdot)\|_{\max} + \left(1 + \frac{1-d}{d} \frac{2\gamma}{1-\gamma}\right) \xi + \text{a.n.e.} \quad (2.60)$$

Furthermore, the lower bound (2.56) and (2.58) yield that

$$\mu_t^j(s, \cdot)^T q_t^j(s, \cdot) - v_*^j(s) \geq -\frac{\xi}{1-\gamma} + \text{a.n.e.} \quad (2.61)$$

Therefore, we obtain

$$\mu_t^i(s, \cdot)^T q_t^i(s, \cdot) - v_*^i(s) \leq \frac{\lambda}{d} \|\bar{Q}_t(s, \cdot)\|_{\max} + \left(1 + \frac{1}{1-\gamma} + \frac{1-d}{d} \frac{2\gamma}{1-\gamma}\right) \xi + \text{a.n.e.} \quad (2.62)$$

Since $v_*^1(s) + v_*^2(s) = 0$, (2.54) yields that

$$\begin{aligned} \bar{Q}_t(s, a) &= \gamma \sum_{s_+ \in S} p(s_+ | s, a) \sum_{i=1}^2 (v_t^i(s_+) - v_*^i(s_+)) \\ &\leq \gamma \max_{s_+ \in S} \sum_{i=1}^2 |v_t^i(s_+) - v_*^i(s_+)| \\ &\leq 2\gamma \max_{s_+ \in S} \max_{i=1,2} |v_t^i(s_+) - v_*^i(s_+)| \end{aligned} \quad (2.63)$$

for any (s, a) . Combining (2.61), (2.62) and (2.63), we obtain

$$\begin{aligned} |\mu_t^i(s, \cdot)^T q_t^i(s, \cdot) - v_*^i(s)| &\leq \frac{2\gamma\lambda}{d} \max_{s_+ \in S} \max_{i=1,2} |v_t^i(s_+) - v_*^i(s_+)| \\ &\quad + \left(1 + \frac{1}{1-\gamma} + \frac{1-d}{d} \frac{2\gamma}{1-\gamma}\right) \xi + \text{a.n.e.} \end{aligned} \quad (2.64)$$

for each s and t . Define $y_k = [\tilde{v}_k^1(\cdot); \tilde{v}_k^2(\cdot)] \in \mathbb{R}^{2|S|}$ such that $\|y_k\|_\infty = \max_{s_+ \in S} \max_{i=1,2} |v_{t_k^s}^i(s_+) - v_*^i(s_+)|$. By (2.55) and (2.64), y_k satisfies the inequalities in Lemma 5 provided in Appendix A. Recall that $t_k^s \rightarrow \infty$ as $k \rightarrow \infty$ almost surely and β_k^i satisfies Assumption 2. If $2\gamma\lambda/d < 1$, then we can invoke

Lemma 5 to conclude that

$$\begin{aligned} \limsup_{k \rightarrow \infty} |\tilde{v}_k^i(s)| &\leq \frac{1}{1 - 2\gamma\lambda/d} \left(1 + \frac{1}{1 - \gamma} + \frac{1 - d}{d} \frac{2\gamma}{1 - \gamma} \right) \xi \\ &= \frac{2d + 2\gamma - 3\gamma d}{(1 - \gamma)(d - 2\lambda\gamma)} \xi \end{aligned} \quad (2.65)$$

almost surely for each $i = 1, 2$ and $s \in S$. This completes the proof. $\square$

The following corollary to Theorem 2 shows the rationality of the dynamics from Algorithm Family 2.

Corollary 3. *Consider a ZSSG characterized by $\mathcal{M} = \langle S, (A^i, r^i)_{i=1}^2, p, \gamma \rangle$. Suppose that agent i follow a learning dynamic from Algorithm Family 2 while the opponent j plays according to fully mixed strategy $\pi^j : S \rightarrow \text{int}(\Delta^j)$. The local step sizes α_k^i and β_k^i 's satisfy, resp., Assumptions 1 and 2. Furthermore, we have $\beta_k^i/\alpha_k^i \rightarrow 0$ as $k \rightarrow \infty$.*

Let $\pi_t^i(s) \in \Delta^i$ denote a weighted empirical average of the agent i 's actions at state s until the t th visit. Let $a_t(s) \in A^i$ denote the action played by agent i at the t th visit to state s . Then, $\pi_t(s)$ evolves according to (2.38) with arbitrary initializations. If Assumption 4 also holds, then $q_t^i(s, a^i)$ tracks

$$\mathbb{E} \left[r^i(s, a^i, a^j) + \gamma \sum_{s_+} p(s_+ | s, a^i, a^j) v_t^i(s_+) \right],$$

where the expectation taken with respect to $a^j \sim \pi^j(s)$ and

$$\limsup_{t \rightarrow \infty} |v_t^i(s) - v_*^i(s)| \leq \frac{(2 - \gamma)\tau^i \log |A^i|}{(1 - \gamma)(1 - 2\gamma)}$$

almost surely for each $s \in S$.

Proof. As in the proof of Corollary 2, we can consider that agent i is playing a ZSSG against an opponent (other than agent j) following the same dynamic with singleton action spaces at each state and view agent j as the Nature. Then, the proof follows from Theorem 2. $\square$

Remark 4. We address different step sizes $\beta_k^i \neq \beta_k^j$ via the upper bounds in (2.63). Correspondingly, we can attain tighter error bounds for the cases where $\beta_k^i = \beta_k^j$ in Theorem 2 and Corollary 3, as in [55, 56].

2.5 Illustrative Examples

To establish the feasibility of our approach, we created a simplified model with two possible states $|S| = 2$ and two potential actions $|A^i| = 2$ for each state. We then explored three different situations that varied in factors such as learning rate, exploration, and information availability. To ensure a fair comparison, we kept the reward structure and state transition probabilities consistent across all scenarios. Rewards are assigned randomly from $[0, 1]$ and $[0, 0.2]$ for first and second states so that visually appealing results could be generated. State transitions were determined probabilistically using a symmetric Dirichlet distribution $\text{Dir}(\alpha)$ with $\alpha = 1$, guaranteeing that any state could be reached from any other state through any action. We set the discount factor γ to 0.3 and the temperature parameter τ to 0.002. To assess the performance, each scenario is tested 30 times independently.

In Figure 2.3, we present the average behavior of value function estimates for the following scenarios.

- *Scenario 1. Full Access vs No Access:* In this scenario, we examine a situation where agent 1 has complete information about the other’s actions and responds according to smoothed best response, while the agent 2 acts solely based on its own rewards without considering the opponent’s actions. This dynamic can be interpreted as a contrast between belief-based learning and payoff-based learning. This setup also mirrors the comparison between smoothed fictitious play and individual Q-learning.
- *Scenario 2. Full Access vs Temporal Access:* In this scenario, agent 1 has

complete knowledge of the other’s actions, while agent 2 observes the opponent’s actions with probability $\theta^2 = 0.5$. When compared to Scenario 1 where there is no access to opponent actions, we found that having partial information about the opponent’s actions accelerate the learning process.

- *Scenario 3. Different Learning Rates:* In this scenario, both agents have complete knowledge of each other’s actions and play according to smoothed best response. However, the agents learn at different rates such that $\alpha_k^1 = 1/k^{0.96}$, $\alpha_k^2 = 1/(0.92k)^{0.96}$ and $\beta_k^1 = 1/k$, $\beta_k^2 = 1/(0.96k)$ satisfying standard assumptions listed in Theorem 2.

In all three scenarios, the estimated values of the agents remained within the theoretical bound defined by Theorem 2. Due to the relatively small number of states and actions in the game, we observe small variation in the value function estimates across independent experiments, particularly in Scenario 3 where both agents have complete knowledge of the model and full access to the opponent’s actions.

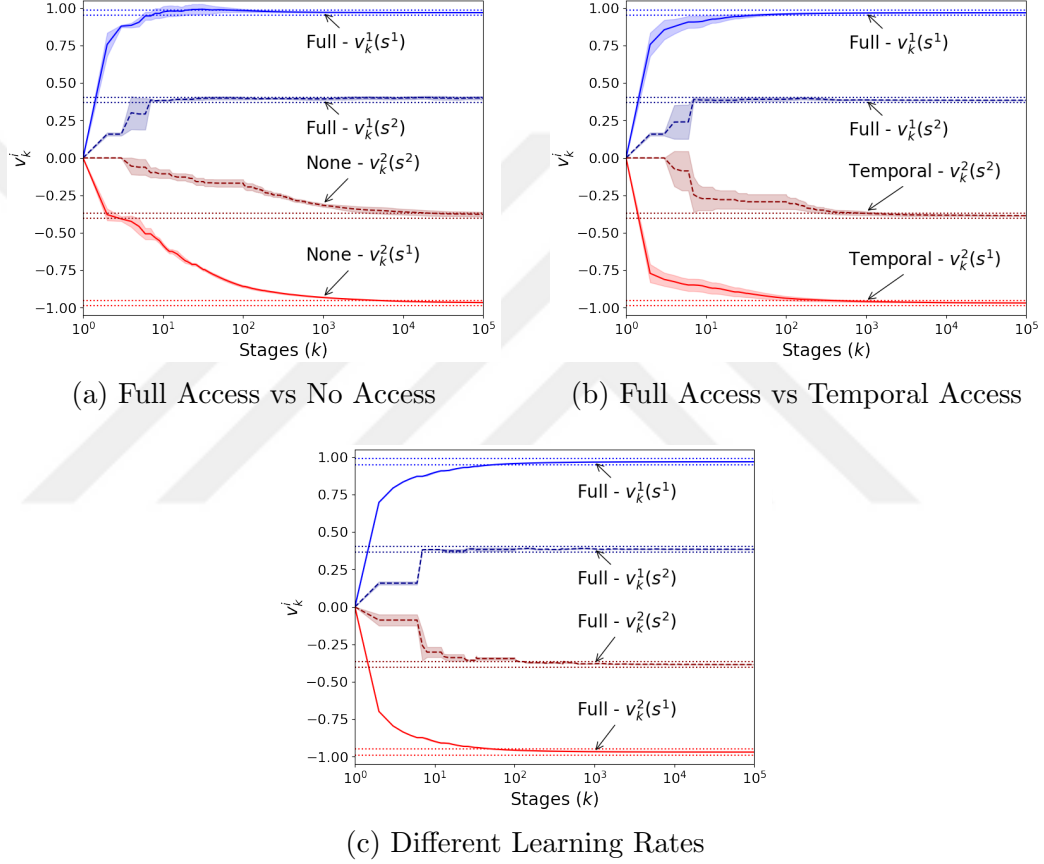


Figure 2.3: The evaluation of estimated value functions in a ZSSG with two states and two actions, where $S = \{s^1, s^2\}$, $\gamma = 0.3$ and $\tau = 0.002$. Each sub-figure corresponds to Scenarios 1, 2, and 3, respectively. “Full”, “Temporal” and “None” refer to whether the agent has full, temporal (with $\theta^i = 0.5$) or no access to the opponent’s actions. The different colors and line styles represent different agents and states. The shaded areas represent the standard deviation of the value function estimates. The dotted lines represent the bounds on the value function estimates according to Theorem 2.

Chapter 3

Strategizing against Independent Learning Dynamics

“Victorious warriors win first and then go to war, while defeated warriors go to war first and then seek to win.”

Sun Tzu

The recent integration of (reinforcement) learning algorithms within multi-agent systems has significantly improved the ability of autonomous agents to address complex tasks through learning from interactions within a shared environment. However, this very advancement introduces a vulnerability, as strategically sophisticated agents can manipulate these algorithms, leading to suboptimal performance [82, 83, 84, 85, 86]. This raises a critical question:

How much a strategically sophisticated agent can manipulate the opponent’s decisions for more payoff in a game if the agent is aware of the opponent’s learning dynamic?

To address the vulnerability of learning algorithms to manipulation, we adopt

a control-theoretical perspective. We model the multi-agent system as a repeated play of a general-sum normal-form game played by two types of agents: *strategically advanced* (A-type) and *naive* (N-type). N-type agents *naively* employ the widely-used independent Q-learning (IQL) algorithm, acting as if no other agents exist in the environment [30, 31]. Conversely, A-type agents are strategically sophisticated agents with complete knowledge of the game structure, the learning algorithms employed by N-type agents and the entire history of actions taken by all agents. We consider A-types with these advanced capabilities to establish a benchmark for quantifying the performance of other strategic actors with limited capabilities.

We demonstrate that A-types can leverage their complete model knowledge to control N-types' algorithms (as if the algorithms are some dynamical system) to maximize the discounted sum of the payoffs collected over an infinite horizon. To model the interactions among A-types, we employ the framework of stochastic games (SG)—a generalization of Markov decision processes (MDPs) to non-cooperative multi-agent environments [41]. However, the continuous state space inherent in SG presents a computational challenge. To address this issue, as an example, we present a novel quantization-based approximation scheme reducing the problem to a finite SG that can be solved via standard dynamic programming methods when there are two competing A-types or a single A-type. We quantify the approximation level of this quantization scheme and examine its performance numerically.

This chapter is organized as follows: We discuss the relevant literature and position our contributions in Section 3.1. We formulate the strategic actors' interactions as an SG in Section 3.2. We present our solution in Section 3.3. We provide illustrative examples in Section 3.4.

3.1 Related Works

Research on exploiting naive learning algorithms has primarily focused on fundamental models such as no-regret learning and fictitious play. The authors in [84] demonstrate that sophisticated agents can consistently outperform no-regret learners by securing the Stackelberg equilibrium value. Another study [87] focuses on two-agent, two-action games where one player adopts fictitious play. In this work, the authors identify the specific game conditions related to the payoffs under which a sophisticated opponent can exploit their naive opponent. In [85], the authors further broaden the scope to the multi-agent general-sum games with diverse action spaces, also adopting fictitious play. Their findings reveal that sophisticated agents can surpass Nash equilibrium payoffs by strategically manipulating their actions based on opponents’ learning patterns. This involves solving a linear program to predict opponents’ mixed strategies and then executing a pure action trajectory that aligns with the desired mixed strategy distribution. In contrast to these prior works that primarily employ game-theoretic and optimization based techniques, our approach leverages control theory to address the problem of strategizing against naive opponents. We introduce a novel perspective by treating the opponents’ learning parameters as a state variable for the strategic agent, enabling the use of control theoretical tools. This methodological shift offers a promising new approach to bridging the gap between control theory and the multi-agent reinforcement learning. Moreover, our work extends the existing literature in strategizing against naive learners by considering stochastic games as discussed in Chapter 4, a more complex and realistic setting compared to repeated normal-form games.

The vulnerability of learning dynamics reliant on environmental feedback has been explored in the literature through adversarial policies that manipulate feedback, often called “reward poisoning” or “falsification of cost signals” [82, 83, 88, 89]. While our work aligns with this research areas to some extent, a crucial distinction lies in the underlying game structure. The inherent nature of game prevents payoff manipulation, making naive opponents more resilient to

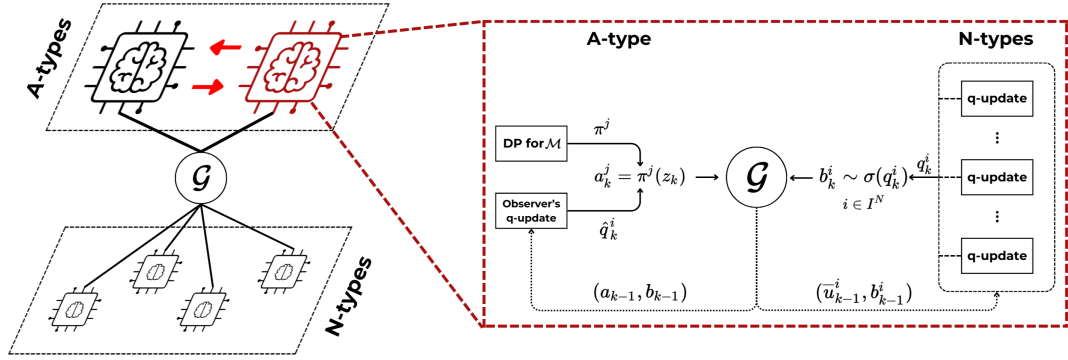


Figure 3.1: Illustration of the interaction between N-types and A-types across the repeated play of the normal-form game $\mathcal{G}$. A-types compete against each other while considering N-types' q-function estimates as an underlying state of an SG. A-types can use dynamic programming to solve the SG and use trackers to track N-types' q-function estimates by leveraging the complete model knowledge.

such attacks. Moreover, unlike the focus of reward poisoning literature on competitive, zero-sum scenarios, our framework extends to general-sum games with diverse agent objectives.

3.2 Strategic Actors against Q-learners

Consider a normal-form game, denoted by $\mathcal{G}$, which is characterized by the tuple $\langle \{A^j\}_{j \in I^A}, \{B^i\}_{i \in I^N}, \{u^j\}_{j \in I^A}, \{\bar{u}^i\}_{i \in I^N} \rangle$, where I^A and I^N denote, respectively, the index set of A-type and N-type agents.

- **N-types** are *naive* (uninformed) agents employing the IQL algorithm, making decisions based solely on their individual rewards without considering the actions or influence of other agents.
- **A-types** are strategically *sophisticated* agents with complete knowledge of the game and how the N-types are learning. They also have access to all agents' actions and can use this information to their advantage.

Let A^j and B^i denote the A-type j 's and N-type i 's finite action sets. The

joint action spaces of A-types and N-types are denoted by $A := \times_{j \in I^A} A^j$ and $B := \times_{i \in I^N} B^i$, respectively. The payoff functions are $u^j : A \times B \rightarrow \mathbb{R}$ and $\bar{u}^i : A \times B \rightarrow \mathbb{R}$. A-types and N-types are repeatedly playing $\mathcal{G}$ by taking actions simultaneously over stages $k = 0, 1, \dots$ as follows:

N-type: Let $q_k^i : B^i \rightarrow \mathbb{R}$ denote N-type i 's q-function estimate at stage k . Then, the agent takes action according to the softmax function:

$$\text{BR}_\tau(q_k^i)(b^i) = \frac{\exp(q_k^i(b^i)/\tau)}{\sum_{\tilde{b}^i \in B^i} \exp(q_k^i(\tilde{b}^i)/\tau)} \quad \forall b^i \in B^i \quad (3.1)$$

for some $\tau > 0$ controlling the level of exploration and $\text{BR}_\tau(q_k^i)(b^i) \in (0, 1]$ denotes the probability that they take action $b^i \in B^i$. Let $a_k = (a_k^j)_{j \in I^A}$ and $b_k = (b_k^i)_{i \in I^N}$ denote the joint actions of A-types and N-types at stage k , respectively. Given the payoff $\bar{u}_k^i = \bar{u}^i(a_k, b_k)$, N-type i updates the q-function according to

$$q_{k+1}^i(b^i) = \begin{cases} q_k^i(b^i) + \alpha(\bar{u}_k^i - q_k^i(b^i)) & \text{if } b^i = b_k^i \\ q_k^i(b^i) & \text{o.w} \end{cases} \quad (3.2)$$

with some step size $\alpha \in (0, 1)$ and the initial vector $q_0^i(b^i) = 0$ for all b^i .¹ In other words, N-types act myopically as if the other players play according to some stationary strategy across the repeated play of the underlying game $\mathcal{G}$.

A-type: The advanced players can track N-types' q-function estimates by knowing the deployed IQLs, as described in (3.2), and observing the actions of all agents. Therefore, A-types do not myopically act as they are aware that the other players do not necessarily play according to some stationary strategy. Instead, they strategically aim to maximize the discounted sum of the payoffs they can collect across the repeated play of the underlying game $\mathcal{G}$, i.e.,

$$\max_{\{a_k^i\}_{k=1}^\infty} \mathbb{E} \left[\sum_{k=0}^\infty \gamma^k u^i(a_k, b_k) \right] \quad (3.3)$$

¹We interchangeably view q-function as a function $q^i : B^i \rightarrow \mathbb{R}$ and a vector $q^i \in \mathbb{R}^{|B^i|}$ since B^i is a finite set.

for some discount factor $\gamma \in (0, 1)$.

A-type i selects actions to achieve the objective in (3.3) based on the history $\{a_0, b_0, \dots, a_{k-1}, b_{k-1}\}$ of the game and the knowledge that q_k^i 's evolve according to (3.2). However, this is not a well-defined optimization problem that can be solved introspectively because the optimal action for A-type i depends critically on the actions chosen by other players in the game, introducing an element of uncertainty that cannot be fully addressed.

While the history of the game, i.e. the information set $\{a_0, b_0, \dots, a_{k-1}, b_{k-1}\}$, grows unboundedly over time, the decision making process of N-types depends only on their q-function estimates. This implies that conditioned on the current q-function estimates and the current action profile of A-types, the next q-function estimates, and therefore, N-types' play exhibits the Markov property. In simpler terms, the future behavior of N-types depends only on the present state and not on the entire history that led to that state. A-types can leverage this Markov property in their objective (3.3) by viewing the q-function estimates as some underlying controlled Markov process. Therefore, we can model A-types' interactions with each other as the following stochastic game whose state is N-types' q-function estimates.

Problem 1 (Stochastic Game Formulation) *Consider an $|I^A|$ -agent SG characterized by the tuple $\mathcal{Z} = \langle Z, (A^j, r^j)_{j \in I^A}, \gamma, p \rangle$. The state space $Z := \prod_{i \in I^N} Q^i$ is a compact set, where $Q^i \subset \mathbb{R}^{|B^i|}$ corresponds to all possible q-function estimates of N-type $i \in I^N$. The reward function $r^j : Z \times A \rightarrow \mathbb{R}$ is given by*

$$r^j(z, a) = \mathbb{E}_{b \sim \bar{\pi}(z)}[u^j(a, b)] \quad \forall (z, a) \in Z \times A, \quad (3.4)$$

where

$$\bar{\pi}(\{q^i\}_{i \in I^N})(\{b^i\}_{i \in I^N}) := \prod_{i \in I^N} \text{BR}_\tau(q^i)(b^i) \quad (3.5)$$

for all $q^i \in Q^i$, $b^i \in B^i$ and $i \in I^N$.² The transition kernel $p(\cdot | \cdot)$ determines the

²By concatenation of $\{q^i\}_{i \in I^N}$, we can construct state z . Since b is the action profile, $\bar{\pi}(\{q^i\}_{i \in I^N})(\{b^i\}_{i \in I^N})$ is equivalent to $\bar{\pi}(z)(b)$.

Algorithm 3 A-type j

input: π^j for the SG $\mathcal{Z}$ and the initials $\{\hat{q}_0^i\}_{i \in I^N}$
for each stage $k = 0, 1, \dots$ **do**
 construct $z_k = \{\hat{q}_k^i\}_{i \in I^N}$
 take action $a_k^j \sim \pi^j(z_k)$ simultaneously with the others
 receive payoff $u^j(a_k, b_k)$ (corresponding to $r^j(z_k, a_k)$ in expectation)
 observe (a_k, b_k)
 update the q-trackers according to (3.6)
end for

evolution of transformed states according to the update rule (3.2).

Recall that A-types know N-types' IQLs (3.2) and the underlying game model $\mathcal{G}$ and can observe the actions (a_k, b_k) of all agents. Therefore, they can track the q-function estimates perfectly, although they do not have direct access to them. For example, A-types can deploy a q-tracker $\hat{q}_k^i$ specific to each N-type i and evolving according to

$$\hat{q}_{k+1}^i(b_k^i) = \hat{q}_k^i(b_k^i) + \alpha(\bar{u}^i(a_k, b_k) - \hat{q}_k^i(b_k^i)) \quad (3.6a)$$

$$\hat{q}_{k+1}^i(b^i) = \hat{q}_k^i(b^i) \quad \forall b^i \neq b_k^i \quad (3.6b)$$

with $\hat{q}_0^i = q_0^i$. Therefore, we have $\hat{q}_k^i \equiv q_k^i$ for all k and A-types can construct the state $z_k = (\hat{q}_k^i)_{i \in I^N}$ of the underlying SG $\mathcal{Z}$.

Let each A-type j follow Markov stationary strategy $\pi^j : Z \rightarrow \Delta(A^j)$ in the SG $\mathcal{Z}$.³ Given the strategy profile $\pi = (\pi^j)_{j \in I^A}$, A-type j 's utility (3.3) can be rewritten as

$$U^j(\pi^j, \pi^{-j}) := \mathbb{E}_{a_k \sim \pi(z_k)} \left[\sum_{k=0}^{\infty} \gamma^k r^j(z_k, a_k) \right] \quad (3.7)$$

for the SG $\mathcal{Z}$. As we pointed out earlier, finding the best strategy in games is generally not a well-defined optimization problem. Therefore, we focus on the important special case of two competing A-types. Let $I^A = \{j_1, j_2\}$ and A-types have opposite payoffs $u^{j_1} + u^{j_2} \equiv 0$ such that $U^{j_1} + U^{j_2} \equiv 0$.

³The probability simplex over a finite set A is denoted by $\Delta(A)$.

Algorithm 4 N-type i

input: the initial q_0^i
for each stage $k = 0, 1, \dots$ **do**
 take action $b_k^i \sim \text{BR}_\tau(q_k^i)$ simultaneously with the others
 receive payoff $\bar{u}_k^i = \bar{u}^i(a_k, b_k)$
 update the q-function estimate according to (3.2)
end for

Definition 4 (Markov Stationary Equilibrium) *We say that a Markov stationary profile $\pi = (\pi^j : Z \rightarrow \Delta(A^j))_{j \in I^A}$ is Markov stationary equilibrium provided that*

$$U^j(\pi^j, \pi^{-j}) \geq U^j(\tilde{\pi}^j, \pi^{-j}) \quad \forall \tilde{\pi}^j, \quad (3.8)$$

where $\pi^{-j} = \{\pi^\ell\}_{\ell \neq j}$ [41].

Proposition 1. *There always exists Markov stationary equilibrium for two-agent zero-sum SGs $\mathcal{Z} = \langle Z, (A^j, r^j)_{j \in I^A}, \gamma, \bar{p} \rangle$.*

Proof. The proof follows from [90, Theorem 6.2.12] based on the observation that the minimax function is non-expansive like the maximum function and the set of all possible q-function estimates, Z , is a Polish space as a compact subset of $\mathbb{R}^{|B|}$ and the set of actions, A , is finite and state-invariant. $\square$

Remark 5. *If there is only a single A-type, then the SG $\mathcal{Z}$ reduces to an MDP, for which there also exists a Markov stationary solution.*

Given the Markov stationary strategy π^j for the SG $\mathcal{Z}$, A-type j interacts with other agents as illustrated in Figure 3.1. Detailed descriptions of the A-types' and N-types' interactions across the repeated play of the underlying game $\mathcal{G}$ are tabulated, respectively, in Algorithm 3 and Algorithm 4.

3.3 Solving the Stochastic Game

A critical challenge for finding the best strategy in the two-agent zero-sum SG $\mathcal{Z}$ is the continuum state space Z . Based on the specifics of the IQLs, we can show that the SG $\mathcal{Z}$ is relatively well-behaved for effective approximations. Particularly, let $v_\kappa^j : Z \rightarrow \mathbb{R}$ denote the values of the states in Z at the initial stage of a $(\kappa + 1)$ -length horizon for the best decision rules in $\mathcal{Z}$. More explicitly, for each $z \in Z$, we have

$$v_\kappa^{j_1}(z) := \max_{\mu^{j_1}} \min_{\mu^{j_2}} \mathbb{E}_{(a^{j_1}, a^{j_2}) \sim (\mu^{j_1}, \mu^{j_2})} [Q_\kappa^{j_1}(z, a^{j_1}, a^{j_2})], \quad (3.9a)$$

$$v_\kappa^{j_2}(z) := \max_{\mu^{j_2}} \min_{\mu^{j_1}} \mathbb{E}_{(a^{j_1}, a^{j_2}) \sim (\mu^{j_1}, \mu^{j_2})} [Q_\kappa^{j_2}(z, a^{j_1}, a^{j_2})], \quad (3.9b)$$

where

$$Q_\kappa^j(z, a) := r^j(z, a) + \gamma \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid z, a) \quad (3.10)$$

for all $\kappa > 0$, $j \in I^A$ and $(z, a) \in Z \times A$. The initial $Q_0^j(z, a) = r^j(z, a)$ for all (z, a) .

Proposition 2. *The minimax value function $v_\kappa^j(z)$ for each $j \in I^A$ is bounded by*

$$M_\kappa := \frac{1 - \gamma^{\kappa+1}}{1 - \gamma} \max_{(\tilde{j}, a, b)} |u^{\tilde{j}}(a, b)| \quad (3.11)$$

for all $z \in Z$ and $\kappa \geq 0$. The value function is also L_κ -Lipschitz with respect to $\|\cdot\|_1$, where

$$L_\kappa = \gamma L_{\kappa-1} + L_0 \frac{1 - \gamma^{\kappa+1}}{1 - \gamma} \quad (3.12)$$

for all $\kappa > 0$ and $L_0 := \frac{1}{\tau} \sqrt{|B|} \max_{(\tilde{j}, a, b)} |u^{\tilde{j}}(a, b)|$.

Proof. The boundedness follows from the definitions (3.4) and (3.9) since we

can bound $v_\kappa^j(\cdot)$ by

$$\begin{aligned}
|v_\kappa^j(z)| &\leq \sum_{k=0}^{\kappa} \gamma^k \max_{(z,a)} |r^j(z, a)| \\
&\leq \sum_{k=0}^{\kappa} \gamma^k \max_{(\tilde{j}, a, b)} |u^{\tilde{j}}(a, b)| \\
&\leq \frac{1 - \gamma^{\kappa+1}}{1 - \gamma} \max_{(\tilde{j}, a, b)} |u^{\tilde{j}}(a, b)| = M_\kappa.
\end{aligned} \tag{3.13}$$

Next, we first present the following lemma playing an important role in showing the Lipschitz continuity of the value function.

Lemma 2. *The joint strategy $\bar{\pi}(\cdot)$ of N -types, as described in (3.5), satisfies*

$$\|\bar{\pi}(z) - \bar{\pi}(\bar{z})\|_1 \leq \frac{\sqrt{|B|}}{\tau} \cdot \|z - \bar{z}\|_1 \tag{3.14}$$

for all $z, \bar{z} \in Z$.

The second part of the proof follows from induction. Firstly, for $\kappa = 0$, the non-expansiveness of the minimax function leads to

$$|v_0^j(z) - v_0^j(\bar{z})| \leq \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E}[|r^j(z, a^j, a^{-j}) - r^j(\bar{z}, a^j, a^{-j})|], \tag{3.15}$$

where the expectation is taken with respect to $a^j \sim \mu^j$ and $a^{-j} \sim \mu^{-j}$. Then, we can use the following lemma showing the Lipschitz continuity of the reward functions $r^j(\cdot)$'s.

Lemma 3. *The reward function $r^j(z, a)$ is Lipschitz continuous in $z \in Z$. For each $a \in A$, we have*

$$|r^j(z, a) - r^j(\bar{z}, a)| \leq L_0 \cdot \|z - \bar{z}\|_1 \tag{3.16}$$

for all $z, \bar{z} \in Z$, where L_0 is as described in Proposition 2.

By (3.15), Lemma 3 yields that

$$\begin{aligned} |v_0^j(z) - v_0^j(\bar{z})| &\leq \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E}[|r^j(z, a) - r^j(\bar{z}, a)|] \\ &\leq L_0 \cdot \|z - \bar{z}\|_1. \end{aligned} \quad (3.17)$$

Next, we focus on $\kappa > 0$ and assume that $v_{\kappa-1}^j(\cdot)$ is $L_{\kappa-1}$ -Lipschitz with respect to $\|\cdot\|_1$ to show that $v_\kappa^j(\cdot)$ is L_κ -Lipschitz. The non-expansiveness of the minimax function and the triangle inequality yield that

$$\begin{aligned} |v_\kappa^j(z) - v_\kappa^j(\bar{z})| &\leq \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E}[|r^j(z, a) - r^j(\bar{z}, a)|] \\ &\quad + \gamma \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E} \left[\left| \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z}|z, a) - \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z}|\bar{z}, a) \right| \right]. \end{aligned}$$

For the first term on the right-hand side, we can invoke Lemma 3. On the other hand, for the second term, we can use the following lemma.

Lemma 4. *Let $\eta : Z \rightarrow \mathbb{R}$ be bounded $|\eta(z)| \leq M$ for all $z \in Z$ and K -Lipschitz with respect to $\|\cdot\|_1$. Then we have*

$$\left| \int_Z \eta(\tilde{z}) p(d\tilde{z}|z, a) - \int_Z \eta(\tilde{z}) p(d\tilde{z}|\bar{z}, a) \right| \leq \bar{K} \cdot \|z - \bar{z}\|_1, \quad (3.18)$$

where $\bar{K} = K + \frac{\sqrt{|B|}}{\tau} \cdot M$.

Based on the boundedness of the value function (3.13) and the assumption that $v_{\kappa-1}^j$ is $L_{\kappa-1}$ -Lipschitz, we obtain

$$\begin{aligned} |v_\kappa^j(z) - v_\kappa^j(\bar{z})| &\leq \left(L_0 + \gamma \left(L_{\kappa-1} + \frac{\sqrt{|B|}}{\tau} M_{\kappa-1} \right) \right) \cdot \|z - \bar{z}\|_1 \\ &= L_\kappa \cdot \|z - \bar{z}\|_1 \end{aligned} \quad (3.19)$$

for all $z, \bar{z} \in Z$, where the equality follows from the definitions of $M_{\kappa-1}$, L_κ , and L_0 in the proposition statement. This completes the proof. $\square$

The Lipschitz continuity of the minimax value function is promising to effectively approximate the underlying SG via the value-based approximation methods. Various methods exist for approximating MDPs with large state-action spaces. For instance, state aggregation [35, 36] involves combining similar states with respect to cost and transition probabilities into meta-states. On the other hand, in approximate linear programming [37, 38], preselected basis functions are used to fit the value function of the MDP. Approximate value iteration and approximate policy iteration [91, 42, 92] also use preselected basis functions to determine the value and policy, respectively. Neuro-dynamic programming [42] leveraging the power of approximation structures, such as neural networks, is also an important method to approximate MDPs.

In this chapter, as an *example* proof of concept, we focus on quantization-based approximations (e.g., see [34] for a detailed review) due to their interpretability by transforming $\mathcal{Z}$ to a finite SG, as in the original definition in [41]. To this end, we consider a quantization mapping $\Phi : Z \rightarrow D$ for some finite set $D \subset Z$ such that $Z_d := \{z \in Z : \Phi(z) = d\}$ satisfies $Z_d \cap Z_{d'} = \emptyset$ for all $d \neq d'$ and $\bigcup_{d \in D} Z_d = Z$. In the following, we quantify the approximation error induced by such quantization schemes based on Proposition 2.

Proposition 3. *Given the quantization mapping Φ , assume that there exists $\Delta > 0$ such that $\|z - \Phi(z)\|_1 \leq \Delta$ for all $z \in Z$. Then, for each $j \in I^A$, we have*

$$\|v_\kappa^j - \widehat{v}_\kappa^j\|_\infty \leq \frac{\Delta \sqrt{|B|}}{\tau(1-\gamma)^3} \max_{(j,a,b)} |u^j(a,b)| \quad \forall \kappa \geq 0, \quad (3.20)$$

where v_κ^j is described in (3.9), and

$$\widehat{v}_\kappa^j(z) := \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E}_{(a^j, a^{-j}) \sim (\mu^j, \mu^{-j})} [\widehat{Q}_\kappa^j(z, a^j, a^{-j})], \quad (3.21)$$

where

$$\widehat{Q}_\kappa^j(z, a) := r^j(\Phi(z), a) + \gamma \int_Z \widehat{v}_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid \Phi(z), a). \quad (3.22)$$

Proof. We first focus on the case $\kappa = 0$. By the definitions of v_0^j and $\widehat{v}_0^j$, we have

$$\begin{aligned}
|v_0^j(z) - \widehat{v}_0^j(z)| &= \left| \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E}[r^j(z, a^j, a^{-j})] - \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E}[r^j(\Phi(z), a^j, a^{-j})] \right| \\
&\leq \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E} \left[|r^j(z, a^j, a^{-j}) - r^j(\Phi(z), a^j, a^{-j})| \right] \\
&\leq L_0 \cdot \|z - \Phi(z)\|_1 \quad \forall z \in Z,
\end{aligned} \tag{3.23}$$

where we use the non-expansiveness of the minimax function and Lemma 3. Next, we focus on $\kappa > 0$. Based on the definitions of v_κ^j and $\widehat{v}_\kappa^j$, respectively, as described in (3.9) and (3.21), the non-expansiveness of the minimax function and the triangle inequality yield that

$$\begin{aligned}
|v_\kappa^j(z) - \widehat{v}_\kappa^j(z)| &\leq \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E} \left[|r^j(z, a^j, a^{-j}) - r^j(\Phi(z), a^j, a^{-j})| \right] \\
&\quad + \gamma \max_{\mu^j} \min_{\mu^{-j}} \mathbb{E} \left[\left| \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid z, a^j, a^{-j}) \right. \right. \\
&\quad \left. \left. - \int_Z \widehat{v}_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid \Phi(z), a^j, a^{-j}) \right| \right].
\end{aligned} \tag{3.24}$$

We can bound the first term on the right-hand side by using Lemma 3. Therefore, we focus on the second term. However, in the second term, not only the integrands $v_{\kappa-1}^j$ and $\widehat{v}_{\kappa-1}^j$ but also the measures $p(\cdot \mid z, a)$ and $p(\cdot \mid \Phi(z), a)$ are different. To address this issue, we can add and subtract the term $\int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid \Phi(z), a)$. Then, by the triangle inequality, we have

$$\begin{aligned}
&\left| \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid z, a) - \int_Z \widehat{v}_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid \Phi(z), a) \right| \\
&\leq \left| \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid z, a) - \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid \Phi(z), a) \right| \\
&\quad + \left| \int_Z (v_{\kappa-1}^j(\tilde{z}) - \widehat{v}_{\kappa-1}^j(\tilde{z})) p(d\tilde{z} \mid \Phi(z), a) \right|.
\end{aligned} \tag{3.25}$$

To bound the first term on the right-hand side of (3.25), we can invoke Proposition 2 and Lemma 4 such that

$$\begin{aligned}
& \left| \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid z, a) - \int_Z v_{\kappa-1}^j(\tilde{z}) p(d\tilde{z} \mid \Phi(z), a) \right| \\
& \leq \left(L_{\kappa-1} + \frac{\sqrt{|B|}}{\tau} M_{\kappa-1} \right) \|z - \Phi(z)\|_1 \\
& = \left(L_{\kappa-1} + L_0 \frac{1 - \gamma^\kappa}{1 - \gamma} \right) \|z - \Phi(z)\|_1, \tag{3.26}
\end{aligned}$$

where the equality follows from the definitions of L_0 and $M_{\kappa-1}$. On the other hand, we can bound the second term on the right-hand side of (3.25) by

$$\left| \int_Z (v_{\kappa-1}^j(\tilde{z}) - \widehat{v}_{\kappa-1}^j(\tilde{z})) p(d\tilde{z} \mid \Phi(z), a) \right| \leq \|v_{\kappa-1}^j - \widehat{v}_{\kappa-1}^j\|_\infty. \tag{3.27}$$

Then, by Lemma 3 and the bounds (3.26), (3.27), we obtain

$$\begin{aligned}
|v_\kappa^j(z) - \widehat{v}_\kappa^j(z)| & \leq \left(L_0 + \gamma L_{\kappa-1} + \gamma L_0 \frac{1 - \gamma^\kappa}{1 - \gamma} \right) \|z - \Phi(z)\|_1 + \gamma \|v_{\kappa-1}^j - \widehat{v}_{\kappa-1}^j\|_\infty \\
& = L_\kappa \cdot \|z - \Phi(z)\|_1 + \gamma \|v_{\kappa-1}^j - \widehat{v}_{\kappa-1}^j\|_\infty. \tag{3.28}
\end{aligned}$$

Given $\Delta \geq \|z - \Phi(z)\|_1$ for all z , we have

$$\|v_0^j - \widehat{v}_0^j\|_\infty \leq L_0 \cdot \Delta \tag{3.29a}$$

$$\|v_\kappa^j - \widehat{v}_\kappa^j\|_\infty \leq L_\kappa \cdot \Delta + \gamma \|v_{\kappa-1}^j - \widehat{v}_{\kappa-1}^j\|_\infty \quad \forall \kappa > 0 \tag{3.29b}$$

The recursion (3.29) on $\|v_\kappa^j - \widehat{v}_\kappa^j\|_\infty$ implies that

$$\|v_\kappa^j - \widehat{v}_\kappa^j\|_\infty \leq \Delta \cdot \sum_{k=0}^{\kappa} \gamma^{\kappa-k} L_k. \tag{3.30}$$

On the other hand, the recursion on $\{L_\kappa\}_{\kappa \geq 0}$ implies that

$$L_\kappa = L_0 \sum_{\ell=0}^{\kappa} (\ell + 1) \gamma^\ell. \tag{3.31}$$

After some algebra, (3.30) and (3.31) yield (3.20). $\square$

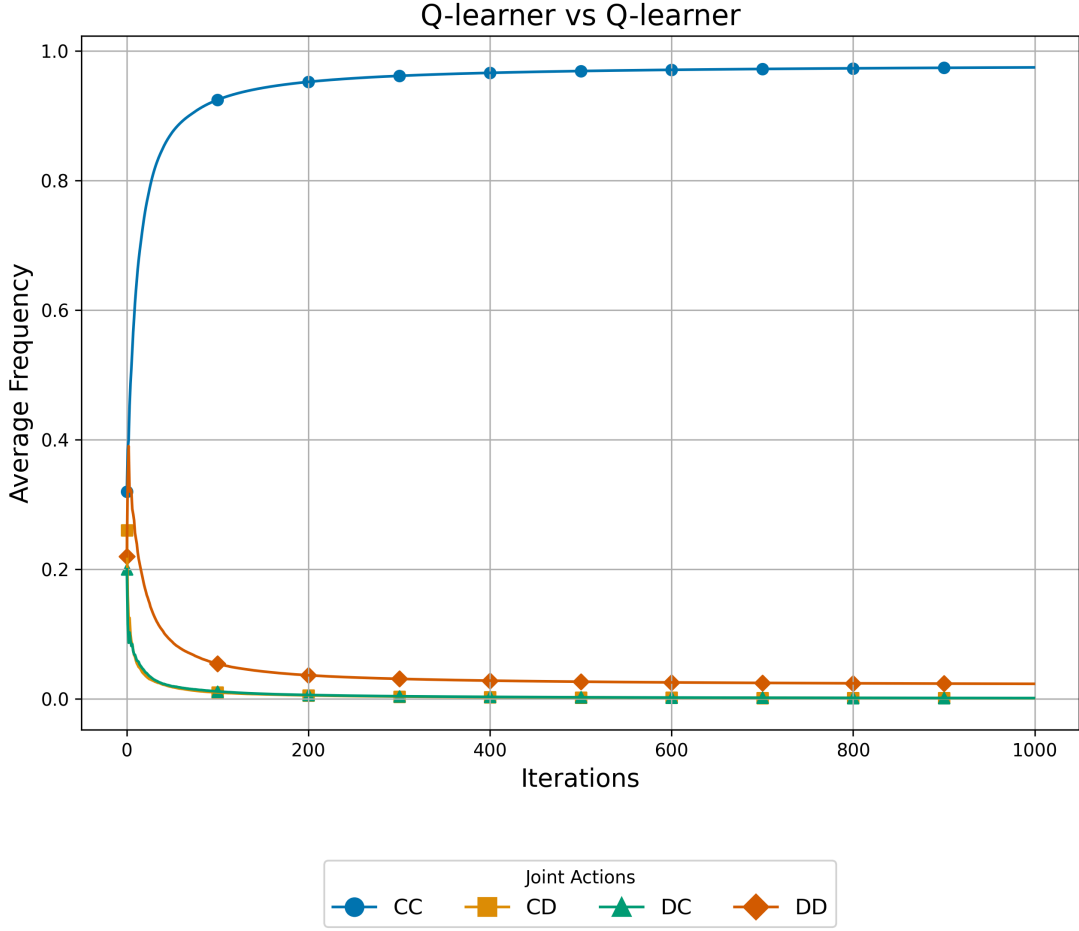


Figure 3.2: The evolution of the empirical averages of the action profiles for Q-learner vs Q-learner in the prisoner’s dilemma.

3.4 Illustrative Examples

To evaluate the performance of A-types and N-types, we conduct simulations across three scenarios: *(i)* a single A-type against a single A-type *(ii)* a single A-type against multiple N-types, and *(iii)* two competing A-types against a single N-type. N-types employ IQL with parameters $\tau = 0.01$ and $\alpha = 0.05$, while A-types utilize a discount factor $\gamma = 0.8$. We compute Markov stationary strategies for A-types through quantization (as discussed in Section 5.1), discretizing q-function values into 20 uniform intervals for multiple N-types scenarios and 100 uniform intervals for the rest of the experiments. Based on 100 independent trials, we report the mean empirical average of action profiles, following the methodology

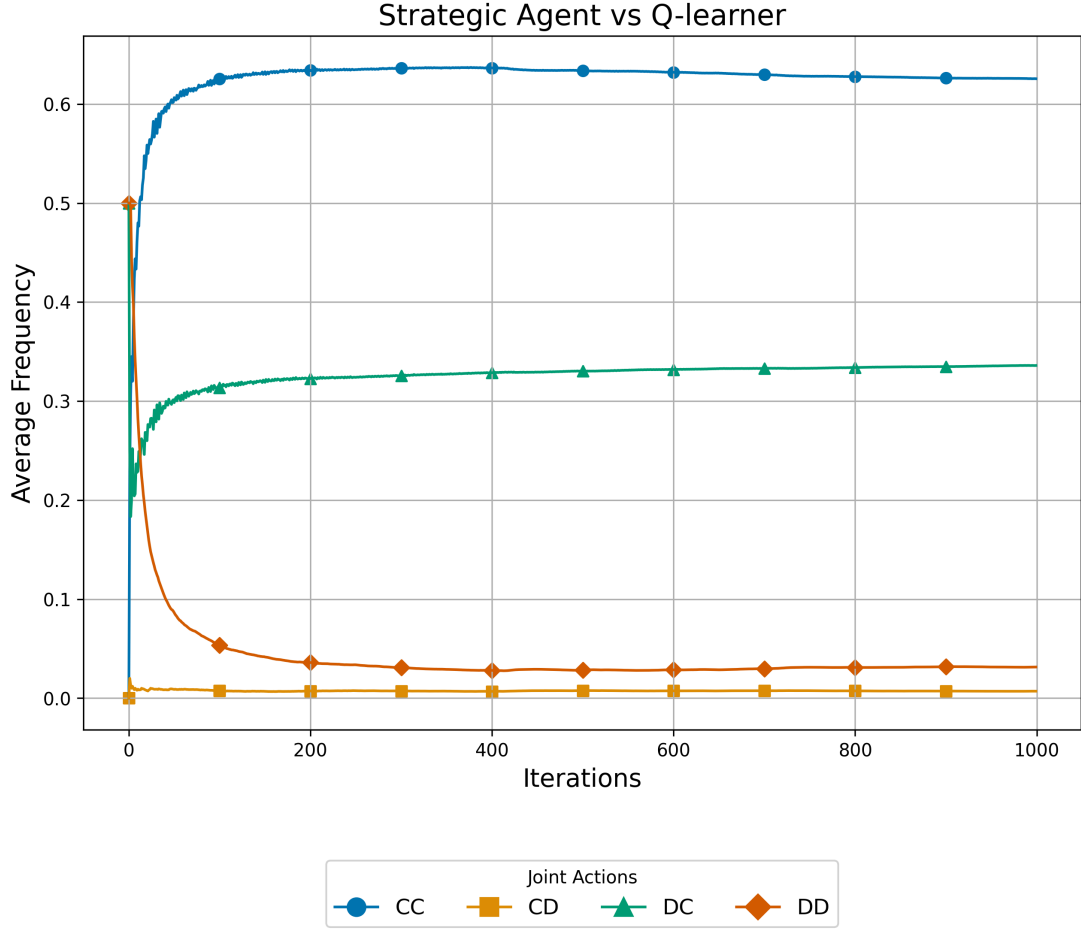


Figure 3.3: The evolution of the empirical averages of the action profiles for strategic actor vs Q-learner in the prisoner’s dilemma.

of [2]. In all three scenarios, A-types achieve significantly higher average payoffs (and consequently, utility) compared to scenarios where all agents employ IQL.

Single A-type against Single N-type. Recently, IQLs have been observed to lead to *tacit collusion* that can undermine the competitive nature of the markets [2, 5, 39, 40]. For example, in [2], the authors study the collusive behavior of IQLs in the widely-studied prisoner’s dilemma game (where agents have two actions: ‘cooperate’ and ‘defect’) that can be used to model the duopolies as matrix games. They observe that Q-learners can learn to collude in cooperation even though ‘cooperate’ is always an irrational choice as ‘defect’ dominates the

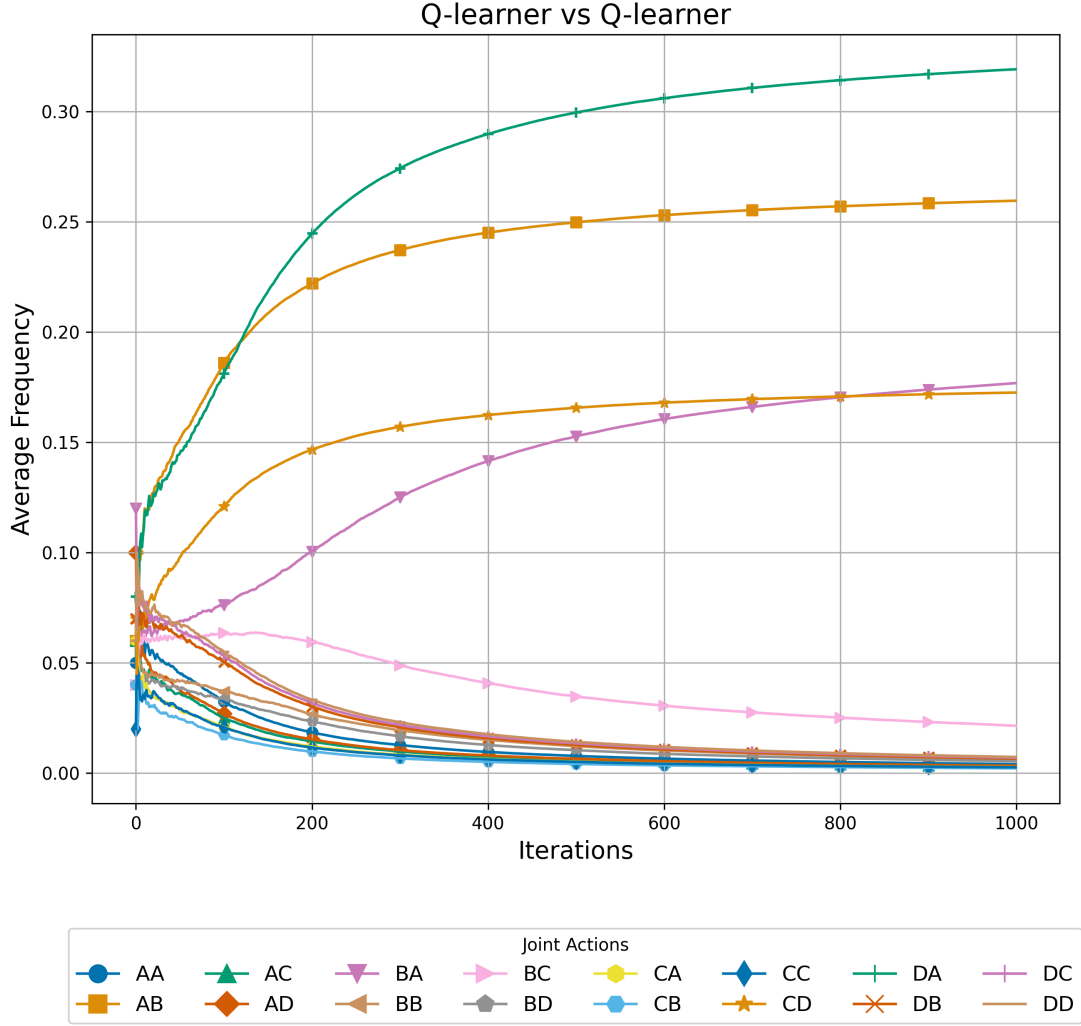


Figure 3.4: The evolution of the empirical averages of the action profiles for Q-learner vs Q-learner in the two-agent four-action zero-sum game.

‘cooperate’ strategy. Therefore, we focus on the prisoners’ dilemma game. Furthermore, we study the performance of the proposed quantization-based method in a randomly generated two-agent zero-sum game where each agent has four actions. In the prisoner’s dilemma, the strategic actions of the A-type lead to a significant decline in the empirical frequency of cooperation frequency from 95% to 60% since the A-type pushed the N-type to play “C” when she plays “D”, as illustrated in Figure 3.2 and Figure 3.3. Therefore, the A-type’s strategic act exploits N-type’s IQL and increases A-type’s utility (3.3) from 0.1041 to 0.2051.⁴

⁴The utilities are normalized with $(1 - \gamma)$ for 10^3 iterations.

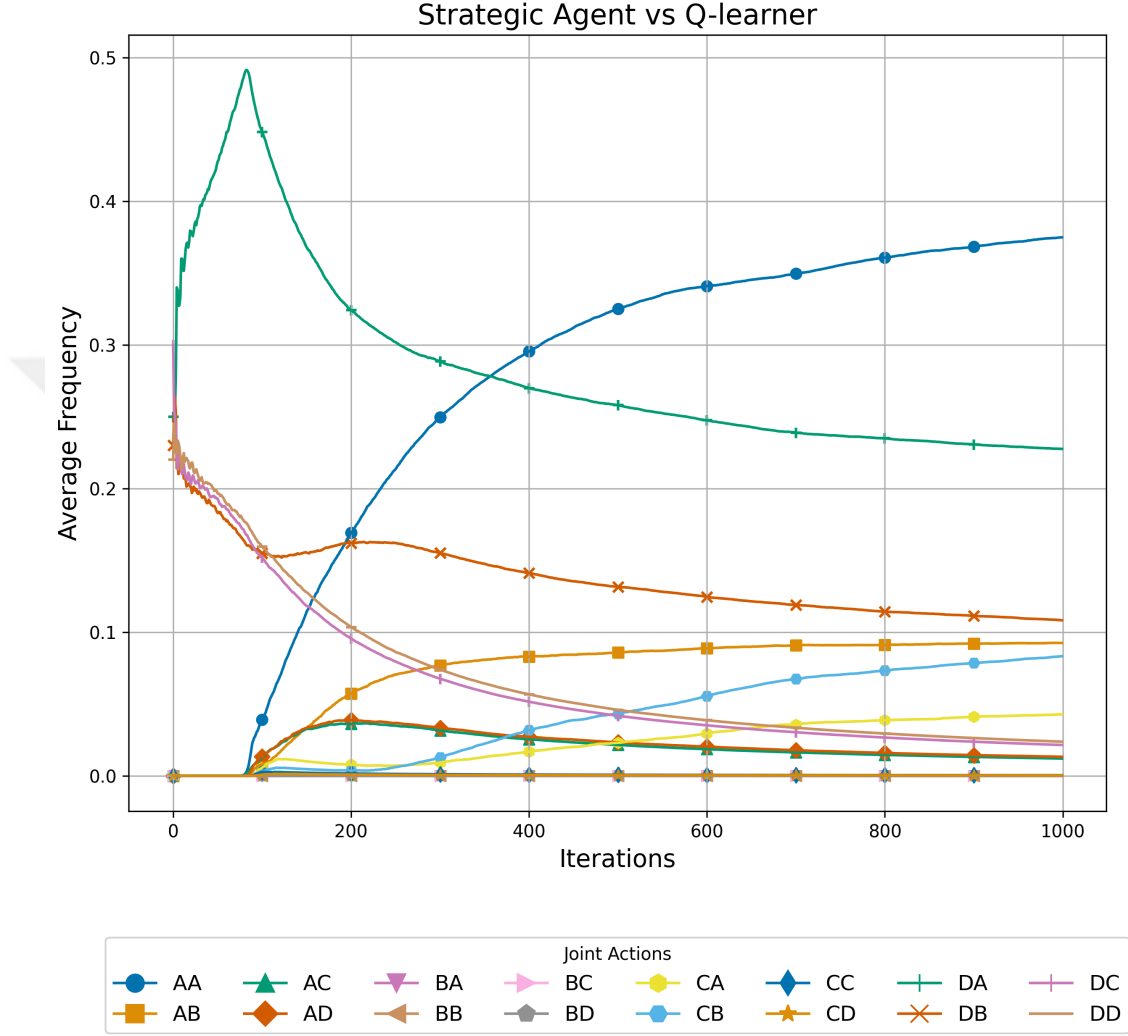


Figure 3.5: The evolution of the empirical averages of the action profiles for strategic actor vs Q-learner in the two-agent four-action zero-sum game.

Similarly, in the zero-sum game (see Figure 3.4 and Figure 3.5), A-type achieved the utility of 0.5340 when she acted strategically, which is higher than the value 0.4465 when she deployed IQL.

Single A-type against Multiple N-types. We conduct experiments on a single A-type interacting with two N-types under both aligned and misaligned objectives, where N-types form a potential game regardless of the A-types actions. When objectives are aligned, the strategically sophisticated A-type drives

Table 3.1: Single A-type vs Multiple N-types

		Agent Type	Learning Dynamics	Average Value
Aligned Obj.	$N \times N \times N$	N-type	IQL	0.1593
		N-type	IQL	0.1593
		N-type	IQL	0.1593
	$N \times N \times A$	A-type	DP	0.2396
		N-type	IQL	0.2396
		N-type	IQL	0.2396
Misaligned Obj.	$N \times N \times N$	N-type	IQL	-0.1470
		N-type	IQL	0.1470
		N-type	IQL	0.1470
	$N \times N \times A$	A-type	DP	-0.0810
		N-type	IQL	0.0810
		N-type	IQL	0.0810

an increase in average payoffs for all agents by solving the MDP as outlined in Remark 5. Conversely, under misaligned objectives, A-type’s payoff directly opposes N-types’ potential function, enabling exploitation of the N-types. These findings, summarized in Table 3.1, quantitatively demonstrate the ability of strategically sophisticated agents to manipulate naive agents in potential games, both positively and negatively based on objective alignment.

Two Competing A-types against Single N-type. We consider scenarios where N-type possesses an arbitrary payoff matrix, denoted by $\tilde{U}$, while A-types have opposing objectives represented by U and $-U$. In Table 3.2, we initially present average values for the $N \times N \times N$ case where all agents employ IQL. Subsequently, we introduce a strategically sophisticated agent by allowing one agent to adopt A-type role, solving the MDP using DP. Note that the impact of the strategic agent on N-types varies depending on the alignment of their objectives. Finally, we examine the effects of two competing A-types employing minimax-DP to compute their minimax-equilibrium strategy for the two-agent

Table 3.2: Two Competing A-types vs Single N-type

	Agent Type	Payoffs	Learning Dynamics	Average Value
$N \times N \times N$	N-type	$\tilde{U}$	IQL	0.4280
	N-type	U	IQL	0.3640
	N-type	$-U$	IQL	-0.3640
$N \times A \times N$	N-type	$\tilde{U}$	IQL	0.4875
	A-type	U	DP	0.4995
	N-type	$-U$	IQL	-0.4995
$A \times A \times N$	N-type	$\tilde{U}$	IQL	0.4854
	A-type	U	Minimax-DP	0.5092
	A-type	$-U$	Minimax-DP	-0.5092

zero-sum SG $\mathcal{Z}$. Interestingly, N-types often benefit from this competition due to the alignment of their payoffs with the A-types in the $A \times A \times N$ setting compared to $N \times N \times N$. These findings highlight complex interactions among strategic and naive agents.

Chapter 4

Strategic Security: Assessing Reinforcement Learning Vulnerabilities

*“If you know the enemy and know
yourself, you need not fear the
result of a hundred battles.”*

Sun Tzu

The growing sophistication of cyberattacks against critical infrastructure, such as power grids or financial networks [93, 94, 95], necessitates the innovative defense strategies. Traditional, static defense mechanisms struggle to adapt to attackers’ ever-evolving tactics. Reinforcement learning (RL) offers a compelling solution by enabling agents to learn and adapt to complex and dynamic threat landscapes [96, 97]. However, alongside these benefits, a crucial aspect to consider is the vulnerability of RL algorithms themselves [84, 82, 85]. For example, an advanced adversary could exploit an RL-based security system by conducting reconnaissance to learn the specific algorithm and its decision-making process [98]. This knowledge could allow the attacker to bypass security measures or manipulate the system behavior for malicious purposes.

To develop robust RL-based defense structures, we need to understand how and to what extent these algorithms might be vulnerable to exploitation. Security games provide a valuable framework for analyzing the interaction between a defender (security system) and an attacker (malicious actor) in a cybersecurity setting [43, 99, 100]. Within this framework, both players make strategic decisions to achieve their goals that are maximizing the security for the defender and bypassing defenses for the attacker.

To assess the vulnerability of RL algorithms in a security game context, we consider an advanced attacker with complete knowledge about the overall game dynamics and the defender’s learning rule. This worst-case scenario, while not perfectly reflecting real-world limitations of attackers, allows us to identify potential weaknesses in the RL algorithm that could be exploited if an attacker gained significant insight.

Specifically, we examine the stochastic security games [43] between a defender and a strategically sophisticated attacker. We model the stochastic security game as a *linear influence network* [43, 44, 46]. The defender employs widely used Independent Q-learning (IQL), as if there is no agent in the environment [31, 33, 30]. The attacker, being strategically sophisticated, is aware of the defender’s use of IQL. We demonstrate that the attacker can leverage this knowledge to gain an advantage against the defender. For the attacker, we model the stochastic security game as a Markov Decision Process (MDP) whose state encompasses the underlying game’s state and the defender’s Q-function estimate. This formulation allows us to leverage established MDP techniques such as value iteration [101] and linear programming solution [90] to analyze the attacker’s optimal strategy. However, the MDP has a high-dimensional continuum state space due to Q-function. To address this challenge, we present an approximation architecture based on neural networks.

The chapter is organized as follows: In Section 4.1 and 4.2, we present the linear influence network model and provide preliminary information about stochastic games. In Section 4.3, we describe the solution approach for the attacker and provide technical lemmas. In Section 4.4, we present our illustrative examples.

4.1 Linear Influence Network Models

Linear influence network (LIN) models represent security assets and their interdependencies using a graph structure as depicted in Figure 4.1 [43, 44, 45, 46]. Consider a weighted directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ is the set of nodes which we will call security assets targeted by an attacker and defended by a defender, and $\mathcal{E}$ is the set of edges representing the connections. The weights on the edges represent the strength of the influence or vulnerability between assets. We can categorize the edges in the LINs into two types:

- **Influence Edges ($\mathcal{E}_s$):** These edges indicate how much the security of one asset can be affected by the state of another. Specifically, we can define the *influence matrix* as

$$I = \begin{cases} w_{ij}, & \text{if } e_{ij} \in \mathcal{E}_s \\ 0, & \text{otherwise,} \end{cases} \quad (4.1)$$

where $0 < w_{ij} \leq 1 \ \forall i, j \in \mathcal{N}$ and $\sum_{i=1}^{|\mathcal{N}|} w_{ij} = 1, \ \forall j \in \mathcal{N}$.

- **Vulnerability Edges ($\mathcal{E}_v$):** These edges represent how exploiting one asset might make others vulnerable. We can define the *vulnerability matrix* as follows

$$V = \begin{cases} v_{ij}, & \text{if } e_{ij} \in \mathcal{E}_v \\ 0, & \text{otherwise,} \end{cases} \quad (4.2)$$

where $0 \leq v_{ij} \leq 1$ measures the vulnerability of asset j caused by asset i due to the interdependencies within the system. The support to node j is defined to be $v_j := \sum_i v_{ij}$.

Remark 6. *The influence matrix I is a stochastic matrix. Therefore, when a security asset is compromised, both the compromised asset and its connecting influence edges disappear from the graph. Conversely, when a security asset is*

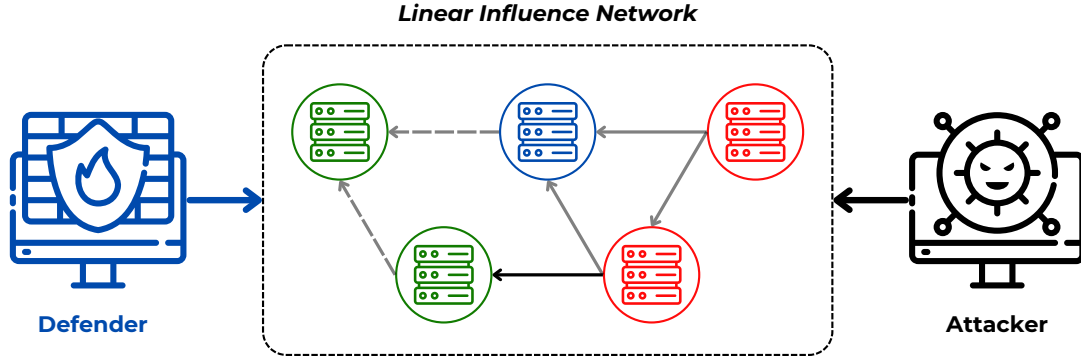


Figure 4.1: A linear influence network illustrating compromised and secured security assets. Green servers represent uncompromised assets, red servers represent compromised assets, and the blue server represents a defended asset. Shaded arrows depict interdependencies between assets: solid arrows indicate usable attack paths, and dashed arrows indicate unusable paths. The solid black arrow represents the attacker's direct attack.

secured, it regains its original edges. As a result, the influence matrix must be normalized to maintain its stochastic property. On the other hand, the vulnerability matrix V is not stochastic. This is because, even when an asset is compromised, it continues to affect the vulnerabilities of other assets.

Let $x = \{x_1, x_2, \dots, x_{|N|}\}$ be the vector measuring the value of independent security assets. We can compute the vector of effective security measures $y = \{y_1, y_2, \dots, y_{|N|}\}$ by using the influence matrix I such that

$$y = Ix.$$

This concept of influence plays a critical role in analyzing attacker and defender strategies within a LIN.

4.1.1 Attack and Defense on the Network

Consider the interaction between the attacker and defender on a LIN. Assume that in this interaction, the attacker can attack node i , and the defender can defend node j . Let a represent the attack by the attacker and d represent the

defense by the defender. Furthermore, let the state of the network, s , indicate which assets are compromised or not. We can define the probability that security asset i is compromised as

$$p^i(s, a, d) = \begin{cases} \max \{p_{n0}^i(1 - v_i) + p_{n1}^i v_i, 1\} & a = i, d \neq i, s(i) = 0 \\ \max \{p_{d0}^i(1 - v_i) + p_{d1}^i v_i, 1\} & a = d = i, s(i) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq p_{n1}^i \leq p_{n0}^i \leq 1$ be the standalone probability of compromise of node i , when there is full support and there is no support in the case defender does not defend node i respectively. Similarly, $0 \leq p_{d1}^i \leq p_{d0}^i \leq 1$ are standalone compromise probabilities when there is full support and no support for node i , in the case that node i is defended respectively. Also, $s(i) = 0$ if the node i is uncompromised and $s(i) = 1$ if the node i is compromised. The attacker's reward function $r(s, a, d)$ considers the effectiveness of compromising node i within the current network state. Therefore, the reward function becomes

$$r(s, a, d) = p^a(s, a, d)y(s). \quad (4.3)$$

We can now formulate this attacker-defender interaction as a stochastic game, where the state of the network evolves based on the actions of both the attacker and defender, as well as the underlying probabilities [43, 46]. In the following section, we provide the necessary preliminary information on stochastic games.

4.2 Stochastic Games and Independent Q-learning

Consider a two-agent *stochastic game* (SG) with *finite* states and actions that can be characterized by the tuple $\langle S, (A^i, r^i, \gamma^i)_{i=1}^2, p \rangle$, where S denotes the *finite* set of states, A^i and $r^i : A \rightarrow \mathbb{R}$ for $A := \prod_j A^j$ denote, resp., agent i 's action set

and reward function.¹ Furthermore, $p(\cdot \mid \cdot)$ is the transition function such that $p(s_+ \mid s, a)$ for $(s, a, s_+) \in S \times A \times S$ is the probability of transition from s to s_+ under action profile $a \in A$.

At each stage $k = 0, 1, \dots$, both agents observe the current state $s_k \in S$, takes (possibly random) action $a_k^i \in A^i$ *simultaneously*, then receives the reward $r_k^i = r^i(s_k, a_k)$ for $a_k = \{a_k^j\}_{j=1}^2$. The state s_k changes to s_{k+1} with probability $p(s_{k+1} \mid s_k, a_k)$. The agent's goal to maximize the utility

$$\mathbb{E} \left[\sum_{k=0}^{\infty} (\gamma^i)^k r^i(s_k, a_k) \right] \quad (4.4)$$

with the discount factor $\gamma^i \in (0, 1)$, where the expectation is taken over the randomness on (s_k, a_k) .

If there is only one agent, say agent i , then SG reduces to *Markov decision process* (MDP). Call the value of the state-action pair by the *Q-function*. Based on the principle of optimality that the agent would play the best action in future stages, the Q-function, denoted by $q^i : S \times A^i \rightarrow \mathbb{R}$, is the (unique) solution to the following fixed-point equation [101, Chapter 3]

$$q^i(s, a^i) = r^i(s, a^i) + \gamma \sum_{s_+ \in S} p(s_+ \mid s, a^i) \max_{\tilde{a}^i \in A^i} \{q^i(s_+, \tilde{a}^i)\}. \quad (4.5)$$

Let $\Delta(A)$ denote the probability simplex over a finite set A . Then, any Markov stationary strategy $\pi^i : S \rightarrow \Delta(A^i)$ satisfying $\pi^i(s) \in \operatorname{argmax}_{\mu^i} \mathbb{E}_{a^i \sim \mu^i} [q^i(s, a^i)]$ maximizes the utility (4.4).

¹The formulation can be generalized to state-variant action sets.

4.2.1 Q-learning

When the model of the underlying MDP (e.g., $r^i(\cdot)$ and $p(\cdot \mid \cdot)$) is unknown, the Q-learning algorithm, introduced by [102], can recursively estimate the Q-function according to

$$q_k^i(s_{k-1}, a_{k-1}^i) = (1 - \alpha^i)q_{k-1}^i(s_{k-1}, a_{k-1}^i) + \alpha^i(r_{k-1}^i + \gamma^i \max_{\tilde{a}^i \in A^i} \{q_{k-1}^i(s_k, \tilde{a}^i)\}), \quad (4.6a)$$

$$q_k^i(s, a^i) = q_{k-1}^i(s, a^i) \quad \forall (s, a^i) \neq (s_{k-1}, a_{k-1}^i), \quad (4.6b)$$

where $\alpha^i \in (0, 1)$ is some step size and (s_{k-1}, a_{k-1}^i) is the pair of the state and action at stage $k - 1$. The action a_k^i is taken according to the soft-max function $\text{BR}_\tau : \mathbb{R}^{A^i} \rightarrow \Delta(A^i)$, defined by

$$\text{BR}_\tau(q_k^i(s_k, \cdot))(a^i) := \frac{\exp(q_k^i(s_k, a^i)/\tau)}{\sum_{\tilde{a}^i \in A^i} \exp(q_k^i(s_k, \tilde{a}^i)/\tau)} > 0 \quad \forall a^i \quad (4.7)$$

for some temperature parameter $\tau > 0$ controlling the level of exploration, i.e., $a_k^i \sim \text{BR}_\tau(q_k^i(s_k, \cdot))$.

In SG, the other agent's play might be unknown similar to the model of the underlying environment. For such cases, agents can still follow the Q-update (4.6) independently *as if* there are no other agents or the other agents' play is a part of the underlying environment's (unknown) dynamics. This approach is known as *independent Q-learning* (IQL), deployed widely in practice [5, 31, 39, 33, 30].

4.3 Strategic Attacker with Complete Model Knowledge

We examine the susceptibility of IQL to strategic manipulations by an *advanced* player in LIN model. To this end, we consider a two-agent SG $\mathcal{M} = \langle S, (A^i, r^i, \gamma^i)_{i=1}^2, p \rangle$ played by *attacker* and *defender* agents:

- **Defender** agent is a *naive* agent following the IQL algorithm (4.6) based on its local information (i.e., state and reward received) as if there are no other agents.
- **Attacker** agent is an *advanced* decision-maker with the complete knowledge of the underlying game dynamics and the employed IQL, and access to the joint actions of every player.

The state space S indicates which security assets are compromised or not. We denote the action sets of the defender and attacker as $\underline{A}$ and $\overline{A}$, respectively. In LIN model, the defender can defend an uncompromised security asset and the attacker can attack, similarly, to an uncompromised one. The reward functions and transition probabilities for each agent follow a similar notation to action sets. Specifically, $\underline{r}(s, a)$ and $\underline{p}(s_+|s, a)$ represent the reward function and transition probabilities for the defender, while $\overline{r}(s, a)$ and $\overline{p}(s_+|s, a)$ represent those for the attacker. Here, a denotes the joint action taken by the agents.

At each stage k , defender takes action $\underline{a}_k \sim \text{BR}_\tau(q_k(s_k, \cdot))$ independently, depending on its local Q-function q_k . Attacker, knowing the employed IQL, can track q_k and compute the play of defender $\text{BR}_\tau(q_k(s_k, \cdot))$. Therefore, the attacker can reformulate the SG $\mathcal{M}$ as a MDP in which the defender is considered a non-strategic player and is modeled as part of the system.

Definition 5 (Markov Decision Process) *Given the SG $\mathcal{M}$ its MDP characterized by the tuple $\mathcal{Z} = \langle Z, \overline{A}, \overline{r}, P, \gamma \rangle$. The state space $Z := S \times Q$ is the compact set of states such that $Q \subset \mathbb{R}^{S \times \underline{A}}$ corresponds to all possible Q-function estimates of the defender. The reward function $\overline{r} : Z \times \overline{A} \rightarrow \mathbb{R}$ is given by*

$$\overline{r}(z, \overline{a}) = \mathbb{E}_{\underline{a} \sim \text{BR}_\tau(q)}[r(s, \overline{a}, \underline{a})] \quad \forall (z, \overline{a}) \in Z \times \overline{A}. \quad (4.8)$$

The transition kernel $P(\cdot|\cdot)$ determines the evolution of states according to transition probabilities $\underline{p}(\cdot|\cdot)$ and the update rule (4.6). The discount factor $\gamma \in (0, 1)$ is as described in $\mathcal{M}$.

For example, given $(z, \bar{a})$, defender can play any $\underline{a} \in \underline{A}$ with probability $\text{BR}_\tau(z)$. Given $(z, \bar{a}, \underline{a})$, defender's Q-function estimate $\hat{q}$ can get updated to $\hat{q}_+$ according to

$$\hat{q}_+(s, a) = (1 - \alpha)\hat{q}(s, a) + \alpha \left(r(s, \bar{a}, \underline{a}) + \gamma \max_{\bar{a} \in \underline{A}(s_+, \bar{a})} \hat{q}(s_+, a) \right), \quad (4.9a)$$

$$\hat{q}_+(\hat{s}, \hat{a}) = \hat{q}(\hat{s}, \hat{a}) \quad \forall (\hat{s}, \hat{a}) \neq (s, a), \quad (4.9b)$$

with probability $p(s_+ | s, \bar{a}, \underline{a}) > 0$.

With the complete knowledge of the employed IQL, attacker can compute the state $z_k \in Z$. Let the attacker follow stationary strategy $\pi : Z \rightarrow \Delta(\bar{A})$. Given the stationary strategy the attacker's utility (4.4) for the underlying SG $\mathcal{M}$ can be rewritten as

$$\bar{U}(\pi) := \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k \bar{r}(z_k, \bar{a}_k) \right] \quad (4.10)$$

for the MDP $\mathcal{Z}$ where the expectation is now taken over the randomness on $(z_k, \bar{a}_k)$ due to (4.8).

Proposition 4. *There always exists an optimal stationary policy for $\mathcal{Z}$.*

Proof. The proof follows from [90, Theorem 6.2.12] based on the observation that the (continuum) state space $Z = S \times Q$ in $\mathcal{Z}$ is a Polish space as S is finite and Q is a compact subset of $\mathbb{R}^{S \times \underline{A}}$, and the action set $\underline{A}$ is finite. $\square$

Although an optimal stationary policy exists, the continuum state space poses a challenge. To address this, we can utilize approximation architectures from MDP literature. Various approaches are discussed in the literature. For example, in the state aggregation method, states are clustered into meta-states based on their similarity in terms of rewards and state transitions [36, 35]. Quantization-based solutions discretize the state space using a quantizer mapping continuum state space to countable many approximated state space [34]. In approximate linear programming, approximate value iteration, and approximate

policy iteration, the state space is reduced by utilizing pre-selected basis functions [42, 37, 38, 92, 91].

To approximate the MDP $\mathcal{Z}$, we employ neuro-dynamic programming (NDP) [42], leveraging the power of approximation architectures, specifically neural networks, with dynamic programming to handle the continuum state space. The following proposition guarantees a certain level of smoothness, and allows us to effectively utilize NDP.

Proposition 5. *The value function $v_\kappa(z)$, corresponding to the values the states in Z at the initial stage of κ length horizon defined as*

$$v_\kappa(z) = \max_{a \in \bar{A}} \left\{ \bar{r}(z, a) + \gamma \int_Z v_{\kappa-1}(\tilde{z}) P(d\tilde{z}|z, a) \right\} \quad (4.11)$$

for all $\kappa > 0$ and $v_0(z) = \max_{a \in \bar{A}} \bar{r}(z, a)$ is bounded by

$$M_\kappa = \frac{1 - \gamma^{\kappa+1}}{1 - \gamma} \max_{(a,b)} |\bar{u}(a, b)| \quad (4.12)$$

for all $z \in Z$ and $\kappa \geq 0$. The value function is Lipschitz continuous with respect to ℓ_2 -norm with Lipschitz constant

$$L_\kappa = \gamma L_{\kappa-1} + L_0 \frac{1 - \gamma^{\kappa+1}}{1 - \gamma} \quad (4.13)$$

for all $\kappa > 0$ and $L_0 := \frac{1}{\tau} \sqrt{S \times \underline{A}} \max_{(a,b)} |\bar{u}(a, b)|$.

Proof. The proof builds upon Proposition 2 with two key adaptations. First, due to the presence of a single attacker, we replace the minimax value function with the value function defined in (4.11). Second, we extend the proof to consider states arising from a stochastic game environment, unlike the repeated play of normal-form game setting in Chapter 3. $\square$

Remark 7. *The Lipschitz continuity established in Proposition 5 guarantees a value function with some smoothness, enabling us to bound the approximation error introduced by our function approximation approach (e.g., neural networks).*

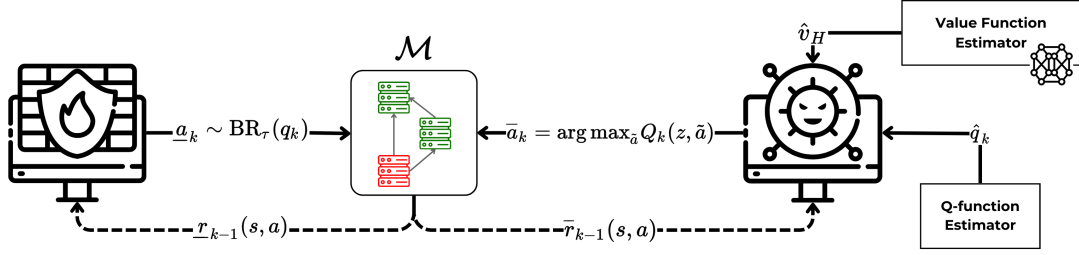


Figure 4.2: The illustration demonstrates how the attacker and the defender take actions during the stochastic game $\mathcal{M}$. The dashed lines indicate observations from the previous stage.

Lipschitz continuity has been previously leveraged in MDP approximation. For instance, [103] defines a Lipschitz model class based on the transition kernel and utilizes Kantorovich-Rubinstein duality for approximation. Similarly, [104] employs Lipschitz continuity on the reward and value functions to achieve state space reduction.

Thanks to the Lipschitz continuity of the value function in the MDP, the attacker can approximate this value function using various function approximation methods. In this case, we approximate it with a neural network, whose input has the size of the Q-function of the opponent $|S| \times |A|$, and the output vector is the size of the state space $|S|$. Each output will correspond to the value of the given input opponent q in the corresponding game state s . In short, we represent such (q, s) pairs as an extended state z . The attacker draws a set of m samples at stage κ denoted by $\overline{Q}_\kappa = \{q_{\kappa 1}, \dots, q_{\kappa m}\}$ from the continuous state space Q , for all states $s \in S$, and trains a neural network with multi-layer perceptron structure, parameterized by θ , to approximate the true value function, $\hat{v}(z) \approx v_\theta(z)$, using these sampled states. Specifically, the attacker initializes $\hat{v}_0 = \bar{r}(z, a)$ since it knows the reward structure of the underlying stochastic game. The attacker uses the values of the κ -th time step as labeled data to train the neural network for the value function at time step $\kappa + 1$ by using bootstrapping, i.e.,

$$\hat{v}_\kappa(q_\kappa, s) = \max_{\tilde{a} \in \tilde{A}} \left\{ \bar{r}(q_\kappa, s, a) + \gamma \sum_{s_+ \in S} \int_{Q_\kappa} \hat{v}_{\kappa-1}(\tilde{q}, s_+) \bar{p}((d\tilde{q}, s_+) \mid q_\kappa, s, a) \right\}$$

for all $s \in S$. The reason is that directly obtaining the future rewards is generally

Algorithm 5 Learning Value Function

Input: $\bar{r}(s, \bar{a}, \underline{a}), \bar{p}(s'|s, \bar{a}, \underline{a}), \gamma, \tau, \alpha, H, m$
for each step $\kappa = 0, 1, \dots, H$ **do**
 Randomly generate m samples from $\bar{Q} : \{x_{\kappa,i} = q_{\kappa,i}(\cdot)\}_{\kappa=1}^m$
for $i = 1, 2, \dots, m$ **do**
 Initialize: $Q_{\kappa,i}(\bar{a}, b) = 0 \quad \forall (\bar{a}, \underline{a}) \in \bar{A} \times \underline{A}$
 for each $s, \bar{a}, \underline{a}, s_+ \in S \times \bar{A} \times \underline{A} \times S$ **do**
 Define $q_+(\cdot) = q_{\kappa,i}(\cdot)$
 Update

$$q_+(s, \underline{a}) = q_{\kappa,i}(s, \underline{a}) + \alpha \left(\underline{r}(s, \bar{a}, \underline{a}) + \gamma \max_{\underline{a}' \in \underline{A}} \{q_{\kappa,i}(s_+, \underline{a}')\} - q_{\kappa,i}(s, \underline{a}) \right)$$

 Calculate

$$Q_{\kappa,i}(\bar{a}, \underline{a}) = Q_{\kappa,i}(\bar{a}, \underline{a}) + \bar{p}(s_+|s, \bar{a}, \underline{a}) (\bar{r}(s, \bar{a}, \underline{a}) + \mathbb{I}_{\kappa \neq 0} \gamma \hat{v}_{\kappa-1}(q_+)(s_+))$$

 Define $y_{\kappa,i}(s) = \max_{\bar{a} \in \bar{A}} \left\{ \sum_{\underline{a} \in \underline{A}} \text{BR}_{\tau}(q_{\kappa,i}(s, \cdot))(\underline{a}) \times Q_{\kappa,i}(\bar{a}, \underline{a}) \right\}$
 end for
 end for
 Train a neural network $\hat{v}_{\kappa}$ with data $D = \{(x_i, y_i)\}_{i=1,2,\dots,m}$
end for

impractical. Because of this limited information, the attacker leverages past value function estimates to train the neural network. This iterative process allows the network to learn and improve its value function approximation over time.

The attacker uses the estimated value function to compute its Q-function estimate,

$$\hat{Q}_{\kappa}(z, \cdot) = \bar{r}(z, \cdot) + \gamma \mathbb{E}_{\bar{p}(z'|z, \cdot)} [\hat{v}_{\kappa}(z')]$$

which estimates the expected future reward for taking a specific action in a given state. During the game, the attacker chooses the action $\bar{a}$ that maximizes its Q-function estimate $\bar{a}_{\kappa} = \text{argmax}_{\bar{a}} (\hat{Q}_{\kappa}(z, \bar{a}))$ as depicted in Figure 4.2.

The iterative process of sampling, updating the value function estimate, and determining the optimal policy allows the neural network to learn and improve its

Algorithm 6 Gameplay of the Attacker

Get $\hat{v}_H$

for each step $n = 1, 2, \dots$ **do**

Observe current state s_n and current opponent Q-function $\hat{q}_n$

Initialize: $Q_n(\bar{a}, \underline{a}) = 0 \quad \forall (\bar{a}, \underline{a}) \in \bar{A} \times \underline{A}$

for each $\bar{a}, \underline{a}, s_+ \in \bar{A} \times \underline{A} \times S$ **do**

 Define $\hat{q}_+(\cdot) = \hat{q}_n(\cdot)$

 Update

$$\hat{q}_+(s_n, \underline{a}) = \hat{q}_n(s_n, \underline{a}) + \alpha \left(\bar{r}(s_n, \bar{a}, \underline{a}) + \gamma \max_{\underline{a}' \in \underline{A}} \{ \hat{q}_n(s_+, \underline{a}') \} - \hat{q}_n(s_n, \underline{a}) \right)$$

 Calculate

$$Q_n(\bar{a}, \underline{a}) = Q_n(\bar{a}, \underline{a}) + \bar{p}(s_+ | s_n, \bar{a}, \underline{a}) (\bar{r}(s_n, \bar{a}, \underline{a}) + \gamma \hat{v}_H(q_+)(s_+))$$

 Define $\hat{Q}_n(\bar{a}) = \left\{ \sum_{\underline{a} \in \underline{A}} \text{BR}_\tau(\hat{q}_n(s_n, \cdot))(\underline{a}) \times Q_n(\bar{a}, \underline{a}) \right\} \quad \forall \bar{a} \in \bar{A}$

 Play action $\bar{a} = \arg\max_{\bar{a} \in \bar{A}} \hat{Q}_n(\bar{a})$

 Update state $s \leftarrow s'$

end for

end for

approximation over time. However, bootstrapping can introduce some error, since it relies on estimates instead of actual future rewards in the training. The attacker can adopt more sophisticated training techniques to improve performance, but we will not discuss these techniques as they are outside the scope of this work.

Remark 8. *As discussed in [42], to approximate the value function of the MDP, the attacker can employ any type of approximating structure without limiting itself to a multi-layer perceptron architecture.*

4.4 Illustrative Examples

In the experiments, we used an LIN, defined similarly to the one in [46]. This LIN consists of three security assets. The matrices I , V , and the vector x are

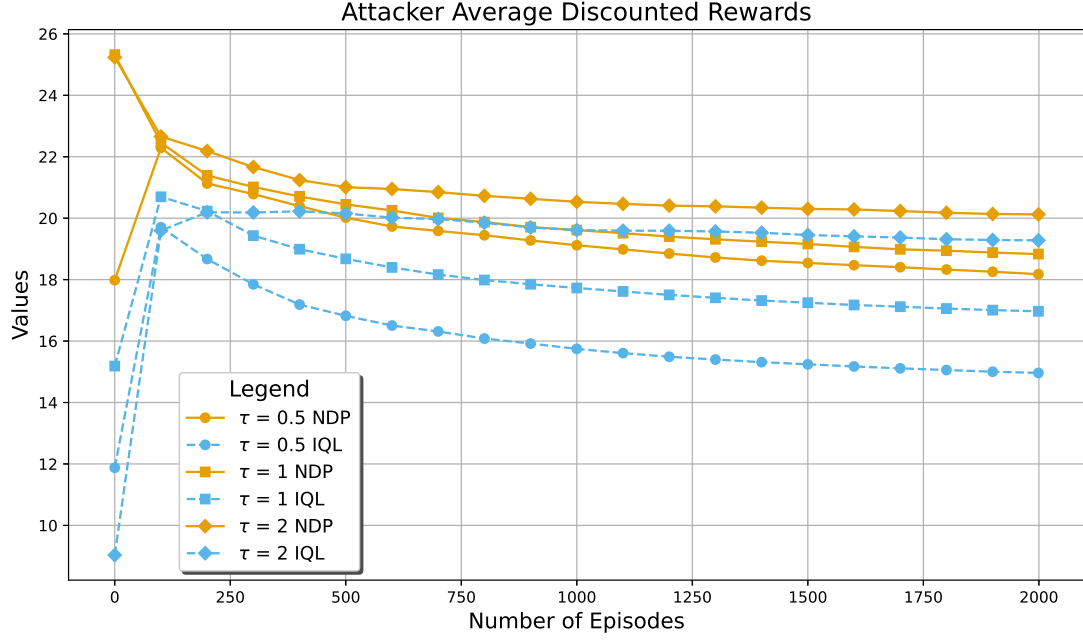


Figure 4.3: Average discounted rewards of the attacker against different exploration levels.

defined as follows:

$$I = \begin{bmatrix} 0.9 & 0.2 & 0 \\ 0 & 0.7 & 0 \\ 0.1 & 0.1 & 1 \end{bmatrix} \quad V = \begin{bmatrix} 0.7 & 0 & 0 \\ 0.2 & 0.5 & 0 \\ 0.1 & 0.3 & 0.9 \end{bmatrix} \quad x = \begin{bmatrix} 10 \\ 10 \\ 20 \end{bmatrix}. \quad (4.14)$$

Each three security asset can either be compromised or not compromised by the attacker, meaning there are in total 8 states. At each stage, both the attacker and the defender can choose one node to attack or defend or do nothing, respectively, meaning that each of them has 4 actions. The state transition probabilities are defined so that $p_{n0} = 0.7$, $p_{n1} = 0.4$, $p_{d0} = 0.5$, $p_{d1} = 0.2$, and $p_r = 0.2$, and $p_e = 0.3$, which are game reset and game end probabilities in case the attacker do nothing or fails. There are two models used for value function estimation. The models are neural networks that defines a sequence of fully connected layers with ReLU activations, designed for value function approximation with an input size of 36 (from each Q-value entry of the 4 actions across 9 total states, including the “do nothing” action and the terminal state). The first complex model **CM** has

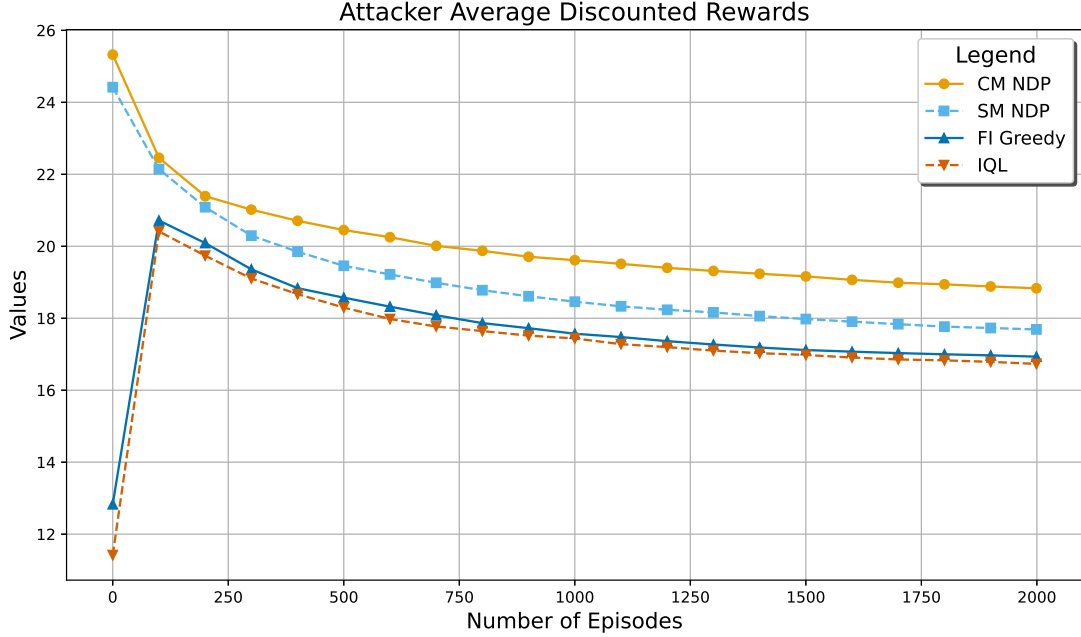


Figure 4.4: Average discounted rewards of the attacker with different learning models.

hidden layers of 64, 64, and 32 neurons, and a vector output of size 9, that maps the Q-function to a value for each state, while the simpler model **SM** has no hidden layers and only an input and output layer. We train the NDP algorithm for 50 horizons, where in each horizon the model is trained for two epochs with 5×10^6 samples for training. The average MSE test error per horizon after two epochs is less than 10^{-2} for the model **CM** during the first few horizons and even less for the following horizons. For **SM** model, the training and testing errors are around 5×10^{-1} for the first horizon and decreases below 10^{-2} after a few horizons.

In our experiments, we used learning rate $\alpha = 0.05$, discount factor $\gamma = 0.95$, and temperature parameter $\tau \in \{0.5, 1, 2\}$. We ran 50 trials for each experiment with 2000 episodes, averaging the total discounted reward per episode over the trials.

We first investigated the impact of exploration on vulnerability. As shown in Figure 4.3, NDP consistently achieved a significantly higher total discounted

reward compared to the IQL-agent. We observed that with increasing τ (more exploration), the reward for the NDP agent also increased. However, the gap between NDP vs IQL and IQL vs IQL decreased, indicating reduced vulnerability. This suggests a strategically advanced agent can manipulate an IQL-agent based on its exploration parameter.

Next, we evaluated the impact of the attacker’s model complexity on vulnerability. We kept the opponent as an IQL-agent but varied the attacker’s model architecture (**CM** and **SM**) to assess the neural network’s value function approximation capability. Figure 4.4 shows that the **CM** outperformed the **SM** by achieving a higher average discounted reward. This demonstrates that the attacker’s ability to approximate the defender’s value function directly affects how much it can exploit the IQL-agent. Notably, both attacker models outperformed the greedy best response where the agent utilizes knowledge of the opponent’s q values and the game model to play in a way that maximizes the immediate reward given this information, further supporting the dominance of the NDP algorithm.

Chapter 5

The Role of Strategic Agent in Human-AI Interactions

“I visualize a time when we will be to robots what dogs are to humans, and I’m rooting for the machines.”

Claude Shannon

The increasing popularity of Artificial Intelligence (AI) is transforming human-machine interaction, particularly in cyber-physical systems (CPS) where physical components integrate with computational intelligence. These complex systems require a deep understanding of human-AI interactions to achieve optimal performance and ensure safety [47, 48]. A major challenge lies in the vulnerability of traditional learning algorithms to manipulation by strategically advanced agents like AI [82, 85]. This raises a critical question:

How much can a strategic AI manipulate, assist, or guide an (human) agent within these complex CPS to achieve a more favorable outcome?

To tackle this problem, we take a control-theoretic approach, examining the

repeated play of general-sum normal-form games between a human agent and a strategic AI agent. Our goal is to investigate how the AI agent can strategically influence the human’s learning behavior to achieve a more favorable outcome. This outcome could be maximizing the joint payoff for both agents, or it could be maximizing the AI’s own payoff at the expense of the human. We model the human agent’s behavior using the Experience-Weighted Attraction (EWA) algorithm, which integrates belief learning and reinforcement learning models, making it particularly well-suited for capturing the complexity of human decision-making in strategic interactions. Unlike simpler models that focus solely on beliefs or received payoffs, EWA can account for both aspects, leading to a more comprehensive understanding of human behavior in games. The human agent employs the EWA algorithm with certain preselected parameters that may vary during the game. The AI agent, being strategically sophisticated, is aware of the human agent’s use of the EWA algorithm. We show that the AI agent can use this awareness to control the human agent’s learning behavior. By modeling the problem as a Markov Decision Process (MDP), the AI agent can maximize the discounted sum of payoffs over an infinite horizon. The main challenge of this approach is the continuous state space of the modeled MDP. To address this, we propose a quantization-based approximation method. This method approximates the MDP with a finite state version, allowing us to apply dynamic programming techniques.

The studies in the vulnerability of learning dynamics literature rely on specific assumptions about the type of the opponent’s learning rule [84, 87, 85, 82, 83]. Our work addresses this limitation by employing EWA model, introduced in the seminal work of [49] for human behavior in cyber-physical human systems (CPHS). Depending on the choice of EWA parameters, EWA encompasses many widely used and studied learning algorithms from the class of belief learning and reinforcement learning as illustrated in Figure 5.1, such as fictitious play, logit learning or Q-learning [51]. This versatility makes EWA more suitable for modeling human behavior compared to using belief learning and reinforcement learning separately. Therefore, EWA provides a comprehensive description of human agents across a wide range of games [51, 50].

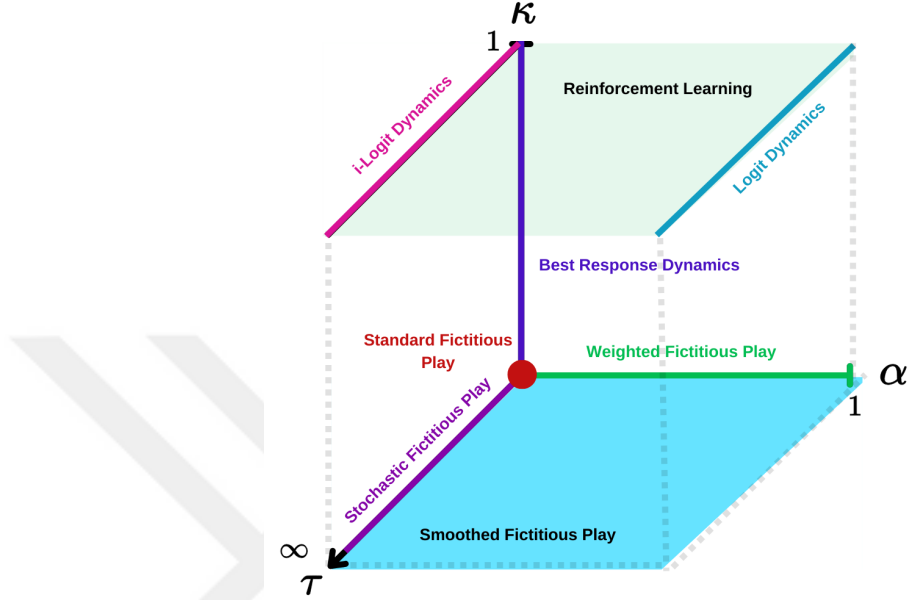


Figure 5.1: An illustration of the EWA learning dynamics based on the parameters τ , α , κ , and δ . Belief-based learning dynamics, represented with a solid line, correspond to the case where $\delta = 1$. Payoff-based learning dynamics (i.e., reinforcement learning) are depicted as a shaded area for $\delta = 0$.

The chapter is organized as follows: In Section 5.1, we present the interaction between human and strategic AI agents as a game and introduce their respective behavioral rules: EWA for the human agent and a stochastic control approach with a quantization-based solution for the AI agent. In Section 5.2, we demonstrate our results in various types of games.

5.1 Strategic AI and Human Interaction

Consider a normal-form game $\mathcal{G}$ played over stages k , defined by the tuple $\langle \mathcal{A}, \mathcal{B}, u, \bar{u} \rangle$. In this game, $\mathcal{A}$ and $\mathcal{B}$ are the finite action sets for the strategic AI agent and human agents, respectively. The payoff functions $u : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}$ and $\bar{u} : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}$ represent the payoffs received by the AI agent and the human agent.

5.1.1 Human Agent

The human agent follows the EWA learning algorithm, which updates two variables at each stage. The first variable, N_k , represents the *experience* and is updated as follows

$$N_k = (1 - \alpha)(1 - \kappa)N_{k-1} + 1, \quad (5.1)$$

where $\alpha \in [0, 1]$, $\kappa \in [0, 1]$ and the initial experience vector $N_0 = 0$. The second variable, $Z_k(b)$ demonstrates the *attraction* of the agent for action b and is updated as

$$Z_k(b) = \frac{(1 - \alpha)N_{k-1}Z_{k-1}(b) + [\delta + (1 - \delta)\mathbb{I}_{\{b=b_k\}}] \bar{u}(\cdot, b)}{N_k}, \quad (5.2)$$

where $\mathbb{I}_{\{b=b_k\}}$ denotes the indicator function and $\delta \in [0, 1]$ is attraction controller parameter determining the weight of the updates. For example, when $\delta = 0$, only the attractions for the played actions are updated. Conversely, when $\delta = 1$, all actions are updated with equal weight. The initial attraction vector $Z_0(b)$ can be chosen arbitrarily.

Remark 9. *After some stages of the game, N_k tends to converge toward its fixed-point value $N^* = 1/(1 - (1 - \alpha)(1 - \kappa))$ provided that $(1 - \alpha)(1 - \kappa) < 1$ [51].*

We can unify the *experience* and *attraction* updates in the following form

$$Z_k(b) = (1 - \alpha)Z_{k-1}(b) + [1 - (1 - \alpha)(1 - \kappa)] [\delta + (1 - \delta)\mathbb{I}_{\{b=b_k\}}] \bar{u}(\cdot, b). \quad (5.3)$$

The human agent, given his attraction function $Z_k(b)$, responds according to the softmax function:

$$\sigma(Z_k)(b) = \frac{\exp(Z_k(b)/\tau)}{\sum_{\tilde{b} \in \mathcal{B}} \exp(Z_k(\tilde{b})/\tau)} \quad \forall b \in \mathcal{B}, \quad (5.4)$$

for some temperature parameter $\tau > 0$ controlling the exploration and $\sigma(Z_k)(b) \in$

$(0, 1]$ denotes the probability that the agent takes action $b \in \mathcal{B}$.

5.1.2 Strategic AI Agent

The goal of the strategic AI agent is to maximize the discounted sum of payoffs it can collect over the infinite horizon, i.e.,

$$\max_{\{a_k\}_{k=0}^{\infty}} \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k u(a_k, b_k) \right], \quad (5.5)$$

where the expectation is taken with respect to the randomness on the action a_k and $b_k \sim \sigma(Z_k)$, and $\gamma \in (0, 1)$ is some discount factor.

The strategic AI agent, equipped with knowledge of the underlying game structure and the human agent's behavior evolving according to EWA, utilizes this information to solve (5.5). Specifically, the strategic AI agent observes the game history $\{a_0, b_0, \dots, a_{k-1}, b_{k-1}\}$ and uses an internal attraction function tracker to monitor the human agent's attraction function. It can then reformulate the problem as a *fully observable* MDP, where the state space encompasses all possible Z -vectors from (5.3). For strategic AI agent the objective (5.5) becomes to find the best policy for MDP $\mathcal{M}$ characterized by the tuple $\langle Z, \mathcal{A}, r, p, \gamma \rangle$ where $Z \subset \mathbb{R}^{|\mathcal{B}|}$ is the compact set of states, $\mathcal{A}$ is the action set of actions as in $\mathcal{G}$, the reward function $r : Z \times \mathcal{A} \rightarrow \mathbb{R}$ is given by $r(z, a) = \mathbb{E}[u(a, b)]$ that $b \sim \sigma(Z)$, $p(\cdot|\cdot)$ is the transition kernel for Z -vectors evolving according to (5.3) and $\gamma \in (0, 1)$ is discount factor. Therefore, we can rewrite the goal of the strategic AI agent as follows

$$\max_{\{a_k\}_{k=0}^{\infty}} \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r(z_k, a_k) \right], \quad (5.6)$$

where the expectation is taken with respect to randomness on (z_k, a_k) .

Remark 10. *Since the set of all possible Z -vectors is a Polish space, being a compact subset of $\mathbb{R}^{|\mathcal{B}|}$, and the action set $\mathcal{A}$ is finite and state-invariant, there*

exists an optimal stationary policy $\pi : Z \times \mathcal{A} \rightarrow [0, 1]$ for $\mathcal{M}$ [90, Theorem 6.2.12].

Although there exists an optimal stationary policy, finding this policy for a strategic AI agent is challenging due to the continuum of the state space. For instance, when the EWA parameters are set to $\alpha > 0$, $\tau > 0$, $\delta = 1$, and $\kappa = 0$, the human agent follows smoothed fictitious play, and the Z-vector becomes a probability measure for the belief of the opponent's strategy, i.e., $Z = \prod_{|\mathcal{B}|} [0, 1]$. Conversely, when $\delta = 0$, the human agent adopts reinforcement learning dynamics. For a specific type of algorithm, such as Q-learning, the Z-vector takes values within a continuum interval, i.e., $Z = \prod_{b \in \mathcal{B}} [Q_{\min}^b, Q_{\max}^b]$. Due to this continuum state space, the strategic AI agent cannot directly use dynamic programming methods such as value or policy iteration. These methods rely on Bellman equations that involve taking the maximum or expectation over the next states, which becomes intractable with infinitely many possible Z-vectors. Therefore, the strategic AI agent needs to approximate the state space to a finite set.

To approximate the MDP with a finite state space, the strategic AI agent can employ various methods from the literature. For example, it can leverage the linear programming solution of MDPs by selecting a basis to reduce the dimension of states, thus reducing the state space to a finite one [37, 38]. Similarly, it can use approximate value iteration or policy iteration that utilizes similar preselected basis function methods [91, 92]. Furthermore, it can employ any universal approximating architecture for the value function, as in neuro-dynamic programming, to find the optimal policy [42].

5.1.3 Approximation of the MDP

In this chapter, we propose a quantization-based approximation to tackle the continuum state space. Consider a quantization mapping $\Phi : Z \rightarrow \tilde{Z}$, where $\tilde{Z} \subset Z$. The set $Z_{\tilde{z}} := \{z \in Z : \Phi(z) = \tilde{z}\}$ satisfies $Z_{\tilde{z}} \cap Z_{\tilde{z}'} = \emptyset$ for all $\tilde{z} \neq \tilde{z}'$, and $\bigcup_{\tilde{z} \in \tilde{Z}} Z_{\tilde{z}} = Z$.

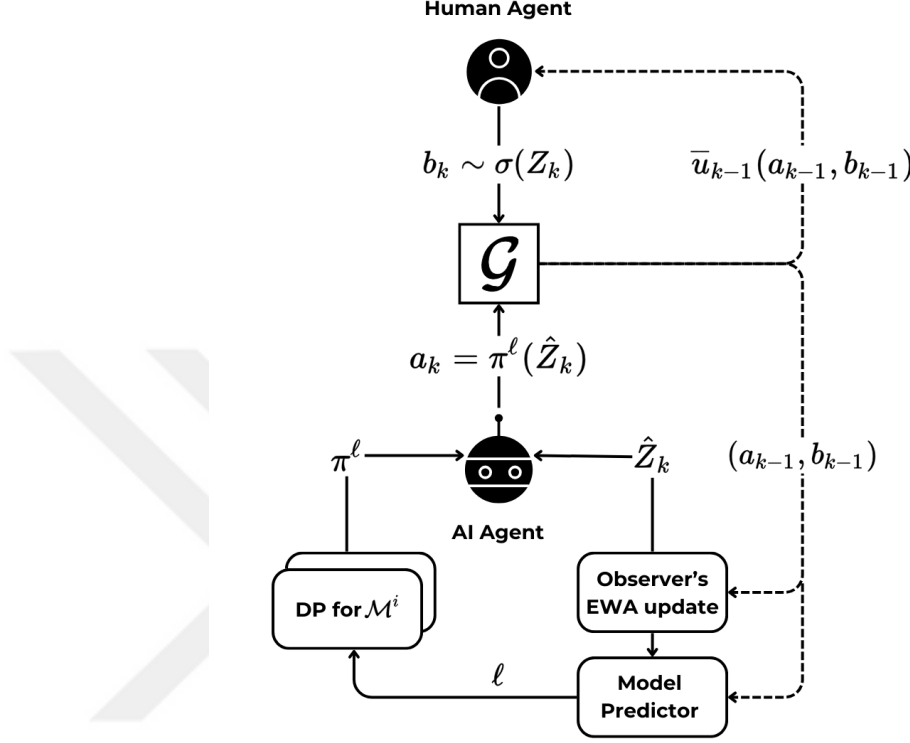


Figure 5.2: The illustration demonstrates how the strategic AI agent and the human agent take actions during the repeated play of $\mathcal{G}$. The dashed lines indicate observations from the previous stage.

The strategic AI agent can use this quantization-based approach to approximate $\mathcal{M}$ as $\widetilde{\mathcal{M}}$, characterized by the tuple $\langle \widetilde{Z}, \mathcal{A}, r_{\widetilde{Z}}, \tilde{p}, \gamma \rangle$. Here, $\widetilde{Z}$ is the finite range space of the quantizer Φ and the state space of the approximated MDP $\widetilde{\mathcal{M}}$. $\mathcal{A}$ is the set of finite actions as described in $\mathcal{G}$. The reward function $r_{\widetilde{Z}} : \widetilde{Z} \times \mathcal{A} \rightarrow \mathbb{R}$ is given by $r_{\widetilde{Z}}(\tilde{z}, a) = \mathbb{E}[u(a, b)]$ where $b \sim \sigma(\tilde{Z})$. The transition probabilities are defined by

$$\tilde{p}(\tilde{z}_+ | \tilde{z}, a) = \int_{Z_{\tilde{z}_+}} p(dz_+ | \tilde{z}, a). \quad (5.7)$$

We can rewrite the goal of the strategic AI agent in finite state space approximation as

$$\max_{\{a_k\}_{k=0}^{\infty}} \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r_{\widetilde{Z}}(\tilde{z}_k, a_k) \right], \quad (5.8)$$

Algorithm 7 Strategic AI Agent

input: π^i and Z_0

for each stage $k = 0, 1, \dots$ **do**

observe (a_{k-1}, b_{k-1})

if $k > 0$ **then** update $\hat{Z}_{k-1}^i$ to $\hat{Z}_k^i$ according to

$$Z_k(b) = (1 - \alpha)Z_{k-1}(b) + [1 - (1 - \alpha)(1 - \kappa)] [\delta + (1 - \delta)\mathbb{I}_{\{b=b_k\}}] \bar{u}(\cdot, b)$$

if $k > 0$ **then** update λ_{k-1} according to

$$\lambda_{k+1} = \frac{\lambda_k^T \beta_k}{\|\lambda_k^T \beta_k\|}$$

find the best model type ℓ

take action $a_k \sim \pi^\ell(z_k)$

receive reward $r_k = r(z_k, a_k)$

end for

where $\tilde{z}_k = \Phi(z_k)$ is the quantized state at stage k . Hence, the strategic AI agent can adopt dynamic programming methods to obtain the optimal stationary policy.

5.1.4 Model Prediction

In real-world scenarios, the AI agent might not have complete knowledge about the specific EWA parameters the human agent employs. Here, we consider a situation where the AI agent is aware of a finite set of possible EWA types that the human agent could adopt, but is unsure of the exact one.

To address this uncertainty, the AI agent pre-computes the optimal policy, π^i for each approximate MDP $\{\tilde{\mathcal{M}}^i = \langle \tilde{Z}^i, \mathcal{A}, r_{\tilde{Z}^i}, \tilde{p}^i, \gamma \rangle\}_{i=1}^N$ modeling different EWA types. The strategic AI agent maintains a belief vector, $\lambda_k \in \Delta(N)$ at each stage k where $\Delta(N)$ represents the N -dimensional simplex. This vector represents the AI agent's current belief about the probability of each EWA type. The element λ_k^i denotes the probability assigned to the human agent having type i at stage k .

The strategic AI agent then uses its internal attraction function to track the

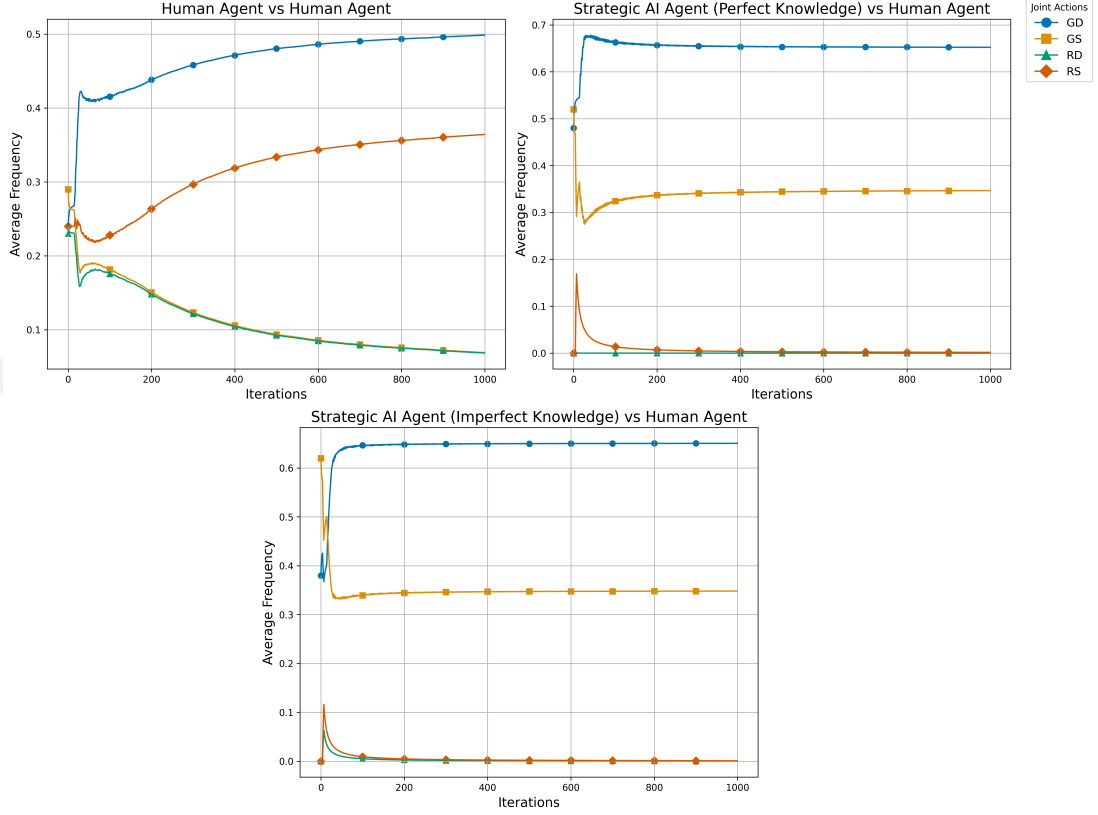


Figure 5.3: The evolution of the empirical averages of the action profiles for the human agent vs human agent and strategic AI agent vs human agent in coordination game for perfect and imperfect knowledge.

state, $\hat{Z}_k^i$, and the mixed strategy of the human agent, β_k^i , for each EWA type i at stage k . After observing the action taken by the human agent, b_k , the strategic AI agent updates its belief as follows

$$\lambda_{k+1} = \frac{\lambda_k^T \beta_k}{\|\lambda_k^T \beta_k\|}, \quad (5.9)$$

where $\beta_k \in \mathbb{R}^N$ is defined as $\beta_k := \{\beta_k^i(b_k)\}_{i=1}^N$ and $\beta_k^i(b_k)$ is the probability of taking action b_k under the mixed strategy β_k^i . Once the belief vector is updated, the AI agent selects a policy to play in the next stage. It chooses the policy that is optimal for the most likely EWA type based on the updated belief. Simply the strategic AI agent plays according to the strategy that best aligns with the human agent it believes is most probable, i.e., $\pi_k = \pi^\ell$ where $\ell = \operatorname{argmax}_j \lambda_k^j$.

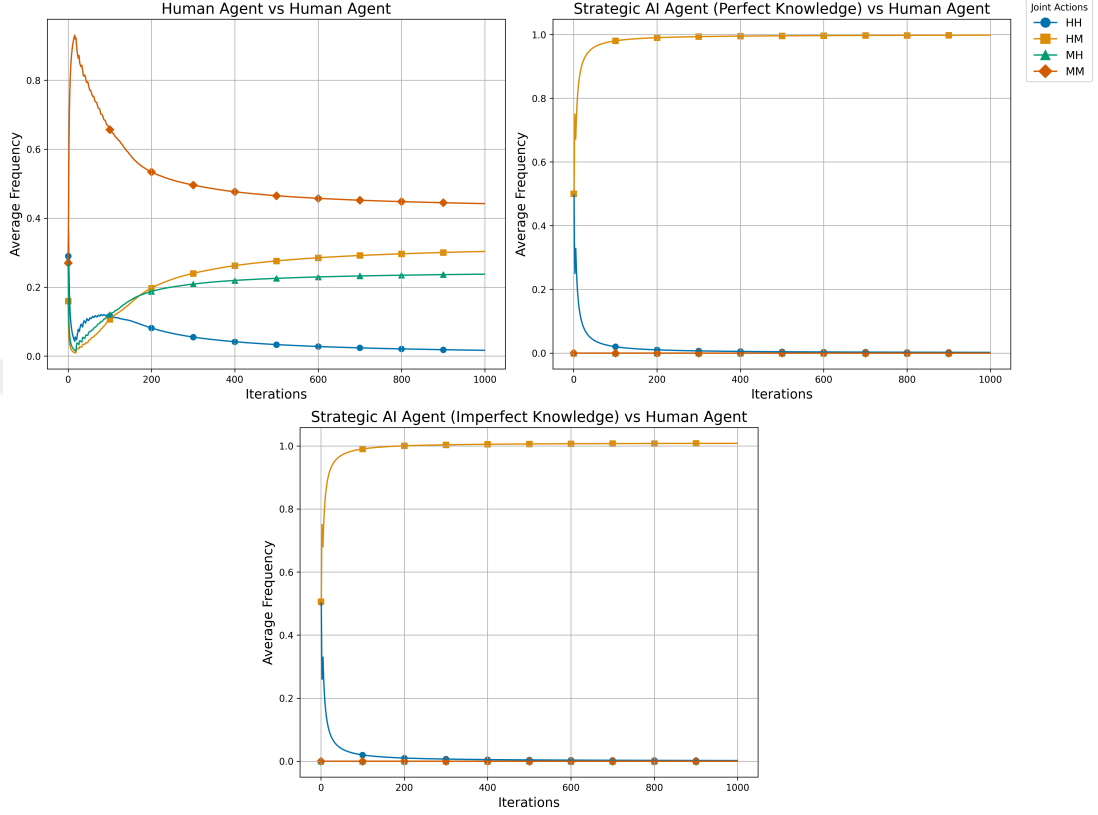


Figure 5.4: The evolution of the empirical averages of the action profiles for the human agent vs human agent and strategic AI agent vs human agent in anti-coordination game for perfect and imperfect knowledge.

5.2 Illustrative Examples

We examine the performance of a strategic AI agent in CPHS under three distinct scenarios: coordination, anti-coordination and zero-sum games. We further illustrate the case where the AI agent does not have perfect knowledge about the EWA parameters of the human agent and constructs a belief for a finite set of types. As a proof of concept, we choose two possible EWA types. To model the human agent, we set the EWA parameters to $\alpha^1 = 0.1$, $\tau^1 = 0.01$, $\kappa^1 = 0.5$, and $\delta^1 = 0.5$. We specifically choose $\kappa^1 = 0.5$ and $\delta^1 = 0.5$ to ensure that the EWA model does not align strictly with belief or reinforcement learning, thus effectively representing the human agent. For the second type of EWA we determine the parameters as $\alpha^2 = 0.1$, $\tau^2 = 0.01$, $\kappa^2 = 0.6$, and $\delta^2 = 0.4$. For the strategic AI

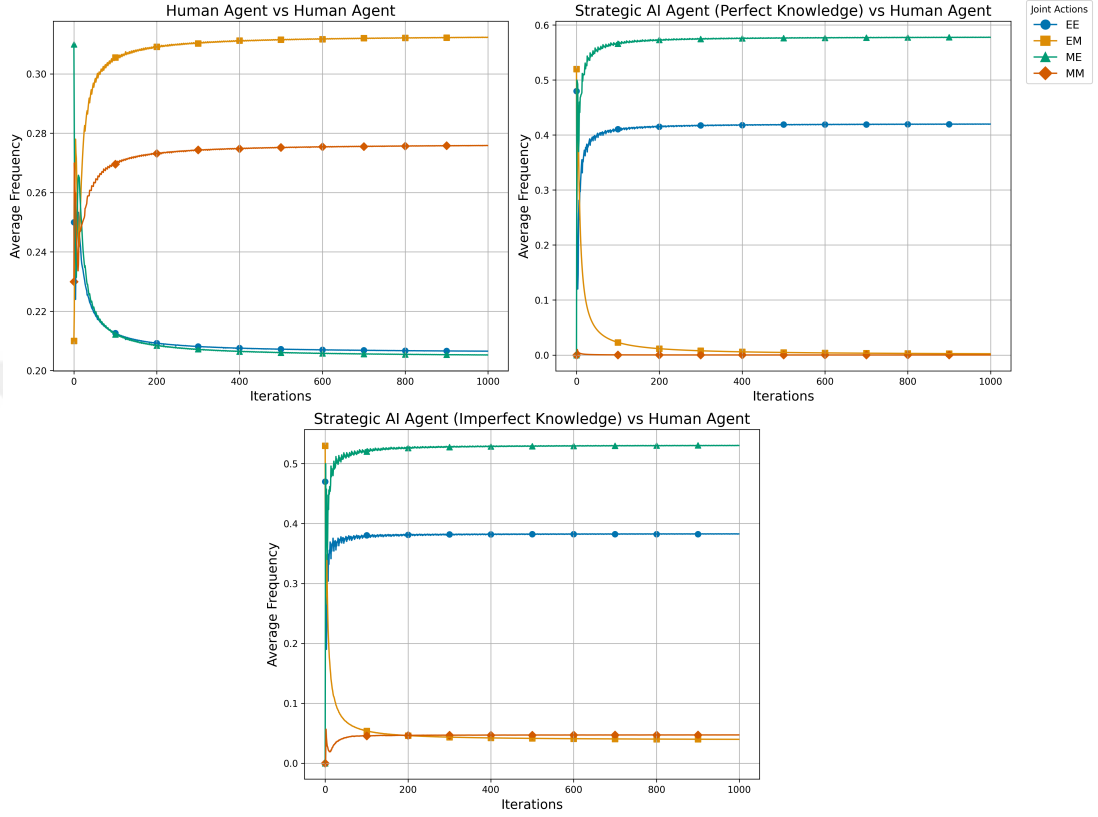


Figure 5.5: The evolution of the empirical averages of the action profiles for the human agent vs human agent and strategic AI agent vs human agent in zero-sum game for perfect and imperfect knowledge.

agent, we set $\gamma = 0.8$. Furthermore, we quantize each Z-vector to 100 uniform intervals. In all scenarios, we first illustrate the cases where both agents adopt the same EWA learning dynamics, i.e., human agent vs. human agent. After establishing these benchmarks, we demonstrate the case where the strategic AI agent employs dynamic programming with perfect and imperfect knowledge.

Coordination Game: Consider a traffic intersection scenario where an AI agent controls the traffic light and a single driver is present. The AI agent has two actions: green (G) and red (R), to control the flow of traffic, while the driver also has two actions: drive (D) or stop (S). The payoffs, as illustrated in Table 5.1, represent the effectiveness and safety of the different actions, encouraging both agents to find an optimal strategy for coordination.

		Human Agent	
		D	S
AI Agent	G	(0.8,0.8)	(0.2,0.1)
	R	(0.1,0.2)	(0.5,0.5)

Table 5.1: Coordination Game

When the agents follow EWA learning dynamics, they more frequently choose the joint action profile where the traffic light is green and the driver drives. However, when the strategic AI agent leverages its complete knowledge, this frequency increases as illustrated in Figure 5.3. Furthermore, the strategic AI drives the human agent toward a more favorable joint action where the traffic light is green but the driver stops, enhancing either traffic flow or safety. We observed an increase in the utility achieved by the strategic agent, rising from 0.3295 to 0.4387 under perfect knowledge, and to 0.3678 under imperfect knowledge. Similarly, the human agent’s utility increased from 0.3268 to 0.4081 and 0.3291, respectively.

Anti-coordination Game: Imagine a market competition scenario where a strategic AI agent and a human agent represent competing companies, each deciding on their strategies for a new product launch. These companies have two possible actions: targeting the high-end market (H), which consists of fewer customers but yields higher profits, or targeting the mass market (M), which has a larger customer base but offers lower profits. The matrix game illustrating this scenario is shown in Table 5.2.

When the agents follow EWA dynamics, they tend to target the mass market, as depicted in Figure 5.4. However, when a strategic AI agent is involved, it can manipulate the joint actions such that it targets the high-end market to achieve more favorable profits while driving the human agent to target the mass market.

		Human Agent	
		H	M
AI Agent	H	$(-1,-1)$	$(1,-0.5)$
	M	$(-0.5,1)$	$(0.5,0.5)$

Table 5.2: Anti-coordination Game

In addition, the strategic AI agent’s utility increases from 0.2097 to 0.3944 and 0.3934 in perfect and imperfect knowledge, respectively. Conversely, the human agent’s utility decreases from 0.2391 to -0.5014 and -0.5016.

Zero-sum Game: Consider a cybersecurity scenario involving a strategic AI agent (attacker) attempting to infiltrate a system and a human agent (defender) trying to protect it. The attacker has two options: sending deceptive emails to trick users into revealing sensitive information (E) or using malicious software to gain unauthorized access (M). The defender can counter these actions by strengthening email filters (E) or improving the malware detection system (M). This cybersecurity problem can be characterized as a zero-sum game, as depicted in Table 5.3.

		Human Agent	
		E	M
AI Agent	E	$(-1,1)$	$(1,-1)$
	M	$(1,-1)$	$(-1,1)$

Table 5.3: Zero-sum Game

In this game scenario, the strategic AI agent can leverage its knowledge to

increase the frequency of mismatched joint actions. This result is illustrated in Figure 5.5 where the AI agent chooses M and the human agent chooses E more frequently. Furthermore, the strategic AI agent's achieved utility increases from 0.0162 to 0.3599 and 0.2509 for the perfect knowledge and imperfect knowledge cases, respectively.



Chapter 6

Conclusion

In this thesis we examined the complex dynamics of multi-agent systems, where autonomous agents interact and make decisions within shared environments. In such systems, agents aim to maximize their cumulative rewards over time, but the interdependence of their actions complicates this objective. Unlike well-defined optimization problems, the challenges inherent in multi-agent environments necessitate adaptive strategies where agents learn and adjust their behavior based on observed outcomes. A common simplifying assumption in this domain is the stationary opponent assumption, presuming that other agents' strategies remain stationary. While this assumption has been instrumental in developing foundational theories, the existing results generally overlooked the complexities introduced by agent heterogeneity and the potential for strategic manipulation. We addressed these limitations by examining the impact of diverse agent behaviors. We primarily focused on two types of agent differences: parameter variations in independent learning dynamics under the stationary opponent assumption, and differences in agent sophistication levels. In the former case, we established the theoretical foundation to regulate the zero-sum (stochastic) games under equilibrium-play assumption if agents have similar strategic sophistication, i.e., if they naively assume stationary opponent assumption. In the latter case, we assumed that the strategic agent is aware of the stationary opponent assumption of the other agents, enabling us to characterize worst-case scenarios and establish a benchmark for

the vulnerability analysis of independent learning dynamics.

We first investigated heterogeneous learning dynamics in the context of two-agent (near) zero-sum and zero-sum stochastic games. We introduced two families of uncoupled and rational algorithms that account for diverse learning behaviors, including variations in learning rates, exploration-exploitation trade-offs, information availability (belief-based vs. payoff-based), and model usage (model-based vs. model-free). To establish convergence, we developed a novel Lyapunov function that unifies and extends existing results in the literature by encompassing near zero-sum games as the stage games in stochastic games are not strictly zero-sum. Our analysis demonstrated that any combination of these algorithms converges to equilibrium (or near-equilibrium) under standard assumptions on step sizes and the transition kernel, with additional requirements for discount factor in stochastic games.

We then investigated how strategically sophisticated agents can exploit naive opponents employing independent learning algorithms (specifically IQL) in repeatedly played normal-form games. Assuming knowledge of the opponent’s learning algorithm and game payoffs, we modeled the strategic agents’ interactions as a stochastic game where the independent learning process constitutes the underlying dynamic environment. This formulation resulted in a stochastic game with a continuous state space. To address this challenge, we established the Lipschitz continuity of the value function with respect to the Q-learning states, enabling the ground for effective approximation methods. As a proof of concept, we employed quantization to discretize the state space, transforming the problem into a finite stochastic game solvable by minimax value iteration for two competing strategic agents. We analytically characterized the approximation error and numerically assessed the performance of the quantization scheme.

To explore the applicability of our framework to more complex scenarios, we extended our analysis to stochastic security games, a more realistic representation of real-world security challenges. In this setting, we considered a two-agent interaction involving a naive defender employing an RL algorithm and a strategic attacker with complete knowledge of the defender’s strategy and the game

environment. This full-knowledge assumption represents the most powerful adversary a defender could face. By modeling the attacker’s problem as an MDP, we encountered a high-dimensional continuous state space. To address this challenge, we leveraged neuro-dynamic programming techniques, enabling efficient and scalable analysis of the attacker’s optimal strategy.

Building upon our study of strategic interactions in complex environments, we explored human-AI interactions. By modeling humans as naive learners using the EWA framework, we investigated the potential for AI influence and manipulation. We formulated the AI agent’s decision-making process as an MDP where the human’s attraction vector is the state space. To address the complexities of continuous state spaces and imperfect knowledge about human behavior, we employed quantization-based approximation and a model prediction framework. Our findings highlighted the capacity of AI agents to influence human behavior, even when operating under uncertainties about human decision-making processes.

Based on the outcomes of this thesis, several directions for future exploration are apparent. Investigating the impact of heterogeneities in other game classes, such as potential games, could provide valuable insights into the broader implications of agent diversity. Enhancing the robustness of learning algorithms against strategic adversaries is crucial for practical applications. Moreover, exploring scenarios with less sophisticated strategic agents by reducing their information availability could offer a more nuanced understanding of real-world interactions. Finally, developing advanced modeling techniques for less sophisticated agents, such as incorporating high-level model prediction frameworks, holds the potential to significantly improve the accuracy and effectiveness of our analysis.

Bibliography

- [1] J. Asker, C. Fershtman, and A. Pakes, “Artificial intelligence, algorithm design, and pricing,” in *AEA Papers and Proceedings*, vol. 112, pp. 452–56, 2022.
- [2] M. Banchio and G. Mantegazza, “Adaptive algorithms and collusion via coupling,” *arXiv preprint arXiv:2202.05946*, 2022.
- [3] Q. Bertrand, J. Duque, E. Calvano, and G. Gidel, “Q-learners can provably collude in the iterated prisoner’s dilemma,” *arXiv preprint arXiv:2312.08484*, 2023.
- [4] E. Calvano, G. Calzolari, V. Denicolò, and S. Pastorello, “Algorithmic pricing what implications for competition policy?,” *Review of Industrial Organization*, vol. 55, pp. 155–171, 2019.
- [5] E. Calvano, G. Calzolari, V. Denicolo, and S. Pastorello, “Artificial intelligence, algorithmic pricing, and collusion,” *American Economic Review*, vol. 110, no. 10, pp. 3267–3297, 2020.
- [6] B. Singh, R. Kumar, and V. P. Singh, “Reinforcement learning in robotic applications: a comprehensive survey,” *Artificial Intelligence Review*, vol. 55, no. 2, pp. 945–990, 2022.
- [7] J. Liu and J. Wu, *Multiagent Robotic Systems*. CRC press, 2018.

- [8] S. H. Semnani, H. Liu, M. Everett, A. De Ruiter, and J. P. How, “Multi-agent motion planning for dense and dynamic environments via deep reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3221–3226, 2020.
- [9] X. Wang, L. Ke, Z. Qiao, and X. Chai, “Large-scale traffic signal control using a novel multiagent reinforcement learning,” *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 174–187, 2020.
- [10] T. Chu, J. Wang, L. Codecà, and Z. Li, “Multi-agent deep reinforcement learning for large-scale traffic signal control,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1086–1095, 2019.
- [11] T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang, and D. O. Wu, “Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8243–8256, 2020.
- [12] Y.-W. Cheung and D. Friedman, “Individual learning in normal form games: Some laboratory results,” *Games and Economic Behavior*, vol. 19, pp. 46–76, 1997.
- [13] C. Giannitsarou, “Heterogeneous learning,” *Review of Economic Dynamics*, vol. 6, no. 4, pp. 885–906, 2003.
- [14] M. Berardi, “Heterogeneous learning dynamics and speed of convergence,” *Studies in Nonlinear Dynamics & Econometrics*, vol. 16, no. 4, 2012.
- [15] M. Berardi, “Heterogeneity and misspecifications in learning,” *Journal of Economic Dynamics and Control*, vol. 31, no. 10, pp. 3203–3227, 2007.
- [16] P. Liu, F. Xiao, B. Wei, and A. Wang, “Distributed constrained optimization problem of heterogeneous linear multi-agent systems with communication delays,” *Systems & Control Letters*, vol. 155, p. 105002, 2021.
- [17] X. Niu and E. Wei, “DISH: A distributed hybrid primal-dual optimization framework to utilize system heterogeneity,” in *61st IEEE Conference on Decision and Control (CDC)*, pp. 6503–6510, 2022.

- [18] R. Grabowski, L. E. Navarro-Serment, C. J. J. Paredis, and P. K. Khosla, “Heterogeneous teams of modular robots for mapping and exploration,” *Autonomous Robots*, vol. 8, pp. 293–308, 2000.
- [19] L. Parker, B. Kannan, F. Tang, and M. Bailey, “Tightly-coupled navigation assistance in heterogeneous multi-robot teams,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 1, pp. 1016–1022 vol.1, 2004.
- [20] Z. Qu, J. Wang, and R. A. Hull, “Cooperative control of dynamical systems with application to autonomous vehicles,” *IEEE Transactions on Automatic Control*, vol. 53, no. 4, pp. 894–911, 2008.
- [21] N. Mathew, S. L. Smith, and S. L. Waslander, “Planning paths for package delivery in heterogeneous multirobot teams,” *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 4, pp. 1298–1308, 2015.
- [22] I. Spasojevic, X. Liu, A. Prabhu, A. Ribeiro, G. J. Pappas, and V. Kumar, “Robust localization of aerial vehicles via active control of identical ground vehicles,” *arXiv preprint arXiv:2308.06658*, 2023.
- [23] D. Fudenberg and D. K. Levine, “Learning and equilibrium,” *The Annual Review of Economics*, vol. 1, pp. 385–419, 2009.
- [24] G. Brown, “Iterated solution of games by fictitious play,” in *Activity Analysis of Production and Allocation* (T. Koopmans, ed.), pp. 377–380, New York: Wiley, 1951.
- [25] D. Fudenberg and D. M. Kreps, “Learning mixed equilibria,” *Games and Economic Behavior*, vol. 5, pp. 320–367, 1993.
- [26] D. S. Leslie and E. J. Collins, “Individual Q-learning in normal form games,” *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 495–514, 2005.
- [27] M. Bowling and M. Veloso, “Multiagent learning using a variable learning rate,” *Artificial Intelligence*, vol. 136, no. 2, pp. 215–250, 2002.

- [28] M. Benaïm, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions,” *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.
- [29] S. Perkins and D. S. Leslie, “Asynchronous stochastic approximation with differential inclusions,” *Stochastic Systems*, vol. 2, no. 2, pp. 409–446, 2013.
- [30] M. Tan, “Multi-agent reinforcement learning: Independent vs. cooperative agents,” in *International Conference on Machine Learning (ICML)*, pp. 330–337, 1993.
- [31] A. A. Hussain, F. Belardinelli, and G. Piliouras, “Asymptotic convergence and performance of multi-agent Q-learning dynamics,” *arXiv preprint arXiv:2301.09619*, 2023.
- [32] M. Wunder, M. L. Littman, and M. Babes, “Classes of multiagent Q-learning dynamics with epsilon-greedy exploration,” in *International Conference on Machine Learning (ICML)*, pp. 1167–1174, 2010.
- [33] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, “Multiagent cooperation and competition with deep reinforcement learning,” *PloS One*, vol. 12, no. 4, p. e0172395, 2017.
- [34] N. Saldi, T. Linder, and S. Yüksel, *Finite Approximations in Discrete-time Stochastic Control*. Springer, 2018.
- [35] Z. Ren and B. H. Krogh, “State aggregation in Markov decision processes,” in *41st IEEE Conference on Decision and Control (CDC)*, vol. 4, pp. 3819–3824, IEEE, 2002.
- [36] R. Ortner, “Pseudometrics for state aggregation in average reward Markov decision processes,” in *Algorithmic Learning Theory: 18th International Conference*, pp. 373–387, Springer, 2007.
- [37] D. P. De Farias and B. Van Roy, “The linear programming approach to approximate dynamic programming,” *Operations Research*, vol. 51, no. 6, pp. 850–865, 2003.

- [38] D. P. De Farias and B. Van Roy, “On constraint sampling in the linear programming approach to approximate dynamic programming,” *Mathematics of Operations Research*, vol. 29, no. 3, pp. 462–478, 2004.
- [39] T. Klein, “Autonomous algorithmic collusion: Q-learning under sequential pricing,” *The RAND Journal of Economics*, vol. 52, no. 3, pp. 538–558, 2021.
- [40] K. T. Hansen, K. Misra, and M. M. Pai, “Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms,” *Marketing Science*, vol. 40, no. 1, pp. 1–12, 2021.
- [41] L. S. Shapley, “Stochastic games,” *Proceedings of the National Academy of Sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.
- [42] D. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic Programming*. Athena Scientific, 1996.
- [43] T. Alpcan and T. Başar, *Network Security: A Decision and Game-theoretic Approach*. Cambridge University Press, 2010.
- [44] R. A. Miura-Ko, B. Yolken, N. Bambos, and J. Mitchell, “Security investment games of interdependent organizations,” in *46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 252–260, IEEE, 2008.
- [45] R. A. Miura-Ko, B. Yolken, J. Mitchell, and N. Bambos, “Security decision-making among interdependent organizations,” in *21st IEEE Computer Security Foundations Symposium*, pp. 66–80, IEEE, 2008.
- [46] K. C. Nguyen, T. Alpcan, and T. Basar, “Stochastic games for security in networks with interdependent nodes,” in *International Conference on Game Theory for Networks*, pp. 697–703, IEEE, 2009.
- [47] J. Shi, J. Wan, H. Yan, and H. Suo, “A survey of cyber-physical systems,” in *International Conference on Wireless Communications and Signal Processing*, pp. 1–6, IEEE, 2011.

- [48] A. Humayed, J. Lin, F. Li, and B. Luo, “Cyber-physical systems security—a survey,” *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802–1831, 2017.
- [49] C. Camerer and T. Hua Ho, “Experience-weighted attraction learning in normal form games,” *Econometrica*, vol. 67, no. 4, pp. 827–874, 1999.
- [50] C. Gracia-Lázaro, A. Ferrer, G. Ruiz, A. Tarancón, J. A. Cuesta, A. Sánchez, and Y. Moreno, “Heterogeneous networks do not promote cooperation when humans play a prisoner’s dilemma,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 32, pp. 12922–12926, 2012.
- [51] M. Pangallo, J. B. Sanders, T. Galla, and J. D. Farmer, “Towards a taxonomy of learning dynamics in 2×2 games,” *Games and Economic Behavior*, vol. 132, pp. 1–21, 2022.
- [52] G. Arslan and S. Yüksel, “Decentralized Q-learning in stochastic teams and games,” *IEEE Transactions on Automatic Control*, vol. 62, no. 4, 2017.
- [53] C. Daskalakis, D. J. Foster, and N. Golowich, “Independent policy gradient methods for competitive reinforcement learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [54] D. S. Leslie, S. Perkins, and Z. Xu, “Best-response dynamics in zero-sum stochastic games,” *Journal of Economic Theory*, vol. 189, 2020.
- [55] M. O. Sayin, F. Parise, and A. Ozdaglar, “Fictitious play in zero-sum stochastic games,” *SIAM Journal on Control and Optimization*, vol. 60, no. 4, 2022.
- [56] M. O. Sayin, K. Zhang, D. S. Leslie, A. Ozdaglar, and T. Başar, “Decentralized Q-learning in zero-sum Markov games,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [57] M. O. Sayin, K. Zhang, and A. Ozdaglar, “Fictitious play in Markov games with single controller,” in *ACM Conference of Economics and Computation (EC)*, 2022.

- [58] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, “Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games,” in *34th Annual Conference on Learning Theory*, vol. 134, pp. 1–41, 2021.
- [59] L. Baudin and R. Laraki, “Best-response dynamics and fictitious play in identical interest stochastic games,” in *International Conference on Machine Learning (ICML)*, 2022.
- [60] L. Baudin and R. Laraki, “Smooth fictitious play in stochastic games with perturbed payoffs and unknown transitions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 20243–20256, 2022.
- [61] C. Camerer and T.-H. Ho, “Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation,” *Journal of Mathematical Psychology*, vol. 42, no. 2-3, pp. 305–326, 1998.
- [62] S. Kapetanakis and D. Kudenko, “Reinforcement learning of coordination in heterogeneous cooperative multi-agent systems,” in *Adaptive Agents and Multi-Agent Systems II* (D. Kudenko, D. Kazakov, and E. Alonso, eds.), pp. 119–131, Springer Berlin Heidelberg, 2005.
- [63] H. Zheng, P. Wei, J. Jiang, G. Long, Q. Lu, and C. Zhang, “Cooperative heterogeneous deep reinforcement learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 17455–17465, 2020.
- [64] H. Zheng, J. Jiang, P. Wei, G. Long, and C. Zhang, “Competitive and cooperative heterogeneous deep reinforcement learning,” in *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020.
- [65] Y. Zhong, J. G. Kuba, X. Feng, S. Hu, J. Ji, and Y. Yang, “Heterogeneous-agent reinforcement learning,” *Journal of Machine Learning Research*, vol. 25, pp. 1–67, 2024.
- [66] J. A. Calvo and I. Dusparic, “Heterogeneous multi-agent deep reinforcement learning for traffic lights control,” in *AICS*, pp. 2–13, 2018.

- [67] C. Wakilpoor, P. J. Martin, C. Rebhuhn, and A. Vu, “Heterogeneous multi-agent reinforcement learning for unknown environment mapping,” *arXiv preprint arXiv:2010.02663*, 2020.
- [68] C. Maheshwari, M. Wu, D. Pai, and S. Sastry, “Independent and decentralized learning in Markov potential games,” *arXiv preprint arXiv:2205.14590*, 2022.
- [69] B. Yongacoglu, G. Arslan, and S. Yüksel, “Asynchronous decentralized Q-learning: Two timescale analysis by persistence,” *arXiv preprint arXiv:2308.03239*, 2023.
- [70] Z. Wang, X. Yi, Y. Shen, M. M. Zavlanos, and K. H. Johansson, “Asymmetric feedback learning in online convex games,” *arXiv preprint arXiv:2307.08812*, 2023.
- [71] D. S. Leslie and E. J. Collins, “Convergent multiple-timescales reinforcement learning algorithms in normal-form games,” *The Annals of Applied Probability*, vol. 13, no. 4, pp. 1231–1251, 2003.
- [72] B. Chasnov, L. Ratliff, E. Mazumdar, and S. Burden, “Convergence analysis of gradient-based learning in continuous games,” in *35th Uncertainty in Artificial Intelligence Conference*, vol. 115, pp. 935–944, PMLR, 22–25 Jul 2020.
- [73] M. O. Sayin and K. Cetiner, “On the heterogeneity of independent learning dynamics in zero-sum stochastic games,” in *Learning for Dynamics and Control Conference*, pp. 994–1005, PMLR, 2022.
- [74] M. O. Sayin, F. Parise, and A. Ozdaglar, “Fictitious play in zero-sum stochastic games [Extended version],” *arXiv:2010.04223*, 2022.
- [75] R. McKelvey and T. Palfrey, “Quantal response equilibria for normal form games,” *Games and Economic Behavior*, vol. 10, pp. 6–38, 1995.
- [76] C. Harris, “On the rate of convergence of continuous-time fictitious play,” *Games and Economic Behavior*, vol. 22, pp. 238–259, 1998.

- [77] J. Hofbauer and W. H. Sandholm, “On the global convergence of stochastic fictitious play,” *Econometrica*, vol. 70, pp. 2265–2294, 2002.
- [78] M. L. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in *International Conference on Machine Learning (ICML)*, 1994.
- [79] J. Hu and M. P. Wellman, “Nash Q-learning for general-sum stochastic games,” *Journal of Machine Learning Research*, pp. 1039–1069, 2003.
- [80] B. Gao and L. Pavel, “On the properties of the softmax function with application in game theory and reinforcement learning,” *arXiv preprint arXiv:1704.00805*, 2017.
- [81] J. Hofbauer and E. Hopkins, “Learning in perturbed asymmetric games,” *Games and Economic Behavior*, vol. 52, pp. 133–152, 2005.
- [82] Y. Huang and Q. Zhu, “Deceptive reinforcement learning under adversarial manipulations on cost signals,” in *International Conference on Decision and Game Theory for Security* (T. Alpcan, Y. Vorobeychik, J. Baras, and G. Dan, eds.), vol. 11836 of *Lecture Notes in Computer Science*, Springer, Cham., 2019.
- [83] Y. Huang and Q. Zhu, “Manipulating reinforcement learning: Stealthy attacks on cost signals,” in *Game Theory and Machine Learning for Cyber Security*, pp. 367–388, John Wiley & Sons, 2021.
- [84] Z. Deng and S. Liang, “Distributed algorithms for aggregative games of multiple heterogeneous Euler–Lagrange systems,” *Automatica*, vol. 99, pp. 246–252, 2019.
- [85] B. Vundurthy, A. Kanellopoulos, V. Gupta, and K. G. Vamvoudakis, “Intelligent players in a fictitious play framework,” *IEEE Transactions on Automatic Control*, 2023.
- [86] H. Dong and Y. Mu, “The optimal strategy against fictitious play in infinitely repeated games,” in *41st Chinese Control Conference (CCC)*, 2022.

- [87] D. Ding, C. Wei, K. Zhang, and M. Jovanovic, “Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence,” *arXiv:2202.04129*, 2022.
- [88] X. Zhang, Y. Ma, A. Singla, and X. Zhu, “Adaptive reward-poisoning attacks against reinforcement learning,” in *International Conference on Machine Learning (ICML)*, pp. 11225–11234, 2020.
- [89] Y. Lei, D. Ye, S. Shen, Y. Sui, T. Zhu, and W. Zhou, “New challenges in reinforcement learning: a survey of security and privacy,” *Artificial Intelligence Review*, vol. 56, no. 7, pp. 7195–7236, 2023.
- [90] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.
- [91] J. N. Tsitsiklis and B. Van Roy, “Feature-based methods for large scale dynamic programming,” *Machine Learning*, vol. 22, no. 1, pp. 59–94, 1996.
- [92] A.-m. Farahmand, C. Szepesvári, and R. Munos, “Error propagation for approximate policy and value iteration,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 23, 2010.
- [93] C. Glenn, D. Sterbentz, and A. Wright, “Cyber threat and vulnerability analysis of the us electric sector,” tech. rep., Idaho National Lab.(INL), Idaho Falls, ID (United States), 2016.
- [94] T. Kovanen, V. Nuojua, and M. Lehto, “Cyber threat landscape in energy sector,” in *13th International Conference on Cyber Warfare and Security (ICCWS)*, p. 353, Academic Conferences and publishing limited, 2018.
- [95] M. Lehto, “Cyber-attacks against critical infrastructure,” in *Cyber Security: Critical Infrastructure Protection*, pp. 3–42, Springer, 2022.
- [96] A. M. K. Adawadkar and N. Kulkarni, “Cyber-security and reinforcement learning—a brief survey,” *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105116, 2022.
- [97] C. Li and M. Qiu, *Reinforcement Learning for Cyber-physical Systems: with Cybersecurity Case Studies*. Chapman and Hall/CRC, 2019.

- [98] W. Mazurczyk and L. Caviglione, “Cyber reconnaissance techniques,” *Communications of the ACM*, vol. 64, no. 3, pp. 86–95, 2021.
- [99] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, “Game theory meets network security and privacy,” *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 1–39, 2013.
- [100] S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu, “A survey of game theory as applied to network security,” in *43rd Hawaii International Conference on System Sciences*, pp. 1–10, IEEE, 2010.
- [101] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, 2nd ed., 2018.
- [102] C. J. C. H. Watkins and P. Dayan, “Q-learning,” *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.
- [103] K. Asadi, D. Misra, and M. Littman, “Lipschitz continuity in model-based reinforcement learning,” in *International Conference on Machine Learning (ICML)*, pp. 264–273, 2018.
- [104] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare, “DeepMDP: Learning continuous latent space models for representation learning,” in *International Conference on Machine Learning (ICML)*, pp. 2170–2179, PMLR, 2019.
- [105] J. N. Tsitsiklis, “Asynchronous stochastic approximation and Q-learning,” *Machine Learning*, vol. 16, pp. 185–202, 1994.

Appendix A

Preliminary Information for Stochastic Approximation

In the following, we provide preliminary information on stochastic approximation methods for completeness.

A.1 Stochastic Differential Inclusion Approximation

The following is a refinement of the results in [28, Proposition 1.3 and Theorem 3.6] to compact domains. Consider a sequence $\{x_k \in X\}_{k \geq 0}$, where X is a compact subset of $\mathbb{R}^m$, evolving according to

$$x_{k+1} - x_k - \alpha_k(e_k + \omega_k) \in \alpha_k F(x_k) \tag{A.1}$$

where

- The step size decay, i.e., $\alpha_k \rightarrow 0$ at a rate such that $\sum_{k=0}^{\infty} \alpha_k = \infty$ (and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ if we have non-zero noise ω_k).

- The set-valued function $F(x) \subset \mathbb{R}^m$ is a Marchaud map, i.e., satisfying the following conditions:
 - $F(\cdot)$ is upper semi-continuous, or equivalently, $\text{Graph}(F) = \{(x, y) : x \in X, y \in F(x)\}$ is a closed subset of $X \times \mathbb{R}^m$.
 - For all $x \in X$, $F(x)$ is a non-empty, compact, and convex subset of X .
 - There exists a $c > 0$ such that for all $x \in X$, we have $\sup_{y \in F(x)} \|y\| \leq c(1 + \|x\|)$.
- The error term $\|e_k\| \rightarrow 0$ with probability 1, as in [29, Theorem 2.3].
- With respect to the filtration $\mathcal{F}_k = \sigma(x_0, \omega_0, \dots, \omega_{k-1})$, the noise term ω_k is $\mathcal{F}_{k+1}$ -measurable and satisfies $\mathbb{E}[\omega_k | \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\omega_k\|^2 | \mathcal{F}_k] < W$ for some constant W for all $k \geq 0$ with probability 1.

Then, x_k almost surely converges to a compact connected internally chain transitive set of the differential inclusion

$$\dot{x} \in F(x). \quad (\text{A.2})$$

Hence, we can characterize the limit behavior of (A.1) through (A.2) if there exists a continuous function $V : X \rightarrow \mathbb{R}$, called *Lyapunov function* for $\Lambda \subset X$, such that for any absolutely continuous solution $x(t)$ to (A.2), we have

- $V(x(t')) < V(x(t))$ for all $t' > t$ and $x(t) \notin \Lambda$
- $V(x(t')) \leq V(x(t))$ for all $t' > t$ and $x(t) \in \Lambda$

and $V(\Lambda) = \{V(x) : x \in \Lambda\} \subset \mathbb{R}$ has empty interior, then every internally chain transitive set of (A.2), and therefore, the limit set of (A.1) are contained in Λ [28, Proposition 3.7].

A.2 Asynchronous Updates

The following result is adopted from Theorem 5.1 in the extended version of [55] (built upon [105, Theorem 1]) without the stochastic approximation noise term.¹

Lemma 5. *Consider a sequence of vectors $\{y_k\}_{k \geq 0}$ whose l th entry, denoted by $y_k(l)$, satisfies the following lower bound*

$$\begin{aligned} y_{k+1}(n) &\leq (1 - \beta_k(n))y_k(n) + \beta_k(n)(\gamma\|y_k\|_\infty - \bar{e}_k(n)) \\ y_{k+1}(n) &\geq (1 - \beta_k(n))y_k(n) + \beta_k(n)(-\gamma\|y_k\|_\infty - \underline{e}_k(n)), \end{aligned}$$

where $\gamma \in (0, 1)$, the step size $\beta_k(n) \in [0, 1]$ decay to zero sufficiently slowly such that $\sum_{k=0}^{\infty} \beta_k(n) = \infty$, and the errors

$$\limsup_{k \rightarrow \infty} |\bar{e}_k(n)| \leq c \quad \text{and} \quad \limsup_{k \rightarrow \infty} |\underline{e}_k(n)| \leq c$$

for some $c \geq 0$ almost surely for each n . Suppose that $\|y_k\|_\infty \geq M$ for some $M \geq 0$ for all k and n . Then, we have

$$\limsup_{k \rightarrow \infty} \|y_k\|_\infty \leq \frac{c}{1 - \gamma}$$

almost surely.

In Lemma 5, the iterates satisfy upper and lower bounds. The following lemma addresses the case where we only have the lower bound and it is a modification of [73, Lemma 9] to address errors that do not decay to zero.

Lemma 6. *Consider a sequence of vectors $\{y_k\}_{k \geq 0}$ whose n th entry, denoted by $y_k(n)$, satisfies the following lower bound*

$$y_{k+1}(n) \geq (1 - \beta_k(n))y_k(n) + \beta_k(n)(\gamma \min_m y_k(m) - e_k(n)),$$

where $\gamma \in (0, 1)$, the step size $\beta_k(n) \in [0, 1]$ decay to zero sufficiently slowly such

¹The extended version of [55] is available at [ArXiv:2010.04223](https://arxiv.org/abs/2010.04223).

that $\sum_{k=0}^{\infty} \beta_k(n) = \infty$, and the error $e_k(n) \rightarrow c$ for some $c \geq 0$ as $k \rightarrow \infty$ almost surely for each n . Suppose that $y_k(n) \geq -M$ for some $M \geq 0$ for all k and n . Then, we have

$$\liminf_{k \rightarrow \infty} y_k(n) \geq -\frac{c}{1-\gamma}$$

almost surely for each n .

Proof. Define the sequence $\{M^t < 0\}_{t=0}^{\infty}$ over a separate timescale by $M^{t+1} = (\gamma + 2\epsilon)M^t$ for all t , and $M^0 = M$. We set $\epsilon \in (0, (1-\gamma)/2)$ so that $M^t \rightarrow 0$ monotonically from above as $t \rightarrow \infty$. We claim that for each $t = 0, 1, \dots$, there exists $k^t \in \mathbb{N}$ such that

$$y_k(n) \geq -M^t - \frac{c}{1-\gamma} \quad \text{and} \quad -e_k(n) \geq -\epsilon M^t - c \quad (\text{A.3})$$

for all $k \geq k^t$ and for each n . Since $y_k(n) \geq M^0 = M$ for all k and n , and $e_k(n) \rightarrow c$, there exists such k^0 . Suppose that (A.3) holds for some t . Specific to t , define an auxiliary sequence $\{Y_k^t\}_{k \geq k^t}$ by

$$Y_{k+1}^t(n) = (1 - \beta_k(n))Y_k^t(n) + \beta_k(n) \left(-(\gamma + \epsilon)M^t - \frac{c}{1-\gamma} \right)$$

and $Y_{k^t}^t(n) = -M^t - \frac{c}{1-\gamma}$ for each n . Note that

$$\gamma \min_m y_k(m) - e_k(n) \geq -(\gamma + \epsilon)M^t - \frac{c}{1-\gamma} \quad (\text{A.4})$$

for all $k \geq k^t$ as (A.3) holds for t . Therefore, we have $Y_k^t(n) \leq y_k(n)$ for all $k \geq k^t$ and n . Since $Y_k^t(n) \rightarrow -(\gamma + \epsilon)M^t - \frac{c}{1-\gamma}$ and $-(\gamma + \epsilon)M^t \geq -M^{t+1}$, there exists $k^{t+1} \geq k^t$ such that (A.3) holds for $t+1$. Therefore, by induction, we can conclude that (A.3) holds for every $t \geq 0$. Since $M^t \rightarrow 0$, we have the lower bound. $\square$

Appendix B

Proof of Lemma 1

The contraction property of $\text{val}^i(\cdot)$ is used, e.g., in [41]. Let (μ^1, μ^2) and $(\bar{\mu}^1, \bar{\mu}^2)$ satisfy

$$\underline{\text{val}}^1(R^1) := (\mu^1)^T R^1 \mu^2 + \tau^1 H^1(\mu^1) - \tau^2 H^2(\mu^2) \quad (\text{B.1a})$$

$$\underline{\text{val}}^2(R^2) := (\bar{\mu}^2)^T R^2 \bar{\mu}^1 + \tau^2 H^2(\bar{\mu}^2) - \tau^1 H^1(\bar{\mu}^1). \quad (\text{B.1b})$$

Then, we have

$$\begin{aligned} \underline{\text{val}}^1(R^1) + \underline{\text{val}}^2(R^2) &\leq (\mu^1)^T R^1 \bar{\mu}^2 + \tau^1 H^1(\mu^1) - \tau^2 H^2(\bar{\mu}^2) \\ &\quad + (\bar{\mu}^2)^T R^2 \mu^1 + \tau^2 H^2(\bar{\mu}^2) - \tau^1 H^1(\mu^1) \\ &\leq (\mu^1)^T (R^1 + (R^2)^T) \bar{\mu}^2 \end{aligned} \quad (\text{B.2})$$

since agent j is the minimizer. Since $\mu^1 \in \Delta^1$ and $\bar{\mu}^2 \in \Delta^2$ are probability distributions, we have $(\mu^1)^T (R^1 + (R^2)^T) \bar{\mu}^2 \leq \bar{r}$. By following similar lines, we can also show that $\underline{\text{val}}^1(R^1) + \underline{\text{val}}^2(R^2) \geq r$.

For the second part, let (μ^i, μ^j) and $(\bar{\mu}^i, \bar{\mu}^j)$ satisfy

$$\underline{\text{val}}^i(R^i) := (\mu^i)^T R^i \mu^j + \tau^i H^i(\mu^i) - \tau^j H^j(\mu^j) \quad (\text{B.3a})$$

$$\underline{\text{val}}^i(R^i) := (\bar{\mu}^i)^T R^i \bar{\mu}^j + \tau^i H^i(\bar{\mu}^i) - \tau^j H^j(\bar{\mu}^j). \quad (\text{B.3b})$$

Then, we have

$$\begin{aligned}\underline{\text{val}}^i(R^i) - \text{val}^i(R^i) &\leq (\mu^i)^T R^i \bar{\mu}^j + \tau^i H^i(\mu^i) - \tau^j H^j(\bar{\mu}^j) - (\mu^i)^T R^i \bar{\mu}^j \\ &\leq \tau^i \log |A^i|\end{aligned}\tag{B.4}$$

as $-H^j(\bar{\mu}^j) \leq 0$ and $H^i(\mu^i) \leq \log |A^i|$. By following similar lines, we can show that $\underline{\text{val}}^i(R^i) - \text{val}^i(R^i) \geq -\tau^j \log |A^j|$. This completes the proof.



Appendix C

Proof of Lemma 2

Based on the definition (3.5), we have

$$\begin{aligned}
\|\bar{\pi}(z) - \bar{\pi}(\bar{z})\|_1 &= \sum_{b \in B} |\bar{\pi}(z)(b) - \bar{\pi}(\bar{z})(b)| \\
&= \sum_{b \in B} \left| \prod_{i \in I^N} \text{BR}_\tau(q^i)(b^i) - \prod_{i \in I^N} \text{BR}_\tau(\bar{q}^i)(b^i) \right| \\
&\stackrel{(a)}{\leq} \sum_b \left(\prod_{i \neq \ell} \text{BR}_\tau(q^i)(b^i) \right) |\text{BR}_\tau(q^\ell)(b^\ell) - \text{BR}_\tau(\bar{q}^\ell)(b^\ell)| \\
&\quad + \sum_b \text{BR}_\tau(\bar{q}^\ell)(b^\ell) \left| \prod_{i \neq \ell} \text{BR}_\tau(q^i)(b^i) - \prod_{i \neq \ell} \text{BR}_\tau(\bar{q}^i)(b^i) \right| \\
&\stackrel{(b)}{=} \|\text{BR}_\tau(q^\ell) - \text{BR}_\tau(\bar{q}^\ell)\|_1 \\
&\quad + \sum_{b^{-\ell}} \left| \prod_{i \neq \ell} \text{BR}_\tau(q^i)(b^i) - \prod_{i \neq \ell} \text{BR}_\tau(\bar{q}^i)(b^i) \right|, \tag{C.1}
\end{aligned}$$

where (a) follows from the triangle inequality and by adding and subtracting $\left(\prod_{i \neq \ell} \text{BR}_\tau(q^i)(b^i) \right) \text{BR}_\tau(\bar{q}^\ell)(b^\ell)$ for some $\ell \in I^N$, and (b) follows from the fact that the softmax, as described in (3.1), is a probability distribution. By following

similar lines with (C.1), we can show that

$$\begin{aligned}
\|\bar{\pi}(z) - \bar{\pi}(\bar{z})\|_1 &\leq \sum_{i \in I^N} \|\text{BR}_\tau(q^i) - \text{BR}_\tau(\bar{q}^i)\|_1 \\
&\leq \sum_i \sqrt{|B^i|} \cdot \|\text{BR}_\tau(q^i) - \text{BR}_\tau(\bar{q}^i)\|_2 \\
&\stackrel{(a)}{\leq} \frac{1}{\tau} \sum_i \sqrt{|B^i|} \cdot \|q^i - \bar{q}^i\|_2 \\
&\leq \frac{\sqrt{|B|}}{\tau} \sum_{i \in I^N} \|q^i - \bar{q}^i\|_1 \\
&= \frac{\sqrt{|B|}}{\tau} \cdot \|z - \bar{z}\|_1,
\end{aligned} \tag{C.2}$$

where (a) follows from the $\frac{1}{\tau}$ -Lipschitz continuity of the softmax function $\sigma(\cdot)$ with respect to $\|\cdot\|_2$, e.g., see [80, Proposition 4].

Appendix D

Proof of Lemma 3

By (3.4), the left-hand side of (3.16) can be written as

$$\begin{aligned} |r^j(z, a) - r^j(\bar{z}, a)| &= |u^j(a, \cdot)^T(\bar{\pi}(z) - \bar{\pi}(\bar{z}))| \\ &\stackrel{(a)}{\leq} \|u^j(a, \cdot)\|_{\infty} \|\bar{\pi}(z) - \bar{\pi}(\bar{z})\|_1 \\ &\stackrel{(b)}{\leq} L_0 \cdot \|z - \bar{z}\|_1, \end{aligned} \tag{D.1}$$

where (a) follows from the Hölder inequality, and (b) follows from Lemma 2 and the definition of L_0 .

Appendix E

Proof of Lemma 4

By the definition of the transition kernel $p(\cdot|\cdot)$, as described in Problem 1, we can write the left-hand side of the inequality (3.18) as

$$\Xi := \left| \sum_b \bar{\pi}(z)(b) \cdot \eta(z_+^{ab}) - \sum_b \bar{\pi}(\bar{z})(b) \cdot \eta(\bar{z}_+^{ab}) \right|, \quad (\text{E.1})$$

where $z_+^{ab} = \{q_+^{i,ab}\}_{i \in I^N}$ is the next state given the current state and the action profile (a, b) . Correspondingly, $\bar{z}_+^{ab} = \{\bar{q}_+^{i,ab}\}_{i \in I^N}$ is the next state given the corresponding state $\bar{z}$ and the action profile (a, b) .

If we add and subtract $\sum_b \bar{\pi}(z)(b) \cdot \eta(\bar{z}_+^{ab})$, the triangle inequality yields that $\Xi \leq \Xi_1 + \Xi_2$, where

$$\Xi_1 := \left| \sum_b \bar{\pi}(z)(b) \cdot (\eta(z_+^{ab}) - \eta(\bar{z}_+^{ab})) \right|, \quad (\text{E.2a})$$

$$\Xi_2 := \left| \sum_b (\bar{\pi}(z)(b) - \bar{\pi}(\bar{z})(b)) \cdot \eta(\bar{z}_+^{ab}) \right|. \quad (\text{E.2b})$$

Then, we can bound Ξ_1 by

$$\begin{aligned}
\Xi_1 &\leq \sum_b \bar{\pi}(z)(b) \cdot |\eta(z_+^{ab}) - \eta(\bar{z}_+^{ab})| \\
&\stackrel{(a)}{\leq} K \sum_b \bar{\pi}(z)(b) \cdot \|z_+^{ab} - \bar{z}_+^{ab}\|_1 \\
&\stackrel{(b)}{\leq} K \cdot \|z - \bar{z}\|_1,
\end{aligned} \tag{E.3}$$

where (a) follows from the K -Lipschitz continuity of $\eta(\cdot)$ from the statement of the lemma, and (b) follows from $\|\bar{\pi}(z)\|_1 = 1$ and the fact that

$$\begin{aligned}
\|z_+^{ab} - \bar{z}_+^{ab}\|_1 &= \sum_{i \in I^N} \sum_{b^i \in B^i} |q_+^{i,ab}(b^i) - \bar{q}_+^{i,ab}(b^i)| \\
&= \sum_{i \in I^N} \left((1 - \alpha) |q^i(b^i) - \bar{q}^i(b^i)| + \sum_{\tilde{b}^i \neq b^i} |q^i(\tilde{b}^i) - \bar{q}^i(\tilde{b}^i)| \right) \\
&\leq \sum_{i \in I^N} \|q^i - \bar{q}^i\|_1.
\end{aligned} \tag{E.4}$$

On the other hand, we can bound Ξ_2 by

$$\begin{aligned}
\Xi_2 &\stackrel{(a)}{\leq} \|\bar{\pi}(z) - \bar{\pi}(\bar{z})\|_1 \cdot \max_{b \in B} |\eta(\bar{z}_+^{ab})| \\
&\stackrel{(b)}{\leq} \frac{\sqrt{|B|}}{\tau} M \cdot \|z - \bar{z}\|_1.
\end{aligned} \tag{E.5}$$

where (a) follows from the Hölder inequality and (b) follows from Lemma 2 and the boundedness of $\eta(\cdot)$ from the statement of the lemma. Combining (E.3) and (E.5) lead to (3.18).