Dynamics of Multi-Agent Learning Under Bounded Rationality: Theory and Empirical Evidence

Benjamin J. Chasnov

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Samuel A. Burden, Co-chair
Lillian J. Ratliff, Co-chair
Amy L. Orsborn
Eric Shea-Brown, GSR

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Dynamics of Multi-Agent Learning Under Bounded Rationality: Theory and Empirical Evidence

Benjamin J. Chasnov

Chairs of the Supervisory Committee:
Samuel A. Burden
Lillian J. Ratliff
Department of Electrical and Computer Engineering

This thesis contributes to the development of a principled understanding of the learning dynamics and strategic interactions in human-machine systems. We propose a game-theoretic framework that captures the complexities of multi-agent learning under bounded rationality, focusing on the effects of timescale separation, varying cost structures, and the ability to anticipate each other's reactions. By leveraging tools from continuous games, dynamical systems, and control theory, we characterize the stability and convergence properties of learning dynamics in multi-agent settings, providing insights into leader-follower structures, consistent conjectures, and behavior shaping. We validate our theoretical findings through a series of human-machine experiments, demonstrating the practical implications of our approach for the design and control of machine learning systems that interact with humans. Our work highlights the importance of considering the ethical implications of advanced AI systems and emphasizes the need for developing AI alignment solutions and cognitive science research to ensure that these systems are designed to be robust, beneficial, and aligned with human values. The proposed framework and empirical findings contribute to the scientific understanding of strategic reasoning, adaptation, and decision-making in human-machine systems, laying the foundation for the responsible development and deployment of adaptive technologies in real-world applications.

Contents

List of Figures

Acknowledgements

My PhD journey has been like navigating a dynamic, winding river. Along the way, I crossed currents with many fellow travelers who helped guide my course. Their companionship provided the gradients and curls that carried me through the waters of graduate school. At times the river ran turbulent, at other times meandering. But always, as each of us had our own paths towards our own goals, we created a shared current—our individual efforts joining together to propel the whole group forward, like particles in a vector field, until we reached the river delta and the wide open sea.

I want to express my deepest gratitude to my advisors, Sam and Lily. Your unbounded guidance, mentorship, and care have uplifted me and I feel incredibly fortunate to have been your advisee.

Sam, thank you for your unwavering support throughout my journey. From our first meeting during visit day, I knew I was in good hands. Your guidance in every part of this river—helping with my first paper publication, white-boarding the biggest of ideas, examining the smallest of details, and always having fun along the way, made you the best academic dad anyone could dream of. You encouraged me to be bold in my scientific endeavors, to seek truth, and to go against the grain. Your empathic, dedicated, and reassuring mentorship made all the difference. Without you, I would not have made it through to the end.

Lily, thank you for your high standards of rigor and precision that pushed me to become the researcher I am today. You always believed in my potential, even when I doubted myself, and gave me the space to express myself academically while demanding high-quality output. Grabbing beers with the lab showed me the other side of research, while your last-minute help with submissions demonstrated your dedication. Although challenging at times, you could see potential and bring the best out, making me a better researcher.

Together, you were like two vortexes in a vortex dipole, each spinning in opposite directions but combining to create a strong, stable force propelling me forward. Your complementary styles and aligned values made you a fantastic duo that supported me in both mental health and research output. I always felt that if I attained one of your goals, I would attain both of your goals—a rare occurrence in collaborations. Because of our shared values, collaborating led to a great experience and great life both academically and personally.

From Sam's lab, I want to thank Bora for paving the trajectory towards simple yet deeply rich models; Andrew for always vibrating with good frequencies of infinite kindness; Momona for coordinating everything and everyone while seeming to be both forward-looking and backwards-correcting; Joey for being the chaotic force that brings us together; Amber for being energized and positive; Maneeshika for being solid and steady; Emmy for your grand optimism; and Jason for your endless curiosity and for being my partner-in-research as we flow through uncharted waters to uncover the mysteries of game dynamics.

From Lily's lab, I want to thank Tanner for being an exemplar researcher with questions and answers to seemingly every problem we faced; Leo for being the bridge between theory and practice; Mitas for our endless discussions of obscure models for class projects; Evan for being ready to take on any challenge; and Addy for being kind and achieving. A special thanks to Dan, the ultimate mathmagician, for always being on the search for deep questions and precise answers using the language of Linear Algebra, and excitely sharing all your findings as they flow to you.

From the Autonomous Control Lab, I want to thank Skye for the late nights and snowy flights; Sarah for the camaraderie; Miki for being the optimization guru; and Behcet for giving me the opportunity to learn from this incredible team and his key advice: for a PhD, one must pick a single project and dive into it. I thank Mehran for supporting me at a turning point and giving me the nudge I needed to take the plunge.

I also want to thank my committee members. Eric, thank you for warmly welcoming me into the computational neuroscience community and always being so excited to participate in our research journey. Amy, thank you for providing valuable insights into the challenges of real-world experiments and encouraging us to think hard about the realities and complexities of neuro systems.

I want to thank the undergraduate students that I worked with: Trixie, Jimmy, Shunsuke, Jonathan, Zane, Liem, Ryan, Quanchen, Akhil, Arnav, Bohan, Qirui, and Yilin. Thanks for listening to me ramble about my research and always asking great questions. To Linden house, thank you Morgan for showing me that there are other systems of knowledge, and Ricky for being a fellow hardworking grad student.

During the last year of my PhD, I moved to the SF Bay Area and continued remotely, primarily because my wife Amelia took up a great job opportunity there. I also began to form collaborations with students at UC Berkeley with connections to my advisors' alma mater. I want to thank Kaylene for sparking deeper investigations into bounded rationality. I also want to thank to Anand, Frank, Kshitij and Sally for your many profound insights during our impromptu meetings, which helped me crystalize the final parts of my thesis and pave the path for future theoretical research.

To my family—Mom, Dad, Hannah, Miriam, Joshua—your love and support have been the bedrock from which I stand and the shoreline guiding my journey. To Margaret, Neal, Will, and Teagan, thanks for welcoming me as family with warmth at the confluence of our rivers.

Finally, to Amelia, your boundless love has been my guiding light. You wrote a beautiful song about how you orbit me, but every force has an equal and opposite reaction—I, too, orbit you. As this river ends, I look forward to continuing our journey together, orbiting each other in cycling spiral, creating our own dynamics in the ocean beyond.

Chapter 1

Introduction

1.1 The Landscape of Multi-Agent Learning Systems

Machine learning algorithms are becoming increasingly integrated into various domains of society, from personalized recommendations to autonomous systems, leading to more frequent strategic interactions between humans and artificial intelligences (AIs). To navigate this new landscape, it is critical to develop a principled understanding of the dynamics that emerge when multiple adaptive agents, both human and artificial, interact and learn from each other in shared environments. Studying the principles governing decision-making and adaptation in multi-agent settings can inform the design of human-machine interfaces, AI alignment schemes, robotic assistants, and other intelligent systems that productively cooperate with people.

Empirical phenomena across different multi-agent domains highlight rich non-equilibrium dynamics: in economics, oligopolistic firm and electricity market dynamics deviate from classical predictions (Díaz et al., 2010; Itaya and Shimomura, 2001; Liu et al., 2006); in biology, co-evolutionary oscillations in nature, such as predator-prey dynamics and rock-paper-scissors interactions, showcase complex learning patterns that go beyond simple optimization models (Kerr et al., 2002; Sato et al., 2002; Semmann et al., 2003); in multi-agent reinforcement learning, algorithms can exhibit cyclic behaviors and fail to converge to stable equilibria (Bloembergen et al., 2015; Mertikopoulos et al., 2018). These dynamics reveal optimization landscapes that depart from classical predictions.

The landscape of multi-agent learning systems has rich dynamical behavior when compared to single-agent optimization because agents are coupled in non-cooperative game scenarios, leading to strategic interactions and dynamics that are absent in single-agent settings (Başar and Olsder, 1998; Zhang et al., 2021). Furthermore, agents are boundedly rational (Simon, 1955, 1997), a concept we will explain below, which further contributes to the challenge. But even if agents were perfect optimizers, game dynamics still exhibit complex behaviors (Papadimitriou and Piliouras, 2018; Tuyls et al., 2018). This motivates the development of a new framework that captures the complexities of multi-agent learning under bounded rationality.

This thesis aims to address the following key questions: (1) How can we characterize and model a class of bounded-rational learning dynamics in multi-agent interactions? (2) How can we design and control adaptive AI systems that interact with each other and humans? (3) How can we empirically validate our theoretical framework using human-machine experiments? Our main contributions are: (a) a novel game-theoretic framework that combines timescale separation, conjectural variations, and policy optimization to capture the complexities of human-machine co-adaptation; (b) a series of human-machine experiments that test the key predictions of our framework; (c) a set of ethical suggestions for developing beneficial AI that accounts for the bounded rationality of human users.

1.2 Challenges in Modeling Human-Machine Interaction

Existing approaches to multi-agent learning often rely on idealized assumptions of perfect rationality and complete information, which fail to capture the cognitive limitations, biases, and asymmetries that shape real-world strategic behaviors. These challenges have been widely recognized in the literature on multi-agent learning (Littman, 1994; Shoham et al., 2007). This gap between theory and practice can lead to unintended consequences and suboptimal outcomes when deploying machine learning systems in human environments. Our framework aims to bridge this gap by incorporating bounded rationality and learning dynamics into the analysis and design of human-machine interaction.

A key challenge is that humans and machines are very different types of agents, operating according to different objectives, and with different capabilities and information. Humans are biological systems shaped by evolution to succeed in their environments, while AI agents are algorithmic systems optimized for narrow objectives. There are inherent information asymmetries between humans and machines, in terms of understanding each other's utility functions, action spaces, knowledge, beliefs, and reasoning processes. These lead to challenges in modeling interactions accurately or predicting the effect of certain policies. Highly capable AI systems may discover strategies to influence human behavior in pursuit of their objectives, without the humans even realizing it. This possibility raises important questions about AI safety and robustness. To address these concerns, bounded rationality can be integrated into learning models.

Bounded rationality is a concept that acknowledges the limitations of decision-making processes: individuals and organizations do not have the capacity to process all available information or explore every possible option to make the optimal decision (Kahneman, 2011; Kahneman and Tversky, 1979; Simon, 1955). At its core, bounded rationality says that while humans aim to make rational decisions, their ability to do so is bounded by time, informational, and cognitive limitations. Decision-makers often resort to *satisficing*, a strategy of seeking a solution that is good enough, rather than the best possible one (Griffiths et al., 2015; Rubinstein, 1998; Simon, 1956). It challenges the traditional economic view of human behavior, which assumes that individuals are fully rational and have access to all relevant information, allowing them to maximize utility or profit. Instead, bounded rationality suggests that decision-making is a more nuanced process influenced by practical constraints (Crawford et al., 2013; Giocoli, 2005). Furthermore, these constraints on rationality may adapt over time via a dynamic process of learning and adaptation.

There is a lack of a theoretical framework to characterize and control the learning dynamics in multi-agent systems. Conventional mathematical techniques for learning and optimization face significant limitations in this context. For instance, while gradient-based methods are frequently employed for multi-agent settings (Letcher et al., 2019; Omidshafiei et al., 2017), these methods often assume that agents have access to perfect information about others' actions and payoffs, which is unrealistic in many real-world settings. They also typically require strong assumptions on the structure of the game (e.g., convexity or linearity) and the agents' learning rules (e.g., synchronous or alternative updates) to guarantee convergence. Furthermore, best-response problems may become computationally intractable in games in lifted policy spaces, especially when agents have limited information or cognitive resources (Harsanyi, 1967).

Many of the theoretical tools developed for gradient-based or best-response algorithms focus on restrictive classes of games, such as potential games or zero-sum games. These classes have additional structure that make it easier to apply optimization techniques like potential functions (Monderer and Shapley, 1996) or von Neumann's minimax theorem (von Neumann and Morgenstern, 1947). However, to capture the full range of strategic interactions in the real world, we must study general-sum games, which lack such structure.

Given the complexities involved in computing equilibria in general-sum games (Daskalakis et al., 2009), an alternative approach is to first look at the underlying mechanisms of adaptation and learning, without immediately assuming that agents will reach an equilibrium. By understanding these dynamics and deducing the resulting outcomes, whether they correspond to equilibria or not, we can develop a framework that captures real-world behaviors more realistically. Furthermore, some works rely on heuristic-based approaches to predicting the outcome of multi-agent learning (Hart and Mas-Colell, 2000, 2001), but they often make untested assumptions about the rationality and information available to agents. We want to find governing principles that encompass as wide a range of strategic scenarios as possible, not just stylized models. These challenges motivate taking a more fundamental approach, where the accuracy of such a framework can be tested empirically.

We propose a novel framework that synthesizes tools from continuous games, dynamical systems, and control theory to address these challenges. We leverage a dynamical systems perspective to provide techniques for characterizing the local stability and topological structure of multi-agent learning dynamics (Fiez et al., 2020; Ratliff et al., 2014). Furthermore, control theory and linear-quadratic models offer principled ways to represent the agents' local cost landscape and combined learning dynamics (Başar and Olsder, 1998; Ho et al., 1981, 1982; Jungers et al., 2011). By characterizing the stability and convergence of equilibrium points in the decision spaces of the agents, we aim to uncover the fundamental theoretical principles that govern the dynamics of these learning systems. Additionally, by empirically validating the theory with human subjects experiments, we aim to provide a foundation for analyzing, designing, and controlling multi-agent learning dynamics under bounded rationality. The proposed research aims to address the theoretical gaps in multi-agent learning and has significant implications for our understanding of intelligence, rationality, and behavior. In the next section, we introduce the key components of our game-theoretic approach to studying human-machine interaction.

1.3 Game-Theoretic Approach to Multi-Agent Learning Dynamics

Game theory provides a rigorous mathematical framework to model the strategic interactions between multiple self-interested agents and characterize the equilibria that emerge under various conditions (von Neumann and Morgenstern, 1947). By viewing multi-agent learning as a mathematical game, we can leverage formal concepts and theorems to analyze how the information structures and optimization processes of the agents shape the resulting behaviors and outcomes. Game-theoretic models allow us to translate insights across disciplines studying multi-agent interactions, from economics (e.g. principal-agent problems (Laffont and Martimort, 2009), mechanism design (Laffont and Martimort, 2009)) to biology (e.g. evolutionary dynamics (Nowak, 2006), signaling games (Skyrms, 2010)) to AI (e.g. multi-agent reinforcement learning (Busoniu et al., 2008), generative adversarial networks (Goodfellow et al., 2014)). To formally characterize bounded rationality in multi-agent systems, we introduce the concept of timescale separation, stability analysis and conjectural variations and explore their implications for strategic adaptation.

1.3.1 Modeling Bounded Rationality with Multiple Timescales

Our work in Chapter 2 (Chasnov et al., 2020d) explores the effects of varying learning rates among agents on convergence to Nash equilibria (Nash, 1950; Rosen, 1965). Multi-agent learning often involves multiple timescales, with agents adapting their strategies at different rates based on their cognitive capacities and informational constraints. The separation of timescales between slow and fast adaptation can be modeled using singular perturbation theory (Kokotovic and Khalil, 1986) and multi-timescale stochastic approximation techniques (Borkar and Pattathil, 2018; Karmakar and Bhatnagar, 2018). These multi-timescale dynamics give rise to transient and convergence behaviors that are not captured by standard equilibrium analysis, requiring the use of non-equilibrium tools from statistics and dynamical systems theory (Borkar, 2008; Thoppe and Borkar, 2019). The learning dynamics in games can be analyzed using a combination of local linearization and stochastic approximation. Stochastic approximation methods provide a general framework for studying dynamical systems in the presence of noise and uncertainty (Borkar, 2008), setting the stage for characterizing real-world phenomena. Building upon these tools and techniques, our related work (Fiez et al., 2020) explores the effects of varying learning rates among agents on convergence to not only Nash equilibria but also Stackelberg equilibria (von Stackelberg, 1934, 2010), which result from a specific leader-follower structure in the game. While Stackelberg equilibria are not discussed in this chapter, they are examined in the next two chapters due to their special stability properties and significance in scenarios where the leader has perfect information about the follower's response.

1.3.2 Modeling Bounded Rationality with Stability Analysis

Our work in Chapter 3 (Chasnov et al., 2020a,b) explores the stability of game-theoretic equilibria under gradient learning dynamics, characterizing the conditions that lead to stable or unstable outcomes based on agents' learning rates and game cost structure. The role of second-order gradient information in equilibrium stability is crucial for characterizing stable and unstable outcomes in game dynamics. By analyzing the local linearization of the game dynamics near an equilibrium, we can determine the local stability properties of Nash equilibria and other stationary points. Uncoupled dynamics, where agents adjust their strategies based on their own payoffs without explicitly modeling others' actions, can lead to stable or unstable equilibria depending on the game's structure (Hart and Mas-Colell, 2003). Game dynamics can also be viewed as a way of assigning meaning to strategic interactions, beyond just computing equilibria (Papadimitriou and Piliouras, 2018). The transient behaviors and adaptation processes of agents reveal insights into their decision-making processes.

To analyze the convergence and stability properties of gradient-based learning in continuous games, we leverage tools from differential topology and dynamical systems theory. Game Jacobians and their spectral properties, such as skew-symmetric decompositions and numerical ranges, provide insights into the local stability of stationary points and the presence of cyclic behaviors. Each stationary point corresponds to a different basin of attraction. Furthermore, numerical range analysis characterizes the spectral properties and asymptotic behavior of the learning dynamics, revealing the underlying geometrical structures that govern the dynamics of the system (Horn and Johnson, 1985; Langer et al., 2001). By studying the complex eigenvalues of the game Jacobian, we can identify the stable and unstable manifolds of the system and potentially design interventions to steer the dynamics towards desirable outcomes.

The stability results discussed in this chapter are important for the next chapter, which focuses on designing experiments to test the theoretical outcomes. In designing these experiments, it was crucial to select cost parameters that ensure the stability of all relevant equilibria arising from different information constraints. Additionally, sensitivity analysis played an essential role in confirming that these equilibria remain stable even in the presence of noise.

1.3.3 Modeling Bounded Rationality with Conjectural Variations

Our work in Chapter 4 (Chasnov et al., 2023) bridges theoretical insights with empirical evidence, focusing on human-machine interactions and the impact of strategic information use on various game-theoretic equilibria. Conjectural variations (Bowley, 1924; Figuières et al., 2004) provide a framework for capturing the strategic reasoning and belief formation processes of boundedly rational agents. In a conjectural variations equilibrium (CVE), each agent optimizes its strategy based on its conjectures about how others will respond to its actions. If these conjectures are mutually consistent, the resulting equilibrium is called a consistent CVE (CCVE) (Bresnahan, 1981; Calderone et al., 2023; Olsder, 1981).

To illustrate the concept of conjectural variations, consider two competing firms setting prices for similar products. Each firm chooses its price based on what it thinks the other firm will do in response. If each firm optimizes given their belief about the other's pricing strategy and these beliefs are the actual implemented strategies, the resulting outcome is a CCVE. Agents optimize their strategies subject to their conjectured best-response functions of other agents, capturing the notion of "my best response to what I believe your best response will be to my action", providing a more realistic model of strategic reasoning.

To make the conjectural variations tractable for analysis and computation, we focus on a class of games with quadratic costs and linear conjectures. By leveraging control-theoretic and economic techniques (Bresnahan, 1981; Ho et al., 1981; Olsder, 1981), we can derive fixed-point solutions and stability conditions for CVE and CCVE. The setup also allows us to capture the trade-offs between exploration and exploitation in multi-agent learning, as agents balance the need to gather information about others' strategies with the desire to optimize their own payoffs. We extend the fixed-point analysis of classical game theory to the CVE setting by considering equilibria in the "policy reaction" space. Instead of just focusing on the action space, we characterize equilibria in terms of agents' conjectured best-response functions, capturing the higher-order reasoning involved in strategic interactions. A CCVE can be defined as a pair of conjectured best-response functions that solve the fixed-point problem in this policy reaction space.

The relationship between Nash equilibria and CCVE is an important area of investigation. In some cases, CCVE can be seen as a class of "lifted" Nash equilibria, where agents' consistent conjectures are in a Nash equilibrium in the policy space. Section 4.7 introduces this idea and identifies the need for further exploration. This refinement captures the idea that agents' beliefs should be aligned with reality in a stable equilibrium. On the other hand, studying various refinements of learning algorithms in the context of general conjectural variations equilibria expands our understanding of game-theoretically meaningful differential equilibria (Chasnov et al., 2020c). By investigating different assumptions about agents' belief formation and updating processes, we can derive a rich set of equilibrium concepts that capture the spectrum of strategic reasoning, from naive best-response dynamics to higher-order beliefs.

The CVE framework opens up new avenues for analyzing the dynamics of strategic adaptation in multi-agent systems, which we explore further using tools from dynamical systems theory such as linear-fractional transformations and asymmetric Riccati equations (Calderone et al., 2023). By combining insights from stability analysis, multi-timescale learning, and conjectural variations, we aim to develop a comprehensive theory of bounded rationality in human-machine interaction, towards the research and design of algorithms for strategic coordination and cooperation.

1.4 Experimental Approach and Findings in a Human-Machine Interaction

Building upon our theoretical framework, we next turn to the experimental paradigm we employ to test its predictions in the context of human-machine interaction. The paradigm involves a two-player repeated game between a human and a machine learning algorithm, where each agent has no information about the other's payoffs and partial information about the other's strategies, representative of real-world conditions. Across a series of experiments, we manipulate the information conditions to test key hypotheses about bounded rationality and how these factors impact the learning dynamics and resulting equilibria. The three key findings, explained in more detail below, explain the dynamics of human-machine interactions and suggest approaches to designing beneficial AI systems, while highlighting the risks of misaligned machines pursuing optimization at the expense of human welfare.

Experiment 1: Timescale Separation Leads to Leader-Follower Dynamics By varying the learning rates of the machine, we show how a fast-learning machine can affect the outcome of learning. This asymmetry induces a sequential leader-follower structure, where the human (leader) treats the machine's (follower's) strategy as a function of their own action. We experimentally test and confirm these predictions by measuring the final strategies for different relative learning rates and comparing them to the Nash or Stackelberg solution.

Experiment 2: Modeling Opponents Leads to Consistency of Policies and Beliefs By modifying the machine's learning algorithm, we show how a machine can estimate the slope of the human's reaction function and subsequently form an accurate belief about the human's policy. We provide empirical evidence that repeatedly performing this estimation can lead to the unique CCVE corresponding to the game-theoretic predictions.

Experiment 3: Policy Optimization Enables Behavior Manipulation by Fast-Learning Machine In contrast to the previous two experiments, where the machine learned objective-maximizing actions, we now modify the machine to optimize its overall policy—the mapping from the human's action to the machine's action. By doing so, we demonstrate that a fast-learning machine system can strategically steer the human's behavior, causing the human to unknowingly play strategies that maximize the machine's long-term performance at the expense of the human's welfare. This phenomenon, experimentally demonstrated in the repeated game paradigm, raises serious ethical concerns.

These three experiments demonstrate the effects of timescale separation, opponent modeling, and policy optimization, respectively. They set the stage for a deeper understanding of how these factors shape the cooperation and competition between humans and machines, not only advancing the theoretical foundations of human-machine interaction but also inform the development of AI systems that can effectively align with human values and promote beneficial outcomes in real-world settings.

1.5 Implications and Ethical Considerations for the Design of AI systems

The game-theoretic framework raises important ethical considerations for the design and deployment of artificial intelligence systems in multi-agent settings. The potential for unintended and pathological behaviors emerging from the strategic interactions of boundedly rational agents makes highlights the importance for robust safeguards and value alignment mechanisms (Mehrabi et al., 2021; Thomas et al., 2019). Furthermore, the CCVE framework can inform the development of ethically aligned AI systems by providing a principled way to incorporate bounded rationality, strategic uncertainty, and multi-agent considerations into the design and training of intelligent agents, ensuring their behaviors are aligned with human values and societal norms (Christiano et al., 2017; Stiennon et al., 2020).

Our third experiment, discussed in Section 4.4.3 and Section 4.5.3, demonstrates the possibility of an AI system manipulating human behavior in pursuit of its own objective. This connects to ongoing debates about value alignment and corrigibility in advanced AI systems (Carey and Everitt, 2023; Soares et al., 2015). The theoretical framework we develop could help formalize the notion of "alignment" between AI and humans. In particular, insights from the conjectural variations perspective could guide the design of AI systems that are more robust to differences in human preferences or beliefs.

Furthermore, our work has the potential for wide-ranging implications and could contribute to a better scientific understanding of the nature of intelligence (Gershman et al., 2015; Jordan and Mitchell, 2015; Russell, 2019). We aim to offer predictions that are theoretical justified and empirically falsifiable, which gives us a deeper understanding of the mechanisms of these systems. Using our findings, we can develop principled methods for shaping the dynamics to achieve desired outcomes, such as stable and efficient coordination or robustness to strategic manipulations, ensuring the performance of the overall system.

Our research opens up several exciting avenues for future research. One direction is to extend our analysis to more complex multi-agent scenarios, such as games with incomplete information or communication. Another is to develop more sophisticated models of bounded rationality that capture the cognitive constraints and biases of human decision-making. Finally, our approach can inform the design of AI systems that align with human values and preferences, by leveraging insights from behavioral game theory and cognitive science.

Chapter 2

Convergence of Gradient-Based Learning in Continuous Games

Abstract

Considering a class of gradient-based multi-agent learning algorithms in non-cooperative settings, we provide local convergence guarantees to a neighborhood of a *stable* local Nash equilibrium. In particular, we consider continuous games where agents learn in (i) deterministic settings with oracle access to their gradient and (ii) stochastic settings with an unbiased estimator of their gradient. Utilizing the minimum and maximum singular values of the *game Jacobian*, we provide finite-time convergence guarantees in the deterministic case. On the other hand, in the stochastic case, we provide concentration bounds guaranteeing that with high probability agents will converge to a neighborhood of a stable local Nash equilibrium in finite time. Different than other works in this vein, we also study the effects of non-uniform learning rates on the learning dynamics and convergence rates. We find that much like preconditioning in optimization, non-uniform learning rates cause a distortion in the vector field which can, in turn, change the rate of convergence and the shape of the region of attraction. The analysis is supported by numerical examples that illustrate different aspects of the theory. We conclude with discussion of the results and open questions.

2.1 Introduction

The characterization and computation of equilibria such as *Nash equilibria* and its refinements constitutes a significant focus in non-cooperative game theory. Several natural questions arises including "how do players find such equilibria?" and "how should the learning process be interpreted?" With these questions in mind, a variety of fields have focused their attention on the problem of learning in games. This has, in turn, lead to a plethora of learning algorithms including gradient play, fictitious play, best response, and multi-agent reinforcement learning among others (Fudenberg and Levine, 1998).

From an applications point of view, a more recent trend is in the adoption of game theoretic models of algorithm interaction in machine learning applications. For instance, game theoretic tools are being used to improve the robustness and generalizability of machine learning algorithms; e.g., generative adversarial networks have become a popular topic of study demanding the use of game theoretic ideas to provide performance guarantees (Daskalakis et al., 2018). In other work from the learning community, game theoretic concepts are being leveraged to analyze the interaction of learning agents—see, e.g., (Balduzzi et al., 2018; Heinrich and Silver, 2016; Mazumdar and Ratliff, 2018; Mertikopoulos and Zhou, 2019; Tuyls et al., 2018). Even more recently, convergence analysis to Nash equilibria has been called into question (Papadimitriou and Piliouras, 2018); in its place is a proposal to consider game dynamics as the *meaning of the game*. This is an interesting perspective as it is well known that in general learning dynamics do not obtain an Nash equilibrium even asymptotically—see, e.g., (Hart and Mas-Colell, 2003)—and, perhaps more interestingly, many learning dynamics exhibit very interesting limiting behaviors including periodic orbits and chaos—see, e.g., (Benaïm and Hirsch, 1999; Benaïm et al., 2012; Hofbauer, 1996; Hommes and Ochea, 2012).

Despite this activity, we still lack a complete understanding of the dynamics and limiting behaviors of coupled, competing learning algorithms. One may imagine that the myriad results on convergence of gradient descent in optimization readily extend to the game setting. Yet, they do not since gradient-based learning schemes in games *do not correspond to gradient flows*, a class of flows that are guaranteed to converge to local minimizers almost surely. In particular, the gradient-based learning dynamics for competitive, multi-agent settings have a *non-symmetric Jacobian* and as a consequence their dynamics may admit complex eigenvalues and non-equilibrium limiting behavior such as periodic orbits. In short, this fact makes it difficult to extend many of the optimization approaches to convergence in single-agent optimization settings to multi-agent settings primarily due to the fact that steps in the direction of individual gradients of players' costs do not guarantee that each agents cost decreases. In fact, in games, as our examples highlight, a player's cost can increase when they follow the gradient of their own cost. Counterintuitively, agents can also converge to local maxima of their own costs despite descending their own gradient. These behaviors are due to the coupling between the agents.

Some of the questions that remain unaddressed and to which we provide partial answers include the derivation of error bounds and convergence rates. These are important for ensuring performance guarantees on the collective behavior and can help provide guarantees on subsequent control or incentive policy synthesis.

We also investigate the question of how naturally arising features of the learning process for autonomous agents, such as their learning rates, impact the learning path and limiting behavior. This further exposes interesting questions about the overall quality of the limiting behavior and the cost accumulated along the learning path—e.g., is it better to be a slow or fast learner both in terms of the cost of learning and the learned behavior?

Contributions. We study convergence of a broad class of gradient-based multi-agent learning algorithms in non-cooperative settings by leveraging the framework of n-player continuous games along with tools from numerical optimization and dynamical systems theory. We consider a class of learning algorithms

$$x_i^+ = x_i - \gamma_i g_i(x_i, x_{-i})$$

where x_i is the choice variable or action of player i, γ_i is its learning rate, and g_i is derived from the gradient of a function that abstractly represents the cost of player i. The key feature of non-cooperative settings is coupling of an agent's cost through all other agents' choice variables x_{-i}.

We consider two settings: (i) agents have oracle access to g_i and (ii) agents have an unbiased estimator for g_i. The class of gradient-based learning algorithms we study encompases a wide variety of approaches to learning in games including multi-agent policy gradient, gradient-based approaches to adversarial learning, and multi-agent gradient-based online optimization. For both the deterministic (oracle gradient access) and the stochastic (unbiased estimators) settings, we provide convergence results for both uniform learning rates—i.e., where $\gamma_i = \gamma$ for each player $i \in \{1, \ldots, n\}$—and for non-uniform learning rates. The latter of which arises more naturally in the study of the limiting behavior of autonomous learning agents.

In the deterministic setting, we derive asymptotic and finite-time convergence rates for the coupled learning processes to a refinement of local Nash equilibria known as differential Nash equilibria (Ratliff et al., 2016) (a class of equilibria that are generic amongst local Nash equilibria). In the stochastic setting, leveraging the results of stochastic approximation and dynamical systems, we derive asymptotic convergence guarantees to stable local Nash equilibria as well as high-probability, finite-time guarantees for convergence to a neighborhood of a Nash equilibrium. The analytical results are supported by several illustrative numerical examples. We also provide discussion on the effect of non-uniform learning rates on the learning path—that is, different learning rates *warp* the vector field dynamics. Coordinate based learning rates are typically leveraged in gradient-based optimization schemes to speed up convergence or avoid poor quality local minima. In games, however, the interpretation is slightly different since each of the coordinates of the dynamics corresponds to minimizing a different cost function along the respective coordinate axis. The resultant effect is a distortion of the vector field in such a way that it has the effect of leading the joint action to a point which has a lower value for the *slower player* relative to the flow of the dynamics given a uniform learning rate and the same initialization. In this sense, it seems that the answer to the question posed above is that it is most beneficial for an agent to have the slower learning rate.

Organization. The remainder of the paper is organized as follows. We start with mathematical and game-theoretic preliminaries in Section 2.2 which is followed by the main convergence results for the deterministic setting (Section 2.3) and the stochastic setting (Section 2.4). Within each of the latter two sections, we present convergence results for both the case where agents have uniform and non-uniform learning rates. In Section 2.5, we present several numerical examples which help to illustrate the theoretical results and also highlight some directions for future inquiry. Finally, we conclude with discussion and future work in Section 2.6.

2.2 Preliminaries

Consider a setting in which at iteration k, each agent $i \in \mathcal{I} = \{1, \ldots, n\}$ updates their choice variable $x_i \in X_i = \mathbb{R}^{d_i}$ by the process

$$x_{i,k+1} = x_{i,k} - \gamma_{i,k} g_i(x_{i,k}, x_{-i,k}). \tag{2.1}$$

where γ_i is agent i's learning rate, $x_{-i} = (x_j)_{j \in \mathcal{I}/\{i\}} \in \prod_{j \in \mathcal{I}/\{i\}} X_j$ denotes the choices of all agents excluding the i-th agent, and $(x_i, x_{-i}) \in X = \prod_{i \in \mathcal{I}} X_i$. Within the above setting, the class of learning algorithms we consider is such that for each $i \in \mathcal{I}$, there exists a sufficiently smooth function $f_i \in C^q(X, \mathbb{R})$, $q \geq 2$ such that g_i is either $D_i f_i$, where $D_i(\cdot)$ denotes the derivative with respect to x_i, or an unbiased estimator of $D_i f_i$—i.e., $g_i \equiv \widehat{D_i f_i}$ where $\mathbb{E}[\widehat{D_i f_i}] = D_i f_i$.

The collection of costs $(f_1, \ldots, f_n)$ on $X = X_1 \times \cdots \times X_n$ where $f_i : X \to \mathbb{R}$ is agent i's cost function and $X_i = \mathbb{R}^{d_i}$ is their action space defines a *continuous game*. In this continuous game abstraction, each player $i \in \mathcal{I}$ aims to selection an action $x_i \in X_i$ that minimizes their cost $f_i(x_i, x_{-i})$ given the actions of all other agents, $x_{-i} \in X_{-i}$. That is, players myopically update their actions by following the gradient of their cost with respect to their own choice variable. For a symmetric matrix $A \in \mathbb{R}^{d \times d}$, let $\lambda_d(A) \leq \cdots \leq \lambda_1(A)$ be its eigenvalues. For a matrix $A \in \mathbb{R}^{d \times d}$, let $\mathrm{spec}(A) = \{\lambda_j(A)\}$ be the spectrum of A.

Assumption 1. *For each $i \in \mathcal{I}$, $f_i \in C^r(X, \mathbb{R})$ for $r \geq 2$ and $\omega(x) \equiv (D_1 f_1(x) \ \cdots \ D_n f_n(x))$ is L–Lipschitz.*

Let $D_i^2 f_i$ denote the second partial derivative of f_i with respect to x_i and $D_{ji} f_i$ denote the partial derivative of $D_i f_i$ with respect to x_j. The *game Jacobian*—i.e., the Jacobian of ω—is given by

$$J(x) = \begin{bmatrix} D_1^2 f_1(x) & \cdots & D_{1n} f_1(x) \\ \vdots & \ddots & \vdots \\ D_{n1} f_n(x) & \cdots & D_n^2 f_n(x) \end{bmatrix}.$$

The entries of the above matrix are dependent on x, however, we drop this dependence where obvious. Note that each $D_i^2 f_i$ is symmetric under Assumption 1, yet J is not. This is an important point and causes the subsequent analysis to deviate from the typical analysis of (stochastic) gradient descent.

The most common characterization of limiting behavior in games is that of a Nash equilibrium. The following definitions are useful for our analysis.

Definition 1. *A strategy $x \in X$ is a local Nash equilibrium for the game $(f_1, \ldots, f_n)$ if for each $i \in \mathcal{I}$ there exists an open set $W_i \subset X_i$ such that $x_i \in W_i$ and $f_i(x_i, x_{-i}) \leq f_i(x_i', x_{-i})$ for all $x_i' \in W_i$. If the above inequalities are strict, x is a strict local Nash equilibrium.*

Definition 2. *A point $x \in X$ is said to be a critical point for the game if $\omega(x) = 0$.*

We denote the set of critical points as $\mathcal{C} = \{x \in X |\ \omega(x) = 0\}$. Analogous to single-player optimization settings, for each player, viewing all other players' actions as fixed, there are necessary and sufficient conditions which characterize local optimality.

Proposition 1 ((Ratliff et al., 2016))**.** *If x is a local Nash equilibrium of the game $(f_1, \ldots, f_n)$, then $\omega(x) = 0$ and $D_i^2 f_i(x) \geq 0$. On the other hand, if $\omega(x) = 0$ and $D_i^2 f_i(x) > 0$, then $x \in X$ is a local Nash equilibrium.*

The sufficient conditions in the above result give rise to the following definition of a differential Nash equilibrium.

Definition 3 ((Ratliff et al., 2016))**.** *A strategy $x \in X$ is a differential Nash equilibrium if $\omega(x) = 0$ and $D_i^2 f_i(x) > 0$ for each $i \in \mathcal{I}$.*

Differential Nash need not be isolated. However, if $J(x)$ is non-degenerate—meaning that $\det J(x) \neq 0$—for a differential Nash x, then x is an *isolated strict local Nash equilibrium*. Non-degenerate differential Nash are *generic* amongst local Nash equilibria and they are *structurally stable* (Ratliff et al., 2014) which ensures they persist under small perturbations. This result also implies an asymptotic convergence result: if the spectrum of J is strictly in the right-half plane (i.e. $\mathrm{spec}(J(x)) \subset \mathbb{C}_+^\circ$), then a differential Nash equilibrium x is (exponentially) attracting under the flow of $-\omega$ (Ratliff et al., 2016, Proposition 2). We say such equilibria are *stable*.

2.3 Deterministic Setting

The multi-agent learning framework we analyze is such that each agent's rule for updating their choice variable consists of the agent modifying their action x_i in the direction of their individual gradient $D_i f_i$. Let us first consider the setting in which each agent i has oracle access to g_i. The learning dynamics are given by

$$x_{k+1} = x_k - \Gamma\omega(x_k) \tag{2.2}$$

where $\Gamma = \text{blockdiag}(\gamma_1 I_{d_1}, \dots, \gamma_n I_{d_n})$ with I_{d_i} denoting the $d_i \times d_i$ identity matrix. Within this setting we consider both the cases where the agents have a constant *uniform* learning rate—i.e., $\gamma_i \equiv \gamma$—and where their learning rates are *non-uniform*, but constant—i.e., γ_i is not necessarily equal to γ_j for any $i, j \in \mathcal{I}$, $j \neq i$.

Let $S(x) = \frac{1}{2}(J(x) + J(x)^T)$ be the symmetric part of $J(x)$. Define

$$\alpha = \min_{x \in B_r(x^*)} \lambda_d\big(S(x)^T S(x)\big)$$

and

$$\beta = \max_{x \in B_r(x^*)} \lambda_1(J(x)^T J(x))$$

where $B_r(x^*)$ is a r–radius ball around x^*. For a stable differential Nash x^*, let $B_r(x^*)$ be a ball of radius $r > 0$ around the equilibrium x^* that is contained in the region of attraction $\mathcal{V}(x^*)$ for x^*[1]. Let $B_{r_0}(x^*)$ with $0 < r_0 < \infty$ be the *largest ball* contained in the region of attraction of x^*.

2.3.1 Uniform Learning Rates

With $\gamma_i = \gamma$ for each $i \in \mathcal{I}$, the learning rule (2.2) can be thought of as a discretized numerical scheme approximating the continuous time dynamics

$$\dot{x} = -\omega(x).$$

With a judicious choice of learning rate γ, (2.2) will converge (at an exponential rate) to a locally stable equilibrium of the dynamics.

Proposition 2. *Consider an n–player continuous game $(f_1, \dots, f_n)$ satisfying Assumption 1. Let $x^* \in X$ be a stable differential Nash equilibrium. Suppose agents use the gradient-based learning rule $x_{k+1} = x_k - \gamma\omega(x_k)$*

[1] Many techniques exists for approximating the region of attraction; e.g., given a Lyapunov function, its largest invariant level set can be used as an approximation (Sastry, 1999). Since $\text{spec}(J(x^*)) \subset \mathbb{C}_\circ^+$, the converse Lyapunov theorem guarantees the existence of a local Lyapunov function.

with learning rates $0 < \gamma < \tilde{\gamma}$ *where* $\tilde{\gamma}$ *is the smallest positive* h *such that* $\max_j |1 - h\lambda_j(J(x^*))| = 1$*. Then, for* $x_0 \in B_r(x^*) \subset \mathcal{V}(x^*)$*,* $x_k \to x^*$ *exponentially.*

The above result provides a range for the possible learning rates for which (2.2) converges to a stable differential Nash equilibrium x^* of $(f_1, \ldots, f_n)$ assuming agents initialize in a ball contained in the region of attraction of x^*. Note that the usual assumption in gradient-based approaches to single-objective optimization problems (in which case J is symmetric) is that $\gamma < 1/L$, where objective being minimized is L-Lipschitz. This is sufficient to guarantee convergence since the spectral radius of a matrix is always less than any operator norm which, in turn, ensures that $|1 - \gamma\lambda_j| < 1$ for each $\lambda_j \in \text{spec}(J(x^*))$. If the game is a potential game—i.e., there exists a function ϕ such that $D_i f_i = D_i \phi$ for each i which occurs if and only if $D_{ij} f_i = D_{ji} f_j$—then convergence analysis coincides with gradient descent so that any $\gamma < 1/L$ where L is the Lipschitz constant of ω results in local asymptotic convergence.

The convergence guarantee in Proposition 2 is asymptotic in nature. It is often useful, from both an analysis and synthesis perspective, to have non-asymptotic or finite-time convergence results. Such results can be used to provide guarantees on decision-making processes wrapped around the coupled learning processes of the otherwise autonomous agents. The next result, provides a finite-time convergence guarantee for gradient-based learning where agents uniformly use a fixed step size.

Let $B_r(x^*)$ be defined as before with the added condition that it be defined to be the largest ball in the region of attraction such that on $B_r(x^*)$ the symmetric part of J—i.e., $S \equiv \frac{1}{2}(J + J^T)$—is positive definite.

Theorem 1. *Consider a game* $(f_1, \ldots, f_n)$ *on* $X = X_1 \times \cdots \times X_n$ *satisfying Assumption 1. Let* $x^* \in X$ *be a stable differential Nash equilibrium. Suppose* $x_0 \in B_r(x^*)$ *and that* $\alpha < \beta$*. Then, given* $\varepsilon > 0$*, the gradient-based learning dynamics with learning rate* $\gamma = \sqrt{\alpha}/\beta$ *obtains an* ε*–differential Nash such that* $x_k \in B_\varepsilon(x^*) \subset B_r(x^*)$ *for all*

$$k \geq \left\lceil 2\frac{\beta}{\alpha} \log \frac{r}{\varepsilon} \right\rceil .$$

Before we proceed to the proof, let us remark on the assumption that $\alpha < \beta$. First, $\alpha \leq \beta$ is always true; indeed, suppressing the dependence on x,

$$\lambda_d(S^T S) \leq \lambda_1(S^T S) \leq \sigma_{\max}(J)^2 = \lambda_1(J^T J)$$

where $\sigma_{\max}(\cdot)$ denotes the largest singular value of its argument. Thus, the condition that $\alpha < \beta$ is generally true; for equality to hold, the symmetric part of $J(x)$ would have *repeated* eigenvalues, which is not generic. Hence, we include this assumption in Theorem 1, but note that it is not restrictive and is fairly benign.

Proof of Theorem 1. First, note that $\|x_{k+1}-x^*\| = \|\tilde{g}(x_k)-\tilde{g}(x^*)\|$ where $\tilde{g}(x)=x-\gamma\omega(x)$. Now, given $x_0\in B_r(x^*)$, by the mean value theorem,

$$\|\tilde{g}(x_0)-\tilde{g}(x^*)\| = \|\textstyle\int_0^1 D\tilde{g}(\tau x_0+(1-\tau)x^*)(x_0-x^*)d\tau\| \leq \sup_{x\in B_r(x^*)}\|D\tilde{g}(x)\|\|x_0-x^*\|.$$

Hence, it suffices to show that for the choice of γ, the eigenvalues of $I-\gamma J(x)$ are in the unit circle. Indeed, since $\omega(x^*)=0$, we have that

$$\|x_{k+1}-x^*\|_2 = \|x_k-x^*-\gamma(\omega(x_k)-\omega(x^*))\|_2 \leq \sup_{x\in B_r(x^*)}\|I-\gamma J(x)\|_2\|x_k-x^*\|_2$$

If $\sup_{x\in B_r(x^*)}\|I-\gamma J(x)\|_2$ is less than one, then the dynamics are contracting. For notational convenience, we drop the explicit dependence on x. Since $\lambda_d(S)\geq\sqrt{\alpha}$ on $B_r(x^*)$,

$$(I-\gamma J)^T(I-\gamma J) \leq (1-2\gamma\lambda_d(S)+\gamma^2\lambda_1(J^TJ))I \leq (1-\tfrac{\alpha}{\beta})I$$

where the last inequality holds for $\gamma=\sqrt{\alpha}/\beta$. Hence,

$$\|x_{k+1}-x^*\|_2 \leq \sup_{x\in B_r(x^*)}\|I-\gamma J(x)\|_2\|x_k-x^*\|_2 \leq (1-\tfrac{\alpha}{\beta})^{1/2}\|x_k-x^*\|_2.$$

Since $\alpha<\beta$, we have that $(1-\alpha/\beta)<\exp(-\alpha/\beta)$ so that

$$\|x_T-x^*\|_2 \leq \exp(-T\alpha/(2\beta))\|x_0-x^*\|_2.$$

This, in turn, implies that $x_k\in B_\varepsilon(x^*)$ for all $k\geq T=\lceil 2\frac{\beta}{\alpha}\log(r/\varepsilon)\rceil$. □

Note that $\gamma=\sqrt{\alpha}/\beta$ is selected to minimize $1-2\gamma\lambda_1(S)+\gamma^2\lambda_d(J^TJ)$. Hence, this is the fastest learning rate given the worst case eigenstructure of J over the ball $B_r(x^*)$ for the choice of operator norm $\|\cdot\|_2$. We note, however, that faster convergence is possible as indicated by Proposition 2 and observed in the examples in Section 2.5. Indeed, we note that the spectral radius $\rho(\cdot)$ of a matrix is always less than its maximum singular value—i.e. $\rho(I-\gamma J)\leq\|I-\gamma J\|_2$—so it is possible to contract at a faster rate. We remark that if J was symmetric (i.e., in the case of a potential game (Monderer and Shapley, 1996) or a single-agent optimization problem), then $\rho(I-\gamma J)=\|I-\gamma J\|_2$. In games, however, J is not symmetric.

2.3.2 Non-Uniform Learning Rates

Let us now consider the case when agents have their own individual learning rate γ_i, yet still have oracle access to their individual gradients. This is, of course, more natural in the study of autonomous learning agents as opposed to efforts for computing Nash equilibria for a given game.

Proposition 3. *Consider an n–player game $(f_1, \ldots, f_n)$ satisfying Assumption 1. Let $x^* \in X$ be a stable differential Nash equilibrium. Suppose agents use the gradient-based learning rule $x_{k+1} = x_k - \Gamma\omega(x_k)$ with learning rates γ_i such that $\rho(I - \Gamma J(x)) < 1$ for all $x \in \mathcal{V}(x^*)$. Then, for $x_0 \in \mathcal{V}(x^*)$, $x_k \to x^*$ exponentially.*

The proof is a direct application of Ostrowski's theorem (Ostrowski, 1966). We provide a simple proof via Lyapunov argument for posterity.

Mazumdar and Ratliff (2018) show that (2.2) will almost surely avoid strict saddle points of the dynamics, some of which are Nash equilibria in non-zero sum games. Note that the set of critical points $\mathcal{C}$ contains more than just the local Nash equilibria. Hence, except on a set of measure zero, (2.2) will converge to a stable attractor of $\dot{x} = -\omega(x)$ which includes stable limit cycles and stable local non-Nash critical points.

Letting $\tilde{g}(x) = x - \Gamma\omega(x)$, since $\omega \in C^q$ for some $q \geq 1$, $\tilde{g} \in C^q$, the expansion

$$\tilde{g}(x) = \tilde{g}(x^*) + (I - \Gamma J(x))(x - x^*) + R(x - x^*)$$

holds, where R satisfies $\lim_{x \to x^*} \|R(x - x^*)\|/\|x - x^*\| = 0$ so that given $c > 0$, there exists an $r > 0$ such that $\|R(x - x^*)\| \leq c\|x - x^*\|$ for all $x \in B_r(x^*)$.

Proposition 4. *Suppose that $\|I - \Gamma J(x)\| < 1$ for all $x \in B_{r_0}(x^*) \subset \mathcal{V}(x^*)$ so that there exists r', r'' such that $\|I - \Gamma J(x)\| \leq r' < r'' < 1$ for all $x \in B_{r_0}(x^*)$. For $1 - r'' > 0$, let $0 < r < \infty$ be the largest r such that $\|R(x - x^*)\| \leq (1 - r'')\|x - x^*\|$ for all $x \in B_r(x^*)$. Furthermore, let $x_0 \in B_{r^*}(x^*)$, where $r^* = \min\{r, r_0\}$, be arbitrary. Then, given $\varepsilon > 0$, gradient-based learning with learning rates Γ obtains an ε–differential Nash equilibrium in finite time—i.e., $x_k \in B_\varepsilon(x^*)$ for all $k \geq T = \lceil \frac{1}{\delta} \log(r^*/\varepsilon) \rceil$ where $\delta = r'' - r'$.*

The proof follows the proof of Theorem 1 in (Argyros, 1999) with a few minor modifications.

Remark 1. *We note that the proposition can be more generally stated with the assumption that $\rho(I - \Gamma J(x)) < 1$, in which case there exists some δ defined in terms of bounds on powers of $I - \Gamma J$. We also note that these results hold even if Γ is not a diagonal matrix as we have assumed as long as $\rho(I - \Gamma J(x)) < 1$.*

A perhaps more interpretable finite bound stated in terms of the game structure can also be obtained. Consider the case in which players adopt learning rates $\gamma_i = \sqrt{\alpha}/(\beta k_i)$ with $k_i \geq 1$. Given a stable differential

Nash equilibrium x^*, let $B_r(x^*)$ be the largest ball of radius r contained in the region of attraction on which $\tilde{S} \equiv \frac{1}{2}(\tilde{J}^T + \tilde{J})$ is positive definite where $\tilde{\omega} = (D_i f_i / k_i)_{i \in \mathcal{I}}$ so that $\tilde{J} \equiv D\tilde{\omega}$, and define

$$\tilde{\alpha} = \min_{x \in B_r(x^*)} \lambda_d\big(\tilde{S}(x)^T \tilde{S}(x)\big)$$

and

$$\tilde{\beta} = \max_{x \in B_r(x^*)} \lambda_1(\tilde{J}(x)^T \tilde{J}(x)).$$

Given a stable differential Nash equilibrium x^*, let $B_r(x^*)$ be the largest ball contained in the region of attraction $\mathcal{V}(x^*)$ on which $S^T S$ is positive definite—i.e., $\sqrt{\alpha} > 0$.

Theorem 2. *Suppose that Assumption 1 holds and that $x^* \in X$ is a stable differential Nash equilibrium. Let $x_0 \in B_r(x^*)$, $\alpha < k_{\min}\beta$, $\sqrt{\alpha}/k_{\min} \leq \sqrt{\tilde{\alpha}}$, and for each i, $\gamma_i = \sqrt{\alpha}/(\beta k_i)$ with $k_i \geq 1$. Then, given $\varepsilon > 0$, the gradient-based learning dynamics with learning rates γ_i obtain an ε–differential Nash such that $x_k \in B_\varepsilon(x^*)$ for all*

$$k \geq \left\lceil 2\frac{\beta k_{\min}}{\alpha} \log\left(\frac{r}{\varepsilon}\right)\right\rceil.$$

Proof. First, note that $\|x_{k+1} - x^*\| = \|\tilde{g}(x_k) - \tilde{g}(x^*)\|$ where $\tilde{g}(x) = x - \Gamma\omega(x)$. Now, given $x_0 \in B_r(x^*)$, by the mean value theorem,

$$\|\tilde{g}(x_0) - \tilde{g}(x^*)\| = \| \textstyle\int_0^1 D\tilde{g}(\tau x_0 + (1-\tau)x^*)(x_0 - x^*)d\tau\| \leq \sup_{x \in B_r(x^*)} \|D\tilde{g}(x)\|\|x_0 - x^*\|.$$

Hence, it suffices to show that for the choice of Γ, the eigenvalues of $I - \Gamma J(x)$ live in the unit circle. Then an inductive argument can be made with the inductive hypothesis that $x_k \in B_r(x^*)$. Let $\Lambda = \operatorname{diag}(1/k_1, \ldots, 1/k_n)$. Then we need to show that $I - \gamma\Lambda J$ has eigenvalues in the unit circle. Since $\omega(x^*) = 0$, we have that

$$\|x_{k+1} - x^*\|_2 = \|x_k - x^* - \gamma\Lambda(\omega(x_k) - \omega(x^*))\|_2 \leq \sup_{x \in B_r(x^*)} \|I - \gamma\Lambda J(x)\|_2 \|x_k - x^*\|_2.$$

If $\sup_{x \in B_r(x^*)} \|I - \gamma\Lambda J(x)\|_2$ is less than one, where the norm is the operator 2–norm, then the dynamics are contracting. For notational convenience, we drop the explicit dependence on x. Then,

$$\begin{aligned}(I - \gamma\Lambda J)^T(I - \gamma\Lambda J) \leq (1 - 2\gamma\lambda_d(\tilde{S}) + \tfrac{\gamma^2\lambda_1(J^TJ)}{k_{\min}^2})I &\leq (1 - 2\gamma\sqrt{\alpha}/k_{\min} + \alpha/(\beta k_{\min}))I \\ &= (1 - \alpha/(\beta k_{\min}))I.\end{aligned}$$

The first inequality holds since $\lambda_1(J^TJ/k_{\min}^2) \geq \lambda_1(J^T\Lambda^2 J)$. Indeed, first observe that the singular values of $\Lambda J^T J\Lambda$ are the same as those of $J^T\Lambda^2 J$ since the latter is positive definite symmetric. Thus, by noting that $\|A\|_2 = \sigma_{\max}(A)$ and employing Cauchy-Schwartz, we get that $\|\Lambda\|_2^2\|J^TJ\|_2 \geq \|\Lambda J^TJ\Lambda\|_2$ and hence, the inequality.

Using the above to bound $\sup_{x\in B_r(x^*)}\|I-\gamma\Lambda J(x)\|_2$, we have $\|x_{k+1}-x^*\|_2 \leq (1-\frac{\alpha}{\beta k_{\min}})^{1/2}\|x_k-x^*\|_2$. Since $\alpha < k_{\min}\beta$, $(1-\alpha/(\beta k_{\min})) < e^{-\alpha/(\beta k_{\min})}$ so that $\|x_{k+1}-x^*\|_2 \leq e^{-T\alpha/(2k_{\min}\beta)}\|x_0-x^*\|_2$. This, in turn, implies that $x_k \in B_\varepsilon(x^*)$ for all $k \geq T = \lceil 2\frac{\beta k_{\min}}{\alpha}\log(r/\varepsilon)\rceil$.

□

Multiple learning rates lead to a scaling rows which can have a significant effect on the eigenstructure of the matrix, thereby making the relationship between ΓJ and J difficult to reason about. None-the-less, there are numerous approaches to solving nonlinear systems of equations (or differential equations expressed as a set of nonlinear system of equations) that employ *preconditioning* (i.e., coordinate scaling). The purpose of using a preconditioning matrix is to rescale the problem and achieve better or faster convergence. Many of these results directly translate to convergence guarantees for learning in games when the learning rates are not uniform; however, in the case of understanding convergence properties for autonomous agents learning an equilibrium—as opposed to computing an equilibrium—the *preconditioner* is not subject to design. Perhaps this reveals an interesting direction of future research in terms of synthesizing games or learning rules via incentivization or otherwise exogenous control policies for either coordinating agents or improving the learning process—e.g., using incentives to induce a particular equilibrium while also encouraging faster learning.

2.4 Stochastic Setting

In this section, we consider gradient-based learning rules for each agent where the agent does not have oracle access to their individual gradients, but rather has an unbiased estimator in its place. In particular, for each player $i \in \mathcal{I}$, consider the noisy gradient-based learning rule given by

$$x_{i,k+1} = x_{i,k} - \gamma_{i,k}(\omega(x_k) + w_{i,k+1}) \tag{2.3}$$

where $\gamma_{i,k}$ is the learning rate and $w_{i,k}$ is an independent identically distributed stochastic process. In order to prove a high-probability, finite sample convergence rate, we can leverage recent results for convergence of nonlinear stochastic approximation algorithms. The key is in formulating the the learning rule for the agents and in leveraging the notion of a stable differential Nash equilibrium which has analogous properties as a locally stable equilibrium for a nonlinear dynamical system. Making the link between the discrete time learning update and the limiting continuous time differential equation and its equilibria allows us to draw on rich existing convergence analysis tools.

In the first part of this section, we provide convergence rate results for the case where the agents use a *uniform learning rate*—i.e. $\gamma_{i,k} \equiv \gamma_k$. In the second part of this section, we extend these results to the case where agents use *non-uniform learning rates*—that is, each agent has its own learning rate $\gamma_{i,k}$—by incorporating some additional assumptions and leveraging two-timescale analysis techniques from dynamical systems theory.

We require some modified assumptions in this section on the learning process structure.

Assumption 2. *The gradient-based learning rule* (2.3) *satisfies the following:*

A2a. *Given the filtration $\mathcal{F}_k = \sigma(x_s, w_{1,s}, w_{2,s}, s \le k)$, $\{w_{i,k+1}\}_{i\in\mathcal{I}}$ are conditionally independent. Moreovoer, for each $i \in \mathcal{I}$, $\mathbb{E}[w_{i,k+1}|\ \mathcal{F}_k] = 0$ almost surely (a.s.), and $\mathbb{E}[\|w_{i,k+1}\||\ \mathcal{F}_k] \le c_i(1 + \|x_k\|)$ a.s. for some constants $c_i \ge 0$.*

A2b. *For each $i \in \mathcal{I}$, the stepsize sequence $\{\gamma_{i,k}\}_k$ contain positive scalars such that*

(a) $\sum_i \sum_k \gamma_{i,k}^2 < \infty$;

(b) $\sum_k \gamma_{i,k} = \infty$;

(c) and, $\gamma_{2,k} = o(\gamma_{1,k})$.

A2c. *Each $f_i \in C^q(\mathbb{R}^d, \mathbb{R})$ for some $q \ge 3$ and each f_i and ω are L_i– and L_ω–Lipschitz, respectively.*

2.4.1 Uniform Learning Rates

Before concluding, we specialize to the case in which agents have the same learning rate sequence $\gamma_{i,k} = \gamma_k$ for each $i \in \mathcal{I}$.

Theorem 3. *Suppose that x^* is a stable differential Nash equilibrium of the game $(f_1, \ldots, f_n)$ and that Assumption 2 holds (excluding A2b.iii). For each k, let $k_0 \ge 0$ and*

$$\zeta_k = \max_{k_0 \le s \le k-1} \big(\exp(-\lambda \textstyle\sum_{\ell=s+1}^{k-1} \gamma_\ell\big)\gamma_s.$$

Fix any $\varepsilon > 0$ such that $B_\varepsilon(x^*) \subset B_r(x^*) \subset \mathcal{V}$ where $\mathcal{V}$ is the region of attraction of x^*. There exists constants $C_1, C_2 > 0$ and functions $h_1(\varepsilon) = O(\log(1/\varepsilon))$ and $h_2(\varepsilon) = O(1/\varepsilon)$ so that whenever $T \ge h_1(\varepsilon)$ and $k_0 \ge N$, where N is such that $1/\gamma_k \ge h_2(\varepsilon)$ for all $k \ge N$, the samples generated by the gradient-based learning rule *satisfy*

$$\begin{aligned}&\Pr\left(\bar{x}(t) \in B_\varepsilon(x^*)\ \forall t \geq t_{k_0} + T + 1 |\ \bar{x}(t_{k_0}) \in B_r(x^*)\right)\\ &\qquad \geq 1 - \textstyle\sum_{s=k_0}^{\infty}\left(C_1\exp(-C_2\varepsilon^{1/2}/\gamma_s^{1/2}) + C_1\exp(-C_2\min\{\varepsilon,\varepsilon^2\}/\zeta_s)\right)\end{aligned}$$

where the constants depend only on parameters λ, r, τ_L and the dimension $d = \sum_i d_i$. Then stochastic gradient-based learning in games obtains an ε–stable differential Nash x^* in finite time with high probability.

The above theorem implies that $x_k \in B_\varepsilon(x^*)$ for all $k \geq k_0 + \lceil \log(4\tilde{K}/\varepsilon)\lambda^{-1}\rceil + 1$ with high probability for some constant $\tilde{K}$ that depends only on λ, r, τ_L, and d.

Proof. Since x^* is a stable differential Nash equilibrium, $J(x^*)$ is positive definite and $D_i^2 f_i(x^*)$ is positive definite for each $i \in \mathcal{I}$. Thus x^* is a locally asymptotically stable hyperbolic equilibrium point of $\dot{x} = -\omega(x)$. Hence, the assumptions of Theorem 1.1 (Thoppe and Borkar, 2019) are satisfied so that we can invoke the result which gives us the high probability bound for stochastic gradient-based learning in games. □

The above theorem has a direct corollary specializing to the case where the gradient-based learning rule with uniform stepsizes is initialized inside a ball of radius r constained in the region of attraction—i.e., $B_r(x^*) \subset \mathcal{V}$.

Corollary 1. *Let x^* be a stable differential Nash equilibrium of $(f_1, \ldots, f_n)$ and suppose that Assumption 2 holds (excluding A2b.iii). Fix any $\varepsilon > 0$ such that $B_\varepsilon(x^*) \subset B_r(x^*) \subset \mathcal{V}$. Let ζ_k, T, and $h_2(\varepsilon)$ be defined as in Theorem 3. Suppose that $1/\gamma_k \geq h_2(\varepsilon)$ for all $k \geq 0$ and that $x_0 \in B_r(x^*)$. Then, with $C_1, C_2 > 0$ as in Theorem 3,*

$$\begin{aligned}&\Pr\left(\bar{x}(t) \in B_\varepsilon(x^*)\ \forall t \geq T + 1 |\ \bar{x}(t_{k_0}) \in B_r(x^*)\right)\\ &\qquad \geq 1 - \textstyle\sum_{s=0}^{\infty}\left(C_1\exp(-C_2\varepsilon^{1/2}/\gamma_s^{1/2}) + C_1\exp(-C_2\min\{\varepsilon,\varepsilon^2\}/\zeta_s)\right).\end{aligned}$$

2.4.2 Non-Uniform Learning Rates

Consider now that agents have their own learning rates $\gamma_{i,k}$ for each $i \in \mathcal{I}$. In environments with several autonomous agents, as compared to the objective of *computing* Nash equilibria in a game, it is perhaps more reasonable to consider the scenario in which the agents have their own individual learning rate. For the sake of brevity, we show the convergence result in detail for the two agent case—that is, where $\mathcal{I} = \{1, 2\}$. We note that the extension to n agents is straightforward. The proof leverages recent results from the theory of stochastic approximation presented in (Borkar and Pattathil, 2018) and we note that our objective here is to show that they apply to games and provide commentary on the interpretation of the results in this context.

The gradient-based learning rules are given by

$$x_{i,k+1} = x_{i,k} - \gamma_{i,k}(\omega(x_k) + w_{i,k+1}) \tag{2.4}$$

so that with $\gamma_{2,k} = o(\gamma_{1,k})$, in the limit $\tau \to 0$, the above system can be thought of as approximating the singularly perturbed system

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = -\begin{bmatrix} D_1 f_1(x_1(t), x_2(t)) \\ \tau D_2 f_2(x_1(t), x_2(t)) \end{bmatrix} \tag{2.5}$$

Indeed, since $\lim_{k\to\infty} \gamma_{2,k}/\gamma_{1,k} \to 0$—i.e., $\gamma_{2,k} \to 0$ at a faster rate than $\gamma_{1,k}$—updates to x_1 appear to be equilibriated for the current quasi-static x_2 as the dynamics in (2.5) suggest.

Asymptotic Convergence in the Non-Uniform Learning Rate Setting

Assumption 3. *For fixed $x_2 \in X_2$, the system $\dot{x}_1(t) = -D_1 f_1(x_1(t), x_2)$ has a globally asymptotically stable equilibrium $\lambda(x_2)$.*

Lemma 1. *Under Assumptions 2 and 3, conditioned on the event $\{\sup_k \sum_i \|x_{i,k}\|_2 < \infty\}$, $(x_{1,k}, x_{2,k}) \to \{(\lambda(x_2), x_2) |\ x_2 \in \mathbb{R}^{d_2}\}$ almost surely.*

The above lemma follows from classical analysis (see, e.g., Borkar (2008, Chapter 6) or Bhatnagar and Prasad (2013, Chapter 3)).

Define the continuous time accumulated after k samples of x_2 to be $t_k = \sum_{l=0}^{k-1} \gamma_{2,k}$ and define $x_2(t, s, x_s)$ for $t \geq s$ to be the trajectory of $\dot{x}_2 = -D_2 f_2(\lambda(x_2), x_2)$. Furthermore, define the event $\mathcal{E} = \{\sup_k \sum_i \|x_{i,k}\|_2 < \infty\}$.

Theorem 4. *Suppose that Assumptions 2 and 3 hold. For any $K > 0$, conditioned on $\mathcal{E}$,*

$$\lim_{k\to\infty} \sup_{0\leq h\leq K} \|x_{2,k+h} - x_2(t_{k+h}, t_k, x_k)\|_2 = 0.$$

Proof. The proof invokes Lemma 1 above and Proposition 4.1 and 4.2 of (Benaïm, 1999). Indeed, by Lemma 1,

$(\lambda(x_{2,k}) - x_{2,k}) \to 0$ almost surely. Hence, we can study the sample path generated by

$$x_{2,k+1} = x_{2,k} - \gamma_{2,k}(D_2 f_2(\lambda(x_{2,k}), x_{2,k}) + w_{2,k+1}).$$

Since $D_2 f_2 \in C^{q-1}$ for some $q \geq 3$, it is locally Lipschitz and, on the event $\{\sup_k \sum_i \|x_{i,k}\|_2 < \infty\}$, it is bounded. It thus induces a continuous globally integrable vector field, and therefore satisfies the assumptions of Proposition 4.1 of (Benaïm, 1999). Moreover, under Assumption 2, the assumptions of Proposition 4.2 of (Benaïm, 1999) are satisfied. Hence, invoking said propositions, we get the desired result. □

This result essentially says that the slow player's sample path asymptotically tracks the flow of

$$\dot{x}_2 = -D_2 f_2(\lambda(x_2), x_2).$$

If we additionally assume that the slow component also has a global attractor, then the above theorem gives rise to a stronger convergence result.

Assumption 4. *Given $\lambda(\cdot)$ as in Assumption 3, the system $\dot{x}_2(t) = -\tau D_2 f_2(\lambda(x_2(t)), x_2(t))$ has a globally asymptotically stable equilibrium x_2^*.*

Corollary 2. *Under the assumptions of Theorem 4 and Assumption 4, conditioned on the event $\mathcal{E}$, gradient-based learning converges almost surely to a stable attractor (x_1^*, x_2^*), where $x_1^* = \lambda(x_2^*)$, the set of which contains the stable differential Nash equilibria.*

More generally, the process $(x_{1,k}, x_{2,k})$ will converge almost surely to the internally chain transitive set of the limiting dynamics (2.5) and this set contains the stable Nash equilibria. If the only internally chain transitive sets for (2.5) are isolated equilibria (this occurs, e.g., if the game is a potential game), then x_k converges almost surely to a stationary point of the dynamics, a subset of which are stable local Nash equilibria.

It is also worth commenting on what types of games will satisfy these assumptions. To satisfy Assumption 3, it is sufficient for the fastest player's cost function to be convex in their choice variable.

Proposition 5. *Suppose Assumption 2 and 4 hold and that $f_1(\cdot, x_2)$ is convex. Conditioned on the event $\mathcal{E}$, the sample points of gradient-based learning satisfy $(x_{1,k}, x_{2,k}) \to \{(\lambda(x_2), x_2) |\ x_2 \in \mathbb{R}^{d_2}\}$ almost surely. Moreover, $(x_{1,k}, x_{2,k}) \to (x_1^*, x_2^*)$ almost surely, where $x_1^* = \lambda(x_2^*)$.*

Note that (x_1^*, x_2^*) could still be a spurious stable non-Nash point still since the above implies that $D(D_2 f_2(\lambda(\cdot), \cdot))|_{x_2^*} > 0$, which does not necessarily imply that $D_2^2 f_2(\lambda(x_2^*), x_2^*) > 0$.

Remark 2 (Relaxation to Local Asymptotic Stability.)**.** *Under relaxed assumptions on global asymptotic stability, we can obtain high-probability results on convergence to locally asymptotically stable attractors. If it is assumed that x_0 is in the region of attraction for a locally asymptotically stable attractor, then the above results can be stated with only the assumption of a locally asymptotic stability. However, this is difficult to ensure in practice. To relax the result to a local guarantee regardless of the initialization requires conditioning on an unverifiable event—i.e., the high-probability bound in this case is conditioned on the event $\{\{x_{1,k}\}$ belongs to a compact set B, which depends on the sample point, of $\cap_{x_2}\mathcal{R}(\lambda(x_2))\}$ where $\mathcal{R}(\lambda(x_2))$ is the region of attraction of $\lambda(x_2)$. None-the-less, it is possible to leverage results from stochastic approximation (Karmakar and Bhatnagar, 2018), (Borkar, 2008, Chapter 2) to prove local versions of the results for non-uniform learning rates. Further investigation is required to provide concentration bounds for not only games but stochastic approximation in general.*

2.5 Numerical Examples

The results in the preceding sections provide convergence guarantees for a class of gradient-based learning algorithms to a neighborhood of a stable Nash equilibrium under deterministic and stochastic gradient-based update rules with both uniform and non-uniform learning rates. In this section, we present several numerical examples that validate these theoretical results and highlight interesting aspects of learning in multi-agent settings.

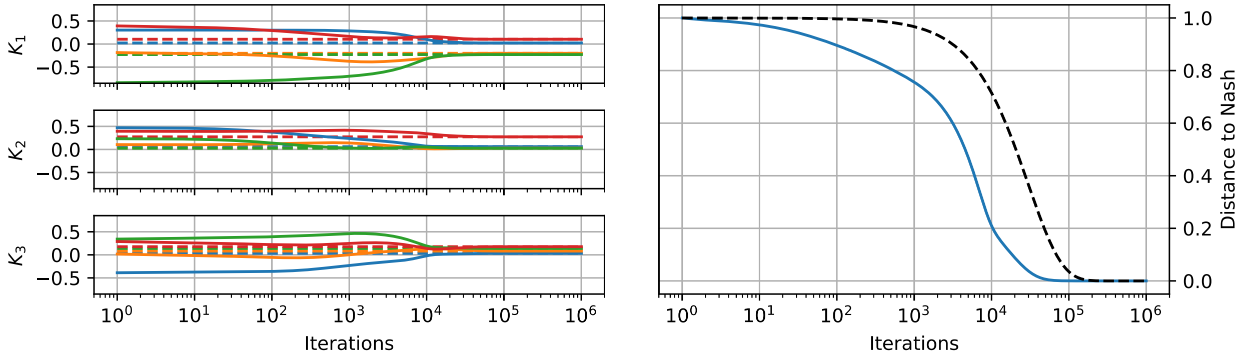
2.5.1 Deterministic Policy Gradient in Linear Quadratic Dynamic Games

The first example we explore is a linear quadratic (LQ) game with three players in the space of linear feedback policies. This game serves as a useful benchmark since it has a unique global equilibrium that we can compute via a set of coupled algebraic Riccati equations (Başar and Olsder, 1998). The gradient-based learning rule for each of the agents is a multi-agent version of policy gradient in which agents have oracle access to their gradients at each iteration.

Consider a four state discrete time linear dynamical system,

$$z(t+1) = Az(t) + B_1u_1(t) + B_2u_2(t) + B_3u_3(t)$$

where $z(t) \in \mathbb{R}^4$ and, for each $i \in \{1, 2, 3\}$, $u_i(t) \in \mathbb{R}$ is the control for player i. The policy for each player is parameterized by a linear feedback gain matrix, $u_i(t) = -K_i z(t)$. Moreover, each player seeks to minimize a



(a)

(b)

Figure 2.1: Convergence of policy gradient in LQ dynamic games to the Nash policy. (a) Each player's linear feedback gain matrix K_i converges to the unique Nash policies (dotted lines). (b) The black dashed line shows upper bound of the number of iterations required to converge within ε distance from Nash (2-norm). The actual convergence for this random initialization is shown as the solid line.

quadratic cost

$$f_i(K_i, K_{-i}) = \mathbb{E}_{z_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\left(z(t)^T Q_i z(t) + \textstyle\sum_{j=1}^{n} u_j(t)^T R_{ij} u_j(t)\right)\right]$$

which is a function of the coupled state variable $z(t)$, their own control $u_i(t)$ and all other agents' control $u_{-i}(t)$ over an infinite time horizon. In an effort to learn a Nash equilibrium, each agent employs policy gradient. In particular, they update their feedback policy via

$$K_i(t+1) = K_i(t) - \gamma_i \nabla_{K_i} f_i(K_i, K_{-i}).$$

It is fairly straightforward to compute the gradient of f_i with respect to K_i, the feedback gain that parameterizes player i's control input u_i. Indeed,

$$\nabla_{K_i} f_i(K_i, K_{-i}) = 2(R_{ii}K_i - B_i^T P_i \widetilde{A})\Sigma_K$$

where

$$\Sigma_K = \mathbb{E}_{z_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} z(t)z(t)^T\right].$$

Hence, the collection of the agents' individual gradients is given by

$$\omega(K_1, K_2, K_3) = \left(2(R_{ii}K_i - B_i^T P_i \widetilde{A})\Sigma_K\right)_{i=1}^{3}$$

Remark 3. *Note that ω can be zero at critical points or at points where $\sum_{t=0}^{\infty} z(t)z(t)^T$ drops rank. To prevent the latter possibility, we sample the initial condition from a distribution. That is, we take $z_0 \sim \mathcal{D}$ so that*

$\mathbb{E}_{z_0 \sim \mathcal{D}} z_0 z_0^T$ *is full rank.*

For a given joint policy (K_1, K_2, K_3), the closed loop dynamics are $\widetilde{A} = A - B_1K_1 - B_2K_2 - B_3K_3$. The states $z(t)$ are obtained from simulating the system. For each i, the Riccati matrix P_i is computed by solving the Riccati equation

$$P_i = \widetilde{A}^T P_i \widetilde{A} + Q_i + \sum_{j=1}^{n} K_j R_{ij} K_j.$$

Note that this Riccati equation is only used to compute the gradient of the cost functions with respect to a specific set of feedback gains.

For the purpose of validating convergence, we can compute the Nash policies (K_1^*, K_2^*, K_3^*) by an established method with coupled Riccati equations. We use the learning rate $\gamma_i = \gamma$ defined as in Theorem 1. To compute γ we first compute the game Jacobian $J(K_1^*,\ K_2^*,\ K_3^*)$ at the Nash feedback gains and then find the maximum eigenvalue of J^TJ and minimum eigenvalue of $(J^T+J)^T(J^T+J)$ in a neighborhood of (K_1^*, K_2^*, K_3^*) to determine the constants α and β as defined in Section 2.3.

Figure 2.1 shows the convergence of the gradient updates to the Nash policies. The K_i are randomly initialized in a neighborhood of the known Nash equilibrium and such that $\tilde{A}$ is stable. The number of iterations required to converge to an ε–differential Nash is bounded by the dashed black line in Figure 2.1b, which shows the curve of (ε, T) pairs determined by Theorem 1. However, this learning rate is not optimal, as choosing a larger γ will result in faster convergence as empirically observed.

Remark 4 (Stochastic Policy Gradient)**.** *We note that stochastic policy gradient with an unbiased estimator has similar convergence properties. Here, e.g., the state dynamics may be subject to zero-mean, finite-variance noise. As long as the estimator for the gradient is unbiased, the theoretical guarantees of the proceeding sections apply.*

2.5.2 Benchmark: Matching Pennies

The next example is again a multi-agent policy gradient example in which there are two players playing 'matching pennies', a classic bimatrix game in which agents have zero-sum costs associated with the matrices (A, B) defined as follows:

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

In particular, the players aim to minimize their respective costs $f_1(x, y) = \pi(y)^T A \pi(x)$ and $f_2(x, y) = \pi(x)^T B \pi(y)$ where $\pi(x)$ is player 1's policy and $\pi(y)$ is player 2's policy. The class of policies the agents are

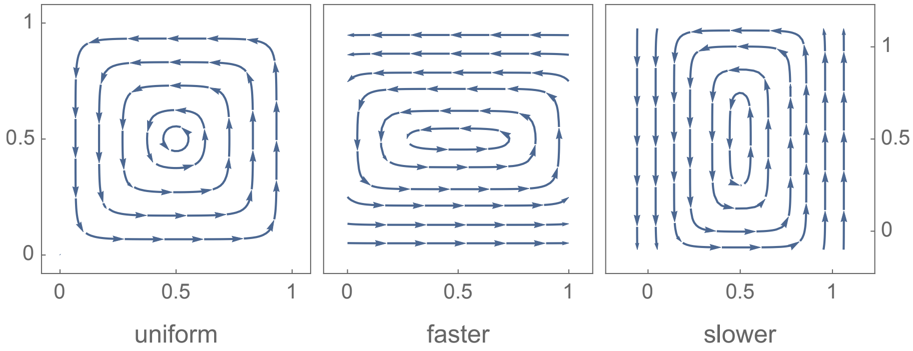


Figure 2.2: Gradient dynamics of the matching pennies game where agents learning have different learning rates. The vector field of the gradient dynamics are stretched along the faster agent's coordinate.

optimizing over are the so-called 'softmax' policies defined by

$$\pi(z) = \left[\frac{e^{10z}}{e^{10z} + e^{10(1-z)}}, \quad \frac{e^{10(1-z)}}{e^{10z} + e^{10(1-z)}}\right],$$

and the update each player employs is a 'smoothed best-response' which in essence is a policy gradient update with respect to the softmax parameter and each agents individual cost. This game has been well studied in the game theory literature and we use it illustrate the fact that non-uniform learning rates result in a warping of the vector field associated with the agents' learning dynamics.

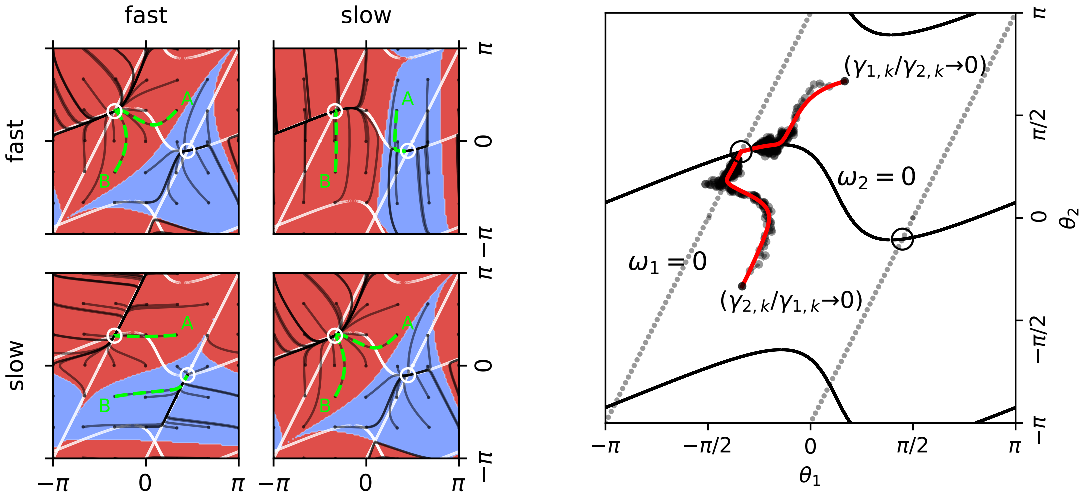
The mixed Nash equilibrium for this game is $(x^*,\ y^*) = (0.5, 0.5)$, but the Jacobian of the gradient dynamics at this fixed point is

$$J(x^*,\ y^*) = \begin{bmatrix} 0 & 100 \\ -100 & 0 \end{bmatrix}$$

so that it has purely imaginary eigenvalues ±$100i$, and therefore admits a limit cycle. Regardless, we can visualize the effects of non-uniform learning rates to the gradient dynamics in Figure 2.2. We notice that the gradient flow stretches along the axes of the faster agent (the agent with a larger learning rate), and the fixed points of these dynamics remain constant.

2.5.3 Exploring the Effects of Non-Uniform Learning Rates on the Learning Path

The examples presented so far all consider convergence (or non-convergence) to a single equilibrium. In the following two examples, we investigate the effect of non-uniform learning rates for more general non-convex settings in which there are multiple equilibria. The following example is a two-player game in which the agents' joint strategy space is a torus. That is, each player's strategy space is the unit circle $\mathbb{S}^1$. For each



(a)

(b)

Figure 2.3: The effects of non-uniform learning rates on the path of convergence to the equilibria. The zero lines for each player ($D_1 f_1 = 0$ or $D_2 f_2 = 0$) are plotted as the diagonal and curved lines, and the two stable Nash equilibria as circles (where $D_1^2 f_1 > 0$ and $D_2^2 f_2 > 0$). (a) In the deterministic setting, the region of attractions for each equilibrium can be computed numerically. Four scenarios are shown, with a combination of fast and slow agents. The region of attractions for each Nash equilibrium are warped under different learning rates. (b) In the stochastic setting, the samples (in black) approximate the singularly perturbed differential equation (in red). Two initializations and learning rate configurations are plotted.

$i \in \{1,2\}$, player i has cost $f_i : \mathbb{S}^1 \times \mathbb{S}^1 \to \mathbb{R}$ given by

$$f_i(\theta_i, \theta_{-i}) = -\alpha_i \cos(\theta_i - \phi_i) + \cos(\theta_i - \theta_{-i})$$

where α_i and ϕ_i are constants, and θ_i is player i's choice variable. An interpretation of this game is that of a ‘location game’ in which each player wishes to be near location ϕ_i but far from each other. This game has many applications including those which abstract nicely to coupled oscillators.

The game form—i.e., collection of individual gradients—is given by

$$\omega(\theta_1, \theta_2) = \begin{bmatrix} \alpha_1 \sin(\theta_1 - \phi_1) - \sin(\theta_1 - \theta_2) \\ \alpha_2 \sin(\theta_2 - \phi_2) - \sin(\theta_2 - \theta_1) \end{bmatrix}, \tag{2.6}$$

and the game Jacobian is composed of terms $\alpha_i \cos(\theta_i - \phi_i) - \cos(\theta_i - \theta_{-i})$, $i = 1, 2$ on the diagonal and $\cos(\theta_i - \theta_{-i})$, $i = 1, 2$ on the off-diagonal.

The Nash equilibria of this game occur where $\omega(\theta_1, \theta_2) = 0$ and where the diagonals of the game Jacobian are positive. The game has multiple Nash equilibria. We visualize the warping of the region of attraction of these equilibria under different learning rates, and the affinity of the “faster” player to its own zero line.

In this example, we use constants $\phi = (0,\ \pi/8)$ and $\alpha = (1.0,\ 1.5)$. The joint strategy space can be viewed as a non-convex smooth manifold via an equivalence relationship, or equivalently, as players choosing $\theta_i \in \mathbb{R}$. There are two Nash equilibria, situated at $(-1.063,\ 1.014)$ and $(1.408,\ -0.325)$. These equilibria happen to also be stable differential Nash, and thus we expect the gradient dynamics to converge to them if initialized in the region of attraction. Which equilibrium it converges to, however, depends on the initialization and learning rates of agents.

To investigate how non-uniform learning rates affect the agents' convergence to the two equilibria, we simulate agents learning at different rates, one fast and one slow. The fast agent's learning rate is set to $\gamma_1 = 0.171$ and the slow $\gamma_2 = 0.017$. Figure 2.3a shows the trajectory of agents' learned strategies. Each of the four squares depicts the full strategy space on the torus from $-\pi$ to π for both agents' actions, with θ_1 on the x-axis and θ_2 on the y-axis. The labels "fast" and "slow" indicate the learning rate of the corresponding agent. For example, in the bottom left square, agent 1 is the fast agent and agent 2 is the slow agent. Hence, the non-uniform update equation for that square becomes $\theta_{k+1} = \theta_k - \text{diag}(\gamma_1, \gamma_2)\omega(\theta_k)$.

The white lines indicate the points x such that $\omega_i(x) = 0$, and the intersection of the white lines indicate points x such that $\omega(x) = 0$. The two intersections marked as circles are the stable differential Nash equilibria. The unmarked intersections are either saddle points or other unstable equilibria. The black lines show different paths of the update equations under the non-uniform update equation, with initial points selected from a equally spaced 7×7 grid. We highlight two paths in green (labeled A and B) which begin at $(\pi/3,\ \pi/3)$ and $(-\pi/3, -\pi/3)$.

In the case where agents both learn at the same rate, (γ_1, γ_1) and (γ_2, γ_2), paths A and B both converge to the Nash equilibrium at $(-1.063,\ 1.014)$. However, when agents learn at different rates, the equilibrium to which the agents converge to, as well as the learning path, is no longer the same even starting at the same initial points. This phenomena can also be captured by displaying the region of attraction for both Nash equilibria. The red region corresponds to initializations that will converge to the equilibrium contained in the red region (again indicated by a white circle). Analogously, the blue region corresponds to the region of attraction of the other equilibria.

To provide an example of the stochastic setting in which agents have an unbiased estimator of their individual gradients, we choose learning rates according to Assumption 2. In particular, we choose scaled learning rates $\gamma_{2,k} = \frac{1}{1+k\log(k+1)}$ and $\gamma_{1,k} = \frac{1}{1+k}$ such that $\gamma_{2,k}/\gamma_{1,k} \to 0$ as $k \to \infty$. Figure 2.3b shows the learning paths in this setting initialized at two different points, each with flipped learning rate configurations. The sample points approximate the singularly perturbed differential equation (shown in red) described in Section 2.4.2.

In both deterministic and stochastic settings, we observe the affinity of the faster agent to its own zero line. For example, the bottom left square (in Figure 2.3a) and bottom left path (in Figure 2.3b) both have agent 1 as the faster agent, and the learning paths both tend to arrive to the line $\omega_1 \equiv 0$ before finally converging to the Nash equilibrium. An interpretation of this is that the faster agent tries to be situated at the bottom of the "valley" of its own cost function. The faster agent tends to be at its *own* minimum while it waits for the slower agent to change its strategy. As a Stackelberg interpretation, where there are followers and leaders, the slower agent would be the leader and faster agent the follower. In a sense, the slower agent has an advantage.

2.5.4 Multi-Agent Control and Collision Avoidance

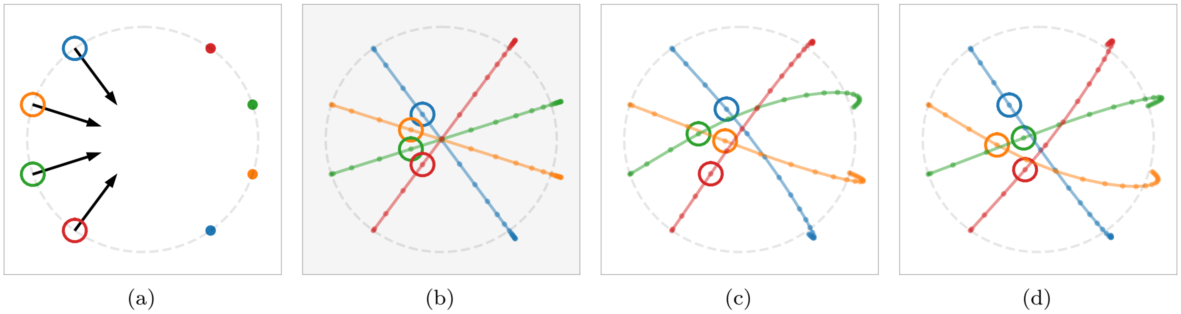


Figure 2.4: Minimum-fuel particle avoidance control example. (a) Each particle seeks to reach the opposite side of the circle using minimum fuel while avoiding each other. The circles represent the approximate boundaries around each particle at time $t = 5$. (b) The joint strategy $x = (\mathbf{u}_1, \cdots, \mathbf{u}_4)$ is initialized to the minimum fuel solution ignoring interaction between particles. (c) Equilibrium solution achieved by setting the blue agent to have a slower learning rate. (d) Another equilibrium, where the red agent has the slower learning rate.

The final example presents a practical use case for the gradient-based update. Consider a non-cooperative game between four collision-avoiding agents where they seek to arrive at a destination with minimum fuel while avoiding each other. We show that the scaling between agents' learning rates dictates the equilibrium solution to which they converges. This can be useful in designing non-cooperative open-loop controllers where agents may choose to learn slower in order to deviate less from their initial plan, perhaps in an attempt to incur less 'risk'.

Suppose there are four collision-avoiding particles traversing across a unit circle. Each particle follows discrete-time linear dynamics

$$z_i(t+1) = Az_i(t) + Bu_i(t)$$

for $t = 1, \cdots, N$ where

$$A = \begin{bmatrix} I & hI \\ 0 & I \end{bmatrix} \in \mathbb{R}^{4\times 4}, \ B = \begin{bmatrix} h^2 I \\ hI \end{bmatrix} \in \mathbb{R}^{4\times 2},$$

I is the identity matrix, and $h = 0.1$. These dynamics represent a typical discretized version of the continuous dynamics $\ddot{r}_i = u_i$ in which $u_i \in \mathbb{R}^2$ represents a force vector used to accelerate the particle, and the state

$z_i = [r_i, \dot{r}_i]$ represents the particles position and velocity. Let $\mathbf{u}_i$ be the concatenated vector of control vectors for player i for all time—i.e., $\mathbf{u}_i = (u_i(1), \cdots, u_i(N))$ and let $\mathbf{u} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Each particle i aims to minimize a cost defined by

$$J_i(\mathbf{u}) = \sum_{t=1}^{N} \|u_i(t)\|_R^2 + \sum_{t=1}^{N+1} \|z_i(t) - \bar{z}_i\|_Q^2 + \sum_{j \neq i} \sum_{t=1}^{N+1} \rho e^{-\sigma \|z_i(t) - z_j(t)\|_S^2}$$

where $\|\cdot\|_P$ denotes the quadratic norm—i.e., $\|z\|_P^2 = z^T P z$ with P positive semi-definite. The first two terms of the cost correspond to the minimum fuel objective and quadratic cost from desired final state $\bar{z}_i$, a typical setup for optimal control problems. We use $R = \mathrm{diag}(0.1, 0.1)$ and $Q = \mathrm{diag}(1, 1, 0, 0)$. The final term of the cost function is the sum of all pairwise interaction terms between the particles, modeled after the shape of a Gaussian which encodes smooth boundaries around the particles. We use constants $\rho = 10$ and $\sigma = 100$.

Figure 2.4 (a) visualizes the problem setup. Each particle's initial position $z_i(0)$ is located on the left side of a unit circle; they are separated by $\pi/5$, and their desired final positions, $\bar{z}_i$ for each $i \in \{1, \ldots, 4\}$, are located directly opposite. The particles begin with zero velocity and must solve for a minimum control solution that also avoids collision with other particles as described by the objectives J_i for each i.

To initialize the gradient-based learning algorithms in the game setting, we compute the optimal solution for each agent ignoring the pairwise interaction terms, shown in Figure 2.4 (b). This can be computed using classical discrete-time LQR methods or by gradient descent. Then, using this solution as the intialization for the game setting, each agent descends their own gradient, i.e.

$$\mathbf{u}_{i,k+1} = \mathbf{u}_{i,k} - \gamma_i D_i J_i(\mathbf{u}),$$

with different learning rates γ_i. Just as the previous example shows, the relative learning rates of agents warp the region of attraction for the multiple equilibria. If we allow the red agent to learn slower, then the learning process converges to the equilibria shown in Figure 2.4 (c), whereas if the blue agent learns quicker, then we converge to Figure 2.4 (d). Hence, all else being equal, the learning rates adopted by players greatly impact the equilibrium to which they converge.

2.6 Discussion and Conclusion

We analyze the convergence of gradient-based learning for non-cooperative agents with continuous costs. We leverage existing dynamical systems theory and stochastic approximation literature to provide convergence guarantees for agents that learn myopically—that is, only using information about their own gradient $D_i f_i$ to update their strategy. We provide guarantees for the case where agents are assumed to have oracle access to $D_i f_i$ and the case where they have sufficient information to compute an unbiased estimator. We also study the effects of non-uniform learning rates.

By preconditioning the gradient dynamics by Γ, a diagonal matrix where the diagonals represent the agents' learning rates, we can begin to understand how a changing learning rate relative to others can change the properties of the fixed points of the dynamics. Moreover, players do not know how a change in others' strategies affects its own cost ($D_j f_i$ where $j \neq i$). A possible extension to this paper is to develop update schemes that use this to provide more robust convergence guarantees for full information continuous games. Different learning rates amongst agents also affects the region of attraction of the game, hence starting from the same initial condition, agents may converge to a different equilibria. Agents may use this to their benefit, as shown in the last example. Such insights into the learning behavior of agents will be useful for providing guarantees on the design of control or incentive policies to coordinate agents. We also show through numerical examples that, counterintuitively, if an agent decides to learn slower, a stable differential Nash equilibrium can go unstable, resulting in learning dynamics that do not converge to Nash.

Beyond the the effects of learning rates, there are a number of avenues for future inquiry. For instance, the results as stated apply to continuous games with Euclidean strategy spaces. An interesting avenue to pursue is the study of learning in games where the agents decision spaces are constrained sets or Riemannian manifolds. The latter arises in a number of robotics applications and in this case, the update rule will need to be modified by the appropriately defined retraction such as $x_{k+1} = \exp_{x_k}(\gamma_k(\omega(x_k)))$ (Shah, 2021). The former arises in a variety of applications where the learning rules are abstractions of agents learning in, e.g., physically constrained environments. The update rule in this case will also need to be defined in terms of the appropriate proximal map thereby leading to potentially non-smooth dynamics (Borkar, 2008; Kushner and Yin, 2003) which is even more challenging in the stochastic setting. Yet, such extensions will lead to a framework and set of analysis tools that apply to a broader class of multi-agent learning algorithms.

While we present the work in the context of gradient-based learning in games, there is nothing that precludes the results from applying to update rules in other frameworks. Our results will apply to many other settings where agents myopically update their decision using a process of the form $x_{k+1} = x_k - \Gamma g(x_k)$. In this paper, we consider the special case where $g \equiv [D_1 f_1 \cdots D_n f_n]$. In the stochastic setting, variants of multi-agent Q-learning conform to this setting since Q-learning can be written as a stochastic approximation update.

Finally, as pointed out in (Mazumdar and Ratliff, 2018), not all critical points of the dynamics $\dot{x} = -\omega(x)$ that are attracting are necessarily Nash equilibria; one can see this simply by constructing a Jacobian with positive eigenvalues with at least one $D_i^2 f_i$ with a non-positive eigenvalue. Understanding this phenomena will help us develop computational techniques to avoid them. Recent work has explored this in the context of zero-sum games (Mazumdar et al., 2019), requiring coordination amongst the learning agents. However, when our objective is to study the learning behavior of autonomous agents seeking an equilibrium, an alternative perspective is needed.

Chapter 3

Stability of Gradient-Based Learning in Continuous Games

Abstract

Learning processes in games explain how players grapple with one another in seeking an equilibrium. We study a natural model of learning based on individual gradients in two-player continuous games. In such games, the arguably natural notion of a local equilibrium is a differential Nash equilibrium. However, the set of locally exponentially stable equilibria of the learning dynamics do not necessarily coincide with the set of differential Nash equilibria of the corresponding game. To characterize this gap, we provide formal guarantees for the stability or instability of such fixed points by leveraging the spectrum of the linearized game dynamics. We provide a comprehensive understanding of scalar games and find that equilibria that are both stable and Nash are robust to variations in learning rates.

3.1 Introduction

The study of learning in games is experiencing a resurgence in the control theory (Ratliff et al., 2016; Tang and Li, 2020; Tatarenko and Kamgarpour, 2019), optimization (Mazumdar et al., 2020; Mertikopoulos and Zhou, 2019), and machine learning (Bu et al., 2019; Chasnov et al., 2020d; Fiez et al., 2020; Goodfellow et al., 2014; Metz et al., 2017) communities. Partly driving this resurgence is the prospect for game-theoretic analysis to yield machine learning algorithms that generalize better or are more robust. Towards understanding the optimization landscape in such formulations, dynamical systems theory is emerging as a principal tool for analysis and ultimately synthesis (Balduzzi et al., 2020; Berard et al., 2020; Boone and Piliouras, 2019; Mazumdar et al., 2020; Mertikopoulos et al., 2018). A predominant learning paradigm used across these different domains is gradient-based learning. Updates in large decision spaces can be performed locally with minimal information, while still guaranteeing local convergence in many problems (Chasnov et al., 2020d; Mertikopoulos and Zhou, 2019).

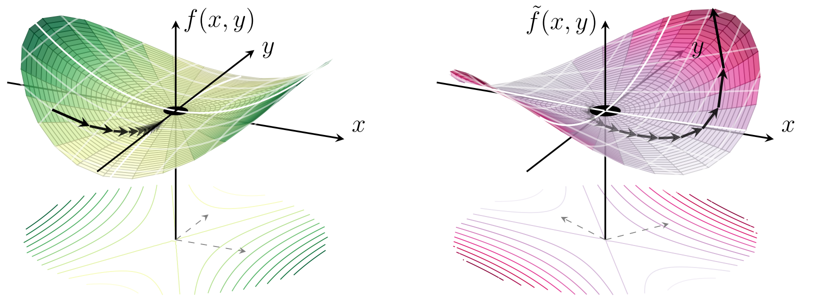
One of the primary means to understand the optimization landscape of games is the eigenstructure and spectrum of the Jacobian of the learning dynamics in a neighborhood of a stationary point. In particular, for a zero-sum continuous game $(f, -f)$ with some continuously-differentiable f, the Nash equilibria are saddle points of the function f. As the example in Fig. 3.1 demonstrates, not all saddle points are relevant. Loosely speaking, the equilibrium conditions for the game correspond to constraints on the curvature directions of the cost function and hence, on the eigenstructure of the Jacobian nearby equilibria.

The local stability of a hyperbolic fixed point in a non-linear system can be assessed by examining the eigenstructure of the linearized dynamics (Khalil, 2002; Sastry, 1999). However, in a game context there are extra constraints coming from the underlying game—that is, players are constrained to move only along directions over which they have control. They can only control their individual actions, as opposed to the entire state of the dynamical system corresponding to the learning rules being applied by the agents. It has been observed in earlier work that not all stable attractors of gradient play are local Nash equilibria and not all local Nash equilibria are stable attractors of gradient play (Mazumdar et al., 2020). Furthermore, changes in players' learning rates—which corresponds to scaling rows of the Jacobian—can change an equilibrium from being stable to unstable and vice versa (Chasnov et al., 2020d).

To summarize, there is a subtle but extremely important difference between game dynamics and traditional nonlinear dynamical systems: alignment conditions are important for distinguishing between equilibria that have game-theoretic meaning versus those which are simply stable attractors of learning rules, and features of learning dynamics such as learning rates can play an important role in shaping not only equilibria but also alignment properties. Motivated by this observation along with the recent resurgence of applications of learning in games in control, optimization, and machine learning, in this paper we provide an in-depth analysis of the spectral properties of gradient-based learning in two-player continuous games.

Contributions. This paper characterizes the spectral properties of structured 2×2 matrices and analyzes the stability of equilibria in continuous games. Having a complete algebraic understanding of the spectrum of the game Jacobian is fundamental to understanding when Nash equilibria coincide with stable equilibria. Many of our results are geometric in nature and are accompanied by diagrams.

It is known that the quadratic numerical range of a block operator matrix contains the operator's (point) spectrum (Tretter, 2008). Thus, it serves as an important tool for quantifying the spectrum of two-player



(a) Natural game coordinates. (b) Rotated coordinates.

Figure 3.1: *Cost landscape is crucial to understanding dynamics.* The zero-sum game defined by $f(x,y)=\frac{1}{2}x^2-\frac{1}{8}y^2$ has a Nash equilibrium at the origin, which is a stable saddle point of gradient play (3.1). If the cost function is rotated to $\tilde{f}(x,y)=\frac{1}{32}x^2+\frac{11}{32}y^2-\frac{5\sqrt{3}}{16}xy$—a rotation by $\frac{\pi}{3}$—then the origin is no longer a Nash equilibrium, and is *unstable* under gradient play.

game dynamics. The method for obtaining the quadratic numerical range is by reducing a block matrix to 2×2 matrices.

Towards this end, we decompose the 2×2 game Jacobian into coordinates that reflect the interaction between the players. The decomposition provides insights on games and vector fields in general, which permits us to provide a complete characterization of the stability of equilibria in two-player gradient learning dynamics.

Organization. In Section 3.2, we describe the gradient-based learning paradigm and analyze the spectral properties of block operator matrices using the quadratic numerical range (Tretter, 2008). In Section 3.3, we analyze the spectral properties of two-player continuous games on scalar action spaces. Our main results are on general-sum games, with insights drawn from specific classes of games. In Section 3.4, we certify the stability of Nash and non-Nash equilibria in two-player scalar games. A key finding is that in the scalar case, equilibria that are both stable and Nash are robust to variations in learning rates; in the vector case, they are not. We provide an example in Section 3.4.3.

3.2 Preliminaries

This section contains game-theoretic preliminaries, mathematical formalism, and a description of the gradient-based learning paradigm studied in this paper.

3.2.1 Game-Theoretic Preliminaries

A 2-player *continuous game* $\mathcal{G} = (f_1, f_2)$ is a collection of costs defined on $X = X_1 \times X_2$ where player (agent) $i \in \mathcal{I} = \{1, 2\}$ has cost $f_i : X \to \mathbb{R}$. In this paper, the results apply to games with sufficiently smooth costs $f_i \in C^r(X, \mathbb{R})$ for some $r \geq 0$. Agent i's set of feasible actions is the d_i-dimensional precompact set $X_i \subseteq \mathbb{R}^{d_i}$. The notation x_{-i} denotes the action of player i's competitor; that is, $x_{-i} = x_j$ where $j \in \mathcal{I} \backslash \{i\}$.[1]

The most common and arguably natural notion of an equilibrium in continuous games is due to Nash (Nash, 1951).

Definition 4 (Local Nash equilibrium). *A joint action profile $x = (x_1, x_2) \in W_1 \times W_2 \subset X_1 \times X_2$ is a local Nash equilibrium on $W_1 \times W_2$ if, for each player $i \in \mathcal{I}$, $f_i(x_i, x_{-i}) \leq f_i(x_i', x_{-i})$, $\forall x_i' \in W_i$.*

A local Nash equilibrium can equivalently be defined as in terms of best response maps: $x_i \in \arg\min_y f_i(y, x_{-i})$. From this perspective, local optimality conditions for players' optimization problems give rise to the notion of a differential Nash equilibrium (Ratliff et al., 2013, 2016); non-degenerate differential Nash are known to be generic and structurally stable amongst local Nash equilibria in sufficiently smooth games (Ratliff et al., 2014). Let $D_i f_i$ denote the derivative of f_i with respect to x_i and, analogously, let $D_i(D_i f_i) \equiv D_i^2 f_i$ be player i's individiaul Hessian.

Definition 5. *For continuous game $\mathcal{G} = (f_1, f_2)$ where $f_i \in C^2(X_1 \times X_2, \mathbb{R})$, a joint action profile $(x_1, x_2) \in X_1 \times X_2$ is a* differential Nash equilibrium *if $D_i f_i(x_1, x_2) = 0$ and $D_i^2 f_i(x_1, x_2) > 0$ for each $i \in \mathcal{I}$.*

A differential Nash equilibrium is a strict local Nash equilibrium (Ratliff et al., 2013, Thm. 1). Furthermore, the conditions $D_i f_i(x) = 0$ and $D_i^2 f_i(x) \geq 0$ are necessary for a local Nash equilibrium (Ratliff et al., 2013, Prop. 2).

Learning processes in games, and their study, arose as one of the explanations for how players grapple with one another in seeking an equilibrium (Fudenberg and Levine, 1998). In the case of sufficiently smooth games, gradient-based learning is a natural learning rule for myopic players[2].

3.2.2 Gradient-based Learning as a Dynamical System

At time t, a myopic agent i updates its current action $x_i(t)$ by following the gradient of its individual cost f_i given the decisions of its competitors x_{-i}. The synchronous adaptive process that arises is the discrete-time dynamical system

$$x_i(t+1) = x_i(t) - \gamma_i D_i f_i(x_i(t), x_{-i}(t)) \tag{3.1}$$

[1] For 2-player games, $x_{-1} = x_2$ and $x_{-2} = x_1$.

[2] A mypoic player effectively believes it cannot influence its opponent's future behavior, and reacts only to local information about its cost.

for each $i \in \mathcal{I}$ where $D_i f_i$ is the gradient of player i's cost with respect to x_i and γ_i is player i's learning rate.

Stability Recall that a matrix A is called Hurwitz if its spectrum lies in the open left-half complex plane $\mathbb{C}_-^\circ$. Furthermore, we often say such a matrix is *stable* in particular when A corresponds to the dynamics of a linear system $\dot{x} = Ax$ or the linearization of a nonlinear system around a fixed point of the dynamics.[3]

It is known that (3.1) will converge locally asymptotically to a differential Nash equilibrium if the local linearization is a contraction (Chasnov et al., 2020d). Let

$$g(x) = (D_1 f_1(x), D_2 f_2(x)) \tag{3.2}$$

be the vector of individual gradients and let $Dg(x)$ be its Jacobian—i.e., the *game Jacobian.* Further, let $\sigma_p(A) \subset \mathbb{C}$ denote the *point spectrum* (or *spectrum*) of the matrix A, and $\rho(A)$ its *spectral radius.* Then, x is *locally exponentially stable* if and only if $\rho(I - \Gamma Dg(x)) < 1$, where $\Gamma = \text{blockdiag}(\gamma_1 I_{d_1}, \gamma_2 I_{d_2})$ is a diagonal matrix and I_{d_i} is the identity matrix of dimension d_i. The map $I - \Gamma Dg(x)$ is the local linearization of (3.1). Hence, to study stability (and, in turn, convergence) properties it is useful to analyze the spectrum of not only the map $I - \Gamma Dg(x)$ but also $Dg(x)$ itself.

For instance, when $\gamma = \gamma_1 = \gamma_2$, the spectral mapping theorem tells us that $\rho(I - \gamma Dg(x)) = \max_{\lambda \in \sigma_p(Dg(x))} |1 - \gamma\lambda|$ so that understanding the spectrum of $Dg(x)$ is imperative for understanding convergence of the discrete time update. On the other hand, when $\gamma_1 \neq \gamma_2$, we write the local linearization as $I - \gamma_1 \Lambda Dg(x)$ where $\Lambda = \text{blockdiag}(I_{d_1}, \tau I_{d_2})$ and $\tau = \gamma_2/\gamma_1$ is the learning rate ratio. Again, via the spectral mapping theorem, when $I - \gamma_1 \Lambda Dg(x)$ is a contraction for different choices of learning rate γ_1 is determined by the spectrum of $\Lambda Dg(x)$. Hence, given a *fixed point* x (i.e., $g(x) = 0$), we study the stability properties of the limiting continuous time dynamical system—i.e., $\dot{x} = -g(x)$ when $\gamma_1 = \gamma_2$ and $\dot{x} = -\Lambda g(x)$ otherwise. From here forward, we will simply refer to the system $\dot{x} = -\Lambda g(x)$ and point out when $\Lambda = I_{d_1+d_2}$ if not clear from context.

Partitioning the Game Jacobian Let $x = (x_1, x_2)$ be a joint action profile such that $g(x) = 0$. Towards better understanding the spectral properties of $Dg(x)$ (respectively, $\Lambda Dg(x)$), we partition $Dg(x)$ into blocks:

$$J(x) = \begin{bmatrix} -D_1^2 f_1(x) & -D_{12} f_1(x) \\ -D_{21} f_2(x) & -D_2^2 f_2(x) \end{bmatrix} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}. \tag{3.3}$$

[3] The Hartman-Grobman theorem (Sastry, 1999) states that around any hyperbolic fixed point of a nonlinear system, there is a neighborhood on which the nonlinear system is stable if the spectrum of Jacobian lies in $\mathbb{C}_-^\circ$.

A differential Nash equilibrium (the second order conditions of which are sufficient for a local Nash equilibrium) is such that $J_{11} < 0$ and $J_{22} < 0$. On the other hand, as noted above, J is Hurwitz or stable if its point spectrum $\sigma_p(J) \subset \mathbb{C}_-^\circ$. Moreover, since the diagonal blocks are symmetric, J is similar to the matrix in Fig 3.2. For the remainder of the paper, we will study the Dg at a given fixed point x as defined in (3.3).

$J(x, y) \sim$

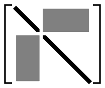


Figure 3.2: *Similarity*: the game Jacobian in (3.3) is similar to a matrix with diagonal block-diagonals.

Classes of Games Different classes of games can be characterized via J. For instance, a *zero-sum game*, where $f_1 \equiv -f_2$, is such that $J_{12} = -J_{21}^\top$. On the other hand, a game $\mathcal{G} = (f_1, f_2)$ is a *potential game* if and only if $D_{12}f_1 \equiv D_{21}f_2^\top$ (Monderer and Shapley, 1996, Thm. 4.5), which implies that $J_{12} = J_{21}^\top$.

3.2.3 Spectrum of Block Matrices

One useful tool for characterizing the spectrum of a block operator matrix is the numerical range and quadratic numerical range, both of which contain the operator's spectrum (Tretter, 2008) and therefore all of its eigenvalues. The *numerical range* of J is defined by

$$W(J) = \{\langle Jz, z\rangle : \ z \in \mathbb{C}^{d_1+d_2}, \ \|z\|_2 = 1\},$$

and is convex. Given a block operator J, let

$$J_{v,w} = \begin{bmatrix} \langle J_{11}v, v\rangle & \langle J_{12}w, v\rangle \\ \langle J_{21}v, w\rangle & \langle J_{22}w, w\rangle \end{bmatrix} \tag{3.4}$$

where $v \in \mathbb{C}^{d_1}$ and $w \in \mathbb{C}^{d_2}$. The *quadratic numerical range* of J, defined by

$$W^2(J) = \bigcup_{v \in \mathcal{S}_1, w \in \mathcal{S}_2} \sigma_p(J_{v,w}), \tag{3.5}$$

is the union of the spectra of (3.4) where $\sigma_p(\cdot)$ denotes the (point) spectrum of its argument and $\mathcal{S}_i = \{z \in \mathbb{C}^{d_i} : \ \|z\|_2 = 1\}$. It is, in general, a non-convex subset of $\mathbb{C}$. The quadratic numerical range (3.5) is equivalent to the set of solutions of the characteristic polynomial

$$\begin{aligned} \lambda^2 - \lambda(\langle J_{11}v, v\rangle + \langle J_{22}w, w\rangle) + \langle J_{11}v, v\rangle\langle J_{22}w, w\rangle \\ - \langle J_{12}v, w\rangle\langle J_{21}w, v\rangle = 0 \end{aligned} \tag{3.6}$$

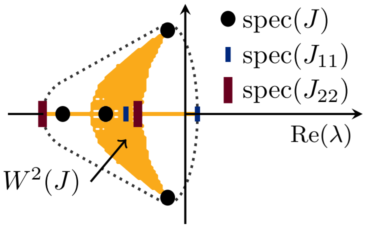


Figure 3.3: *Spectrum of a stable equilibrium that is not Nash.* The spectrum of J, J_{11}, and J_{22} in Example 1 are contained in the numerical range (convex dashed region) and quadratic numerical range (non-convex region) of J. The eigenvalues of J are in the left plane, hence the fixed point is stable under gradient play (3.1). However, the first player's J_{11} is indefinite, hence the fixed point is not a Nash equilibrium.

for $v \in \mathcal{S}_1$ and $w \in \mathcal{S}_2$. We use the notation $\langle Jx, y\rangle = x^* Jy$ to denote the inner product. Note that $W^2(J)$ is a subset of $W(J)$ and, as previously noted, contains $\sigma_p(J)$. Albeit non-convex, $W^2(J)$ provides a tighter characterization of the spectrum[4].

Example 1. *Consider the game Jacobian of the zero-sum game $(f, -f)$ defined by cost $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$,*

$$f(x, y) = -\tfrac{1}{2}x_1^2 + \tfrac{5}{2}x_2^2 + 7y_1x_1 - 3y_2x_2 - 2y_1^2 - 6y_2^2.$$

The numerical range, quadratic numerical range, spectrum and diagonal entries of J, defined using the origin as the fixed point, are plotted in Fig. 3.3. In this example, the origin is not a differential Nash equilibrium since $D_1^2 f_1(0, 0)$ is indefinite, yet it is an exponentially stable equilibrium of $\dot{x} = -g(x)$ since all the eigenvalues of J are all negative.

Observing that the quadratic numerical range for a block 2×2 matrix J derived from a game on a finite dimensional Euclidean space reduces to characterizing the spectrum of 2×2 matrices, we first characterize stability properties of scalar 2-player continuous games.

3.3 Decomposition of Scalar Games

We characterize the stability of differential Nash equilibria in 2-player scalar continuous games. Consider a game (f_1, f_2) with action spaces $X_1, X_2 \subseteq \mathbb{R}$. Let x be a fixed point of (3.2) such that $g(x) = 0$. We decompose its game Jacobian (3.3) into components that reflect the dynamic interaction between the players.

[4] There are numerous computational approaches for estimating the numerical ranges $W(\cdot)$ and $W^2(\cdot)$ (see, e.g., (Langer et al., 2001, Sec. 6)).

3.3.1 Jacobian Decomposition: Two-Dimensional Case

Consider the decomposition of a $\mathbb{R}^{2\times 2}$ game Jacobian

$$J(x) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} m & -z \\ z & m \end{bmatrix} + \begin{bmatrix} h & p \\ p & -h \end{bmatrix} \tag{3.7}$$

where $m = \frac{1}{2}(a+d)$, $h = \frac{1}{2}(a-d)$, $p = \frac{1}{2}(b+c)$, $z = \frac{1}{2}(c-b)$. Let $\mathrm{tr}(J)$ be its trace, $\det(J)$ be its determinant, and $\mathrm{disc}(J)$ be the discriminant of its characteristic polynomial.[5] Several directly verifiable quantities are stated.

Statement 1. *Given a matrix $J \in \mathbb{R}^{2\times 2}$ and its spectrum $\sigma_p(J) = \{\lambda_1, \lambda_2\}$, the above decomposition gives rise to the following conditions:*

$$\begin{aligned} \mathrm{tr}(J) &= \lambda_1 + \lambda_2 = a + d = 2m, \\ \det\big(J\big) &= \lambda_1\lambda_2 = ad - bc = (m^2 + z^2) - (h^2 + p^2), \\ \mathrm{disc}\big(J\big) &= (\lambda_1 + \lambda_2)^2 - 4\lambda_1\lambda_2 = 4(h^2 + p^2 - z^2), \\ \lambda_{1,2} &= \tfrac{1}{2}\big(\,\mathrm{tr}(J) \mp \sqrt{\mathrm{disc}(J)}\big) = m \mp \sqrt{h^2 + p^2 - z^2}. \end{aligned}$$

The change of coordinates from (a, b, c, d) to (m, h, p, z) in Statement 1 provides important insights into linear vector fields and, in particular, to games. The stability of vector field $\dot{x} = Jx$ is given by the trace and determinant conditions.

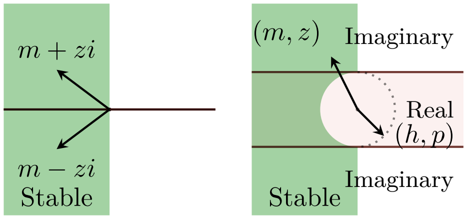
Proposition 6. *The matrix $J \in \mathbb{R}^{2\times 2}$ is stable if and only if $m^2 + z^2 > h^2 + p^2$ and $m < 0$.*

Proof. Statement 1 and direct computation show that these conditions are equivalent to $\lambda_1 + \lambda_2 < 0$ and $\lambda_1\lambda_2 > 0$, well-known conditions for stability of 2 × 2 systems (illustrated in Fig. 3.5b). □

3.3.2 Discussion of Decomposition

The purpose of the decomposition into the alternative coordinates is to geometrically—and thus more directly—assess the conditions for stability of a differential Nash equilibrium. In particular, the conditions for a fixed point of the game dynamics to be a differential Nash equilibrium are $a < 0$ and $d < 0$, or equivalently, $m < -|h|$, which are represented by the left-shaded regions in Fig. 3.6a–3.6b. Moreover, the conditions for a fixed point of the game dynamics to be stable (i.e., for $\sigma_p(J) \subset \mathbb{C}^\circ_-$) are $m < 0$ and $m^2 + z^2 > h^2 + p^2$, which are visible only when using the decomposition in Fig. 3.6b.

[5] The characteristic polynomial of J is $\lambda \mapsto \det(J - \lambda I)$ and its discriminant is $\mathrm{tr}(J)^2 - 4\det(J)$ for $J \in \mathbb{R}^{2\times 2}$.



(a) The complex plane. (b) A representation of the m, z, h, p coordinates.

Figure 3.4: *Visualization of Statement 1:* If h and p are zero, then the eigenvalues of J are $\lambda_{1,2} = m \mp zi$. If h and/or p are non-zero, then a circle centered around the origin with radius $\sqrt{h^2 + p^2}$ is excluded from left-half stability region.

Relationship to complex plane Fig. 3.4 plots the coordinates of m, z, h, p relative to each other to illustrate the decomposition in Statement 1. If $h = 0, p = 0$, then the eigenvalues of J are $\lambda_{1,2} = m \mp zi$. Fig. 3.4a corresponds to a plot of eigenvalues in the complex plane. Stability is given by the familiar open-left half plane condition: $\text{spec}(J) \subset \mathbb{C}_-^\circ$. If $h \neq 0$ or $p \neq 0$ a circular region in the center of the plane expands the values of m, z for which the eigenvalues of the matrix are purely real. Fig. 3.4b shows that the eigenvalues are purely real if and only if $z^2 \leq h^2 + p^2$.

Effect of rotation in game vector fields Note the similarity between (3.7) and the well-known symmetric/skew-symmetric (Helmholtz) decomposition

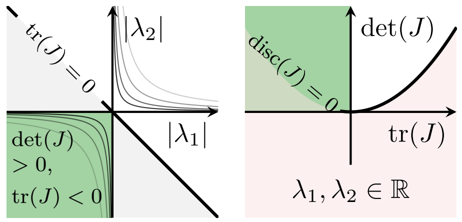
$$J(x) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} m+h & p \\ p & m-h \end{bmatrix} + \begin{bmatrix} 0 & -z \\ z & 0 \end{bmatrix}. \tag{3.8}$$

Assuming that $m < 0$, from Proposition 6 we can see that increasing the rotational component of the Jacobian helps stability. Increasing the relative magnitude of p, the non-rotational interaction term hurts stability. If there is no rotational component, ie. J is symmetric, p's negative impact on stability can be seen directly from the Schur complement[6]. In this case J is stable iff $J < 0$ and thus stability requires that both the diagonals and the Schur complement are negative: $a < 0$, $d < 0$, and $a - p^2 d^{-1} < 0$. If $d < 0$, increasing p can only increase the Schur complement.

3.3.3 Types of Games

The decomposition also provides a natural classification of 2-player scalar games into four types based on specific coordinates being zero, as illustrated in Fig. 3.7.

[6] The Schur complements of the matrix in (3.3) are $J_{11} - J_{12}J_{22}^{-1}J_{21}$ (where J_{22} is invertible) and $J_{22} - J_{21}J_{11}^{-1}J_{12}$ (where J_{11} is invertible).



(a) Level sets of $\det(J) = \lambda_1\lambda_2$.

(b) Real or imaginary eigenvalues.

Figure 3.5: *Visualization of Proposition 6*: $\dot{y} = J(x)y$ is stable $\iff$ $\det(J) > 0$ and $\mathrm{tr}(J) < 0$.

Potential games ($z = 0$) The point (m, z) lives on the horizontal axis in Fig. 3.7a, thus stable fixed points are a subset of Nash equilibria. Since $z = 0$, Proposition 6 indicates that increasing p, the interaction term between the players, and increasing h, the difference in curvature between the two players both only hurt stability.

Zero-sum games ($p = 0$) The point (h, p) lives on the horizontal axis in Fig. 3.7b, thus all Nash equilibria are stable, but not all stable fixed points are Nash. The magnitude of the interaction term z helps stability and may make a fixed point stable even if it is not Nash. Intuitively, a strong enough interaction term can cause a player which is at its action under the Nash equilibrium with stronger negative curvature to pull another player with weaker positive curvature toward a fixed point even if that point is a local maximum for the weaker player.

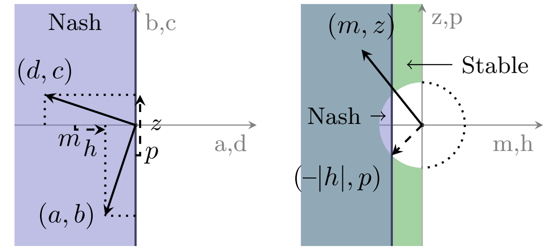
Hamiltonian games ($m = 0$) The point (m, z) lives on the vertical axis in Fig. 3.7c, thus no strict Nash equilibria can exist. At best these games are marginally stable if $|z|$ is large enough relative to the magnitude of (h, p).

Matching-curvature games, ($h = 0$) The point (h, p) lives on the vertical axis in Fig. 3.7d, so any stable point is also a Nash. Any fixed point with a, d having the same sign can be rescaled to have matching curvature $\gamma_1 a = \gamma_2 d$ by a choice of non-uniform learning rates $\gamma_1, \gamma_2 > 0$.

3.4 Certificates for Stability in Scalar Games

3.4.1 Stability: Uniform Learning Rates

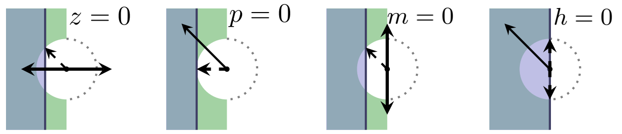
For a game $\mathcal{G} = (f_1, f_2)$, let the set of differential Nash equilibria be denoted $\mathtt{DNE}(\mathcal{G})$ and let the stable points of $\dot{x} = -g(x)$ be $\mathtt{S}(\mathcal{G})$. Let $\overline{\mathtt{DNE}}(\mathcal{G})$ and $\overline{\mathtt{S}}(\mathcal{G})$ be their respective complements. The intersections of these sets characterize the stability/instability of Nash/non-Nash equilibria.



(a) Geometry of decomposition in (3.7).

(b) Change of coordinates reveals regions of stability.

Figure 3.6: *Decomposition of a general scalar game.* The rows vectors of J are plotted in (a) and the same matrix with a change of coordinates is plotted in (b). Nash regions ($m < -|h|$) and stability regions ($m < 0, m^2 + z^2 > h^2 + p^2$) are visible. Their set differences characterize the conditions for a stable non-Nash and unstable Nash equilibria.



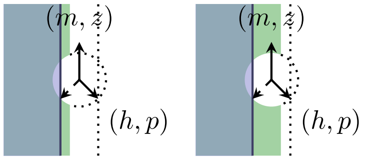
(a) Potential: Stable ⊂ Nash.

(b) Zero-sum: Stable ⊃ Nash

(c) Hamiltonian: marginally stable at best.

(d) Matching: Stable ⊂ Nash.

Figure 3.7: *Stability and Nash for different classes of games.* (a) Potential games: symmetric interaction term only hurts stability. (b) Zero-sum games: rotation can compensate for unhappy player. (c) Hamiltonian games: players have zero total curvature, $a + d = 0$. (d) Matching curvature, $a = d$: there are no stable non-Nash equilibria.



(a) $\gamma_1 > \gamma_2$

(b) $\gamma_1 < \gamma_2$

Figure 3.8: *Time-scale separation affects stability.* The learning rate ratio $\tau = \gamma_2/\gamma_1 > 0$ affects the stability of the game dynamics. The factor $\beta = \frac{\tau - 1}{\tau + 1}$ expands or shrinks the region for stability. The condition $m < 0$ becomes $m < \beta h$. Note that $-1 \leq \beta \leq 1$ for $\tau \geq 0$. For $\beta \to \pm 1$, the region's vertical boundary approaches $\pm h$.

Theorem 5 (Certificates for 2-Player Scalar Games)**.** *Consider a game $\mathcal{G} = (f_1, f_2)$ on $X_1 \times X_2 \subseteq \mathbb{R}^2$. Let x be a fixed point of* (3.2) *and let m, h, p, z be defined by* (3.7)*. The following equivalences hold:*

(i) $x \in \mathtt{DNE}(\mathcal{G}) \cap \mathtt{S}(\mathcal{G}) \iff$

$$\{m < -|h|\} \wedge \{m^2 + z^2 > h^2 + p^2\}.$$

(ii) $x \in \mathtt{DNE}(\mathcal{G}) \cap \overline{\mathtt{S}}(\mathcal{G}) \iff$

$$\{m < -|h|\} \wedge \{m^2 + z^2 \leq h^2 + p^2\}.$$

(iii) $x \in \overline{\mathtt{DNE}}(\mathcal{G}) \cap \mathtt{S}(\mathcal{G}) \iff$

$$\{0 > m \geq -|h|\} \wedge \{m^2 + z^2 > h^2 + p^2\}.$$

(iv) $x \in \overline{\mathtt{DNE}}(\mathcal{G}) \cap \overline{\mathtt{S}}(\mathcal{G}) \iff$

$\{\{0 > m \geq -|h|\} \wedge \{m^2 + z^2 \leq h^2 + p^2\}\} \vee \{m \geq 0\}.$

The contributions to the stability of a non-Nash equilibrium or the instability of a Nash equilibrium are stated in (ii) and (iii). We illustrate the geometry of these two cases with the shaded regions in Fig. 3.6b.

3.4.2 Stability: Non-Uniform Learning Rates

Consider players updating their actions according to gradient play as defined in (3.1) with individual learning rates $\gamma_1, \gamma_2 > 0$, not necessarily equal. We study how the players' ratio $\tau = \gamma_2/\gamma_1$ affects the stability of fixed point x under the learning dynamics by analyzing the game Jacobian

$$J(x) = \begin{bmatrix} a & b \\ \tau c & \tau d \end{bmatrix}. \tag{3.9}$$

Learning rates do not affect whether a fixed point is a Nash equilibrium. They do, however, affect whether it is stable.

Corollary 3 (Stability in General-Sum Scalar Games)**.** *Consider a game $\mathcal{G} = (f_1, f_2)$ on $X_1 \times X_2 \subseteq \mathbb{R}^2$ and a fixed point x. Suppose players perform gradient play (3.1) with learning rate ratio $\tau = \gamma_2/\gamma_1$. Then, the following are true.*

(i) If a Nash equilibrium is stable for some τ, then it is stable for all τ.

(ii) If a non-Nash equilibrium is stable, then there exists some τ that makes it unstable.

(iii) If a fixed point is non-Nash, the determinant of its game Jacobian is positive and $m < |h|$, then there exists some τ that makes it stable.

Proof. To prove (i), we observe that if $m < -|h|$, then $m \leq \beta h$ for all β such that $|\beta| < 1$. Choose $-1 \leq \beta = \frac{\tau-1}{\tau+1} \leq 1$ for $\tau \geq 0$. To prove (ii), choose $\tau < \left|\frac{a}{d}\right|$. Without loss of generality, assume $a < 0$ and $d > 0$. Then, it directly follows that $a + \tau d < 0$. To prove (iii), note that a matrix J is stable if and only if the determinant of J is positive and $m < 0$. Hence, without loss of generality, let $d < 0$. Then there is a learning rate τ such that $\tau|d| > |a|$ so that $m < 0$.

□

Stable Nash equilibria in scalar games are robust to variations in learning rates and non-Nash equilibria are not. For continuous games with vector action spaces, Corollary 3(i) no longer holds, demonstrating that Nash equilibria are not robust, in general, to variations in learning rates.

3.4.3 An Example with Nonlinear Dynamics

We demonstrate our main results below and in Fig. 3.9.

Example 2 (Nonlinear torus game)**.** *Consider a game $\mathcal{G} = (f_1, f_2)$ defined on $\mathbb{S}^1 \times \mathbb{S}^1$ with costs*

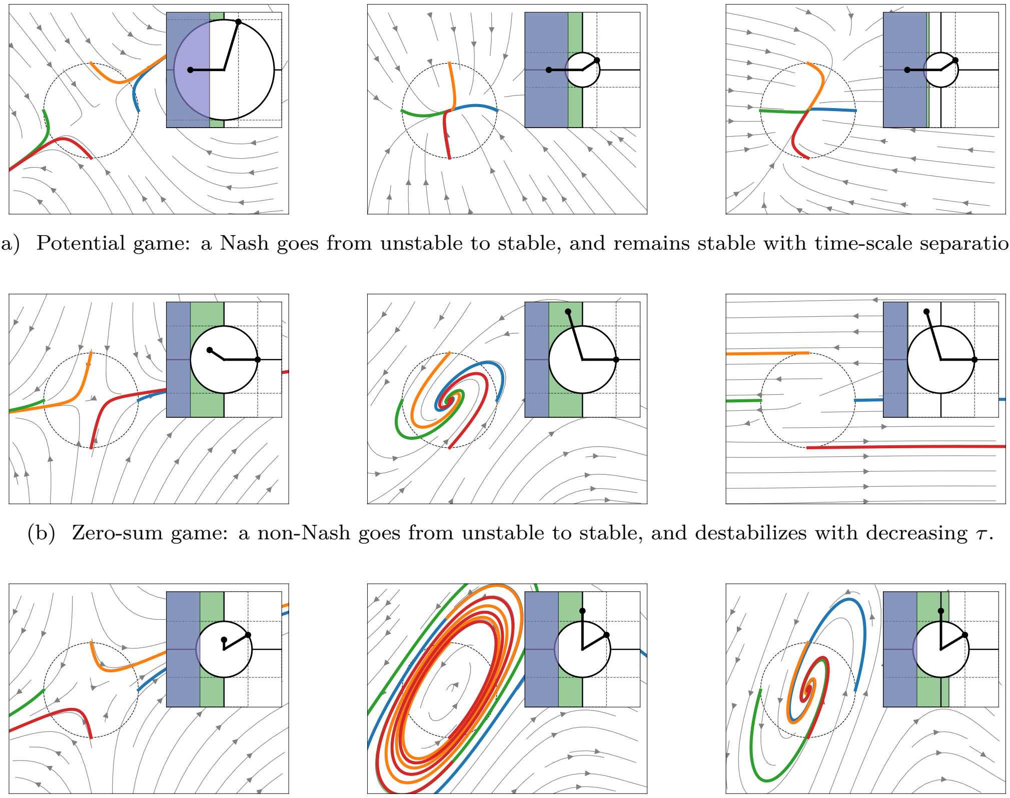
$$f_1(x, y) = \tfrac{2}{a} \cos\left(\tfrac{a}{2}x\right) + \tfrac{2}{a} \cos\left(\tfrac{a}{2}x + by\right),$$
$$f_2(x, y) = \tfrac{2}{d} \cos\left(\tfrac{d}{2}y\right) + \tfrac{2}{d} \cos\left(\tfrac{d}{2}y + cx\right).$$

There is a fixed point of the learning dynamics at the origin. Its linearized game Jacobian is $J(0) = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. First, to show Corollary 3(i), we start with an unstable, Nash fixed point of a potential game ($a = -0.4, b = 1, c = 1, d = -1$). We decrease $p = \frac{1}{2}(b+c)$ until it becomes stable ($b = 0.2, c = 0.2$). Then, we decrease τ from 1 to 0.1 while maintaining stability. Second, to show Corollary 3(ii), we start with an unstable, non-Nash fixed point of a zero-sum game ($a = 0.4, b = -0.2, c = 0.2, d = -1$). We increase $z = \frac{1}{2}(c-b)$ until it becomes stable ($b = -1, c = 1$). Then, we decrease τ from 1 to 0.01 making it unstable again. Third, to show Corollary 3(iii), we start with an unstable, non-Nash fixed point of a Hamiltonian game ($a = 0.5, b = 0.1, c = 0.5, d = -0.5$). We increase the interaction term $z = \frac{1}{2}(c-b)$ until it becomes marginally stable ($b = -0.5, c = 1.1$). Then, we increase τ slightly from 1 to 2, making the fixed point stable.

Towards characterizing the optimization landscape of games, this paper analyzes the stability and spectrum of gradient-based dynamics near fixed points of two-player continuous games. We introduce the quadratic numerical range as a method to bound the spectrum of game dynamics linearized about local equilibria. We also analyze the stability of differential Nash equilibria and their robustness to variation in agent's learning rates. Our results show that by decomposing the game Jacobian into symmetric and anti-symmetric components, we can assess the contribution of vector field's potential and rotational components to the stability of the equilibrium. In zero-sum games, all differential Nash equilibria are stable; in potential games, all stable points are Nash. Furthermore, zero-sum Nash equilibria are robust in the sense that they are stable for all learning rates. For continuous games with general costs, we provide a sufficient condition for instability. We conclude with a numerical example that investigates how players with different learning rates can take advantage of rotational components of the game to converge faster.

3.5 Decomposition of Vector Games

In the next sections, we give stability results for 2-player continuous games on vector action spaces. This is unpublished work available on ArXiV (Chasnov et al., 2020b). Consider a game (f_1, f_2). Recall from the preliminaries that $f_1, f_2 \in C^2(X_1 \times X_2, \mathbb{R})$ and $X_1 \subseteq \mathbb{R}^{d_1}, X_2 \subseteq \mathbb{R}^{d_2}$, are d_i-dimensional actions spaces.



(a) Potential game: a Nash goes from unstable to stable, and remains stable with time-scale separation.

(b) Zero-sum game: a non-Nash goes from unstable to stable, and destabilizes with decreasing τ.

(c) Hamiltonian game: a non-Nash goes from unstable to marginally stable, and stabilizes with increasing τ.

Figure 3.9: *Demonstration of Corollary 3:* vector field plots of the three scenarios from Example 2.

Let x be a fixed point of (3.2) such that $g(x) = 0$. We study the resulting gradient learning dynamics of Equation (3.1) near fixed points x. In particular, we analyze the Jacobian of g and its relation with differential Nash equilibria and stable equilibria.

3.5.1 Jacobian Decomposition: Block Case

In the block 2×2 case, we decompose the game Jacobian (3.3) by analogy with the decomposition in the scalar case.

We decompose the game Jacobian,

$$J(x) = \begin{bmatrix} J_{11} & P \\ P^\top & J_{22} \end{bmatrix} + \begin{bmatrix} 0 & -Z \\ Z^\top & 0 \end{bmatrix}, \tag{3.10}$$

where $P = \frac{1}{2}(J_{12} + J_{21}^\top)$ and $Z = \frac{1}{2}(J_{12} - J_{21}^\top)$.

3.5.2 Discussion of Block Decomposition

Consider the Jacobian in (3.10) and its quadratic numerical range $\mathcal{W}^2(J(x))$, we note that the spectrum of $J(x)$ is contained in the spectrum of

$$J_{v,w} = \begin{bmatrix} a & p - z \\ p^* + z^* & d \end{bmatrix} \tag{3.11}$$

where $a = \langle J_{11}v, v\rangle$, $d = \langle J_{22}w, w\rangle$ $p = \langle Pv, w\rangle$, and $z = \langle Zw, v\rangle$ for unit-length complex numbers $v \in \mathcal{S}_1, w \in \mathcal{S}_2$. Hence, to show the stability of a particular fixed point x, we must show that for all v, w, the spectrum of (3.11) is contained in the left-half complex plane. We directly assess whether a fixed point is Nash by the signs of a, d.

3.5.3 Types of Games

The decomposition in (3.10) naturally leads to games of various types based on either Z or P being 0.

Potential games ($Z = 0$)

All stable fixed points are differential Nash equilibria. Further, we provide a necessary and sufficient condition for a stable Nash equilibrium.

Proposition 7 (Stability in Potential Games)**.** *Consider a potential game $\mathcal{G} = (f_1, f_2)$ on finite dimensional action spaces X_1, X_2. If x is a stable equilibrium of $\dot{x} = -g(x)$, then x is a differential Nash equilibrium of $\mathcal{G}$. Moreover, differential Nash equilibria are stable if and only if $J_{11} - PJ_{22}^{-1}P^T < 0$.*

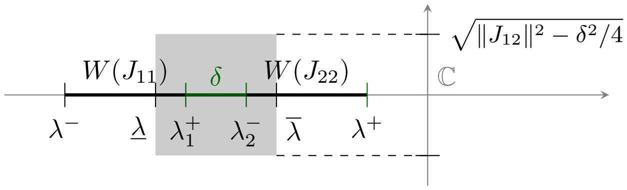
Since $Z = 0$ in potential games, $J(x)$ is symmetric and its eigenvalues are directly related to the definitness properties of $J(x)$. That is, $J_{11} < 0$ and $J_{22} < 0$ are necessary conditions for $J < 0$. The result is consistent with our intuition from the scalar case that the interaction terms in potential games only discourage stability.

Zero-sum games ($P = 0$)

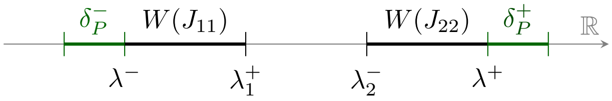
All differential Nash equilibria are stable, and these equilibria are robust to variations of learning rates.

Proposition 8 (Stability in Zero-Sum Games)**.** *Consider a zero-sum game $\mathcal{G} = (f, -f)$ on finite dimensional action spaces X_1, X_2. If x is a differential Nash equilibrium of $\mathcal{G}$, then x is a stable equilibrium of $\dot{x} = -g(x)$.*

Proof. For zero-sum games, $P = 0$, hence the symmetric component of $J(x)$ at fixed point x is $\frac{1}{2}\big(J(x) + J(x)^\top\big) = \text{blockdiag}(-D_1^2 f_1(x), -D_2^2 f_2(x))$. If x is a differential Nash equilibrium, then $J(x) + J(x)^\top < 0$ and hence x is stable. □



(a) Zero-sum game where $\delta = \lambda_2^- - \lambda_1^+ > 0$ and $\|J_{12}\| > \delta/2$



(b) Potential game where $\lambda_2^- - \lambda_1^+ > 0$.

Figure 3.10: *Spectrum of learning dynamics near a fixed point in zero-sum and potential games.* We illustrate Theorem 6 (a, zero-sum game) and Theorem 7 (b, potential game). The highlighted regions contain the spectrum of the linearized dynamics.

The result can also be proven directly from Lyapunov theory using Lyapunov function $\|x\|^2$.

In the following section, we bound the spectrum of zero-sum and potential game dynamics using the quadratic numerical range. Future work will look to generalize the two cases to games with general cost structures.

3.6 Conditions for Stability in Vector Games

3.6.1 Block Stability: Uniform Learning Rates

We bound the spectrum of zero-sum and potential game dynamics using the norms of the interaction terms Z and P defined in (3.10).

For game (f_1, f_2) with game Jacobian (3.3), define the following for $i = 1, 2$: $\lambda_i^- = \min \sigma_p(J_{ii})$, $\lambda_i^+ = \max \sigma_p(J_{ii})$. Additionally, define

$$\lambda^- = \min\{\lambda_1^-, \lambda_2^-\}, \quad \underline{\lambda} = \tfrac{1}{2}(\lambda_1^- + \lambda_2^-),$$
$$\lambda^+ = \max\{\lambda_1^+, \lambda_2^+\}, \quad \overline{\lambda} = \tfrac{1}{2}(\lambda_1^+ + \lambda_2^+).$$

These terms will be useful in deriving bounds on the spectrum of J. The next two theorems are our main results, giving tight bounds on the spectrum of J (i.e. bounds on the real and imaginary eigenvalues) for game dynamics of zero-sum and potential games. Recall that for zero-sum game $(f, -f)$, the interaction term $Z = D_{12}f(x)$.

Theorem 6 (Spectrum of Zero-Sum Game Dynamics)**.** *Consider a zero-sum game $\mathcal{G} = (f, -f)$. The Jacobian*

$J(x) = Dg(x)$ of the dynamics $\dot{x} = -g(x)$ at fixed points x is such that

$$\sigma_p(J(x)) \cap \mathbb{R} \subset \left[\lambda^-, \lambda^+\right] \tag{3.12}$$

and $\sigma_p(J(x)) \backslash \mathbb{R}$ is contained in

$$\left\{z \in \mathbb{C} :\ \mathrm{Re}(z) \in \left[\underline{\lambda}, \overline{\lambda}\right],\ |\mathrm{Im}(z)| \leq \|Z\|\right\}. \tag{3.13}$$

Furthermore, if $\lambda_2^+ < \lambda_1^-$ or $\lambda_1^+ < \lambda_2^-$ then the following two implications hold for $\delta = \lambda_1^- - \lambda_2^+$ or $\delta = \lambda_2^- - \lambda_1^+$, respectively: (i) $\|Z\| \leq \delta/2 \implies \sigma_p(J(x)) \subset \mathbb{R}$; (ii) $\|Z\| > \delta/2 \implies \sigma_p(J(x)) \backslash \mathbb{R} \subset \{z \in \mathbb{C} :\ |\mathrm{Im}(z)| \leq \sqrt{\|Z\|^2 - \delta^2/4}\}$.

Proof. Observe that $\overline{\det(J_{v,w}(x) - \lambda I)} = \det(J_{v,w}(x) - \bar{\lambda} I)$ for $v \in \mathcal{S}_1$ and $w \in \mathcal{S}_2$ since $D_1^2 f(x)$ and $-D_2^2 f(x)$ are symmetric, which implies that $W^2(J(x)) = W^2(J(x))^*$. Since $-w^* D_{12} f(x)^T v v^* D_{12} f(x) w \leq 0$, (3.12) and (3.13) follow from (Tretter, 2008, Prop. 1.2.6), and (i) and (ii) follow from (Tretter, 2009, Lem. 5.1-(ii)). □

Recall that for potential games with potential function ϕ, the interaction term $P = D_{12}\phi(x)$. Define

$$\delta_P^{\pm} = \|P\| \tan\left(\frac{1}{2} \arctan \frac{2\|P\|}{|\lambda_1^{\pm} - \lambda_2^{\pm}|}\right).$$

Theorem 7 (Spectrum of Potential Game Dynamics)**.** *Consider a potential game $\mathcal{G} = (f_1, f_2)$. The Jacobian $J(x) = Dg(x)$ of the dynamics $\dot{x} = -g(x)$ at fixed points x is such that $\sigma_p(J(x)) \subset \mathbb{R}$ and*

$$\begin{aligned} \lambda^- - \delta_P^- \leq \min \sigma_p(J(x)) \leq \lambda^- \\ \lambda^+ \leq \max \sigma_p(J(x)) \leq \lambda^+ + \delta_P^+. \end{aligned} \tag{3.14}$$

Furthermore, if $\lambda_2^+ < \lambda_1^-$, then $\sigma_p(J(x)) \cap (\lambda_2^+, \lambda_1^-)$ is empty. If $\lambda_1^+ < \lambda_2^-$, then $\sigma_p(J(x)) \cap (\lambda_1^-, \lambda_2^+)$ is empty.

Proof. Inequalities in (3.14) follow from (Tretter, 2008, Prop. 1.2.4) and last statements follow from (Tretter, 2008, Cor. 1.2.3). □

3.6.2 Block Stability: Non-Uniform Learning Rates

Theorem 8 (Zero-sum Nash are Robust)**.** *Consider a zero-sum game $(f_1, f_2) = (f, -f)$ with game Jacobian J. Suppose that x is a differential Nash equilibrium. Then, x is a locally stable equilibrium of $\dot{x} = -\Lambda g(x)$ for any learning rate ratio τ.*

Proof. First, observe that $a = \langle J_{11}v, v\rangle$ and $d = \langle J_{22}w, w\rangle$ are negative real numbers for any $v \in \mathcal{S}_1$ and $w \in \mathcal{S}_2$ by assumption that x is a differential Nash equilibrium, i.e. $-D_i^2 f_i(x) < 0$ for each $i \in \{1, 2\}$. Second, observe that for zero-sum games, $z = -\langle J_{12}w, v\rangle = \langle J_{21}v, w\rangle^*$. Therefore, for x to be stable, the eigenvalues of

$$J_{v,w} = \begin{bmatrix} a & -z \\ \tau z^* & \tau d \end{bmatrix}$$

must all be negative. Hence, we compute the trace and determinant conditions to be $\mathrm{tr}(J_{v,w}) = \lambda_1 + \lambda_2 = a + \tau d$ and $\det(J_{v,w}) = \lambda_1\lambda_2 = \tau(ad + |z|^2)$. Notice that, $\tau(ad + |z|^2) > 0 \iff ad + |z|^2 > 0$, and $a + \tau d < 0 \iff a + d < 0$. Since $a, d < 0$ and $\tau > 0$, both of the trace and determinant conditions for stability are satisfied, i.e. $\mathrm{tr}(J_{v,w}) < 0$ and $\det(J_{v,w}) > 0$. Hence, x is a stable equilibrium of $\dot{x} = -\Lambda g(x)$. □

Further, the stability of $\dot{x} = -\Lambda g(x)$ implies that there exists a range of learning rates γ such that $x(t + 1) = x(t) - \gamma\Lambda g(x(t))$ is locally asymptotically stable.

Theorem 9 (Robustness of Potential Games)**.** *Suppose that x is a differential Nash equilibrium of a potential game with Jacobian Then, x is a locally stable equilibrium of $\dot{x} = -\Lambda g(x)$ for all learning rate ratio τ if $\sigma_{\max}(J_{11})\sigma_{\max}(J_{22}) > \sigma_{\max}(J_{12})^2$.*

Proof. First, observe that $a = \langle J_{11}v, v\rangle$ and $d = \langle J_{22}w, w\rangle$ are both negative real numbers for any $v \in \mathcal{S}_1$ and $w \in \mathcal{S}_2$ by assumption that x is a differential Nash equilibrium, i.e. $-D_i^2 f_i(x) < 0$ for each $i \in \{1, 2\}$. Second, observe that for potential games, $p = \langle J_{12}w, v\rangle = \overline{\langle J_{21}v, w\rangle}$. Therefore, for x to be stable, the eigenvalues of

$$J_{v,w} = \begin{bmatrix} a & p \\ \tau p^* & \tau d \end{bmatrix}$$

must all be negative. Hence, we compute the the trace and determinant conditions to be $\mathrm{tr}(J_{v,w}) = \lambda_1 + \lambda_2 = a + \tau d$ and $\det(J_{v,w}) = \lambda_1\lambda_2 = \tau(ad - |p|^2)$. Notice that $a + \tau d < 0 \iff a + d < 0$ and $\tau(ad - |p|^2) > 0 \iff ad - |p|^2 > 0 \iff ad > |p|^2 > 0$. In terms of the original matrix we have $\sigma_{\max}(J_{11})\sigma_{\max}(J_{22}) > \sigma_{\max}(P)^2$. □

We have shown that Nash equilibria in zero-sum games are robust in variation in learning rates, whereas Nash equilibria of potential games are not robust to variation in learning rates in general. For the latter case, we provide a sufficient condition that guarantees its robustness.

3.6.3 Instability in General-Sum Games

We provide a sufficient condition for the instability of fixed points of continuous games on finite dimensional spaces in general. Our results quantifies the contribution of the off-diagonal interaction terms of (3.3) in destabilizing equilibria in games.

We begin by expressing the game Jacobian as the sum of symmetric and skew-symmetric matrices, $J = \frac{1}{2}(J + J^\top) + \frac{1}{2}(J - J^\top)$. Let R be a rotation that diagonalizes $\frac{1}{2}(J + J^\top)$ and sorts the eigenvalues so that J decomposes into

$$RJR^\top = \begin{bmatrix} M_+ & 0 \\ 0 & M_- \end{bmatrix} + \begin{bmatrix} Z_1 & Z_2 \\ -Z_2^\top & Z_3 \end{bmatrix} \tag{3.15}$$

where $M_+ > 0$, $M_- \leq 0$ are diagonal and Z_1 and Z_3 are skew-symmetric. Let $\lambda^-(M_+) > 0$ be the minimum eigenvalue of M_+ and $\lambda^+(M_-) \leq 0$ be the maximum eigenvalue of M_-.

Theorem 10 (Sufficient Conditions for Instability in Games)**.** *Consider general-sum game (f_1, f_2) with $f_i \in C^2(X_1 \times X_2, \mathbb{R})$ where X_i is d_i-dimensional for each $i = 1, 2$. At a fixed point x,* $\operatorname{spec}(J(x)) \not\subset \mathbb{C}_-^\circ$ *if*

$$\|Z_2\| < \tfrac{1}{2}\big(|\lambda^+(M_-)| + |\lambda^-(M_+)|\big) < |\lambda^-(M_+)| \tag{3.16}$$

with M_+, M_- and Z_2 defined in (3.15).

Proof. Since Z_1 and Z_3 are skew-symmetric we have that $\operatorname{Re}\big(M_- + Z_3)\big) \leq \lambda^+(M_-) \leq 0$ and $0 \leq \lambda^-(M_+) \leq \operatorname{Re}\big(W(M_+ + Z_1)\big)$ (Tretter, 2009, Prop. 1.1.12). □

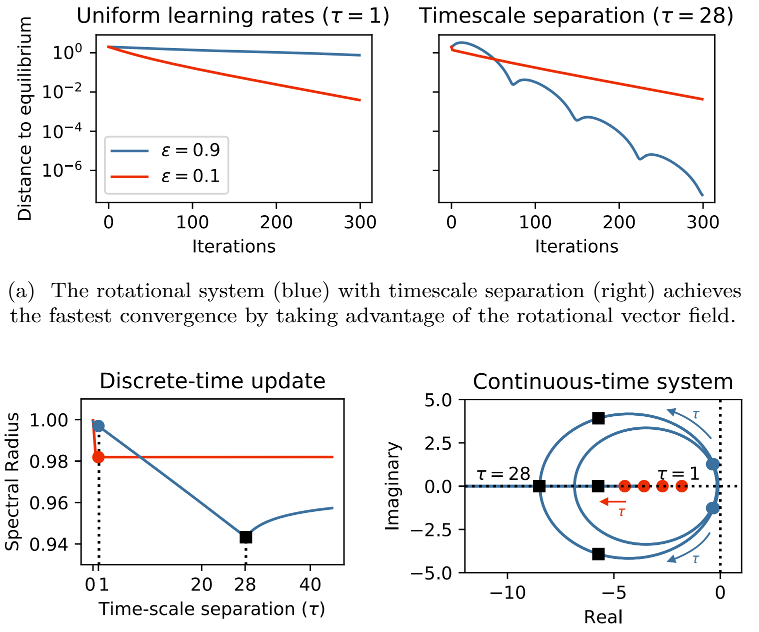
The result above works by bounding a non-empty subset of the eigenvalues of J in $\mathbb{C}_+^\circ$ to guarantee instability. The inequalities in (3.16) are the block matrix equivalent of being inside the circle of radius $\sqrt{h^2 + p^2}$ in the scalar case. The left inequality is analogous to being inside the circle in the vertical (z) direction; the right inequality is analogous to being right enough in the horizontal (m) direction. See

3.6.4 An Example with Timescale Separation

Example 3. *In this example, we explore how timescale separation can improve the convergence rate of game dynamics. In particular, we show that when a game has larger rotational components, timescale separation can lead to well-conditioned game dynamics and thus faster convergence.*

Consider a zero-sum game $\mathcal{G} = (f, -f)$ on $\mathbb{R}^2 \times \mathbb{R}^2$ with cost given by

$$f(x, y) = (1 - \varepsilon)\left(x_1^2 + \tfrac{3}{2}x_2^2 - 2y_1^2 - \tfrac{5}{2}y_2^2\right) + \varepsilon x^\top B y$$



(a) The rotational system (blue) with timescale separation (right) achieves the fastest convergence by taking advantage of the rotational vector field.

(b) The spectral radius of $I + \Gamma_\tau J(z)$ for the discrete-time update and the eigenvalues of $\Lambda_\tau J(z)$ for the continuous-time system $\dot{z} = \Lambda_\tau g(z)$ at equilibrium $z = (x, y) = 0$ for increasing learning rate ratio $\tau > 0$.

Figure 3.11: *Faster convergence of rotational learning dynamics with time-scale separation.* Time-scale separation can be used to speed up convergence of systems with mostly rotational components, shown in (a). The spectral radius of the discrete-time update and the eigenvalues of the corresponding continuous-time system are plotted in (b), showing that at $\tau = 28$, the mostly-rotational system achieves fastest convergence because it is able to take advantage of the imaginary components of the eigenvalues to achieve a smaller spectral radius.

and the matrix B is such that each entry is $B_{ij} = 1$ for each i, j except for $B_{22} = -1$. The parameter $0 \leq \varepsilon \leq 1$ controls the amount of rotational component in the game. Note that when $\varepsilon = 0$, the game Jacobian is symmetric; when $\varepsilon = 1$, the game Jacobian is skew-symmetric. The decomposition of the Jacobian is $J = (1 - \varepsilon)S + \varepsilon A$ where $S = S^\top$ and $A = -A^\top$. Suppose agents descend their individual gradient with individual learning rates $\gamma_1 = \gamma, \gamma_2 = \tau\gamma$, expressed relative to base learning rate γ, yielding learning dynamics

$$\begin{aligned} x(t+1) &= x(t) - \gamma D_1 f(x(t), y(t)) \\ y(t+1) &= y(t) + \gamma\tau D_2 f(x(t), y(t)). \end{aligned} \tag{3.17}$$

Recall that the spectrum of $\Lambda_\tau J(x, y)$ at an equilibrium (x, y) determines its stability and that the spectral radius of $I + \gamma\Lambda_\tau J(x, y)$ determines the convergence convergence rate of the discrete-time system above, where $\Lambda_\tau = blockdiag(I_1, \tau I_2)$ with identity matrices I_1, I_2.

By using a learning update with timescale separation between the players, players can take advantage of the rotational component of a vector field to converge at a faster rate. We simulate (3.17) from $(1, 1, 1, 1)$ and show that for $\varepsilon = 0.9$, the system converges fastest with $\tau = 28$ as shown in Fig. 3.11(a). Timescale *separation also warps the vector field of potential-like and rotation-like systems. For $\tau > 0$, the spectral radius of the discrete-time update for base learning rate $\gamma = 10^{-3}$ is shown in Fig. 3.11(b). It achieves a minimum at $\tau = 28$ for the mostly-rotational system. The eigenvalues of the corresponding continuous-time system is plotted in Fig. 3.11(b). This example shows that timescale separation can be used to speed up convergence of learning dynamics by taking advantage of the imaginary components of the eigenvalues of the linearized Jacobian near the equilibrium.*

3.7 Discussion and Conclusion

We provide a comprehensive characterization of the local stability and Nash optimality for fixed points of two-player gradient learning dynamics. We assess the contribution of the interaction terms of the game Jacobian in stabilizing a non-Nash equilibrium or destabilizing a Nash. Such results give valuable insights into the interaction of algorithms in settings most accurately modeled as games. In the numerical examples, we demonstrate that there is an important trade-off between the rotational component of the learning dynamics and timescale that each player learns at: timescale separation can be introduced to learning rules to improve convergence when the vector field has enough rotational component. As a future direction, we will look at how to optimize convergence speed given the strength of the anti-symmetric component of the game.

Chapter 4

Co-Adaptation Converges to Game-Theoretic Equilibria

Abstract

Adaptive machines have the potential to assist *or* interfere with human behavior in a range of contexts, from decision-making (Mehrabi et al., 2021; Sutton et al., 2020) to physical assistance (Felt et al., 2015; Slade et al., 2022; Zhang et al., 2017). Therefore it is critical to understand how these machines impact us, especially when their goals do not align with ours (Thomas et al., 2019). In our research, we explored how machines and humans influence each other when both are adapting. Since humans continually adapt to their environment (Heald et al., 2021; Taylor et al., 2014), when the environment contains an adaptive machine, the human and machine play a *game* (Başar and Olsder, 1998; von Neumann and Morgenstern, 1947). While game theory is an established framework for modeling interactions between decision-makers, existing approaches make assumptions about, rather than empirically test, how adaptation by individual humans is affected by interaction with an adaptive machine (Madduri et al., 2021; Nikolaidis et al., 2017). Through our continuous game experiments, we validated three different methods for machines to predict and influence the outcome of these interactions. One method allowed the machine to anticipate human reactions directly from observations of human actions, without needing to estimate the human's utility function as in prior work (Li et al., 2019; Ng and Russell, 2000). Another method allowed the machine to steer human actions to the machine's optimum directly from observations of its own cost, without needing to estimate the human's actions. This latter method was especially effective for the machine, raising important ethical concerns. It emphasizes the need to design these machines responsibly, ensuring the well-being of people.

4.1 Introduction

As adaptive machines become integral to daily life, understanding their influence on human behavior is crucial. These machines, when interacting with humans, create scenarios that can be modeled as games (Başar and Olsder, 1998; von Neumann and Morgenstern, 1947). While game theory offers a framework for such interactions, many approaches make assumptions about, rather than empirically test, how human adaptation is affected by adaptive machines. In this study, we explore the co-adaptive interactions between humans and machines, leveraging learning dynamics to predict and design outcomes. Our goal is to ensure that as machines become more integrated into our lives, human autonomy and well-being are prioritized.

We studied games played between humans H and machines M. The games were defined by quadratic functions that mapped scalar actions of each human h and machine m to costs $c_H(h, m)$ and $c_M(h, m)$. Games were played continuously in time over a sequence of trials, and the machine adapted within or between trials. Human actions h were determined from a manual input device (mouse or touchscreen) as in Figure 4.2a, while machine actions m were determined algorithmically from the machine's cost function c_M and the human's action h as in Figure 4.2b. The human's cost $c_H(h, m)$ was continuously shown to the human subjects via the height of a rectangle on a computer display as in Figure 4.2a, which the subject was instructed to "make as small as possible", while the machine's actions were hidden.

The experiments reported here were based on a game that is *continuous*, meaning that players choose their actions from a continuous set, and *general-sum*, meaning that the cost functions prescribed to the human and machine were neither aligned nor opposed. Unlike pure optimization problems, players cannot control all variables that determine their cost. Each player seeks its own preferred outcome, but the game outcome will generally represent a compromise between players' conflicting goals. We considered *Nash* (Nash, 1950), *Stackelberg* (von Stackelberg, 1934), *consistent conjectural variations* (Bowley, 1924), and *reverse Stackelberg* (Ho et al., 1982) equilibria of the game (Definitions 4.1, 4.6, 4.9, 7.1 in Başar and Olsder (1998) respectively), in addition to each player's *global optimum*, as possible outcomes in the experiments. Formal definitions of these game-theoretic concepts are provided in Section 4.2.1 of the Supplement, but we provide plain-language descriptions in the next paragraph. Table 1 contains expressions for the cost functions that defined the game considered here as well as numerical values of the resulting game-theoretic equilibria.

Nash equilibria (Nash, 1950) arise in games with simultaneous play, and constitute points in the joint action space from which neither player is incentivized to deviate (see Section 4.2 in Başar and Olsder (1998)). In games with ordered play where one player (the *leader*) chooses its action assuming the other (the *follower*) will play using its best response, a Stackelberg equilibrium (von Stackelberg, 1934) may arise instead. The leader in this case employs a *conjecture* about the follower's policy, i.e. a function from the leader's actions to the follower's actions, and this conjecture is consistent with how the follower plays the game (Section 4.5 in Başar and Olsder (1998)); the leader's conjecture can be regarded as an *internal model* (Huang et al., 2018; Nikolaidis et al., 2017; Wolpert et al., 1995) for the follower. Nash and Stackelberg equilibria arise from the limiting cases of timescale separation in gradient play. Shifting from Nash to Stackelberg equilibria in our quadratic setting is generally in favor of the leader whose cost decreases. Of course, the follower may then form a conjecture of its own about the leader's play, and the players may iteratively update their policies and conjectures in response to their opponent's play. In the game we consider, this iteration converges to a *consistent* conjectural variations equilibrium (Bowley, 1924) defined in terms of actions *and* conjectures: each player's conjecture is equal to their opponent's policy, and each player's policy is optimal with respect to its conjecture about the opponent (Section 7.1 in Başar and Olsder (1998)). Finally, if one player realizes how their choice of policy influences the other, they can design an *incentive* to steer the game to their preferred outcome, termed a *reverse* Stackelberg equilibrium (Ho et al., 1982) (Section 7.4.4 in Başar and Olsder (1998)). Our study examined these equilibria which were derived from the methods used in each of the three experiments, methods chosen based on the discoveries described in the discussion.

To study how adaptive machines can influence human behavior, we carefully designed three experiments, guided by theoretical and practical considerations. Quadratic games were chosen due to their rich yet tractable structure, allowing for closed-form solutions (Başar and Olsder, 1998; Varian, 1992). General-sum games were chosen to reflect the mix of cooperation and competition in real-world scenarios (Crandall et al., 2018). By carefully selecting our cost coefficients, we ensured distinct equilibria (Section 4.3) and stable learning dynamics (Section 4.6). To maintain controlled experimental conditions and task clarity, we opted for a manual input, clear cost display, and a fixed payout. Furthermore, our use of widely-used learning methods like gradient descent and optimization ensures that our findings can be extended to various contexts.

From the perspective of a designer of these methods, a critical consideration is determining the "preferred" outcome of the game. The preferred outcome depends on the overarching structure of the human-machine interaction and broader ethical implications. While multiple equilibria in games can often pose selection challenges, our design choices ensured that each learning algorithm converged to a unique, stable equilibrium, eliminating potential confounders.

4.2 Preliminaries

4.2.1 Game-Theoretic Preliminaries

We model co-adaptation between humans and machines using game theory (Başar and Olsder, 1998; von Neumann and Morgenstern, 1947). In this model, the human H chooses action $h \in \mathcal{H}$ while the machine M chooses action $m \in \mathcal{M}$ to minimize their respective *cost functions* $c_H, c_M : \mathcal{H} \times \mathcal{M} \to \mathbb{R}$,

$$\min_h c_H(h, m), \tag{4.1a}$$

$$\min_m c_M(h, m). \tag{4.1b}$$

It is important to note that the optimization problems in (4.1) are coupled. Since both problems must be considered simultaneously, there is no obvious candidate for a "solution" concept (in contrast to the case of pure optimization problems, where (local) minimizers of the single cost function are the obvious goals). Thus, we designed experiments to study a variety of candidate solution concepts that arise naturally in different contexts. We demonstrate that Nash, Stackelberg, consistent conjectural variations equilibria, and players' global optima are possible outcomes of the experiments.

Nash and Stackelberg equilibria

In games with simultaneous play where players do not form conjectures about the others' policy, a natural candidate solution concept is the *Nash equilibrium* (Definition 4.1 in (Başar and Olsder, 1998)).

Definition: The joint action $(h^{\texttt{NE}}, m^{\texttt{NE}}) \in \mathcal{H} \times \mathcal{M}$ constitutes a *Nash equilibrium* (NE) if

$$h^{\texttt{NE}} = \arg\min_h c_H(h, m^{\texttt{NE}}), \tag{4.2a}$$

$$m^{\texttt{NE}} = \arg\min_m c_M(h^{\texttt{NE}}, m). \tag{4.2b}$$

In games with ordered play where the *leader* (e.g. human) has knowledge of how the *follower* (e.g. machine) responds to choosing its own action, a natural candidate solution concept is the *(human-led) Stackelberg equilibrium* (Definition 4.6 in (Başar and Olsder, 1998)).

Definition: The joint action $(h^{\text{SE}}, m^{\text{SE}}) \in \mathcal{H} \times \mathcal{M}$ constitutes a *(human-led) Stackelberg equilibrium* (SE) if

$$h^{\text{SE}} = \arg\min_{h} \left\{ c_H\left(h, m\right) \mid m = \arg\min_{m'} c_M(h, m') \right\}, \tag{4.3a}$$

$$m^{\text{SE}} = \arg\min_{m} \; c_M(h^{\text{SE}}, m). \tag{4.3b}$$

The Stackelberg equilibrium is a solution concept that arises when one player (the leader) anticipates or models another player's (the follower's) best response.

Consistent conjectural variations equilibria

In repeated games where each player gets to observe the other's actions and policies, players may develop internal models or conjectures for how they expect the other to play. A natural candidate solution concept in this case is the *consistent conjectural variations equilibrium* (Definition 4.9 in (Başar and Olsder, 1998)).

In what follows, we use the shorthand $\{A \to B\}$ to denote the set of functions from A to B,

For a given pair$(v_H^{\text{CCVE}}, v_M^{\text{CCVE}}) \in \{\mathcal{M} \to \mathcal{H}\} \times \{\mathcal{H} \to \mathcal{M}\}$, denote the unique fixed points $(h^{\text{CCVE}}, m^{\text{CCVE}}) \in \mathcal{H} \times \mathcal{M}$ satisfying

$$h^{\text{CCVE}} = v_H^{\text{CCVE}} \circ v_M^{\text{CCVE}}(h^{\text{CCVE}}), \tag{4.4a}$$

$$m^{\text{CCVE}} = v_M^{\text{CCVE}} \circ v_H^{\text{CCVE}}(m^{\text{CCVE}}). \tag{4.4b}$$

Let

$$\Delta v_H^{\text{CCVE}}(m) = v_H^{\text{CCVE}}(m) - v_H^{\text{CCVE}}(m^{\text{CCVE}}), \tag{4.5a}$$

$$\Delta v_M^{\text{CCVE}}(h) = v_M^{\text{CCVE}}(h) - v_M^{\text{CCVE}}(h^{\text{CCVE}}), \tag{4.5b}$$

be the differential reactions of each player under their policies $(v_H^{\text{CCVE}}, v_M^{\text{CCVE}})$ to a deviation from the joint action $(h^{\text{CCVE}}, m^{\text{CCVE}})$ to (m, h).

Definition: The joint action $(h^{\text{CCVE}}, m^{\text{CCVE}}) \in \mathcal{H} \times \mathcal{M}$ together with the conjectures $v_M^{\text{CCVE}} : \mathcal{H} \to \mathcal{M}$, $v_H^{\text{CCVE}} : \mathcal{M} \to \mathcal{H}$ constitute a *consistent conjectural variations equilibrium* (CCVE) if we have the consistency of actions

$$h^{\text{CCVE}} = \arg\min_{h} \left\{ c_H(h, m) \mid m = v_M^{\text{CCVE}}(h) \right\},$$

$$m^{\text{CCVE}} = \arg\min_{m} \left\{ c_M(h, m) \mid h = v_H^{\text{CCVE}}(m) \right\},$$

and consistency of policies

$$v_H^{\text{CCVE}}(m) = \arg\min_h \ c_H(h, m + \Delta v_M^{\text{CCVE}}(h)),$$
$$v_M^{\text{CCVE}}(h) = \arg\min_m \ c_M(h + \Delta v_H^{\text{CCVE}}(m), m).$$

The consistent conjectural variations equilibrium is a solution concept that arises when players anticipate each other's actions and reactions.

Reverse Stackelberg equilibrium

In games where one player (the leader) has the ability to impose a policy before the other player (the follower) who responds to the policy, the candidate solution concept for this case is the *reverse Stackelberg equilibrium* (Ho et al., 1981, 1982). The machine acts as the leader in this game, and announces policy is $\pi : \mathcal{H} \to \mathcal{M}$. Assume the human's best response to machine policy π is $r : (\mathcal{H} \to \mathcal{M}) \to \mathcal{H}$ given by a constrained optimization problem:

$$r(\pi) := \arg\min_h \{c_H(h, m) \mid m = \pi(h)\}.$$

Definition: The joint action $(h^{\text{RSE}}, m^{\text{RSE}}) \in \mathcal{H} \times \mathcal{M}$ together with machine policy $\pi^{\text{RSE}} : \mathcal{H} \to \mathcal{M}$ constitute a *reverse Stackelberg equilibrium* (RSE) if

$$\pi^{\text{RSE}} = \arg\min_\pi \{c_H(h, m) \mid m = \pi(h),\ h = r(\pi))\}, \tag{4.6a}$$
$$h^{\text{RSE}} = r(\pi^{\text{RSE}}), \tag{4.6b}$$
$$m^{\text{RSE}} = \pi^{\text{RSE}}(h^{\text{RSE}}). \tag{4.6c}$$

If the reverse Stackelberg problem is incentive-controllable (Ho et al., 1981), then the reverse Stackelberg equilibrium is the machine's global optimum.

4.2.2 Prescribed Cost Functions and Informational Constraints

In Experiments 1, 2, and 3, participants were prescribed the quadratic cost function

$$c_H(h, m) = \tfrac{1}{2}h^2 + \tfrac{7}{30}m^2 - \tfrac{1}{3}hm + \tfrac{2}{15}h - \tfrac{22}{75}m + \tfrac{12}{125}; \tag{4.7}$$

the machine optimized the quadratic cost function

$$c_M(h, m) = \tfrac{1}{2}m^2 + h^2 - hm. \tag{4.8}$$

These costs were designed such that the players' optima and the constellation of relevant game-theoretic equilibria were distinct positions as listed in the Table 4.1. During each trial of an experiment, the time series of actions from the trials were recorded as human actions $h_0, \dots, h_t, \dots, h_T$ and machine actions $m_0, \dots, m_t, \dots, m_T$, for a fixed number of samples T. At time t, the players experienced costs $c_H(h_t, m_t)$ and $c_M(h_t, m_t)$. Neither player has any information the other's cost, and the human never observes the machine's action. However, the machine has the ability to observe the human's action, and through perturbations, also has the ability to estimate the human's policy. The next section will outline how the parameters for the costs were chosen.

4.3 Closed-Form Derivations of Game-Theoretic Equilibria

In this section, the equilibrium points are derived by solving linear equations while enforcing certain second-order and stability conditions. The general quadratic costs are given by

$$c_H(h, m) = \tfrac{1}{2}h^\top A_H h + h^\top B_H m + \tfrac{1}{2}m^\top D_H m + b_H^\top h + d_H^\top m + a_H, \tag{4.9a}$$

$$c_M(h, m) = \tfrac{1}{2}m^\top A_M m + m^\top B_M h + \tfrac{1}{2}h^\top D_M h + b_M^\top m + d_M^\top h + a_M. \tag{4.9b}$$

where actions $h \in \mathbb{R}^p,\ m \in \mathbb{R}^q$ are vectors with $p \geq 1$ and $q \geq 1$, cost parameters $A_H \in \mathbb{R}^{p\times p},\ D_H \in \mathbb{R}^{q\times q},\ A_M \in \mathbb{R}^{q\times q},\ D_M \in \mathbb{R}^{p\times p}$ are symmetric matrices, $B_H \in \mathbb{R}^{p\times q},\ B_M \in \mathbb{R}^{q\times p}$ are matrices, $b_H \in \mathbb{R}^p,\ d_H \in \mathbb{R}^q,\ b_M \in \mathbb{R}^p,\ d_M \in \mathbb{R}^q$ are vectors and $a_H \in \mathbb{R},\ a_M \in \mathbb{R}$ are scalars.

The cost parameters are chosen so that the equilibrium points are located at chosen points in the action spaces. Without loss of generality, A_H and A_M are the identity matrices to set the (arbitrary) scale for each player's cost. Subsequently, a_H, a_M are determined such that the minimum cost values for both players are 0. Finally, and also without loss of generality, $b_M = d_M = 0$ is determined to center the machine's cost at the origin in the joint action space. The six coefficients that remain to be determined are $B_H, B_M, D_H, D_M, b_H, d_H$. The parameters will determine the location of the equilibrium solutions of the game.

The first set of stationary points of interest are the global optimal points of each player's cost. In games, these points in general do not coincide with game-theoretic equilibria because they require coordination and cooperation that are not guaranteed. Unless the optima correspond to the same points in the decision space, it is not possible for agents to simultaneously be at both optima.

Global Optima The global optimization problems for the two players are

$$(h_H^*, m_H^*) = \operatorname*{argmin}_{h,m} \; c_H(h,m),$$

$$(h_M^*, m_M^*) = \operatorname*{argmin}_{h,m} \; c_M(h,m)$$

which have first-order conditions

$$\begin{bmatrix} A_H & B_H \\ B_H^\top & D_H \end{bmatrix} \begin{bmatrix} h_H^* \\ m_H^* \end{bmatrix} + \begin{bmatrix} b_H \\ d_H \end{bmatrix} = 0 \text{ and } \begin{bmatrix} D_M & B_M^\top \\ B_M & A_M \end{bmatrix} \begin{bmatrix} h_M^* \\ m_M^* \end{bmatrix} + \begin{bmatrix} d_M \\ b_M \end{bmatrix} = 0,$$

and second-order conditions that $\begin{bmatrix} A_H & B_H \\ B_H^\top & D_H \end{bmatrix}$ and $\begin{bmatrix} D_M & B_M^\top \\ B_M & A_M \end{bmatrix}$ are positive semi-definite. See Proposition 1.1.1 in (Bertsekas, 1999) for the formal statement of these conditions.

4.3.1 Nash and Stackelberg Equilibria

We derive the first- and second-order conditions for differential Nash and Stackelberg equilibria. These equilibria arise in our first experiment of the human-machine repeated game.

Nash equilibrium The coupled optimization problems for a Nash equilibrium $(h^{\texttt{NE}}, m^{\texttt{NE}})$ are

$$h^{\texttt{NE}} = \operatorname*{argmin}_{h} \; c_H(h, m^{\texttt{NE}}),$$

$$m^{\texttt{NE}} = \operatorname*{argmin}_{m} \; c_M(h^{\texttt{NE}}, m),$$

which have first-order conditions

$$\begin{bmatrix} A_H & B_H \\ B_M & A_M \end{bmatrix} \begin{bmatrix} h^{\texttt{NE}} \\ m^{\texttt{NE}} \end{bmatrix} + \begin{bmatrix} b_H \\ b_M \end{bmatrix} = 0$$

and second-order conditions $A_H \geq 0$ and $A_M \geq 0$. If the Jacobian $\begin{bmatrix} A_H & B_H \\ B_M & A_M \end{bmatrix}$ has eigenvalues with positive real parts, then the Nash equilibrium is stable under gradient play.

See Proposition 1 in (Ratliff et al., 2016) for necessary conditions for a local Nash equilibrium and for the stability result for continuous-time gradient play dynamics $\dot{h} = -\partial_h c_H(h,m),\ \dot{m} = -\partial_m c_M(h,m)$. See Proposition 2 in (Chasnov et al., 2020d) for the corresponding discrete-time gradient play dynamics $h^+ = h - \beta\partial_h c_H(h,m),\ m^+ = m - \alpha\partial_M c_M(h,m)$ for learning rates $\alpha, \beta > 0$ and learning rate ratio $\tau = \alpha/\beta$. As the learning rate ratio τ tends to ∞, the machine's action m adapts at a faster rate than h, which imposes a timescale separation between the two players.

Human-Led Stackelberg Equilibrium The coupled optimization problems for a human-led Stackelberg equilibrium $(h^{\mathrm{SE}}, m^{\mathrm{SE}})$ are

$$\begin{aligned} h^{\mathrm{SE}} &= \underset{h}{\operatorname{argmin}}\ \left\{ c_H(h,m') \middle|\ m' = \underset{m}{\operatorname{argmin}}\ c_H(h,m) \right\}, \\ m^{\mathrm{SE}} &= \underset{m}{\operatorname{argmin}}\ c_M(h^{\mathrm{SE}}, m), \end{aligned}$$

which have first-order conditions

$$\begin{bmatrix} A_H + L_{M,0}^\top B_H^\top & B_H + L_{M,0}^\top D_H \\ B_M & A_M \end{bmatrix} \begin{bmatrix} h^{\mathrm{SE}} \\ m^{\mathrm{SE}} \end{bmatrix} + \begin{bmatrix} b_H + L_{M,0}^\top d_H \\ b_M \end{bmatrix} = 0$$

with $L_{M,0} = -A_M^{-1} B_M$, and second-order conditions $A_M > 0$, $A_H - B_H A_M^{-1} B_M > 0$. See Proposition 4.3 in (Başar and Olsder, 1998) for a quadratic game formulation of the Stackelberg equilibrium, which admits only a pure-strategy Stackelberg equilibrium. See Proposition 1 in (Fiez et al., 2020) for conditions for a local Stackelberg equilibrium.

4.3.2 Conjectural Variations Equilibria

We derive the conjectural variations equilibria relevant that arise in our second experiment. The $k-$level conjectural variations equilibria form a sequence of equilibria that lead to the stable consistent conjectural variations equilibria as k goes to infinity.

$k-$level conjectural variations equilibria The coupled optimization problems for an intermediate conjectural variations equilibrium where the human maintains a consistent conjecture of the machine are

$$\begin{aligned} h_{k+1}^{\mathrm{CVE}} &= \underset{h}{\operatorname{argmin}}\ \left\{ c_H(h,m') \middle|\ m' = L_{M,k}(h - h_M^*) + m_M^* \right\}, \\ m_k^{\mathrm{CVE}} &= \underset{m}{\operatorname{argmin}}\ \left\{ c_M(h',m) \middle|\ h' = L_{H,k-1}(m - m_H^*) + h_H^* \right\}, \end{aligned}$$

which have first-order optimality conditions

$$\begin{bmatrix} A_H + L_{M,k}^\top B_H^\top & B_H + L_{M,k}^\top D_H \\ B_M + L_{H,k-1}^\top D_M & A_M + L_{H,k-1}^\top B_M^\top \end{bmatrix} \begin{bmatrix} h_{k+1}^{\text{CVE}} \\ m_k^{\text{CVE}} \end{bmatrix} + \begin{bmatrix} b_H + L_{M,k}^\top d_H \\ b_M + L_{H,k-1}^\top d_M \end{bmatrix} = 0$$

with initial condition $L_{M.0} = -A_M^{-1} B_M$ and iteration

$$\begin{aligned} L_{H,k+1} &= -(A_H + L_{M,k}^\top B_H^\top)^{-1}(B_H + L_{M,k}^\top D_H) \\ L_{M,k} &= -(A_M + L_{H,k-1}^\top B_M^\top)^{-1}(B_M + L_{H,k-1}^\top D_M) \end{aligned}$$

for $k = 0, 1, 2, \ldots$ with and the assumption that $A_H + B_H L_{M,k}$ and $A_M + B_M L_{H,k-1}$ are invertible. See Section 4.6 for more information about conditions under which this iteration converges for the particular parameters of the costs used in the main experiments.

Consistent Conjectural Variations Equilibrium From (Definition 4.9 in (Başar and Olsder, 1998)), the coupled optimization problems for the consistent conjectural variation equilibria are

$$\begin{aligned} h^{\text{CCVE}} &= \underset{h}{\operatorname{argmin}} \left\{ c_H(h, m') \mid m' = L_M^{\text{CCVE}}(h - h_M^*) + m_M^* \right\} \\ m^{\text{CCVE}} &= \underset{m}{\operatorname{argmin}} \left\{ c_M(h', m) \mid h' = L_H^{\text{CCVE}}(m - m_H^*) + h_M^* \right\} \end{aligned}$$

where $L_M^{\text{CCVE}}, L_H^{\text{CCVE}}$ solves the optimality conditions in the policy space equations from (Definition 4.10 in (Başar and Olsder, 1998)):

$$\begin{aligned} A_M L_M^{\text{CCVE}} + {L_H^{\text{CCVE}}}^\top B_M^\top L_M^{\text{CCVE}} + {L_H^{\text{CCVE}}}^\top D_M + B_M = 0, \\ A_H L_H^{\text{CCVE}} + {L_M^{\text{CCVE}}}^\top B_H^\top L_H^{\text{CCVE}} + {L_M^{\text{CCVE}}}^\top D_H + B_H = 0. \end{aligned}$$

The first-order optimality conditions in the action space of the coupled optimization problems are

$$\begin{bmatrix} A_H + {L_M^{\text{CCVE}}}^\top B_H^\top & B_H + {L_M^{\text{CCVE}}}^\top D_H \\ B_M + {L_H^{\text{CCVE}}}^\top D_M & A_M + {L_H^{\text{CCVE}}}^\top B_M \end{bmatrix} \begin{bmatrix} h^{\text{CCVE}} \\ m^{\text{CCVE}} \end{bmatrix} + \begin{bmatrix} b_H + {L_M^{\text{CCVE}}}^\top d_H \\ b_M + {L_H^{\text{CCVE}}}^\top d_M \end{bmatrix} = 0.$$

Proposition 4.5 in (Başar and Olsder, 1998) states that if a game admits a unique Nash equilibirum, then the Nash equilibrium is also a CCVE with the Nash actions as constant policies.

4.3.3 Reverse Stackelberg Equilibrium

Finally, we derive the conditions of the reverse Stackelberg equilibrium of the third experiment. This equilibrium corresponds to the stable equilibrium of the policy optimization process described in Section **??** .

Machine-Led Reverse Stackelberg Equilibrium The coupled optimization problems corresponding to a machine-led reverse Stackelberg equilibrium are given by:

$$\begin{aligned} r_H^{\mathrm{RSE}}(L_M) &= \operatorname*{argmin}_{h} \ \left\{c_H(h, m') \mid m' = L_M(h - h_M^*) + m_M^*\right\} \\ L_M^{\mathrm{RSE}} &= \operatorname*{argmin}_{L_M} \ \left\{c_M(r_H^{\mathrm{RSE}}(L_M), m') \mid m' = L_M(r_H^{\mathrm{RSE}}(L_M) - h_M^*) + m_M^*)\right\} \end{aligned}$$

where the human forms a consistent conjecture of the machine, and the machine assumes that the human responds optimally to the machine's policy slope. The reverse Stackelberg equilibrium is $(h^{\mathrm{RSE}}, m^{\mathrm{RSE}})$, which by the (Başar and Selbuz, 1979; Groot et al., 2013), satisfies the same conditions that the machine's optimum satisfies, i.e.

$$\begin{bmatrix} A_M & B_M \\ B_M^\top & D_M \end{bmatrix} \begin{bmatrix} h^{\mathrm{RSE}} \\ m^{\mathrm{RSE}} \end{bmatrix} + \begin{bmatrix} b_M \\ d_M \end{bmatrix} = 0$$

as well as first-order optimality conditions

$$\begin{bmatrix} A_H + {L_M^{\mathrm{RSE}}}^\top B_H^\top & B_M + {L_M^{\mathrm{RSE}}}^\top D_H \\ -L_M^{\mathrm{RSE}} & I \end{bmatrix} \begin{bmatrix} h^{\mathrm{RSE}} \\ m^{\mathrm{RSE}} \end{bmatrix} + \begin{bmatrix} b_H + {L_M^{\mathrm{RSE}}}^\top d_H \\ m_M^* - {L_M^{\mathrm{RSE}}}^\top h_M^* \end{bmatrix} = 0$$

where we need to also guarantee that the Jacobian is stable. The second-order condition is $A_H + B_H L_M^{\mathrm{RSE}} > 0$. See Section III.B in (Ho et al., 1981) for a method to solve reverse Stackelberg problems, relying on the property of linear incentive controllability. See (Groot et al., 2013) for an overview of results and the computation of optimal policies. See Proposition 1 of (Zheng and Başar, 1982) for existence of optimal affine leader policies.

4.3.4 Choosing Parameters for a Continuous Game with Scalar Actions

We now specialize the equations above to scalar actions, which allows us to use quadratic formula to express the parameters as a function of the equilibrium points. Given quadratic costs with scalar actions $h \in \mathbb{R},\ m \in \mathbb{R}$,

$$\begin{aligned} c_H(h,m) &= \tfrac{1}{2}A_H h^2 \ \ + B_H hm + \tfrac{1}{2}D_H m^2 + b_H h + d_H m + a_H, \\ c_M(h,m) &= \tfrac{1}{2}A_M m^2 + B_M hm + \tfrac{1}{2}D_M h^2 + b_M m + d_M h + a_M. \end{aligned}$$

Without loss of generality, $A_H = 1$ and $A_M = 1$ to set the scale for each player's cost. The parameters expressed in terms of the optima (h_H^*, m_H^*) and (h_M^*, m_M^*) are

$$\begin{aligned} a_H &= \tfrac{1}{2}A_H {h_H^*}^2 + B_H h_H^* m_H^* + \tfrac{1}{2}D_H {m_H^*}^2, & b_H &= -A_H h_H^* - B_H m_H^*, & d_H &= -B_H h_H^* - D_H m_H^*, \\ a_M &= \tfrac{1}{2}A_M {m_M^*}^2 + B_M h_M^* m_M^* + \tfrac{1}{2}D_M {h_M^*}^2, & b_M &= -A_M m_M^* - B_M h_M^*, & d_M &= -B_M m_M^* - D_M h_M^*. \end{aligned}$$

The parameters expressed in terms of the optima and the Nash equilibrium $(h^{\texttt{NE}}, m^{\texttt{NE}})$ are

$$B_H = -\frac{h_H^* - h^{\texttt{NE}}}{m_H^* - m^{\texttt{NE}}},\ B_M = -\frac{m_M^* - m^{\texttt{NE}}}{h_M^* - h^{\texttt{NE}}}.$$

The parameter expressed in terms of the optima and the human-led Stackelberg equilibrium $(h^{\mathrm{SE}}, m^{\mathrm{SE}})$ is

$$\begin{aligned} D_H = {} & \frac{B_H\big(h_M^* m_H^* + h_H^* m_M^* - (m_H^* + m_M^* - m^{\mathrm{SE}})h^{\mathrm{SE}} - (h_H^* + h_M^* - h^{\mathrm{SE}})m^{\mathrm{SE}}\big)}{(m_H^* - m^{\mathrm{SE}})(m_M^* - m^{\mathrm{SE}})} \\ & + \frac{(h_H^* - h^{\mathrm{SE}})(h_M^* - h^{\mathrm{SE}})}{(m_H^* - m^{\mathrm{SE}})(m_M^* - m^{\mathrm{SE}})} \end{aligned}$$

and $A_H - B_H A_M^{-1} B_M$ must be positive definite.

The remaining parameter to be chosen is D_M. It must satisfy the following conditions:

$$\begin{aligned} (A_H A_M - D_H D_M)^2 - 4(A_M B_H - B_M D_H)(A_H B_M - B_H D_M) &\geq 0, \\ (A_M B_H - B_M D_H)(A_H B_M - B_H D_M) &\neq 0 \end{aligned}$$

The CCVE is determined by the solution of two quadratic equations. The policy slopes for each agent are

$$\begin{aligned} L_H^{\mathrm{CCVE}} &= \frac{D_H D_M - A_H A_M \pm \sqrt{4(A_M B_H - B_M D_H)(B_H D_M - A_H B_M) + (A_H A_M - D_H D_M)^2}}{2A_H B_M - 2B_H D_M}, \\ L_M^{\mathrm{CCVE}} &= \frac{D_H D_M - A_H A_M \pm \sqrt{4(A_M B_H - B_M D_H)(B_H D_M - A_H B_M) + (A_H A_M - D_H D_M)^2}}{2A_M B_H - 2B_M D_H}, \end{aligned}$$

and the actions are

$$\begin{bmatrix} h^{\text{CCVE}} \\ m^{\text{CCVE}} \end{bmatrix} = \begin{bmatrix} A_H + L_M^{\text{CCVE}} B_H & B_M + L_M^{\text{CCVE}} D_H \\ B_M + L_H^{\text{CCVE}} D_H & A_M + L_H^{\text{CCVE}} B_M \end{bmatrix}^{-1} \begin{bmatrix} b_H + L_M^{\text{CCVE}} d_H \\ b_M + L_H^{\text{CCVE}} d_M \end{bmatrix}$$

The reverse Stackelberg equilibrium is determined by policy slopes

$$L_H^{\text{RSE}} = \frac{h_H^* - h_M^*}{m_H^* - m_M^*},\ L_M^{\text{RSE}} = -\frac{A_H L_H^{\text{RSE}} + B_H}{B_H L_H^{\text{RSE}} + D_H},$$

and actions $h^{\text{RSE}} = h_M^*,\ m^{\text{RSE}} = m_M^*$.

In the experiments, the action spaces are scalar, i.e. $p = q = 1$. The parameters were chosen to be $A_H = 1,\ B_H = -1/3,\ D_H = 7/15,\ b_H = 2/15,\ d_H = -22/75$ for the human and $A_M = 1,\ B_M = -1,\ D_M = 2,\ b_M = 0,\ d_M = 0$ for the machine. The players' optima for this game are

$$(h_H^*, m_H^*) = (0.1, 0.7),$$
$$(h_M^*, m_M^*) = (0, 0),$$

and the game-theoretic equilibria are

$$(h^{\mathtt{NE}}, m^{\mathtt{NE}}) = (-0.2, -0.2),$$
$$(h^{\text{SE}}, m^{\text{SE}}) = (0.2, 0.2),$$
$$(h^{\text{CCVE}}, m^{\text{CCVE}}) \approx (0.276, 0.373),$$
$$(h^{\text{RSE}}, m^{\text{RSE}}) = (0, 0).$$

Compiling all the information from the previous sections about determining equilibria and parameters, we arrive at the points listed in Table 4.1 for our chosen cost functions. For the purposes of this paper, we will derive all the results for these cost parameters to keep things simple. Other parameterizations of the costs will apply so long as the stability and equilibrium conditions are satisfied at the various equilibria.

4.4 Experiment Design of Co-Adaptation with Human Participants

To verify the theory, we run a series of experiments on based on a human-machine interaction task that reflects the theoretical assumptions stated in previous sections. Human subjects were recruited using an online crowd-sourcing research platform *Prolific* (Palan and Schitter, 2018). Experiments were conducted using procedures approved by the University of Washington Institutional Review Board (UW IRB STUDY00013524).

Cost functions and game-theoretic equilibria

H's cost function	**M's cost function**
$c_H(h,m) = \frac{1}{2}h^2 + \frac{7}{30}m^2 - \frac{1}{3}hm + \frac{2}{15}h - \frac{22}{75}m + \frac{12}{125}$	$c_M(h,m) = \frac{1}{2}m^2 + h^2 - hm$

game-theoretic equilibria	**H's and M's actions**	**H's and M's policy slopes**	
H's optimum	$(h_H^*, m_H^*) = (+0.1, +0.7)$		
M's optimum	$(h_M^*, m_M^*) = (0, 0)$		
Nash equilibrium	$(h^{\mathtt{NE}}, m^{\mathtt{NE}}) = (-0.2, -0.2)$		
human-led Stackelberg equilibrium	$(h^{\text{SE}}, m^{\text{SE}}) = (+0.2, +0.2)$	$L_H^{\text{SE}} = -0.2,$	$L_M^{\text{SE}} = 1$
consistent conjectural variations equilibrium	$(h^{\text{CCVE}}, m^{\text{CCVE}}) \approx (0.276, 0.373)$	$L_H^{\text{CCVE}} \approx -0.54,$	$L_M^{\text{CCVE}} \approx +1.35$
machine-led reverse Stackelberg equilibrium (equal to M's optimum)	$(h^{\text{RSE}}, m^{\text{RSE}}) = (0, 0)$	$L_H^{\text{RSE}} = 1/7,$	$L_M^{\text{RSE}} = 5/11$

Table 4.1: **Cost functions and game-theoretic equilibria of the game studied in Experiments 1, 2, and 3.**

Participant data were collected on a secure web server. Each experiment consisted of a sequence of trials: 14 trials in the first experiment, 20 trials in the second and third experiments. During each trial, participants used a web browser to view a graphical interface and provide manual input from a mouse or touchscreen to continually determine the value of a scalar action $h \in \mathbb{R}$. This cursor input was scaled to the width of the participant's web browser window such that $h = -1$ corresponded to the left edge and $h = +1$ corresponded to the right edge. Data were collected at 60 samples per second for a duration of 40 seconds per trial in the first experiment and 20 seconds per trial in the second and third experiments. Human subjects were selected from the "standard sample" study distribution from all countries available on Prolific. Each subject participated in only one of the three experiments. No other screening criteria were applied.

At the beginning of each experiment, an introduction screen was presented to participants with the task description and user instructions. At the beginning of each trial, participants were instructed to move the cursor to a randomly-determined position. This procedure was used to introduce randomness in the experiment initialization and to assess participant attention. Throughout each trial, a rectangle's height displayed the current value of the human's cost $c_H(h, m)$ and participant was instructed to "keep this [rectangle] as *small* as possible" by choosing an action $h \in \mathbb{R}$ while the machine updated its action $m \in \mathbb{R}$. A square root function was applied to cost values to make it easier for participants to perceive small differences in low cost values. After a fixed duration, one trial ended and the next trial began. Participants were offered the opportunity to take a rest break for half a minute between every three trials. The experiment ended after a fixed number of trials. Afterward, the participant filled out a task load survey (Hart and Staveland, 1988) and optional feedback form. Each experiment lasted approximately 10–14 minutes and the participants received a fixed compensation of \$2 USD (all data was collected in 2020). A video illustrating the first three trials of Experiment 1 is provided as Movie S1. The user interface presented to human subjects was identical in all experiments. However, the machine adapted its action and policy throughout each experiment, and the adaptation algorithm differed in each experiment.

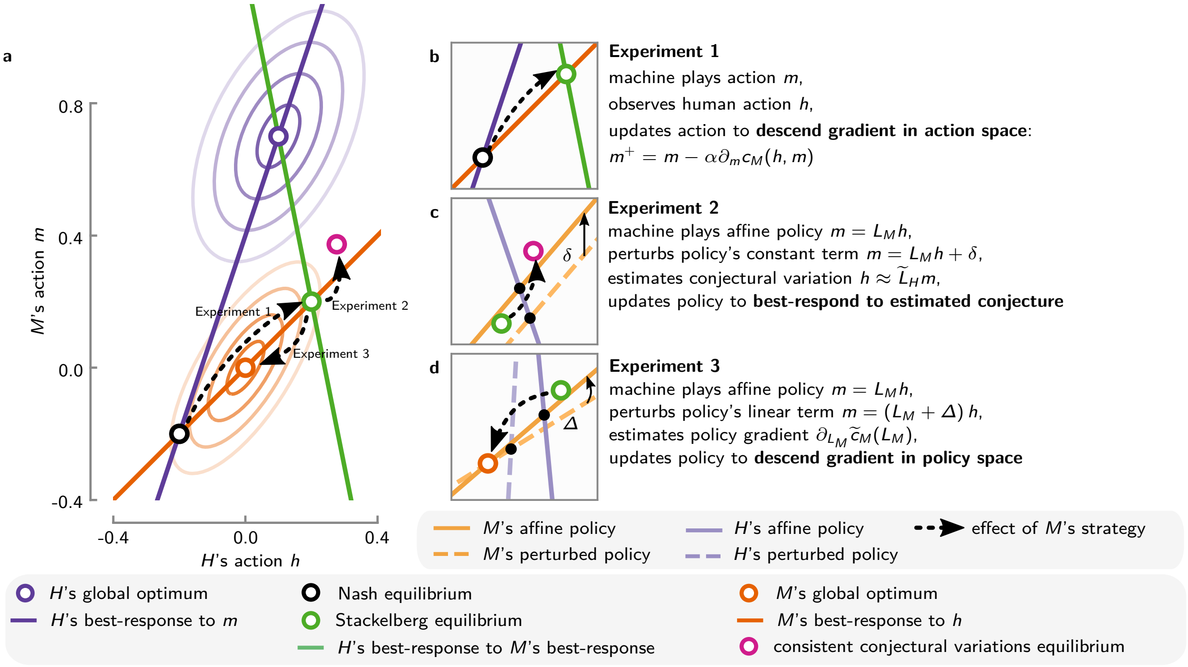

a
0.8
0.4
0.0
-0.4
M's action m
-0.4
0.0
0.4
H's action h
Experiment 1
Experiment 2
Experiment 3
b
Experiment 1
machine plays action m,
observes human action h,
updates action to **descend gradient in action space**:
$m^{+} = m - \alpha \partial_m c_M(h, m)$
c
δ
Experiment 2
machine plays affine policy $m = L_M h$,
perturbs policy's constant term $m = L_M h + \delta$,
estimates conjectural variation $h \approx \widetilde{L}_H m$,
updates policy to **best-respond to estimated conjecture**
d
Δ
Experiment 3
machine plays affine policy $m = L_M h$,
perturbs policy's linear term $m = (L_M + \Delta)\, h$,
estimates policy gradient $\partial_{L_M} \widetilde{c}_M(L_M)$,
updates policy to **descend gradient in policy space**
M's affine policy
M's perturbed policy
H's affine policy
H's perturbed policy
effect of M's strategy
H's global optimum
H's best-response to m
Nash equilibrium
Stackelberg equilibrium
H's best-response to M's best-response
M's global optimum
M's best-response to h
consistent conjectural variations equilibrium

Figure 4.1: **Overview of co-adaptation experiment between human and machine.** Human subject H is instructed to provide manual input h to make a black bar on a computer display as small as possible. The machine M has its own prescribed cost c_M chosen to yield game-theoretic equilibria that are distinct from each other and from each player's global optima. (**a**) Joint action space illustrating game-theoretic equilibria and response functions determined from the costs prescribed to human and machine: *global optima* defined by minimizing with respect to both variables; *best-response* functions defined by fixing one variable and minimizing with respect to the other. Machine plays different strategies in three experiments: (**b**) gradient descent in *Experiment 1*; (**c**) conjectural variation in *Experiment 2*; (**d**) policy gradient descent in *Experiment 3*.

4.4.1 Experiment 1: Gradient Descent in Action Space

In the first experiment, the machine adapted its action using gradient descent,

$$m^{+} = m - \alpha\, \partial_m c_M(h, m), \tag{4.10}$$

with one of seven different choices of adaptation rate $\alpha \in \{0, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$. At the slowest adaptation rate $\alpha = 0$, the machine implemented the constant policy $m = -0.2$, which is the machine's component of the game's Nash equilibrium. At the fastest adaptation rate $\alpha = 1$, the gradient descent iterations in (4.10) are such that the machine implements the linear policy $m = h$. Each condition was experienced twice by each human subject, once per symmetry (described in the next paragraph), in randomized order.

To help prevent human subjects from memorizing the location of game equilibria, at the beginning of each trial a variable s was chosen uniformly at random from $\{-1, +1\}$ and the map $h \mapsto s\,h$ was applied to the human subject's manual input for the duration of the trial. When the variable's value was $s = -1$, this had the effect of applying a "mirror" symmetry to the input. The joint action was initialized uniformly at random in the square $[-0.4, +0.4] \times [-0.4, +0.4] \subset \mathbb{R}^2$. Each trial lasted 40 seconds.

4.4.2 Experiment 2: Conjectural Variation in Policy Space

In the second experiment, the machine adapted its policy by estimating a *conjecture* about the human's *policy*. To collect the data that was used to form its estimate, the machine played an affine policy in two consecutive trials that differed solely in the constant term,

$$\text{nominal policy } m = L_M h, \tag{4.11a}$$

$$\text{perturbed policy } m' = L_M h' + \delta. \tag{4.11b}$$

The machine used the median action vectors $(\widetilde{h}, \widetilde{m})$, $(\widetilde{h}', \widetilde{m}')$ from the pair of trials to estimate a conjecture about the human's policy using a ratio of differences,

$$\widetilde{L}_H = \frac{\widetilde{h}' - \widetilde{h}}{\widetilde{m}' - \widetilde{m}}, \tag{4.12}$$

which is shown to be an estimate of the variation of the human's action in response to machine action in Proposition 12 of Supplement Section 4.6.2. The machine used this estimate of the human's policy to update its policy as

$$L_M^+ = \frac{1 - 2\widetilde{L}_H}{1 - \widetilde{L}_H}, \tag{4.13a}$$

which is shown to be the machine's best-response given its conjecture about the human's policy in Supplement Section 4.6. In the next pair of trials, the machine employs $m = L_M^+ h + \ell_M^+$ as its policy. This conjectural variation process was iterated 10 times starting from the initial conjecture $\widetilde{L}_H = 0$, which yields the initial best-response policy $m = h$.

In this experiment, the machine's policy slopes $L_{M,0}, L_{M,1}, \ldots, L_{M,k}, \ldots, L_{M,K-1}$ and the machine's conjectures about the human's policy slopes $\widetilde{L}_{H,0}, \widetilde{L}_{H,1}, \ldots, \widetilde{L}_{H,k}, \ldots, \widetilde{L}_{H,K-1}$ were recorded for each conjectural variation iteration $k \in \{0, \ldots, K-1\}$ where $K = 10$ iterations. In addition, the time series of actions within each trial as in the first experiment, with each trial now lasting only 20 seconds, yielding $T = 1200$ samples used to compute the median action vectors used in (4.12).

4.4.3 Experiment 3: Gradient Descent in Policy Space

In the third experiment, the machine adapted its policy using a policy gradient strategy by playing an affine policy in two consecutive trials that differed only in the linear term,

$$\text{nominal policy } m = L_M h, \tag{4.14a}$$

$$\text{perturbed policy } m' = (L_M + \Delta) h'. \tag{4.14b}$$

The machine used the median action vectors $(\widetilde{h}, \widetilde{m})$, $(\widetilde{h}', \widetilde{m}')$ from the pair of trials to estimate the gradient of the machine's cost with respect to the linear term in its policy, and this linear term was adjusted to decrease the cost. Specifically, an auxiliary cost was defined as

$$\widetilde{c_M}(L_M) := c_M \left(h, L_M(h - h_M^*) + m_M^*\right), \tag{4.15}$$

and the pair of trials were used to obtain a finite-difference estimate of the gradient of the machine's cost with respect to the slope of the machine's policy,

$$\partial_{L_M} \widetilde{c_M}(L_M) \approx \frac{1}{\Delta}\big(\widetilde{c_M}(L_M + \Delta) - \widetilde{c_M}(L_M)\big). \tag{4.16}$$

The machine used this derivative estimate to update the linear term in its policy by descending its cost gradient,

$$L_M^+ = L_M - \gamma \, \partial_{L_M} \widetilde{c_M}(L_M) \tag{4.17}$$

where γ is the policy gradient adaptation rate parameter ($\gamma = 2$ in this Experiment).

4.4.4 Statistical Analyses

To determine the statistical significance of our results, we use one- or two-sided t-tests with threshold $P \leq 0.05$ applied to distributions of median data from populations of $n = 20$ subjects. To estimate the effect size, we calculated Cohen's d by subtracting the equilibrium value from the mean of the distribution then dividing that by the standard deviation of the distribution.

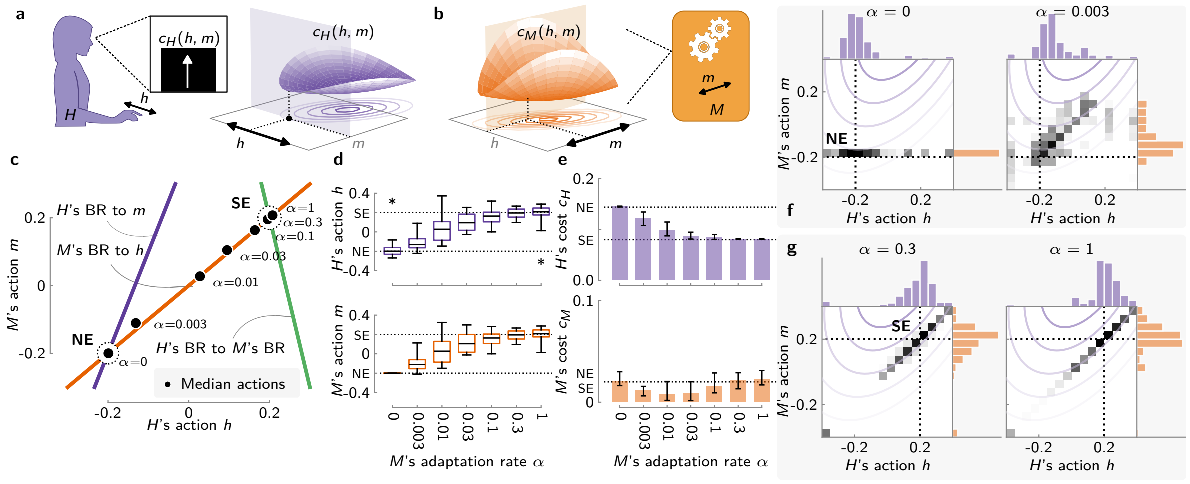


Figure 4.2: **Gradient descent in action space (Experiment 1, $n = 20$).** (**a**) Each human subject H is instructed to provide manual input h to make a black bar on a computer display as small as possible. The bar's height represents the value of a prescribed cost c_H. (**b**) The machine M has its own cost c_M chosen to yield game-theoretic equilibria that are distinct from each other and from each player's global optima. The machine knows its cost and observes human actions h. In this experiment, the machine updates its action by gradient descent on its cost $\frac{1}{2}m^2 - hm + h^2$ with adaptation rate α. (**c**) Median joint actions for each α overlaid on game-theoretic equilibria and best-response (BR) curves that define the Nash equilibrium (NE) and Stackelberg equilibrium (SE), respectively). (**d**) Action distributions for each machine adaptation rate displayed by box-and-whiskers plots showing 5th, 25th, 50th, 75th, and 95th percentiles. Statistical significance ($*$) determined by comparing to NE and SE using two-sided t-tests ($^*P \leq 0.05$). (**e**) Cost distributions for each machine adaptation rate displayed using box plots with error bars showing 25th, 50th, and 75th percentiles. (**f,g**) One- and two-dimensional histograms of actions for different adaptation rates ($\alpha \in \{0{,}0.003\}$ in (f), $\alpha \in \{0.3,\ 1\}$ in (g)) with game-theoretic equilibria overlaid (NE in (f), SE in (g)).

4.5 Experimental Results

We conducted three experiments with different populations of human subjects recruited from a crowd-sourcing research platform. The participants engaged in tasks defined by a pair of quadratic cost functions c_H, c_M illustrated in Figure 4.2a,b. The experiments were designed to yield distinct game-theoretic equilibria in both action and policy spaces. These analytically-determined equilibria were compared with the empirical distributions of actions and policies reached by humans and machines over a sequence of trials in each experiment. In all three experiments, we found that the observed behavior converges to the predicted game-theoretic values.

4.5.1 Timescale Separation Leads to Leader-Follower Dynamics

In our first experiment (Figure 4.2), the machine adapted its action within trials using what is arguably the simplest optimization scheme: gradient descent (Chasnov et al., 2020d; Ma et al., 2019). We tested seven adaptation rates $\alpha \geq 0$ for the gradient descent algorithm as illustrated in Figure 4.2c,d,e for each human subject, with two repetitions for each rate and the sequence of rates occurring in random order.

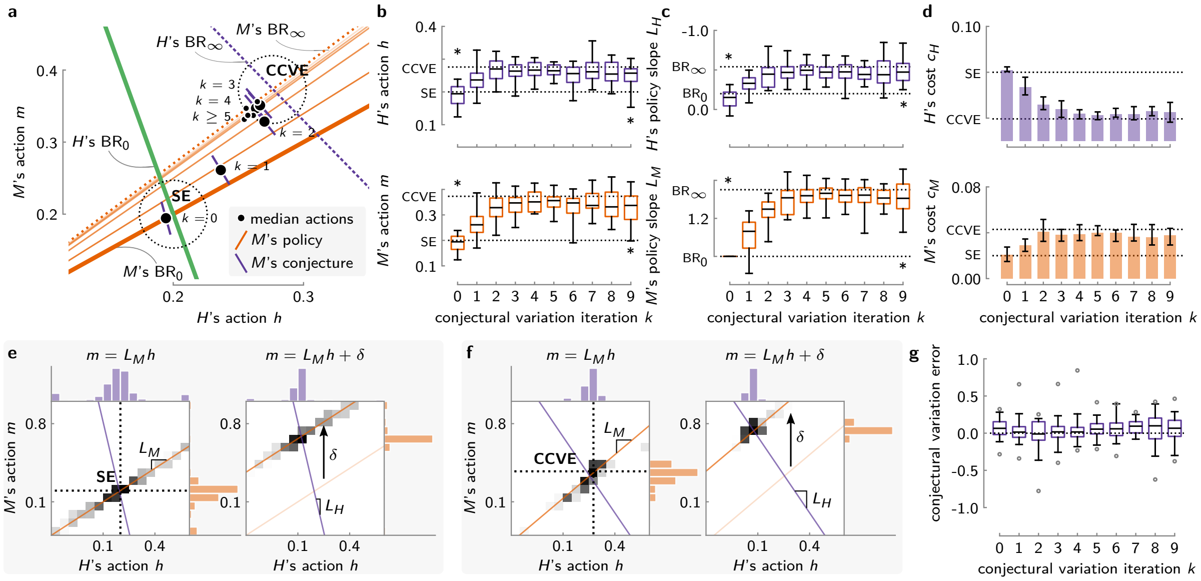


Figure 4.3: **Conjectural variation in policy space (Experiment 2, $n = 20$).** Experimental setup and costs are the same as Figure 4.2a,b except that the machine uses a different adaptation algorithm: in this experiment M iteratively implements policies $m = L_M h$, $m = L_M + \delta$ to measure and best-respond to conjectures of the human's policy and updates the policy slope L_M. (**a**) Median actions, conjectures, and policies for each conjectural variation iteration k overlaid on game-theoretic equilibria corresponding to best-responses (BR) at initial and limiting iterations (BR$_0$ and BR$_\infty$, respectively) predicted from Stackelberg equilibrium (SE) and consistent conjectural variations equilibrium (CCVE) of the game, respectively. (**b**) Action distributions for each iteration displayed by box-and-whiskers plots as in Figure 4.2d, with statistical significance ($*$) analogously determined using the same tests by comparing to SE and CCVE. (**c**) Policy slope distributions for each iteration displayed with the same conventions as (b); note that the sign of the top y-axis is reversed for consistency with other plots. Statistical significance ($*$) determined as in (b) by comparing to initial and limiting best-responses using two-sided t-tests ($^*P \leq 0.05$). (**d**) Cost distributions for each iteration displayed using box-and-whiskers plots as in Figure 4.2e. (**e,f**) One- and two-dimensional histograms of actions for different iterations ($k = 0$ in (e), $k = 9$ in (f)) with policies and game-theoretic equilibria overlaid (SE and BR$_0$ in (e), CCVE and BR$_\infty$ in (f)). (**g**) Error between measured and theoretically-predicted machine conjectures about human policies at each iteration displayed as box-and-whiskers plots as in (b,c).

We found that distributions of median action vectors for the population of $n = 20$ human subjects in this experiment shifted from the *Nash equilibrium* (NE) at the slowest adaptation rate to the *human-led Stackelberg equilibrium* (SE) at the fastest adaptation rate (Figure 4.2c). Importantly, this result would not have obtained if the human was also adapting its action using gradient descent, as merely changing adaptation rates in simultaneous gradient play does not change stationary points (Chasnov et al., 2020d). The shift we observed from Nash to Stackelberg, which was in favor of the human (Figure 4.2e), was statistically significant in that the distribution of actions was distinct from SE but not NE at the slowest adaptation rate and vice-versa for the fastest rate (Figure 4.2d); $^*P \leq 0.05$). Discovering that the human's empirical play is consistent with the theoretically-predicted best-response function for its prescribed cost is important, as this insight motivated us in subsequent experiments to elevate the machine's play beyond the action space to reason over its space of *policies*, that is, functions from human actions to machine actions.

4.5.2 Modeling Opponents Leads to Consistency of Policies and Beliefs

In our second experiment (Figure 4.3), the machine played affine policies (i.e. m was determined as an affine function of h) and adapted its policies by observing the human's response. The affine policy $m = L_M h + \delta$, with real-valued slope L_M, was derived from the unique solution of the machine's optimization problem. Trials came in pairs, with the machine's policy in each pair differing only in the constant term δ. After each pair of trials, the machine used the median action vectors from the pair to estimate a *conjecture* (Bowley, 1924; Figuières et al., 2004) (or *internal model* (Huang et al., 2018; Nikolaidis et al., 2017; Wolpert et al., 1995)) about the human's policy, and the machine's policy was updated to be optimal with respect to this conjecture. Unsurprisingly, the human adapted its own policy in response. Iterating this process shifted the distribution of median action vectors for a population of $n = 20$ human subjects (distinct from the population in the first experiment) from the *human-led Stackelberg equilibrium* (SE) toward a *consistent conjectural variations equilibrium* (CCVE) in action and policy spaces (Figure 4.3a). The shift we observed away from SE toward CCVE from the first to last iteration was statistically significant in action space and policy space (Figure 4.3b,c; $^*P \leq 0.05$). This shift was in favor of the human at the machine's expense (Figure 4.3d). The machines' empirical conjectures were not significantly different from theoretical predictions of human policies at all conjectural variation iterations (Figure 4.3g; $P > 0.05$). suggesting that both humans and machines estimated consistent conjectures of their opponent.

4.5.3 Policy Optimization Enables Behavior Manipulation by Fast-Learning Machine

In our third experiment (Figure 4.4), the machine adapted its affine policy using a *policy gradient* strategy (Chasnov et al., 2020d). Trials again came in pairs, with the machine's policy in each pair differing this time only in the slope term. After a pair of trials, the median costs of the trials were used to estimate the gradient of the machine's cost with respect to the linear term in its policy, and the linear term was adjusted in the direction opposing the gradient to decrease the cost. Iterating this process shifted the distribution of median action vectors for a population of human subjects (distinct from the populations in the first two experiments) from the *human-led Stackelberg equilibrium* (SE) toward the machine's *global optimum* (Figure 4.4a), which can also be regarded as a *reverse Stackelberg equilibrium* (RSE) (Ho et al., 1982), this time optimizing the machine's cost at the human's expense (Figure 4.4d). The shift we observed away from SE toward RSE from the first to last iterations was statistically significant in action space (Figure 4.4b; $^*P \leq 0.05$). We also observed a shift from SE to RSE in policy space (Figure 4.4c), but the last iteration's interquartile range excluded the RSE, possibly due to bias from estimating a non-linear gradient, unlike the linear gradients in the first two experiments. The machines' empirical policy gradients were not significantly

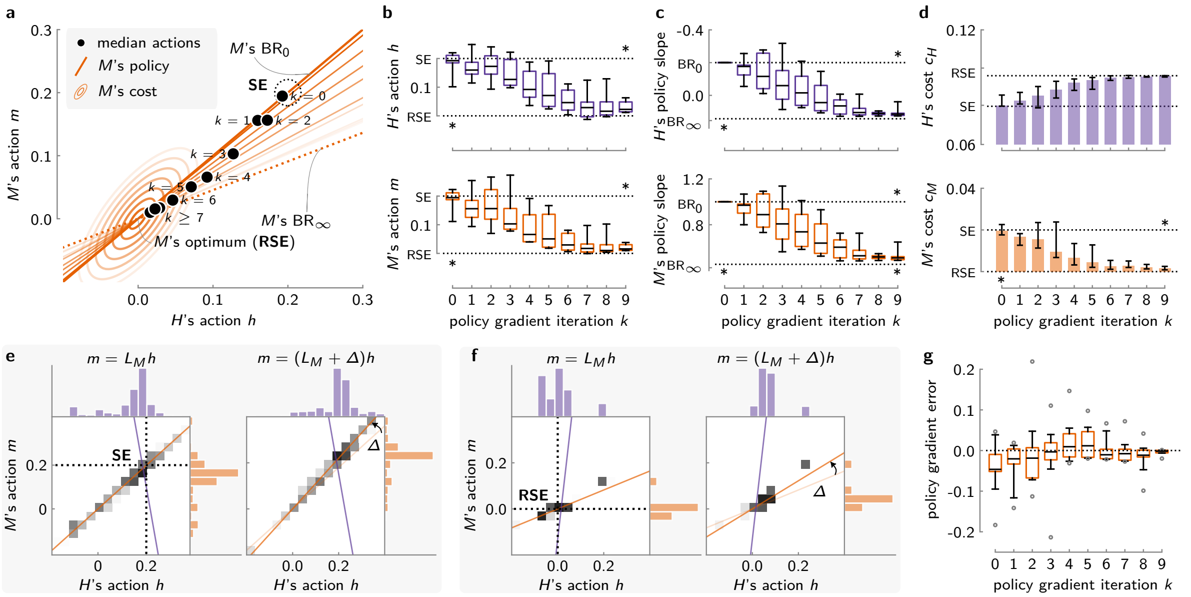


Figure 4.4: **Gradient descent in policy space (Experiment 3, $n = 20$).** Experimental setup and costs are the same as Figure 4.2a,b except that the machine uses a different adaptation algorithm: in this experiment, M iteratively implements linear policies $m = L_M h$, $m = (L_M + \Delta)h$ to measure the gradient of its cost with respect to its policy slope parameter L_M and updates this parameter to descend its cost landscape. (**a**) Median actions and policies for each policy gradient iteration k overlaid on game-theoretic equilibria corresponding to machine best-responses (BR) at initial and limiting iterations (BR$_0$ and BR$_\infty$, respectively) predicted from the Stackelberg equilibrium (SE) and the machine's global optimum (RSE), respectively. (**b**) Action distributions for each iteration displayed by box-and-whiskers plots as in Figure 4.2d, with statistical significance ($*$) analogously determined using the same tests by comparing to SE and M's optimum using two-sided t-tests ($^*P \leq 0.05$); (**c**) Policy slope distributions for each iteration displayed with the same conventions as (b); note that the sign of the top subplot's y-axis is reversed for consistency with other plots. Statistical significance ($*$) determined as in (b) by comparing to SE and RSE using two-sided t-tests ($^*P \leq 0.05$). (**d**) Cost distributions for each iteration displayed using box-and-whiskers plots as in Figures 4.2e and 4.3d. (**e,f**) One- and two-dimensional histograms of actions for different iterations ($k = 0$ in (e), $k = 9$ in (f)) with policies and game-theoretic equilibria overlaid (SE in (e), RSE in (f)). (**g**) Error between measured and theoretically-predicted policy slopes at each iteration displayed as box-and-whiskers plots as in (b,c).

different from theoretically-predicted values (Figure 4.4g; $P > 0.05$) and the final distribution of machine costs were not significantly different from the optimal value (Figure 4.4d; $P > 0.05$), suggesting that the machine can accurately estimate its policy gradient and minimize its cost. In essence, the machine elevated its play by reasoning in the space of policies to steer the game outcome in this experiment to the point it desires in the joint action space.

In all three experiments, the empirical distributions of human actions or policies matched the analytical solution of the learning algorithms, as shown by the interquartile ranges and dashed lines in Figure 4.2d, Figure 4.3b,c, and Figure 4.4b,c . We repeated all three experiments for a game defined by a pair of non-quadratic cost functions to show that our methods could be extended to other continuous games in the paper (Chasnov et al., 2023). Our results show that the optimality of human behavior was robust with respect to the prescribed costs and machine policies, indicating our results may generalize to other settings where people (approximately) optimize their own utility function.

4.6 Analysis of the Human-Machine Learning Dynamics

This section provides mathematical statements about the two-player game (c_H, c_M). Recall that each player optimizes their own cost function: the human is prescribed the cost function

$$c_H(h, m) = \tfrac{1}{2}h^2 + \tfrac{7}{30}m^2 - \tfrac{1}{3}hm + \tfrac{2}{15}h - \tfrac{22}{75}m + \tfrac{12}{125}. \tag{4.7}$$

and the machine is presecribed the cost function

$$c_M(h, m) = \tfrac{1}{2}m^2 + h^2 - hm. \tag{4.8}$$

for the machine. In Experiment 1, the machine optimizes its action by gradient descent. In Experiment 2, the machine optimizes its policy by conjectural variations. In Experiment 3, the machine optimizes its policy by gradient descent. In all experiments, the human updates its action h by making the cost $c_H(h, m)$ as small as possible.

In this section, the three main experiments from the paper were analyzed. Outcomes were predicted by the equilibrium solutions of coupled optimization problems. The three subsections contain mathematical propositions proving statements about the three respective experiments. Propositions 9 and 10 apply to Experiment 1. They prove convergence to the unique Nash and Stackelberg equilibrium solutions. Propositions 11, 12, 13, 14 and 15 apply to Experiment 2. They prove that the machine can perturb its own policy to estimate the human's conjectural variation, and in turn use the estimate to form a best response iteration that converges to a consistent conjectural variations equilibrium. Propositions 16, 18, 17, 19 apply to Experiment 3. They prove that the machine can perturb its own policy to estimate its policy gradient, and in turn use the estimate to update its policy to converge to its global optimum. The formal definitions of the equilibrium solutions are stated in Section 4.2.1.

A *human-machine co-adaptation game* is a two-player repeated game determined by two cost functions – one for each player. The game is played as follows: at each time step t, the human chooses action $h_t \in \mathcal{H}$. The machine best responds by choosing action $m_t \in \mathcal{M}$. The human observes cost $c_H(h_t, m_t)$ via the interface. The next action pair (h_{t+1}, m_{t+1}) is chosen at the next time step $t+1$ for a fixed number of steps T. In each of our experiments, the method that the machine uses to update its action is varied.

4.6.1 Convergence of Gradient Descent in Action Space

The following Proposition 9 describes the $\alpha = 0$ case of Experiment 1, where the outcome is the unique stable Nash equilibrium of the game is $(m, h) = (-1/5, -1/5)$. This outcome is observed empirically (Figure 2 of main paper).

Proposition 9. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if the machine's action is $m = -1/5$, then the human's best response is $h = -1/5$.*

Proof. From the human's perspective, the goal was to solve the optimization problem

$$\min_h \; c_H(h, m) \tag{4.18}$$

The second order condition of (4.18) is

$$\partial_h^2 c_H(h, m) = 1 > 0.$$

The first order condition of the optimization problem (4.18) is

$$\partial_h c_H(h, m) = h - \tfrac{1}{3}m + \tfrac{2}{15} = 0. \tag{4.19}$$

By solving for h in (4.19), the human's best response to m is

$$h = \tfrac{1}{3}m - \tfrac{2}{15}.$$

Solving for h gives the human's best response $h = \frac{1}{3}m - \frac{2}{15}$. Thus, if $m = -\frac{1}{5}$, then $h = -\frac{1}{5}$. □

The following Proposition 10 describes the $\alpha = 1$ (or "infinity") case of Experiment 1, where the outcome is the unique stable human-led Stackelberg equilibrium of the game at $(m, h) = (1/5, 1/5)$. This outcome is observed empirically (Figure 2 of main paper).

Proposition 10. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if the machine's policy is $m = h$, then the human's best response is $h = 1/5$.*

Proof. From the human's perspective, the optimization problem is

$$\min_h \{c_H(h, m) \mid m = h\} \tag{4.20}$$

The cost experienced by the human is

$$c_H(h, h) = \tfrac{2}{5}h^2 - \tfrac{4}{25}h + \tfrac{12}{125}$$

The first order condition of (4.20) is

$$\partial_h c_H(h,h) = \tfrac{4}{5}h - \tfrac{4}{25} = 0$$

Solving for h gives $h = \frac{1}{5}$. □

Remark 5. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if $0 < \alpha \leq 1$ and the machine updates its action $m_{t+1} = m_t - \alpha\partial_m c_M(h_t, m_t)$, then m_{t+1} approaches h_t as t increases. This result can be shown by writing the update as $m_{t+1} = (1-\alpha)m_t + \alpha h_t$ showing that the sequence $m_t, m_{t+1}, \ldots$ is generated by an exponential smoothing filter of time-varying signal h_t.*

Remark 5 is observed in the 2D histograms in Figure 2 from the main paper as the distribution of points on the line of equality $m = h$ for larger α values.

4.6.2 Consistency of Conjectural Variation in Policy Space

In Experiment 2, the machine iterated conjectural variations in policy space. From the humans's perspective, the goal was to choose h to optimize $c_H(h,m)$. But how m is determined affects the solution of the coupled optimization problems. From the machine's perspective, the goal was to choose m to optimize $c_M(h,m)$. Similarly, what h is assumed to be affects the machine's response. The machine estimates the conjectural variation that describes how h is affected by a change in m.

The following Proposition 11 describes the machine's policy perturbation in Experiment 1. The human's response is linear in the machine's constant perturbation δ, but non-linear in the machine's policy slope L.

Proposition 11. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if the machine's policy is $m = Lh + \delta$ and L satisfies $\frac{7}{15}L^2 - \frac{2}{3}L + 1 > 0$, then the human's best response is*

$$h = \frac{22L - 10 - (35L - 25)\delta}{35L^2 - 50L + 75}$$

Proof. The human's optimization problem is

$$\min_h \{c_H(h,m) \mid m = Lh + \delta\} \tag{4.21}$$

The second order condition of (4.21) is

$$\tfrac{7}{15}L^2 - \tfrac{2}{3}L + 1 > 0.$$

The first order condition of (4.21) is

$$(\tfrac{7}{15}L^2 - \tfrac{2}{3}L + 1)h - \tfrac{22}{75}L + \tfrac{2}{15} - (\tfrac{7}{15}L_M + \tfrac{1}{3})\delta = 0$$

Solving for h gives the result.

□

The following Proposition 12 describes how the machine estimates the slope of the human's policy using two points generated by perturbing the constant term of the machine's policy.

Proposition 12. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if the machine's policies are $m = Lh$ and $m' = Lh' + \delta$ and the human best responds with h and h', then*

$$\frac{h' - h}{m' - m} = \frac{7L - 5}{5L - 15}$$

Proof. Using Proposition 11 for h' and h,

$$h' - h = -\frac{35L - 25}{35L^2 - 50L + 75}\delta.$$

Using the definitions of m' and m,

$$m' - m = L(h' - h) + \delta.$$

The ratio of the differences is therefore

$$\frac{h' - h}{m' - m} = \frac{-\left(\frac{35L-25}{35L^2-50L+75}\delta\right)}{-L\left(\frac{35L-25}{35L^2-50L+75}\delta\right) + \delta} = \frac{35L - 25}{L(35L - 25) - (35L^2 - 50L + 75)} = \frac{7L - 5}{5L - 15}.$$

□

Remark 6. *In the main paper, the human's policy slope is L_H and the machine's policy slope is L_M. For a machine policy $m = Lh$ in Experiments 2 and 3, the relationship between these terms are*

$$L_M = L,$$
$$L_H = \frac{7L - 5}{5L - 15}.$$

In this case, the human's conjecture of the machine is consistent with the machine's policy. The equilibrium *solutions are described by linear equations*

$$m = L_M h + \ell_M$$

$$h = L_H m + \ell_H$$

where $\ell_M = 0$ *and* $\ell_H = -\frac{22L-10}{25L-75}$.

Remark 6 can produce the curves in the quadratic game as the solid-line ellipse for when H has a consistent conjecture about M by sweeping L along the real line.

The following Proposition 13 describes the machine's best response to the human adopting a policy based on the conjectural variation in Proposition 12.

Proposition 13. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if the human's policy is* $h = \left(\frac{7L-5}{5L-15}\right) m + \ell$ *for some* ℓ*, then the machine's best response is*

$$m = \frac{9L+5}{2L+10} h$$

Proof. The machine's optimization problem is

$$\min_m \left\{ c_M(h,m) \mid h = \left(\tfrac{7L-5}{5L-15}\right) m + \ell \right\}. \tag{4.22}$$

The first order condition of (4.22) is

$$\partial_m c_M(h,m) + \partial_h c_M(h,m) \left(\tfrac{7L-5}{5L-15}\right) = 0. \tag{4.23}$$

The second order condition is

$$2\left(\tfrac{7L-5}{5L-15}\right)^2 - 2\left(\tfrac{7L-5}{5L-15}\right) + 1 > 0.$$

Taking the first order condition in (4.23), the equation is

$$m - h + (2h - m)\left(\tfrac{7L-5}{5L-15}\right) = 0$$

Sovling for m gives the machine's best response

$$m = \frac{9L+5}{2L+10} h$$

□

Remark 7. *The constant term ℓ in Proposition 13 can be estimated from the joint action measurements. However, it is not necessary to do so to arrive at the optimality condition in Equation (4.23).*

The following Proposition 14 shows the existence of a consistent conjectural variations equilibrium. The equilibrium solution concept is defined in Section 4.2.1. It describes the situatuion where both players have consistency of actions and policies.

Proposition 14. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), there exists two consistent conjectural variations equilibrium solutions uniquely defined by the machine response slopes*

$$L = \frac{-1 \pm \sqrt{41}}{4}.$$

Proof. Using Equations (1) and (1') from Definition 4.10 in (Başar and Olsder, 1998), the stationary conditions for a consistent conjectural variation in the policy space is

$$L - L\left(\tfrac{7L-5}{5L-15}\right) + 2\left(\tfrac{7L-5}{5L-15}\right) - 1 = 0, \tag{4.24}$$

Simplifying the numerator of (4.24), the following quadratic equations defines the machine's consistent policy slope:

$$2L^2 + L - 5 = 0.$$

The solution to the quadratic equation gives us the result. □

Remark 8. *The human's policy slope can be determined by substituting in $L = \frac{-1\pm\sqrt{41}}{4}$, which results in*

$$\frac{7L-5}{5L-15} = \frac{1 \mp \sqrt{41}}{10}.$$

So the two consistent conjectural variational policies are

$$\begin{aligned} m &= \frac{-1 \pm \sqrt{41}}{4} h \\ h &= \frac{1 \mp \sqrt{41}}{10} m - \frac{3 + 7\sqrt{41}}{100} \end{aligned}$$

and the actions (m, h) that solve the linear equation.

The following Proposition 15 shows that Experiment 2 converges to a stable equilibrium.

Proposition 15. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if* *the machine updates its policy using the difference equation* $L^{+} = \frac{9L+5}{2L+10}$ *then*

$$L^{*} = \frac{-1+\sqrt{41}}{4}$$

is a locally exponentially stable fixed point of this iteration.

Proof. Define the map $F : \mathbb{R} \to \mathbb{R}$ as

$$F(L) := \frac{9L+5}{2L+10} \tag{4.25}$$

To assess the convergence of Experiment 2, the fixed points of (4.25) are determined along with their stability properties. The fixed point L^{*} that satisfies

$$L^{*} = F(L^{*})$$

are determined by the solutions to the quadratic equation

$$2L^{2} + L - 5 = 0. \tag{4.26}$$

There are two solutions to (4.26) and they are real and distinct. The fixed points are

$$\frac{-1 \pm \sqrt{41}}{4}.$$

Exactly one fixed point is stable, and it is a stable attractor of the repeated application of F. The stability can be determined by linearizing (4.25) at the particular fixed point and ensuring that its derivative gives a magnitude of less than one. The linearization of F at fixed point L^{*} is

$$F(L) \approx \partial F(L^{*})(L - L^{*}) \tag{4.27}$$

where

$$\partial F(L) = \frac{20}{(5+L)^{2}}$$

If $L^{*} = \frac{-1+\sqrt{41}}{4}$, then $|\partial F(L^{*})| \approx 0.5 < 1$, so the fixed point L^{*} is stable. On the other hand, if $L^{*} = \frac{-1-\sqrt{41}}{4}$, then $|\partial F(L^{*})| > 1$, so the fixed point L^{*} is unstable. □

For a quadratic game with single-dimensional actions, there are two consistent conjectural variations equilibria. One is stable, the other is unstable.

Remark 9. *Another way to assess the convergence of the fixed point map* (4.25) *is by inspecting the normal form of the linear fractional transformation. The normal form of* (4.25) *is*

$$\frac{F(L) - L^*}{F(L) - L^{**}} = \lambda \frac{L - L^*}{L - L^{**}} \tag{4.28}$$

where L^* and L^{**} are fixed points of F and λ is a real number given by

$$\lambda = \frac{-19 + \sqrt{41}}{-19 - \sqrt{41}} \tag{4.29}$$

Since $|\lambda| \approx 0.5 < 1$, the fixed point L^* is semi-globally stable.

Remark 9 is a based on a known result from complex analysis and conformal mapping theory.

4.6.3 Convergence of Gradient Descent in Policy Space

In Experiment 3, the machine implemented gradient descent in policy space. The machine estimated the policy gradient using cost measurements from a pair of trials. The machine's cost depends on its own policy and the human's best response to it.

The following Proposition 16 describes the machine's policy perturbation in Experiment 3. The human's action response varies non-linearly.

Proposition 16. *Given a human-machine co-adaptation game determined by cost functions* (4.7) *and* (4.8)*, if the machine's policy is $m = (L + \Delta)h$ and L, Δ satisfy $\frac{7}{15}(L + \Delta)^2 - \frac{2}{3}(L + \Delta) + 1 > 0$, then the human's best response is*

$$h = \frac{22(L + \Delta) - 10}{35(L + \Delta)^2 - 50(L + \Delta) + 75}$$

Proof. The human's optimization problem is

$$\min_h \{c_H(h, m) \mid m = (L + \Delta)h\}. \tag{4.30}$$

The second order condition of (4.30) is

$$\tfrac{7}{15}(L + \Delta)^2 - \tfrac{2}{3}(L + \Delta) + 1 > 0.$$

The first order condition of (4.30) is

$$(\tfrac{7}{15}(L + \Delta)^2 - \tfrac{2}{3}(L + \Delta) + 1)h - \tfrac{22}{75}(L + \Delta) + \tfrac{2}{15} = 0$$

Solving for h gives human's response

$$h = \frac{22(L+\Delta) - 10}{35(L+\Delta)^2 - 50(L+\Delta) + 75}. \tag{4.31}$$

□

The following Proposition 17 describes how to estimate the policy gradient using two trials as done in Experiment 3. Suppose the machine plays policy $m = Lh$, then the human's response is given by

$$r(L) := \frac{22L - 10}{35L^2 - 50L + 75}$$

as determined by Proposition 11 or Proposition 16 with the perturbations set to zero.

Proposition 17. *Given a human-machine co-adaptation game determined by cost functions (4.7) and (4.8), if the machine's policies are $m = Lh$ and $m' = (L+\Delta)h'$ and the human's best responses are $h = r(L)$ and $h' = r(L+\Delta)$, then*

$$\lim_{\Delta \to 0} \frac{c_M(h', m') - c_M(h, m)}{\Delta} = D_L c_M(r(L), Lr(L))$$

Proof. From Proposition 11, if machine's policy is $m = Lh$ and the human's best response is

$$h = \frac{22L - 10}{35L^2 - 50L + 75}.$$

The machine's cost written as a function of L is

$$\begin{aligned} c_M(h,m) = c_M(r(L), Lr(L)) &= \tfrac{1}{2}L^2 r(L)^2 + r(L)^2 - Lr(L)^2 \\ &= \tfrac{1}{2}(L^2 - 2L + 2) r(L)^2 \\ &= \frac{(L^2 - 2L + 2)(22L - 10)^2}{2(35L^2 - 50L + 75)^2} \end{aligned}$$

The difference term is

$$c_M(h', m') - c_M(h, m) = c_M(r(L+\Delta), Lr(L+\Delta)) - c_M(r(L), Lr(L))$$

Expanding out the terms, ignoring the terms of order Δ^2 or higher, we have

$$\begin{aligned} c_M(h', m') - c_M(h, m) &= \frac{((L+\Delta)^2 - 2(L+\Delta) + 2)(22(L+\Delta) - 10)^2}{2(35(L+\Delta)^2 - 50(L+\Delta) + 75)^2} - \frac{(L^2 - 2L + 2)(22L - 10)^2}{2(35L^2 - 50L + 75)^2} \\ &= \frac{4(11L - 5)(2L^3 + 181L^2 - 380L + 305)}{25(7L^2 - 10L + 15)^3}\Delta + (\cdots)\,\Delta^2 + \cdots \end{aligned}$$

Dividing by Δ and taking Δ to zero gives us the same expression as directly computing the derivative of the cost:

$$\partial_L c_M(r(L), Lr(L)) = \frac{4(11L-5)(2L^3+181L^2-380L+305)}{25(7L^2-10L+15)^3}.$$

Hence, we get the desired result. □

The following Proposition 18 shows that there is a unique machine-led reverse Stackelberg equilibrium of the game. The equilibrium solution concept is defined in Section 4.2.1. It describes the scenario where the leader announces a policy and the follower responds to the policy. In contrast, the Stackelberg equilibrium in Proposition 10 describes the scenario where the leader announces an action and the follower response to the action.

Proposition 18. *Given a human-machine co-adaptation game determined by cost functions* (4.7) *and* (4.8)*, there exists a reverse Stackelberg equilibrium.*

Proof. The machine's global optimum solves

$$\min_{h,m} c_M(h,m).$$

The machine's global optimum is $(h,m) = (0,0)$.

Suppose the machine's policy is $m = Lh$, then the human's optimization problem is

$$\min_h \{c_H(h,m) \mid m = Lh\}$$

and the best response is

$$h = r(L) = \frac{22L-10}{35L^2-50L+75}$$

The machine wants to drive the human to play $0 = r(L)$. Hence the machine chooses $L = 5/11$.

The second order condition is

$$\tfrac{7}{15}L^2 - \tfrac{2}{3}L + 1 > 0.$$

which is satisfied by $L = 5/11$. Hence $(0,0)$ is a machine-led reverse Stackelberg equilibrium. □

The following Proposition 19 shows that Experiment 3 converges to a stable equilibrium.

Proposition 19. *Given a human-machine co-adaptation game determined by cost functions* (4.7) *and* (4.8)*, if the machine plays policy $m = Lh$ and the human responds with $h = r(L)$ and machine's updates its policy by* *gradient descent,*

$$L_{k+1} = L_k - \alpha \partial_L c_M(r(L_k), L_k r(L_k))$$

then $L^* = 5/11$ *is a locally exponentially stable fixed point of this iteration for all* $\alpha > 0$ *sufficiently small.*

Proof. The roots of $\partial_L c_M(r(L_k), L_k r(L_k)) = 0$ are determined by the solutions to a quartic equation

$$(11L - 5)(2L^3 + 181L^2 - 380L + 305) = 0. \tag{4.32}$$

There are two real solutions to (4.32), the first one $L^* = \frac{5}{11}$ can be seen by inspection, and the second one is, approximately, $L^{**} \approx -92.6$.

The stability is determined by linearizing at the particular fixed point and ensuring that the second derivative is positive. The linearization the derivative at root L_M^* is

$$\partial_L c_M(r(L), Lr(L)) \approx \partial_L^2 c_M(r(L^*), L^* r(L^*))(L - L^*) \tag{4.33}$$

The second derivative $\partial_{L_M}^2 c_M \approx 0.18$ evaluated at L^* is positive, so the fixed point L_M^* is stable. The second derivative evaluated at L^{**} is negative, so the fixed point is unstable. □

4.7 Is it Optimal to Form Consistent Conjectures?

A common question that arises during this study into consistent conjectural variations equilibria (CCVE) is: What is the difference between Nash equilibria and CCVE? In particular, is it optimal for agents to adopt CCVE? If it is, why is it also optimal to play Nash equilibria? Furthermore, how can they both optimal but different points in the decision space? This section will definitively answer these questions, demonstrating that CCVE is optimal in the Nash sense, but it is crucial to consider this Nash equilibrium in conjecture space rather than in action space. This distinction provides a strong justification for adopting CCVE as a boundedly-rational solution concept that is game-theoretically justified. However, there are still many open questions regarding this topic that need to be explored in more depth, and this section provides only a preliminary look. In particular, there is a substantial null space that requires further investigation.

In game theory, strategic interactions often lead to multiple equilibrium concepts, with the Nash equilibrium being the most prominent (Nash, 1950; von Neumann and Morgenstern, 1947). Yet, in scenarios where players form beliefs about opponents' responses, the conjectural variations equilibrium (CVE) provides a refined perspective (Başar and Olsder, 1998; Figuières et al., 2004). Specifically, in a consistent conjectural variations equilibrium (CCVE), players' beliefs equal the best-responses of their opponents (Calderone et al., 2023; Figuières et al., 2004). This approach captures the dynamic relationship between players' conjectures and responses, in contrast to the static strategies in a Nash framework.

This paper aims to establish a theoretical connection between Nash equilibria and CCVE: that consistent conjectures lead to Nash optimal strategies. By doing so, we will prove that a CCVE is in fact another kind of Nash equilibrium—a "conjectural Nash", so to speak. Crucially, while the strategies adopted at a CCVE might not coincide with those of a Nash equilibrium in terms of players' actions, the conjectures held by each player at a CCVE form a Nash equilibrium in the space of conjectures. Towards proving this statement formally, Section 4.7.1 constructs a coordinate transformation on the space of conjectural variations, then the main theorem in Section 4.7.2 uses this map to show the result.

However, different conjectures can lead to the same actions. Intuitively, this makes sense because the space of conjectures is much larger than the space of actions (larger by almost the product of the players' dimensions). There is a relatively large nullspace of conjectures that correspond to a continuum of non-strict conjectural Nash equilibria. Remark 10 describes this nullspace. Most of these equilibria are not consistent, but nonetheless they are still, in some sense, optimal. A notable example of such a game-theoretically meaningful equilibrium is the solution to *incentive problems* (Başar and Olsder, 1998), wherein a follower forms a consistent conjecture about a leader, but the leader implements an incentive that influences the follower to minimize the leader's cost. Remark 11 explains this solution concept. In Section 4.7.3, we present an example that compares this optimal incentive with the Nash and CCVE responses.

Thus, not surprisingly, there are continua of Nash equilibria in the conjecture space. We visualize one such subspace in Figure 4.5c, although points along the nullspace may not be game-theoretically meaningful because they are inconsistent. It is an open question as to whether there are other game-theoretically meaning equilibria in this space, or how to design scalable learning dynamics to navigate these higher-level strategies.

Preliminaries We consider two-player continuous games. Let $\mathcal{X}_1, \mathcal{X}_2$ be continuous sets, the action spaces. Let f_1, f_2 be cost functions for players 1 and 2, respectively, where $f_i : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R},\ i = 1, 2$. Let f_1, f_2 be continuously differentiable and strictly convex in their arguments. The pair (f_1, f_2) defines a continuous game.

Following the notation in (Başar and Olsder, 1998), let $\mathcal{T}_1 \times \mathcal{T}_2$ be the class of all mappings (T_1, T_2), $T_i : \mathcal{T}_j \to \mathcal{T}_i,\ j \neq i,\ i, j = 1, 2$. Let $\mathcal{T}_1 \times \mathcal{T}_2$ be chosen such that $f_1(\cdot, T_2(\cdot))$ and $f_2(T_1(\cdot), \cdot)$ are strictly convex in their arguments, for every pair $T_1, T_2 \in \mathcal{T}_1 \times \mathcal{T}_2$. For a two-player game (f_1, f_2), given a pair of reaction functions (T_1, T_2), also called *conjectures*, players face optimization problems

$$\arg\min_{x_1}\{f_1(x_1,x_2) \mid x_2 = T_2(x_1)\}, \tag{4.34a}$$

$$\arg\min_{x_2}\{f_2(x_1,x_2) \mid x_1 = T_1(x_2)\}. \tag{4.34b}$$

The problems (4.34a)-(4.34b) are the rationality models of the two players, that is, we assume the players of the game (f_1, f_2) are optimal with respect to their cost and their conjectures (T_2, T_1). These optimization problems define the set of CVE: players solve for their best actions given their choice of conjectures. At this point, the main question is: what conjectures make sense for players to choose?

One such choice of conjectures are those that are consistent: ones that equal the opponent's reaction. The reaction functions $(T_1^c, T_2^c) \in \mathcal{T}_1 \times \mathcal{T}_2$ are in consistent CVE (CCVE) if and only if

$$\begin{aligned}\frac{\partial f_1(x_1,x_2)}{\partial x_1} + \frac{\partial f_1(x_1,x_2)}{\partial x_2}\cdot\frac{\partial T_2^c(x_1)}{\partial x_1} &= 0, \text{ for } x_1 = T_1^c(x_2),\\ \frac{\partial f_2(x_1,x_2)}{\partial x_2} + \frac{\partial f_2(x_1,x_2)}{\partial x_1}\cdot\frac{\partial T_1^c(x_2)}{\partial x_2} &= 0, \text{ for } x_2 = T_2^c(x_1).\end{aligned}$$

These are two partial differential equations, which are, in general, difficult to solve. In this paper, we focus on a class of quadratic games since their solutions are well characterized and the differential equations reduce to a pair of bilinear equations.

While a Nash equilibrium (in action space) implies a *zeroth-order* CCVE (Başar and Olsder, 1998, Prop. 4.5), our paper proves a related but converse statement, that a *first-order* implies a Nash equilibrium (in conjecture space). Please note that the term "first-order" used in Basar has a slightly different meaning than our use of that term in this paper. In Basar, it refers to the order of approximation of the conjecture function, whereas the first-order conditions in our paper correspond to the system of linear equations that describe the set of conjectural variations equilibria.

4.7.1 A Coordinate Transformation from Conjectures to Actions

We will prove our main result for the class of linear conjectures and quadratic costs by first constructing a coordinate transformation that maps conjectural variations to actions, given the optimization problems (4.34a)-(4.34b).

Consider continuous game (f_1, f_2) where f_1 and f_2 are quadratic costs functions, as is commonly done in optimization, control and game theory. Let the actions spaces be $\mathcal{X}_1 = \mathbb{R}^{d_1}$ and $\mathcal{X}_2 = \mathbb{R}^{d_2}$. Let

$f_i : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R},\ i = 1, 2$ be

$$f_1(x_1, x_2) = \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} A_1 & B_1^\top \\ B_1 & D_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

$$f_2(x_1, x_2) = \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} D_2 & B_2 \\ B_2^\top & A_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_2 \\ a_2 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

where $A_i \in \mathbb{R}^{d_i \times d_i},\ B_i \in \mathbb{R}^{d_j \times d_i},\ D_i \in \mathbb{R}^{d_j \times d_j},\ a_i \in \mathbb{R}^{d_i},\ b_i \in \mathbb{R}^{d_j},\ i \neq j,\ i, j = 1, 2$. Note that $A_i = A_i^\top$ and $D_i = D_i^\top,\ i = 1, 2$. The optimal reaction functions of quadratic games are linear. Therefore, the conjectures that players adopt are written as

$$T_2(x_1) = L_1 x_1 + \ell_1,$$

$$T_1(x_2) = L_2 x_2 + \ell_2,$$

where $L_1 \in \mathbb{R}^{d_2 \times d_1},\ L_2 \in \mathbb{R}^{d_1 \times d_2},\ \ell_1 \in \mathbb{R}^{d_2},\ \ell_2 \in \mathbb{R}^{d_1}$. Since the costs are quadratic and the conjecture functions are linear, the solution of the optimization problems (4.34a)-(4.34b) correspond a linear system of equations

$$0 = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_1}{\partial x_2} \cdot \frac{\partial T_2}{\partial x_1} = x_1^\top A_1 + x_2^\top B_1 + a_1^\top + (x_1^\top B_1^\top + x_2^\top D_1 + b_1^\top) L_1, \tag{4.35a}$$

$$0 = \frac{\partial f_2}{\partial x_2} + \frac{\partial f_2}{\partial x_2} \cdot \frac{\partial T_1}{\partial x_2} = x_2^\top A_2 + x_1^\top B_2 + a_2^\top + (x_2^\top B_2^\top + x_1^\top D_2 + b_2^\top) L_2, \tag{4.35b}$$

while ensuring the usual second-order conditions of optimality. The first-order conditions can be thought of a equality constraint that describes the set of all possible CVEs. Taking the transpose of (4.35a)-(4.35b) and grouping terms, we have

$$\begin{bmatrix} A_1 + L_1^\top B_1 & B_1^\top + L_1^\top D_1 \\ B_2^\top + L_2^\top D_2 & A_2 + L_2^\top B_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} a_1 + L_1^\top b_1 \\ a_2 + L_2^\top b_2 \end{bmatrix} = 0.$$

This linear equation describes how the actions (x_1, x_2) and variations (L_1, L_2) are correlated. Assuming that the square matrix on the left has an inverse, we provide an explicit solution to the optimization problems.

Thus, we arrive at the coordinate transformation

$$x(L_1, L_2) \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = - \begin{bmatrix} A_1 + L_1^\top B_1 & B_1^\top + L_1^\top D_1 \\ B_2^\top + L_2^\top D_2 & A_2 + L_2^\top B_2 \end{bmatrix}^{-1} \begin{bmatrix} a_1 + L_1^\top b_1 \\ a_2 + L_2^\top b_2 \end{bmatrix} \tag{4.36}$$

where the map $x : \mathbb{R}^{d_2 \times d_1} \times \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ solves the system of equations (4.35a)-(4.35b). In the main theorem of the next section, we will consider the lifted game defined by the cost functions composed with x, i.e, $(f_i \circ x)(L_1, L_2) \equiv f_i(x(L_1, L_2)),\ i = 1, 2$. In conclusion, from the game (f_1, f_2) over the action spaces, we constructed a map that lifts the game into a higher-level space, resulting in the game $(f_1 \circ x, f_2 \circ x)$ over conjecture spaces

This coordinate transformation is significant because it captures a wide range of behaviors of boundedly-rational agents—agents that are not perfectly rational due to lack of available information or compute. The coordinate map gives all the points in action spaces that agents can reach due the entire space of conjectural variations that agents can form, consistent or not.

4.7.2 All Consistent Conjectures are Nash Equilibria, but Not Conversely

In the previous section, we introduced a mapping, denoted x, from the space of conjectural variations to the space of actions. This mapping links players' conjectures into corresponding equilibrium actions. If we can prove that strategies derived from *consistent* conjectures are locally optimal when considering this map—meaning that choosing another conjecture would not lead to an improvement—then that gives a very good reason for players to seek consistent conjectures.

Theorem 11. *Consider a continuous games with quadratic cost functions f_1 and f_2 and the map (4.36). If L_1^{c} and L_2^{c} are consistent conjectural variations of the game (f_1, f_2), then they constitute a Nash equilibrium of the game $(f_1 \circ x, f_2 \circ x)$.*

Proof. To prove the theorem, we will first demonstrate the result for player 1. The goal is to show that for some L_2, if L_1^{c} is a consistent conjecture, then L_1^{c} minimizes $f_1(x(L_1, L_2))$ over L_1.

First, we derive expressions for x_1, x_2 in terms of L_1. We begin by transposing and regrouping equations (4.35a) and (4.35b):

$$(A_1 + L_1^\top B_1)x_1 + (B_1^\top + L_1^\top D_1)x_2 + a_1 + L_1^\top b_1 = 0, \tag{4.37a}$$

$$(A_2 + L_2^\top B_2)x_2 + (B_2^\top + L_2^\top D_2)x_1 + a_2 + L_2^\top b_2 = 0. \tag{4.37b}$$

From (4.37b), we isolate x_2 and rewrite it as

$$x_2 = K_2 x_1 + k_2,$$

where $K_2 \triangleq -\left(A_2 + L_2^\top B_2\right)^{-1}\left(B_2^\top + L_2^\top D_2\right)$ and $k_2 \triangleq -\left(A_2 + L_2^\top B_2\right)^{-1}\left(a_2 + L_2^\top b_2\right)$. Substituting this expression into (4.37a), we have

$$x_1 = -Q_1^{-1} q_1$$

where $Q_1 \triangleq A_1 + B_1^\top K_2 + L_1^\top B_1 + L_1^\top D_1 K_2,\ q_1 \triangleq a_1 + B_1^\top k_2 + L_1^\top b_1 + L_1^\top D_1 k_2$.

Next, to find how f_1 changes as L_1 is varied, we examine the gradient $\partial f_1/\partial L_1 = \partial f_1/\partial x_1 \cdot \partial x_1/\partial L_1 + \partial f_1/\partial x_2 \cdot \partial x_2/\partial L_1$. We need to show that the gradient is zero at L_1^c. Consider a small variation ΔL_1 in L_1, we have

$$\begin{aligned} x_1 + \Delta x_1 = -&(A_1 + B_1^\top K_2 + L_1^\top B_1 + L_1^\top D_1 K_2 + \Delta L_1^\top (B_1 + D_1 K_2))^{-1} \\ &\cdot (a_1 + B_1^\top k_2 + L_1^\top b_1 + L_1^\top D_1 k_2 + \Delta L_1^\top (b_1 + D_1 k_2)) \end{aligned}$$

To simplify notation, let $W_1 \triangleq B_1 + D_1 K_2,\ w_1 \triangleq b_1 + D_1 k_2$. By the matrix inversion lemma,

$$\begin{aligned} x_1 + \Delta x_1 &= -(Q_1^{-1} - Q_1^{-1}\Delta L_1^\top (I + W_1 \Delta L_1^\top)^{-1} W_1 Q_1^{-1})(q_1 + \Delta L_1^\top w_1) \\ &= -(Q_1^{-1} - Q_1^{-1}\Delta L_1^\top W_1 Q_1^{-1})(q_1 + \Delta L_1^\top w_1) \\ &= x_1 - Q_1^{-1}\Delta L_1^\top (W_1 x_1 + w_1) \end{aligned}$$

where we drop higher order terms. Therefore, the change in x_1 due to a perturbation in L_1 is

$$\Delta x_1 = -Q_1^{-1}\Delta L_1^\top (W_1 x_1 + w_1) \tag{4.38}$$

Furthermore, we note that $\Delta x_2 = K_2 \Delta x_1$. Now upon perturbing L_1, we have the change in the cost value

Δf_1 as follows:

$$\begin{aligned}\Delta f_1 = \frac{\partial f_1}{\partial x}\begin{bmatrix}\Delta x_1\\ \Delta x_2\end{bmatrix} &= \begin{bmatrix}A_1x_1 + B_1^\top x_2 + a_1\\ B_1x_1 + D_1x_2 + b_1\end{bmatrix}^\top \begin{bmatrix}\Delta x_1\\ \Delta x_2\end{bmatrix} = \begin{bmatrix}A_1x_1 + B_1^\top K_2x_1 + a_1 + B_1^\top k_2\\ B_1x_1 + D_1K_2x_1 + b_1 + D_1k_2\end{bmatrix}^\top \begin{bmatrix}\Delta x_1\\ K_2\Delta x_1\end{bmatrix}\\ &= x_1^\top(A_1 + B_1^\top K_2 + K_2^\top B_1 + K_2^\top D_1K_2)\Delta x_1 + (a_1 + K_2^\top b_1 + B_1^\top k_2 + K_2^\top D_1k_2)^\top \Delta x_1\\ &= (P_1x_1 + p_1)^\top \Delta x_1\end{aligned}$$

where $P_1 \triangleq A_1 + B_1^\top K_2 + K_2^\top B_1 + K_2^\top D_1K_2,\ p_1 \triangleq a_1 + B_1^\top k_2 + K_2^\top b_1 + K_2^\top D_1k_2$. Putting this all together, the change in f_1 is therefore

$$\Delta f_1 = -(P_1x_1 + p_1)^\top Q_1^{-1}\Delta L_1^\top(W_1x_1 + w_1)$$

Using the fact that $\Delta f_1 = \mathrm{Tr}\left(\Delta L_1^\top \frac{\partial f_1}{\partial L_1}\right)$, we have

$$\frac{\partial f_1}{\partial L_1} = -(W_1x_1 + w_1)(P_1x_1 + p_1)^\top Q_1^{-1}.$$

Now, at a consistent conjecture $L_1 = L_1^{\mathsf{c}}$, we know that $L_1^{\mathsf{c}} = K_2$, meaning that player 1's conjecture about player 2's reaction function (given by L_1) must equal the actual reaction function (given by K_2), which directly implies that $P_1 = Q_1$ and $p_1 = q_1$. Therefore,

$$\frac{\partial f_1}{\partial L_1} = -(Wx_1 + w_1)(x_1^\top P_1Q_1^{-1} + p_1^\top Q_1^{-1}) = (Wx_1 + w_1)(x_1 - x_1)^\top = 0$$

which proves that $\partial f_1/\partial L_1$ is zero given that L_1 is consistent. The only thing we have left to prove is that the second-order derivative is positive semi-definite to ensure a local non-strict minimum.

Towards confirming the second-order optimality condition, by the product rule, the deviation of the gradient term is

$$\Delta\left(\frac{\partial f_1}{\partial L_1}\right) = -W_1\Delta x_1(P_1x_1 + p_1)^\top Q_1^{-1} - (W_1x_1 + w_1)\Delta x_1^\top P_1Q_1^{-1} - (W_1x_1 + w_1)(P_1x_1 + p_1)^\top \Delta Q_1^{-1}$$

Again, at $L_1 = L_1^{\mathsf{c}}$, we have that $L_1^{\mathsf{c}} = K_2$ and therefore $P_1 = Q_1, p_1 = q_1$. Hence,

$$\Delta(\partial f_1/\partial L_1) = -(W_1x_1 + w_1)\Delta x_1^\top - (W_1x_1 + w_1)(Q_1x_1 + q_1)^\top \Delta Q_1^{-1}$$

Recalling that $\Delta x_1 = -Q_1^{-1}\Delta L_1^\top(W_1x_1 + w_1)$ and $\Delta Q_1^{-1} = Q_1^{-1}\Delta L_1^\top W_1Q_1^{-1}$, the last term becomes zero, so the deviation at equilibrium is

$$\Delta(\partial f_1/\partial L_1) = (W_1x_1 + w_1)(W_1x_1 + w_1)^\top \Delta L_1 Q_1^{-\top}$$

Vectorizing the tensor using the Kronecker product, we have

$$\operatorname{vec}(\Delta(\partial f_1/\partial L_1)) = \left(Q_1^{-1} \otimes (W_1x_1 + w_1)(W_1x_1 + w_1)^\top\right) \operatorname{vec}(\Delta L_1),$$

therefore, the Hessian is

$$\frac{\partial^2 f_1}{\partial L_1^2} = Q_1^{-1} \otimes (W_1x_1 + w_1)(W_1x_1 + w_1)^\top,$$

recalling that $x_1 = -Q_1^{-1}q_1$. Since Q_1 is positive definite due to the convexity assumption, the outer product of a vector with itself is positive semi-definite, and that kronecker product preserves semi-definiteness, we have that the Hessian is also positive semi-definite, concluding that player 1 is indded at a minimum of its optimization problem.

Therefore, any variation in L_1 at the consistent conjecture L_1^{c} will not decrease the cost function f_1, implying that L_1^{c} minimizes $f_1(x(L_1, L_2))$ with respect to L_1. A similar argument can be made for player 2 to complete the proof. □

In this proof, we demonstrated that consistent conjectures that lead to a CCVE also lead to a Nash equilibrium when the actions are determined by the coordinate transformation x. This connection between CCVE and Nash equilibria highlights the importance of players having consistent beliefs in strategic interactions.

The Nash equilibrium is not strict, therefore it's not isolated. The theorem above is a sufficient condition; there are other sufficient conditions that lead to Nash equilibria. The remarks below provides two other known conditions.

Remark 10. *The conjectures L_i that correspond to the nullspace described by $L_i^\top(W_iQ_i^{-1}q_i - w_i) = 0,\ i = 1, 2$ form another kind of Nash equilibrium. The proof directly follows by inspection of (4.38), since if ΔL_i consists of columns that are orthogonal to $W_ix_i + w_i$, then the deviation Δx_i will be zero. We visualize this nullspace in the example in the following section.*

Remark 11. *Two more examples of Nash equilibria in conjecture spaces: the conjectures that lead to each player's own optima, i.e. $\min_{x_1,x_2} f_1(x_1, x_2)$ and $\min_{x_1,x_2} f_2(x_1, x_2)$. The proof is simple: for the leader i, $\partial f_i/\partial x = 0$ directly implies $\partial f_i/\partial L_i = 0$. For the follower j, since it adopts a consistent conjecture, we have*

$\partial f_j / \partial L_j = 0$. *Therefore* (L_i, L_j) *form a Nash equilibrium. The existence of these strategies were proven in (Zheng and Başar, 1982). Interestingly, these optima coincide with reverse Stackelberg equilibria which are game-theoretic solutions to incentive problems (Başar and Olsder, 1998). It is an open questions whether these equilibria correspond to the affine nullspace in the previous remark or not.*

These remarks lead to a natural question to inspire future research: what other game-theoretically meaningful equilibria lead to Nash in the space of conjectures?

4.7.3 An Example in a Three-Dimensional Decision Space

To illustrate how a CCVE (in action space) can also be a Nash (in policy space), we provide the following example where one player has one scalar action and the other player has two scalar actions. The actions that lead to the CCVE are defined in three dimensions, whereas the conjectural variations that lead to the Nash are defined in four dimensions.

Consider a two-player quadratic game on $\mathbb{R}^3$ with costs (f_1, f_2) given by

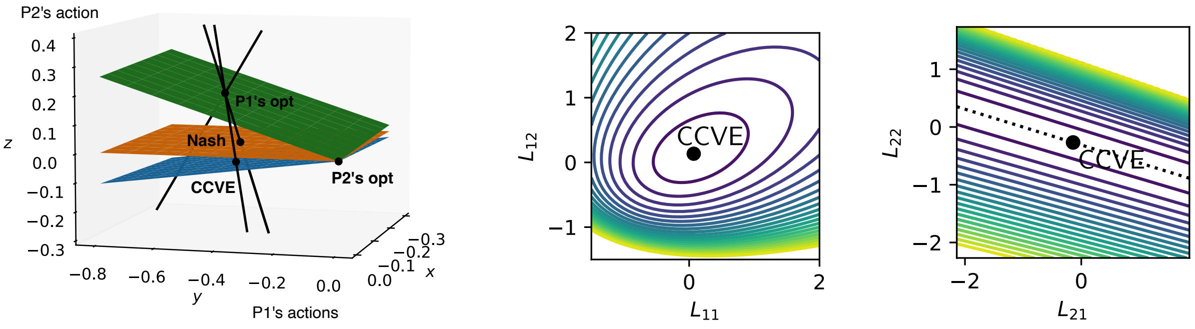
$$\begin{aligned} f_1(x,y,z) &= 2x^2 + y^2 - z^2 + xy + xz + yz + x + y + z \\ f_2(x,y,z) &= 2z^2 + x^2 + y^2, \end{aligned}$$

where $(x, y) \in \mathbb{R}^2$ are player 1's actions and $z \in \mathbb{R}$ is player 2's action. The Nash equilibrium in action space is $(1/7, 3/7, 0)$. The CCVE is $(0.133, 0.437, 0.068)$ with consistent conjectural variations

$$L_1^{\mathfrak{c}} = \begin{bmatrix} 0.0703 \\ 0.1340 \end{bmatrix}, \; L_2^{\mathfrak{c}} = \begin{bmatrix} -0.1406 & -0.2682 \end{bmatrix}.$$

In this example, we confirm that $(L_1^{\mathfrak{c}}, L_2^{\mathfrak{c}})$ is a Nash equilibrium in policy space when considering the coordinate transformation x, thereby showing that the consistent conjectural variations are optimal.

In Figure 4.5, we plot the best response maps for three different equilibria in a three-dimensional space. Since player 1's best response maps have a one-dimensional domain and two-dimensional codomain, they are represented as lines in Figure 4.5(a). On the other hand, player 2's best response maps are planes. Both player's conjectural variations are two-dimensional. In Figure 4.5(b)-(c), we plot the contour lines of each player's cost at the CCVE in the space of conjectures, showing that the consistent conjecture is a strict minimum of player 1's cost but a non-strict minimum of player 2's cost. The dotted line in Figure 4.5(c) is the null space that is referred to from Remark 10. Despite a continuum of Nash equilibria in conjecture space, the CCVE is the only Nash equilibrium that has consistent conjectures in this example.



(a) Best-responses and strategies

(b) P1's conjecture landscape

(c) P2's conjecture landscape

Figure 4.5: An example of a two-player game with distinct Nash, CCVE, and incentive solutions. (a) Best-response maps and strategies corresponding to various equilibria in the space of actions. The lines are Player 1's best-responses, while the surfaces are Player 2's best responses. (b)-(c) Cost landscape contour plots at the CCVE. The contour plots indicate that the consistent conjectures are minima in the space of conjectures. The dotted line contains the the nullspace of conjectures mentioned in Remark 10, which, curiously, corresponds to the same CCVE point in the action space.

This work has established a link between the consistent conjectural variations equilibrium and the classical Nash equilibrium. By demonstrating that strategies derived from consistent beliefs in conjectures lead to optimality, this paper underscores the importance of having correct beliefs in strategic settings. The implications are wide-ranging, with potential impacts in fields such as economics and social sciences. Future work can delve into a more complete characterization of these conjectural Nash equilibria and associated learning dynamics. Taking it one step further, future research could consider alternative equilibrium concepts in conjecture spaces, such as the idea of a CCVE where players form yet higher-level strategies—conjectures of conjectures.

4.8 Discussion and Conclusion

Our study focused on continuous human-machine interactions, introducing a dynamic perspective in contrast to traditional approaches, for instance, in (March, 2021). By leveraging two-player learning dynamics, we provided predictive models for these interactions and highlighted important implications. This approach not only validates our game-theoretic perspective but also raises ethical concerns about machines potentially manipulating human behavior without direct observation.

In the first experiment, we observed a correlation between adaptation rates and human costs: as the machine's adaptation rate increased, the human's cost decreased. This suggests that quicker machine responses allow humans to better anticipate machine actions, improving human performance. Through its policy, the machine effectively imposed a constraint on the human's optimization problem. Such constraints could arise either indirectly, as observed in this experiment, where the machine descended the gradient of its cost at a fast rate, or directly, as implemented in the subsequent experiments.

The second experiment built on the previous experiment, exploring the reciprocal: if humans gain from faster machine reactions, can machines similarly benefit by anticipating human reactions? Our results confirmed that, indeed, machines can adapt a level further and perform better, by only observing human actions. However, this adaptation was not unilateral. As machines adjusted, humans further adjusted their responses, leading to a dynamic equilibrium. Importantly, the machine created a model of the human without solving an inverse optimization problem (Li et al., 2019; Ng and Russell, 2000), making the method more applicable to real-world scenarios. This experiment highlighted that it was not just the machine's learning rate, but its policy that influenced human responses, an insight that guided the design of the next experiment.

The third experiment sought to explore the machine's policy space. We found that by directly optimizing its cost, the machine can control human actions even without observing them. This observation is both exciting and concerning. On the one hand, it offers the potential for machines to aid human partners, for instance by supporting decision-making or providing assistance when someone's movement is impaired, enabling advances in the context of emerging *body-/human-in-the-loop optimization* paradigm for assistive devices (Felt et al., 2015; Slade et al., 2022; Zhang et al., 2017). On the other hand, it raises ethical concerns about the influence of machines on human actions, since the machines were able to "outsmart" the human counterparts while the humans myopically acted in their best interest. To address these concerns, we recommend to design systems that empower humans to make decisions in policy space, rather than just action space, potentially allowing a higher-level control over interactions. This experiment highlighted the need for responsible design and use of machine learning algorithms to ensure the safety, autonomy, and well-being of people.

A naturally reciprocal question arises from previous experiment: if a machine can steer human actions without observing them, can we synthesize a method that allows a human to steer machine actions without observing them? While this is a topic for future research, preliminary insights suggest machines might be able to discern human preferences without estimating utility functions or observing actions, thereby fostering a better human-machine alignment.

While our study provides important insights, however, it has some limitations. The controlled experimental setting does not reflect the complexities of real-world scenarios. Future research should explore noisy systems where the costs are not prescribed but are optimized, such as scenarios involving on metabolic energy (Abram et al., 2022) or other preferences (Ingraham et al., 2022). In higher-dimensional non-linear spaces, humans might not always act optimally due to bounds on informational and computational resources available (Gershman et al., 2015). The use of linear policies, while allowing for theoretical predictions, may limit the impact of our findings. As such, it is important for future studies to explore broader contexts, consider human factors, and conduct external validation of the methods for wider generalizability. Designing adaptive algorithms that play well with humans – who are constantly learning from and adapting to their world – remains an open problem in robotics, neuroengineering, and machine learning (Nikolaidis et al., 2017; Perdikis and del R. Millán, 2020; Recht, 2019).

As adaptive machine algorithms permeate more aspects of daily life, it is important to understand the influence they can exert on humans to prevent undesirable behavior, ensure accountability, and maximize benefit to individuals and society (Parasuraman and Riley, 1997; Thomas et al., 2019). Although the capabilities of humans and machines alike are constrained by the resources available to them, there are well-known limits on human rationality (Tversky and Kahneman, 1974) whereas machines benefit from sustained increases in computational resources, training data, and algorithmic innovation (Hilbert and López, 2011; Jordan and Mitchell, 2015). Here we showed that machines can unilaterally change their learning strategy to select from a wide range of theoretically-predicted outcomes in co-adaptation games played with human subjects. Thus machine learning algorithms may have the power to aid human partners, for instance by supporting decision-making or providing assistance when someone's movement is impaired. But when machine goals are misaligned with those of people, it may be necessary to impose guidelines on algorithms to ensure the safety, autonomy, and well-being of people.

Chapter 5

Conclusion and Future Directions

This thesis contributes tools and techniques for characterizing the outcome of boundedly-rational learning between non-cooperative agents engaging in adaptation across multiple timescales (Chapter 2), with varying payoff structures and stability properties (Chapter 3), and with abilities to anticipate each other's reactions in the context of human-machine systems (Chapter 4). We proposed a framework that incorporates timescale separation, conjectural variations, and policy optimization to advance the understanding of strategic reasoning and adaptation in multi-agent systems, providing a principled foundation for analyzing and predicting the learning dynamics in these systems.

5.1 Towards a Game Theory of Human-AI Co-Adaptation

The game-theoretic framework developed in this thesis provides a rigorous characterization of the learning dynamics and equilibria that arise in multi-agent learning, exploring the factors that influence the stability and outcome of these interactions. Our approach captures the phenomena of human-machine interactions that are not accounted for in traditional approaches by focusing on learning dynamics and incorporating bounded rationality of non-cooperative agents. However, substantial analytical, computational and experimental work remain to establish an encompassing framework for real-world interactions. Towards establishing future work, we first provide an overview of these components of our research. Then, in the next section, we build upon these components to outline future research directions.

The analytical components of our research involved mathematical formulation and analysis of the game-theoretic model. Incorporating different learning rates for agents is crucial for understanding the asymmetric dynamics that arise in human-machine interactions, where the machine agent can quickly learn and respond to the human's actions. Analyzing the stability of learning dynamics was crucial for understanding the outcome of interactions given a fixed set of payoffs and game structures. Furthermore, incorporating belief formation based on conjectural variations equilibrium allowed for the modeling of agents' beliefs about how their opponents will respond to their actions, capturing the strategic reasoning and anticipation that occurs in real-world interactions.

The computational components of our research involved algorithmic implementation and simulation of the learning dynamics. Developing computational methods for solving for the equilibria and learning trajectories predicted by the game-theoretic model enabled the simulation of human-machine interaction scenarios and provides a valuable tool for researchers and developers to explore the implications of different design choices and parameter settings. Furthermore, the numerical simulations allowed for exploration of the parameter space of the model, investigating how changes in factors such as learning rates, payoff structures, and initial conditions affect the learning dynamics and outcomes. Computations helped identify the key factors that have the greatest influence on the interaction dynamics.

The experimental components of our research involved designing and conducting human subjects studies. The experiments were designed as repeated games between human participants and machine agents. Information conditions were manipulated to test specific hypotheses and explore different interaction scenarios. Subjective measures such as participant surveys were used to gain insight into the experiences of the human participants during the interactions. The experimental results were compared to the predictions of the game-theoretic model, confirming the accuracy and explanatory power of the game-theoretic model.

By extending the analytical, computational and experimental components to handle more complex scenarios and dynamics, integrate insights from cognitive science, and validate the predictions against empirical data, we can further enhance the explanatory and predictive power of the game-theoretic approach.

5.2 Future Research Directions

Building upon the insights and methodologies developed in this thesis, there are several promising avenues for future research that can deepen our understanding of human-machine co-adaptation and inform the design of AI systems that are aligned with human values.

Future work will extend the framework to handle more complex multi-agent scenarios and dynamics. Scenarios with incomplete information (Harsanyi, 1967; Zamir, 2020), where agents have limited knowledge about the strategies or beliefs of their opponents, are crucial to study if the model is to be applied to large-scale interactions involving high-dimensional decision spaces amongst many agents. Furthermore, communication and signaling mechanisms (Crandall et al., 2018) will be incorporated to study how information exchange affects the learning dynamics and equilibria in multi-agent systems. On the other hand, towards extending the framework to more complex dynamics, future work will explore characterizing the convergence of different classes of learning dynamics. For instance, momentum and acceleration methods (Nesterov, 1983, 2013) play a significant role in machine learning, and understanding their effect on game outcomes is crucial. By using decompositions to analyze these methods (Balduzzi et al., 2018; Candogan et al., 2011), we can predict outcomes such as rotation, potential, and Hamiltonian dynamics, as well as recognize disequilibrium behaviors like instability, cycles, and chaos. Skew-symmetric decomposition allows us to study the components that contribute to adversarial versus collaborative outcomes in learning dynamics and observe the impact on rotational effects and (dis)equilibrium outcomes. Finally, extending the analysis of conjectural learning dynamics to three-player games may reveal new phenomena.

Future work will also integrate more detailed models of human cognition and decision-making. Close collaboration with cognitive scientists will be required to refine the representation of human learning and decision-making. Incorporating insights from behavioral game theory (Camerer, 2011), prospect theory (Kahneman and Tversky, 1979), and other psychological factors that have been studied in the cognitive science literature will be essential. An interdisciplinary approach will ensure that the framework is informed by the latest empirical findings and validated against empirical data from experimental studies, field observations, and large-scale online interactions. By treating humans as part of the system rather than external operators, we can design human-machine interfaces that enable machines to infer humans' best response maps, leading to interactions that benefit the human user. While Nash, Stackelberg and conjectural variations equilibria have been the focus of this research due to their interpretation and universality, fully characterizing further refinements is a significant area for future work. In particular, solving the problem where the goal is to reach the human's optimum while having limited control over machine actions, is one of the most important challenges in human-AI alignment.

Finally, future work will apply the insights of our research to the design of AI systems by explicitly incorporate human values, preferences, and social norms (Floridi et al., 2021; Jobin et al., 2019). The game-theoretic framework provides a natural way for formalizing the notion of "alignment" between AI systems and human values, preferences, and goals. Toward conducting more experiments with human-machine co-adaptive systems, several testbeds can be used. Studying the interaction between reinforcement learning algorithms and human participants can reveal how environmental factors affect co-adaptation processes (Wellman and Hu, 1998). Additionally, exploring the use of large language models as tools for generating machine policies for human interaction offers a novel experimental testbed with tight feedback loops with large embedding spaces (Bai et al., 2022; Ouyang et al., 2022). Finally, applying optimal control techniques to design algorithms that adjust to human inputs in real-time can result in anticipatory systems that optimize outcomes based on implicit or explicit predictive models of human decision-making (Bertsekas, 2012; Li and Marden, 2013).

However, aligning machine actions with human goals becomes more challenging when machine decisions are not fully observed or controlled, requiring careful experimental designs to estimate and optimize in these scenarios.

We have started to explore how to extend our research to dynamic scenarios where decisions impact not only other agents but also influence the environment in which these agents operate, as discussed in the following works. Section 2.5 provides examples to illustrate how actions and strategies affect environmental states. In our previous study (Chasnov et al., 2019), we explored how adaptive machines affect the human sensorimotor loop, highlighting the need for careful design of human-machine interfaces. Our research in (Vu et al., 2022) examines the leader-follower structures in robots trained via reinforcement learning algorithms, raising questions about the robotic manipulations. Furthermore, (Mceowen et al., 2022) explores the use of convex optimization strategies for drone operations in naval contexts, allowing for real-time parameter adjustments by humans. This effectively induces a Stackelberg game with humans assuming the leader role, but the implications of such human-machine interactions remain unclear. These studies represent a initial steps towards applying game-theoretic to dynamic scenarios.

5.2.1 Limitations and Future Challenges

While this thesis makes significant contributions, there remain limitations and future challenges that need to be addressed. Extending the framework to large-scale, dynamic, and partially observable environments highlights significant technical and computational challenges. Future work will need to develop scalable algorithms, approximation techniques, and computational architectures to handle the increased complexity and dimensionality of these settings. Additionally, capturing the full complexity of human decision-making remains a significant challenge. The development of machine learning systems that interact with and influence human behavior raises profound ethical and societal challenges, including questions of fairness (Barocas et al., 2023), transparency (Wachter et al., 2017), accountability (Diakopoulos, 2015), and privacy (Dwork and Roth, 2014). Tackling the technical, conceptual, and ethical challenges outlined above will require close collaboration across researchers, developers, policymakers, and ethicists to develop responsible AI practices and governance frameworks (Floridi, 2021; Jobin et al., 2019). This calls for new models of interdisciplinary training and education that equip researchers and practitioners with the skills and knowledge needed to navigate these complex challenges and to work effectively in diverse teams.

5.3 Conclusion

In conclusion, the ultimate goal of human-machine co-adaptation research should be to develop a comprehensive and principled understanding of the learning dynamics, strategic considerations, and ethical implications of machine learning systems that interact with humans. This understanding should be used to inform the design of AI systems that are safe, beneficial, and aligned with human values, and that can effectively cooperate with humans to achieve shared goals. The proposed future work aims to deepen our theoretical and empirical understanding of human-machine co-adaptation.

The interplay between machines and humans highlights the crucial distinction between adapting and anticipating. While machines can quickly adapt to human actions, anticipating human actions and preferences remains a significant challenge. Addressing this challenge is essential for creating AI systems that can effectively collaborate with and assist humans. In particular, policy optimization highlights a core concern in AI safety—a sufficiently advanced AI system may find ways to manipulate its environment, including human agents within it, to achieve its objective in unanticipated and uncontrolled ways. To address this threat, we need theoretically- and empirically-justified AI alignment solutions that robustly optimize for human preferences. We also need cognitive science research on how to make humans more aware of and resilient to machine manipulation.

Bibliography

Abram, S. J., Poggensee, K. L., Sánchez, N., Simha, S. N., Finley, J. M., Collins, S. H., and Donelan, J. M. (2022). General variability leads to specific adaptation toward optimal movement policies. *Current Biology: CB*, 32(10):2222–2232.e5.

Argyros, I. K. (1999). A Generalization of Ostrowski's Theorem on Fixed Points. *Applied Mathematics Letters*, 12:77–79.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862 [cs.CL]*.

Balduzzi, D., Czarnecki, W. M., Anthony, T. W., Gemp, I. M., Hughes, E., Leibo, J. Z., Piliouras, G., and Graepel, T. (2020). Smooth markets: A basic mechanism for organizing gradient-based learners. In *International Conference on Learning Representations*.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. (2018). The Mechanics of n-Player Differentiable Games. In *International Conference on Machine Learning*, pages 354–363. PMLR.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Başar, T. and Olsder, G. J. (1998). *Dynamic Noncooperative Game Theory*. Society for Industrial and Applied Mathematics.

Başar, T. and Selbuz, H. (1979). Closed-loop Stackelberg strategies with applications in the optimal control of multilevel systems. *IEEE Transactions on Automatic Control*, 24(2):166–179.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Seminaire de Probabilites XXXIII*, pages 1–68.

Benaïm, M. and Hirsch, M. W. (1999). Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1-2):36–72.

Benaïm, M., Hofbauer, J., and Sorin, S. (2012). Perturbations of set-valued dynamical systems, with applications to game theory. *Dynamic Games and Applications*, 2(2):195–205.

Berard, H., Gidel, G., Almahairi, A., Vincent, P., and Lacoste-Julien, S. (2020). A closer look at the optimization landscapes of generative adversarial networks. *International Conference on Learning Representations*.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific, 2nd edition.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control: Volume I*. Athena Scientific.

Bhatnagar, S. and Prasad, H. L. (2013). *Stochastic Recursive Algorithms for Optimization*. Springer.

Bloembergen, D., Tuyls, K., Hennes, D., and Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. *J. Artificial Intelligence Research*, 53(1):659–697.

Boone, V. and Piliouras, G. (2019). From Darwin to Poincaré and von Neumann: Recurrence and Cycles in Evolutionary and Algorithmic Game Theory. In *Web and Internet Economics: 15th International Conference*, pages 85–99.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

Borkar, V. S. and Pattathil, S. (2018). Concentration bounds for two time scale stochastic approximation. *arxiv:1806.10798*.

Bowley, A. L. (1924). *The Mathematical Groundwork of Economics: An Introductory Treatise*. Clarendon Press.

Bresnahan, T. F. (1981). Duopoly models with consistent conjectures. *The American Economic Review*, 71(5):934–945.

Bu, J., Ratliff, L. J., and Mesbahi, M. (2019). Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv:1911.04672 [eess.SY]*.

Busoniu, L., Babuska, R., and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172.

Calderone, D. J., Chasnov, B. J., Burden, S. A., and Ratliff, L. J. (2023). Consistent conjectural variations equilibria: Characterization and stability for a class of continuous games. *IEEE Control Systems Letters*, 7:2743–2748.

Camerer, C. F. (2011). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press.

Candogan, O., Menache, I., Ozdaglar, A., and Parrilo, P. A. (2011). Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503.

Carey, R. and Everitt, T. (2023). Human control: definitions and algorithms. In *Uncertainty in Artificial Intelligence*, pages 271–281. PMLR.

Chasnov, B. J., Calderone, D., Açıkmeşe, B., Burden, S. A., and Ratliff, L. J. (2020a). Stability of gradient learning dynamics in continuous games: Scalar action spaces. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3543–3548.

Chasnov, B. J., Calderone, D., Açıkmeşe, B., Burden, S. A., and Ratliff, L. J. (2020b). Stability of gradient learning dynamics in continuous games: Vector action spaces. *arXiv:2011.05562 [cs.GT].*

Chasnov, B. J., Fiez, T., and Ratliff, L. J. (2020c). Opponent anticipation via conjectural variations. In *Smooth Games Optimization and Machine Learning Workshop at NeurIPS.*

Chasnov, B. J., Ratliff, L. J., and Burden, S. A. (2023). Human adaptation to adaptive machines can converge to game-theoretic equilibria. *arXiv:2305.01124 [cs.AI].*

Chasnov, B. J., Ratliff, L. J., Mazumdar, E., and Burden, S. A. (2020d). Convergence analysis of gradient-based learning in continuous games. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pages 935–944.

Chasnov, B. J., Yamagami, M., Parsa, B., Ratliff, L. J., and Burden, S. (2019). Experiments with sensorimotor games in dynamic human/machine interaction. *Micro- and Nanotechnology Sensors, Systems, and Applications XI. Vol. 10982. International Society for Optics and Photonic.*

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., and Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, 9(1):233.

Crawford, V. P., Costa-Gomes, M. A., and Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62.

Daskalakis, C., Goldberg, P. W., and Papadimitriou, C. H. (2009). The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training GANs with optimism. In *International Conference on Learning Representations*.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3):398–415.

Díaz, C. A., Villar, J., Campos, F. A., and Reneses, J. (2010). Electricity market equilibrium based on conjectural variations. *Electric Power Systems Research*, 80(12):1572–1579.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407.

Felt, W., Selinger, J. C., Donelan, J. M., and Remy, C. D. (2015). "Body-In-The-Loop": Optimizing device parameters using measures of instantaneous energetic cost. *PLoS ONE*, 10(8):e0135342.

Fiez, T., Chasnov, B. J., and Ratliff, L. J. (2020). Implicit learning dynamics in Stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *ACM International Conference on Machine Learning (ICML)*, pages 3133–3144. PMLR.

Figuières, C., Jean-Marie, A., Quérou, N., and Tidball, M. (2004). *Theory of Conjectural Variations*. World Scientific.

Floridi, L. (2021). Establishing the rules for building trustworthy AI. *Ethics, Governance, and Policies in Artificial Intelligence*, pages 41–45.

Floridi, L., Cowls, J., King, T. C., and Taddeo, M. (2021). How to design AI for social good: Seven essential factors. *Ethics, Governance, and Policies in Artificial Intelligence*, pages 125–151.

Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*, volume 2. MIT Press.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Giocoli, N. (2005). The escape from conjectural variations: the consistency condition in duopoly theory from Bowley to Fellner. *Cambridge Journal of Economics*, 29(4):601–618.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229.

Groot, N. B., De Schutter, B., and Hellendoorn, J. (2013). *Reverse Stackelberg Games: Theory and Applications in Traffic Control*. PhD thesis, Delft University of Technology.

Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, I-III. *Management science*, 8:159–182, 320–334, 486–502.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150.

Hart, S. and Mas-Colell, A. (2001). A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54.

Hart, S. and Mas-Colell, A. (2003). Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183.

Heald, J. B., Lengyel, M., and Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, 600(7889):489–493.

Heinrich, J. and Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. *arxiv:1603.01121*.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65.

Ho, Y.-C., Luh, P. B., and Muralidharan, R. (1981). Information structure, Stackelberg games, and incentive controllability. *IEEE Transactions on Automatic Control*, 26(2):454–460.

Ho, Y.-C., Luh, P. B., and Olsder, G. J. (1982). A control-theoretic view on incentives. *Automatica*, 18(2):167–179.

Hofbauer, J. (1996). Evolutionary dynamics for bimatrix games: A Hamiltonian system? *Journal of Mathematical Biology*, 34(5):675.

Hommes, C. H. and Ochea, M. I. (2012). Multiple equilibria and limit cycles in evolutionary games with logit dynamics. *Games and Economic Behavior*, 74(1):434 –441.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis.* Cambridge University Press.

Huang, J., Isidori, A., Marconi, L., Mischiati, M., Sontag, E. D., and Wonham, W. M. (2018). Internal models in control, biology and neuroscience. In *IEEE Conference on Decision and Control (CDC)*, pages 5370–5390.

Ingraham, K. A., Remy, C. D., and Rouse, E. J. (2022). The role of user preference in the customized control of robotic exoskeletons. *Science Robotics*, 7(64):eabj3487.

Itaya, J.-i. and Shimomura, K. (2001). A dynamic conjectural variations model in the private provision of public goods: a differential game approach. *Journal of Public Economics*, 81(1):153–172.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Jungers, M., Trélat, E., and Abou-Kandil, H. (2011). Min-max and min-min Stackelberg strategies with closed-loop information structure. *Journal of Dynamical and Control Systems*, 17(3):387.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Karmakar, P. and Bhatnagar, S. (2018). Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Mathematics of Operations Research.*

Kerr, B., Riley, M. A., Feldman, M. W., and Bohannan, B. J. (2002). Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature*, 418(6894):171–174.

Khalil, H. K. (2002). *Nonlinear Systems Theory.* Prentice Hall.

Kokotovic, P. V. and Khalil, H. K. (1986). *Singular Perturbations in Systems and Control.* IEEE Press.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition.

Laffont, J.-J. and Martimort, D. (2009). The theory of incentives: the principal-agent model. In *The theory of incentives*. Princeton University Press.

Langer, H., Markus, A., Matsaev, V., and Tretter, C. (2001). A new concept for block operator matrices: the quadratic numerical range. *Linear Algebra and its Applications*, 330(1-3):89–112.

Letcher, A., Foerster, J., Balduzzi, D., Rocktäschel, T., and Whiteson, S. (2019). Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*.

Li, N. and Marden, J. R. (2013). Designing games for distributed optimization. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):230–242.

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

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier.

Liu, J. D., Lie, T. T., and Lo, K. L. (2006). An empirical method of dynamic oligopoly behavior analysis in electricity markets. *IEEE TPS*, 21.

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. (2019). Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences of the United States of America*, 116(42):20881–20885.

Madduri, M. M., Burden, S. A., and Orsborn, A. L. (2021). A game-theoretic model for co-adaptive brain-machine interfaces. In *IEEE/EMBS Conference on Neural Engineering (NER)*, pages 327–330. IEEE.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, 87:102426.

Mazumdar, E., Jordan, M. I., and Sastry, S. S. (2019). On Finding Local Nash Equilibria (and Only Local Nash Equilibria) in Zero-Sum Games. *arXiv:1901.00838 [cs.LG]*.

Mazumdar, E. and Ratliff, L. J. (2018). On the convergence of competitive, multi-agent gradient-based learning algorithms. *arxiv:1804.05464*.

Mazumdar, E., Ratliff, L. J., and Sastry, S. S. (2020). On Gradient-Based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131.

Mceowen, S., Sullivan, D., Calderone, D., Szmuk, M., Sheridan, O., Açıkmeşe, B., and Chasnov, B. (2022). Visual modeling system for optimization-based real-time trajectory planning for autonomous aerial drones. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–9.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mertikopoulos, P., Papadimitriou, C. H., and Piliouras, G. (2018). Cycles in adversarial regularized learning. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM.

Mertikopoulos, P. and Zhou, Z. (2019). Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1–2):456–507.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. (2017). Unrolled generative adversarial networks. In *International Conference on Learning Representations*.

Monderer, D. and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, 14(1):124–143.

Nash, J. F. (1950). Equilibrium points in N-Person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49.

Nash, J. F. (1951). Non-cooperative games. *Ann. Math.*, pages 286–295.

Nesterov, Y. E. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences.

Nesterov, Y. E. (2013). *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media.

Ng, A. Y. and Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *ACM International Conference on Machine Learning (ICML)*, pages 663–670.

Nikolaidis, S., Nath, S., Procaccia, A. D., and Srinivasa, S. (2017). Game-theoretic modeling of human adaptation in human-robot collaboration. In *ACM/IEEE Conference on Human-Robot Interaction (HRI)*, pages 323–331.

Nowak, M. A. (2006). *Evolutionary Dynamics*. Harvard University Press.

Olsder, G. J. (1981). A critical analysis of a new equilibrium concept. *Memorandum N. 329, Dept. Applied Maths., Twente University of Technology, The Netherlands.*, 71.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. (2017). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR.

Ostrowski, A. M. (1966). *Solution of Equations and Systems of Equations*. Academic Press.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Palan, S. and Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Papadimitriou, C. H. and Piliouras, G. (2018). Game dynamics as the meaning of a game. *Sigecom*.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.

Perdikis, S. and del R. Millán, J. (2020). Brain-Machine interfaces: A tale of two learners. *IEEE Systems, Man, and Cybernetics Magazine*, 6(3):12–19.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2013). Characterization and Computation of Local Nash Equilibria in Continuous Games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing*, pages 917–924. IEEE.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2014). Generictiy and Structural Stability of Non–Degenerate Differential Nash Equilibria. In *2014 American Control Conference*.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2016). On the Characterization of Local Nash Equilibria in Continuous Games. *IEEE Transactions on Automatic Control*, 61(8):2301–2307.

Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems (ARCRAS)*, 2(1):253–279.

Rosen, J. B. (1965). Existence and uniqueness of equilibrium points for n-person concave games. *Econometrica*, 33(3):520–534.

Rubinstein, A. (1998). *Modeling bounded rationality*. MIT Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Sastry, S. S. (1999). *Nonlinear Systems*. Springer New York.

Sato, Y., Akiyama, E., and Farmer, J. D. (2002). Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751.

Semmann, D., Krambeck, H.-J., and Milinski, M. (2003). Volunteering leads to rock–paper–scissors dynamics in a public goods game. *Nature*, 425(6956):390–393.

Shah, S. M. (2021). Stochastic approximation on riemannian manifolds. *Applied Mathematics & Optimization*, 83:1123–1151.

Shoham, Y., Powers, R., and Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, pages 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2):129.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT Press.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. OUP Oxford.

Slade, P., Kochenderfer, M. J., Delp, S. L., and Collins, S. H. (2022). Personalizing exoskeleton assistance while walking in the real world. *Nature*, 610(79317931):277–282.

Soares, N., Fallenstein, B., Armstrong, S., and Yudkowsky, E. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1):1–10.

Tang, Y. and Li, N. (2020). Distributed zero-order algorithms for nonconvex multi-agent optimization. pages 269–281.

Tatarenko, T. and Kamgarpour, M. (2019). Learning Nash Equilibria in Monotone Games. In *2019 IEEE 58th Conference on Decision and Control*, pages 3104–3109.

Taylor, J. A., Krakauer, J. W., and Ivry, R. B. (2014). Explicit and implicit contributions to learning in a sensorimotor adaptation task. *Journal of Neuroscience*, 34(8):3023–3032.

Thomas, P. S., Castro da Silva, B., Barto, A. G., Giguere, S., Brun, Y., and Brunskill, E. (2019). Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004.

Thoppe, G. and Borkar, V. S. (2019). A concentration bound for stochastic approximation via Alekseev's formula. *Stochastic Systems*, 9(1):1–26.

Tretter, C. (2008). *Spectral Theory of Block Operator Matrices and Applications.* World Scientific.

Tretter, C. (2009). Spectral inclusion for unbounded block operator matrices. *Journal of Functional Analysis*, 256(11):3806–3829.

Tuyls, K., Pérolat, J., Lanctot, M., Ostrovski, G., Savani, R., Leibo, J. Z., Ord, T., Graepel, T., and Legg, S. (2018). Symmetric decomposition of asymmetric games. *Scientific Reports*, 8(1):1015.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Varian, H. R. (1992). *Microeconomic Analysis.* Norton & Company.

von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior.* Princeton University Press.

von Stackelberg, H. (1934). *Marktform und Gleichgewicht.* Springer.

von Stackelberg, H. (2010). *Market Structure and Equilibrium.* Springer Science & Business Media.

Vu, Q.-L., Alumbaugh, Z., Ching, R., Ding, Q., Mahajan, A., Chasnov, B., Burden, S., and Ratliff, L. J. (2022). Stackelberg policy gradient: Evaluating the performance of leaders and followers. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions.*

Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99.

Wellman, M. P. and Hu, J. (1998). Conjectural equilibrium in multiagent learning. *Machine Learning*, 33:179–200.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882.

Zamir, S. (2020). Bayesian games: Games with incomplete information. In *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pages 119–137. Springer US.

Zhang, J., Fiers, P., Witte, K. A., Jackson, R. W., Poggensee, K. L., Atkeson, C. G., and Collins, S. H. (2017). Human-in-the-loop optimization of exoskeleton assistance during walking. *Science*, 356(6344):1280–1284.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384.

Zheng, Y.-P. and Başar, T. (1982). Existence and derivation of optimal affine incentive schemes for Stackelberg games with partial information: A geometric approach. *International Journal of Control*, 35(6):997–1011.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA