Algorithmic Collusion: Theory & Practice

Patrick Chang
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2025

Acknowledgements

As I review my journey of writing this thesis, I am reminded of how fortunate I am to have Álvaro Cartea as my supervisor, and how fortunate I am to be at the Oxford-Man Institute (OMI).

To Álvaro, I am deeply grateful for your guidance, advice, and support throughout this journey. From our discussions over *one* pint, to polishing my drafting and encouraging me to aim for impactful research questions, your mentorship has profoundly shaped my research and growth.

I would like to thank my examiners Alan Beggs, Emilio Calvano, Xiaowen Dong, Tom Noe, and Leandro Sánchez-Betancourt for their helpful comments and feedback.

To Gabriel García-Arenas, Rob Graumans, Roel Oomen, José Penalva, and Harrison Waldon, I am fortunate to have you as talented coauthors. I have learned a great deal from each of you.

I would like to thank the OMI for funding my degree, and to thank all the members of the OMI for creating a wonderful work environment. Special thanks to Jen Desmond and Anthony Ledford for being a constant pillar of support at the OMI.

To Álvaro Arroyo, Fayçal Drissi, Gerado Durán-Martín, and Marcello Monga I am grateful to have you as friends. Your presence has made this journey exceptionally special.

Finally, I am deeply thankful to my parents and my sister for their unwavering support and encouragement.

Abstract

We develop a framework to analyze the evolution of bounded memory strategies in a repeated game. In this framework, we introduce the algorithmic learning equations, a set of ordinary differential equations which characterizes the finite-time and asymptotic behavior of the stochastic interaction between learning algorithms that learn a bounded memory strategy in a repeated game. Our framework allows us to study repeated games under a variety of monitoring structures, including perfect, public, private, and any of the combinations.

Using this framework, we use a dynamic generalization of smooth fictitious play with bounded m-memory strategies to model learning with bounded rationality that is consistent with learning by algorithms. With this learning model, we prove a Folk theorem when players with bounded rationality learn as they play a repeated potential game. In a repeated potential game with perfect monitoring, we use this learning model to show that for any feasible and individually rational payoff profile, if players have sufficient memory, are sufficiently patient, and best respond with sufficiently few mistakes, then the players have a non-zero probability of learning an m-memory strategy profile that achieves an average payoff close to the specified payoff profile for an appropriate continuation game. Moreover, the strategy profile learned is an m-memory ϵ-subgame perfect equilibrium of the repeated game.

Finally, we examine a case study where high-frequency traders (HFTs) in the European ETF market break the pre-trade anonymity of limit orders by signaling their type in an otherwise anonymous market. We explain the behavior of HFTs with a model that considers competitive and collusive equilibria. The model shows that the behavior of the HFTs is consistent with that in a collusive equilibrium where HFTs signal themselves to avoid sniping each other's limit orders. Signaling enables the HFTs to share the benign flow from retail limit orders, and to share the additional benign flow from impatient investors who otherwise would have traded with a retail investor's limit order.

Contents

Chapter 1

Introduction

1.1 Motivation

Learning Algorithms

Recent advance in learning algorithms (i.e., autonomous artificial agents (AAs)) hold promise to facilitate and improve data-driven sequential decision-making problems in an increasingly digital world. However, in many real world applications, the algorithms make decisions in a multi-agent system whereby multiple algorithms concurrently act and adapt.

Unfortunately, the theoretical foundations for multi-agent learning are lacking with potential opacity at how the algorithms arrive at decisions. Furthermore, learning interactions with other algorithms can be unpredictable and can lead to unintended behaviors. Indeed, several competition authorities such as the OECD, the EU Commissioner for Competition, and the Federal Trade Commission (for a more detailed list and references see Assad et al., 2021) are concerned that AAs may tacitly learn to collude. That is, when algorithms learn to extract rents above the competitive level in a coordinated fashion, even if they have not been specifically instructed to do so and even if they do not communicate with one another. This form of tacit collusion would defy current antitrust policy, which typically targets explicit agreements among would-be competitors, and has many implications for competition policies (see Harrington, 2018).[1]

Multi-agent learning in the simplest form reduces to learning in games, which has been studied by game theorists for decades. Instead of assuming that players are perfectly rational, the focus is on repeated interactions between players who act and adapt based on the history of play. Relaxing the assumptions about knowledge and rationality of players allows one to test the sensitivity of game theoretic results to these typical assumptions. Much of the

[1] Based on recent case law, evidence of overt communication between firms appears to be a prerequisite to successfully prosecute collusive arrangements (see Chassang and Ortner, 2023).

literature in learning in games focuses on establishing positive and negative convergence results of the learning interaction to a solution concept such as a Nash equilibrium.

Over the last few decades, many convergence results have been established for evolutionary models and myopic (stateless) learning algorithms, where the strategies are not conditioned on information sets relating to past actions. This is a stark contrast to the vast literature on repeated games where various Folk theorems are established for different information sets. The key issue is the inherent non-stationarity induced by multiple players learning and adapting their strategies. The changes in the strategies change the actions of the players, which in turn changes the dynamics of the information set with which their strategies are conditioned on. In the repeated games literature, Folk theorems are established with a fixed strategy that players know from the start (how players arrive at such strategies without communication is unclear, see Green et al., 2014), while the premise of learning in games is to *learn* a strategy through repeated interactions.

To study AAs learning to collude, we move beyond action learning and study strategy learning. Learning repeated game strategies is the first step to learn to collude in the full economic sense, where cooperation is sustained through self-policing. That is, the repeated game strategy must contain a reward-punishment mechanism to ensure that players do not deviate for short-term profits because deviations will be punished. Hence, a reward-punishment mechanism ensures that cooperation is an equilibrium outcome sustained by inter-temporal incentives.

Despite these challenges, the extant literature uses simulations to study the learning interaction between AAs and finds that algorithms can learn supracompetitive outcomes (see Waltman and Kaymak, 2008), and also finds that algorithms can learn strategies to sustain supracompetitive outcomes through a behavior that resembles a reward-punishment mechanism (see Calvano et al., 2020, 2021). However, recent literature also uses simulations to show that the strategies learned may or may not be collusive (see Abada and Lambin, 2023; Lambin, 2023).[2]

In this thesis, we use theory to settle whether or not simple AAs can learn to collude. We do so by connecting the learning component of AAs with existing Folk theorems by asking the following question: Does the Folk theorem continue to hold when less than fully rational players (i.e., AAs) learn as they play the game?

[2] Asker et al. (2022, 2023) and Hansen et al. (2021) theoretically show that algorithms can learn supracompetitive outcomes, but these outcomes are not supported by a collusive equilibrium.

Pricing Algorithms

Another concern about algorithmic collusion is that it is not only limited learning algorithms. Advances in compute power has allowed pricing algorithms (that do not learn a strategy) to handle large quantities of data, which has created a shift from mechanically-set prices to prices set by algorithms. Although pricing algorithms have been used for decades (e.g., airlines industry and financial sector), the widespread use of algorithms has also raised concerns of possible anti-competitive behavior as they can make it easier for firms to achieve and sustain collusion without any formal agreement or human interaction (see OECD, 2017).

Unlike AAs learning to collude, which remain theoretical in nature, there are recent case studies that raise concerns for pricing algorithms. The most recent example is the Justice Department filing a civil antitrust lawsuit against RealPage Inc. for its unlawful scheme to decrease competition among landlords.[3] Another example is the online retailer Trod Ltd. and its co-conspirators who agreed to adopt specific pricing algorithms for the sale of certain posters sold on Amazon Marketplace.[4]

In the final chapter of the thesis, we examine a case study where high-frequency traders (HFTs) in the European ETF market break the pre-trade anonymity of limit orders by signaling their type in an otherwise anonymous market.[5] In an anonymous limit order book, revealing yourself to others is a voluntary and costly decision. Since rational agents do not give information for free, the question is then: what are the offsetting benefits for HFTs? What happens when HFTs break the pre-trade anonymity of limit orders and reveal themselves to each other?

1.2 Contribution

Theory of Algorithmic Collusion

There are two steps to address the question of whether simple AAs can learn to collude. First, in Chapter 3, we develop a framework to analyze the evolution of bounded memory strategies in a repeated game. We introduce the algorithmic learning equations (ALEs), a set of ordinary differential equations (ODEs) which characterizes the finite-time and asymptotic behavior of the stochastic interaction between learning algorithms that learn a

[3] See Justice Department Sues RealPage for Algorithmic Pricing Scheme that Harms Millions of American Renters.

[4] See Online Retailer Pleads Guilty for Fixing Prices of Wall Posters.

[5] The nature of high-frequency trading means that decisions are made by algorithms. Human traders are simply unable to compete. The median reaction time of HFTs is 91 microseconds (see Cartea et al., 2025).

bounded memory strategy in a repeated game. Our framework allows us to study repeated games under a variety of monitoring structures, including perfect, public, private, and any of the combinations.

Key to this generality is to define appropriate states of the game to capture all the information sets required by the players' strategies, i.e., the information players condition their action on. This setup allows us to study learning algorithms with behavioral strategies that condition on the (recent) history. Specifically, we have an automaton-like representation of the space of bounded memory strategies used by the players so that we can analyze the evolution of their strategies through learning.

Our derivation of the ALEs relies on a time-rescaling and two-step averaging procedure to apply stochastic approximation methods. In the first step, we re-scale time to compare updates of a discrete time process with a continuous time function. Then, we average the learning rule, conditional on the state process and the parameters that the learning rule updates. In the second step, we average the state process with respect to its stationary distribution to obtain the deterministic mapping characterized by the ALEs to describe the dynamics of the learning algorithm. In this step, the stationary distribution is obtained by considering the state process driven by a fixed strategy profile. When this process is ergodic, it has a unique stationary distribution.

Our first result proves that with high probability, trajectories of the ALEs approximate the evolution of strategies over finite time horizons. That is, over finite time intervals, we obtain uniform convergence in probability between the trajectories of the ALEs and the trajectories of the discrete-time learning algorithms. This result is useful to describe the evolution of the path of the learning interactions, however, the theorem does not characterize asymptotic behavior of the algorithms. Our second result characterizes the asymptotic behavior of the algorithms. First, we prove a notion of asymptotic convergence, i.e., we show that the learning algorithms are asymptotic pseudo-trajectories of the ALEs, in the sense of Benaïm (1999). Second, we prove that under certain conditions, the evolution of the learning algorithms converge to an asymptotically stable rest point of the ALEs.

A key consideration of our result, in the spirit of Ma et al. (1990), is to provide a set of minimal sufficient conditions that are easy to verify. Specifically, the important assumptions to verify are basic characteristics of the learning model, and basic characteristics of the monitoring structure.

Equipped with a framework to study learning bounded memory strategies in a repeated game, we study the connection between learning and Folk theorems in Chapter 4. To this end, we use a dynamic generalization of smooth fictitious play as our model of learning. We use this learning model to prove a subgame perfect Folk theorem from learning without

communication in repeated potential games (which includes Cournot oligopoly with linear demand, see Monderer and Shapley, 1996). Our result is as follows. Suppose that players play a repeated potential game and have perfect monitoring. For any $\epsilon > 0$ and for any feasible and individually rational payoff profile $\mathbf{v}$, if players use bounded m-memory strategies with sufficient memory, are sufficiently patient, and best respond with sufficiently few mistakes, then the players have a non-zero probability of learning (converging to) an m-memory strategy profile that yields an average payoff profile within ϵ of $\mathbf{v}$ for an appropriate continuation game. Moreover, the strategy profile learned is an m-memory ϵ-subgame perfect equilibrium of the repeated game for any continuation game from convergence onwards.

Our focus on a subgame perfect Folk theorem instead of a Nash Folk theorem arises from our interest in collusion. A necessary condition to establish collusion (tacit or explicit) is for the strategy to embody a reward-punishment mechanism (see Harrington, 2018). Subgame perfection ensures that threats are credible, i.e., it is in each player's interests to carry out the punishment when there is a deviation. Given that learning algorithms are bounded in memory, the subgame perfect equilibrium we use is one that only considers m-memory strategies. We refer to these equilibrium strategies as an m-memory subgame perfect equilibrium of the repeated game.

A key aspect of collusion is that players recognize the repeated nature of their interaction. Most learning models in repeated games have players that learn stage game strategies that, by design, do not account for repeated interactions. Hence, players in these learning models cannot learn to collude because they cannot sustain cooperation through intertemporal incentives by using reward-punishment strategies. To overcome this, our analysis is underpinned by a learning model which we call *state-dependent smooth fictitious play*. The learning model is a dynamic generalization of smooth fictitious play with bounded m-memory strategies. In our learning model, players understand they interact repeatedly and therefore optimize their expected discounted stream of future payoffs. Players also understand there is a dynamic relationship between current and future actions and they use the past to learn about how current actions affect future ones. However, players are bounded in rationality because at each round of play they assume this dynamic relationship is fixed and given by their belief, and they assume this dynamic relationship is adequately described by past play.

Concretely, in our learning model, players are infinitely-lived, they know their own payoffs for the game, and they assume that everyone plays according to a stationary m-memory strategy profile.[6] Learning takes place as players update their belief over the

[6]The learning rule is uncoupled because players have no knowledge of their opponents' utility function (see Hart and Mas-Colell, 2003).

stationary m-memory strategy profile based on the empirical frequency of play. The belief is given by a collection of distributions for each player's actions conditional on each m-memory history. At each point in time, players use the belief over the stationary m-memory strategy profile to build a hypothetical world in which players play according to that belief. Within this hypothetical world, and given the current m-memory history, the players compute action values, that is, for each action they compute the expected discounted payoffs associated with the one-shot deviation principle. Then, they play a smoothed best response with respect to these action values. Therefore, at each point in time, the players' behavioral responses are non-myopic and optimize for both the immediate payoff and the expected discounted stream of future payoffs based on their belief.

Our learning model reduces to smooth fictitious play when $m = 0$. Thus, our learning model lies between myopic learning models that learn stage game strategies and rational (Bayesian) learning (see e.g., Kalai and Lehrer, 1993). However, unlike other learning models that also lie between these two extremes (see e.g., Foster and Young, 2003), our learning process remains tractable because we use an automaton-like representation of the space of m-memory strategy profiles.

Using the ALEs, we obtain a system of ODEs that approximate the evolution of the players' beliefs. The ODEs, which we call the state-dependent smoothed best response dynamics, allow us to analyze the evolution of beliefs in the space of stationary m-memory strategy profiles. The ODEs also allow us to prove that play from our learning model converges to a connected subset of m-memory ϵ-subgame perfect equilibria of the repeated game with probability one in repeated potential games. Proving convergence is a necessary step, but it is insufficient to establish a Folk theorem from learning because it does not characterize the equilibria that can be learned. Therefore, we refine the convergence result to show that for any m-memory pure strategy subgame perfect equilibrium of the repeated game, play from our learning model has a non-zero probability of converging to an m-memory strategy profile near the target equilibrium. Moreover, the strategy profile learned approximately achieves the average payoff profile of the target equilibrium.

The Folk theorem from learning follows from our convergence result once we use m-memory pure strategy profiles to obtain a Folk theorem. Barlo et al. (2016) prove a Folk theorem using m-memory pure strategy profiles when monitoring is perfect and without public randomization but without learning. Therefore, by combining their result and our convergence result, we obtain a Folk theorem from learning in repeated potential games. Finally, our Folk theorem from learning does not say which equilibrium will be learned because it only states that there is a non-zero probability of converging to the equilibrium

required to support the payoff profile.[7]

Our final result addresses equilibrium selection. Specifically, for any $\varepsilon > 0$, if the belief lies within a neighborhood of the target equilibrium, then there exists a sufficiently strong belief (i.e., a new empirical observation has little impact when updating the belief) such that play converges to the target equilibrium with probability greater than or equal to $1 - \varepsilon$.

In Chapter 5, we study toy models of a duopoly to visualize the space of one-memory strategies and its evolution through learning. We compare the learning outcomes of state-dependent smooth fictitious play and Q-learning under perfect monitoring and imperfect public monitoring. We find that both learning models can learn to collude, but Q-learning can also learn non-equilibrium strategies.

Rationalizing the Case Study

In Chapter 6, we propose a model of the limit order book that considers competitive and collusive equilibria. Our model rationalizes the behavior observed in Cartea et al. (2025) and highlights the economic forces that underlie the incentives of HFTs to reveal themselves to each other. Our model also answers the following questions: under what conditions would an HFT reveal herself to others? How do HFTs mutually benefit from revealing themselves, and how can they collectively enforce this behavior?

In our model, there are patient and impatient investors, and there are $N \geq 3$ risk-neutral HFTs who play an infinitely repeated trading game. The trading game is a generalization of the two-period trading game in Budish et al. (2024), and it retains the main elements of latency arbitrage and adverse selection. In each trading game, our model considers the possibility that one of the HFTs receives short-lived private information, or that a patient investor arrives and sends a limit order that improves the quoted spread. Informed HFTs trading alongside patient investors introduces strategic ambiguity. Ambiguity arises because a limit order that is sent inside the spread could be benign flow from a patient investor or it could be toxic flow from an informed HFT. In the latter case, the informed HFT pretends to be a patient investor and posts a toxic limit order to fool a sniper into trading with the toxic limit order, i.e., the sniper adversely selects herself when trading with the informed HFT.

This ambiguity creates a decision problem for the HFTs. Do HFTs snipe an incoming spread-improving limit order without knowing who sent it and potentially face the risk of adversely selecting themselves, or do HFTs wait until they can verify that the limit order

[7]The result includes a probabilistic element because of the intrinsic randomness from learning, and because of the multiplicity of equilibria.

was sent by a patient investor before they trade? The choice of HFTs to snipe or to wait-and-verify affects the equilibrium bid-ask spread set by HFTs and it also affects the quoted spread that impatient investors (i.e., liquidity traders) face.

In the competitive equilibrium, if HFTs immediately snipe the limit orders that improve the quoted spread, then the impatient investors always trade at the spread set by the HFTs, i.e., impatient investors cannot trade with spread-improving limit orders posted by patient investors. On the other hand, if HFTs do not snipe immediately, then the impatient investors can occasionally trade with a spread-improving limit order posted by a patient investor. Thus, when HFTs do not snipe immediately, the HFT who provides liquidity has fewer opportunities to provide liquidity to an impatient investor, while the exposure to being adversely selected by other HFTs does not decrease. Therefore, to compensate for a decrease in revenue from benign flow without a decrease in adverse selection costs, HFTs set a wider equilibrium bid-ask spread.

The intuition to determine if HFTs immediately snipe incoming limit orders that improve the quoted spread is straightforward. In the competitive equilibrium, if the benign flow from immediately sniping patient investors is sufficiently profitable, then HFTs immediately snipe any limit order that improves the quoted spread and bear the costs of being fooled by informed HFTs pretending to be patient investors. However, if the cost of being fooled by informed HFTs is too high, then HFTs wait until they can verify that the order was sent by a patient investor before they trade.

With a competitive baseline established, we study a collusive equilibrium to explain the observed behavior of HFTs. To study the effect and rationale of signaling, we focus on collusive strategies that do not cooperate to post spreads that are wider than the competitive spread (see Dutta and Madhavan, 1997). This restriction enables us to focus on the benefits that arise from breaking anonymity. The collusive strategy of interest is a reversion strategy with a cooperation phase and a punishment phase (see Green and Porter, 1984). HFTs cooperate until they observe a deviation or suspect a deviation from cooperation, both of which trigger a punishment phase that reverts to competitive play for T iterations of the trading game, after which, play returns to the cooperation phase.

In the cooperation phase, HFTs signal to each other to avoid trading with each other. This enables the HFTs to share the profitable benign flow of patient investors, and as a byproduct, enables the HFTs to receive additional benign flow from impatient investors who would otherwise have matched with a patient investor's spread-improving limit order. These supracompetitive profits from cooperation provide the necessary incentives for the HFTs to willingly reveal themselves to each other. However, when cooperating, HFTs have

myopic incentives to cheat. The incentive to cheat can be deterred through intertemporal incentives with the threat of reverting to competitive play in the punishment phase.

When the incentives align and cheating can be deterred, we obtain a collusive equilibrium in which supracompetitive profits from cooperation arise from the ability to identify benign limit orders sent by patient investors and from the ability to identify toxic limit orders sent by informed HFTs. The resulting collusive equilibrium is one where: (i) quoted spreads are on average wider than that in the competitive equilibrium because HFTs can safely snipe limit orders sent by patient investors, and (ii) the trading costs for impatient investors are higher because they are forced to trade at the spread set by HFTs (they cannot trade with spread-improving limit orders from a patient investor unless play is in the punishment phase).

Finally, the collusive equilibrium does not always exist. Factors that affect the existence of the collusive equilibrium include the number of HFTs, arrival of private information, and the mispricing of limit orders from patient investors. Theoretically, as the number N of HFTs increases, the possibility of colluding decreases. Furthermore, the collusive equilibrium does not exist if the HFTs do not receive short-lived private information. Lastly, the collusive equilibrium exists only when the limit orders sent by patient investors are sufficiently mispriced.

Relation to Papers

- The contents of Chapter 3 are based on the paper Cartea et al. (2022d).
- The contents of Chapter 4 are based on the paper Cartea et al. (2022c).
- The contents of Chapter 6 are based on the paper Cartea et al. (2025).

Other Papers

Throughout the DPhil, I have co-authored other papers about the unintended consequences of learning algorithms.

- Cartea et al. (2022a,b) study how market makers who use reinforcement learning algorithms can end up quoting supracompetitive prices.
- Cartea et al. (2023a) study how learning algorithms can learn to manipulate the market.

Chapter 2

Related Literature

2.1 Learning in Games

Learning in games is a fundamental field in game theory (see Fudenberg and Levine, 1998, 2009, for an overview), where the focus is to study repeated interactions between players who act and adapt based on the history of play. Within this field, the experimental literature often focuses on establishing similarities between the learning models and how humans learn; whereas the theoretical literature often focuses on establishing positive and negative convergence results of the learning interaction to solution concepts.

Action Learning

There are two distinctive approaches to model how humans learn. These approaches are split into reinforcement learning and belief learning. The models differ on two levels: what information players use and whether players optimize given that information. The canonical reinforcement learning models from Cross (1973), Arthur (1991), and Erev and Roth (1998) aim to replicate human behavior with very limited rationality. Belief learning in the form of fictitious play from Brown (1951) offers a higher level of rationality whereby players optimize myopically with respect to their beliefs. Camerer and Ho (1999) propose experience-weighted attraction learning that blends elements from reinforcement learning and belief learning.

In addition to fitting these learning models to experimental data, theorists have also studied the asymptotic properties of these learning models. Arthur (1993) makes a connection between his learning model and the replicator dynamics from evolutionary game theory. Börgers and Sarin (1997) also makes the connection between Cross' learning model and the replicator dynamics.

Several authors establish convergence results for Erev and Roth's learning model, which is related to the Maynard Smith replicator dynamics. Beggs (2005) proves that strategies converge in constant-sum games with unique equilibria if they are pure or if they are mixed and the game is 2×2. Hopkins and Posch (2005) proves that strategies converge to a pure strategy Nash equilibrium in a re-scaled partnership game. Duffy and Hopkins (2005) proves similar convergence results for the Erev and Roth's learning model with different choice rules. Similarly, several convergence results are established for stochastic fictitious play which is related to the perturbed best response dynamics. Benaïm and Hirsch (1999a) proves convergence results in a 2×2 game. Hofbauer and Sandholm (2002) establish global convergence results for four classes of games: games with an interior ESS, zero sum games, potential games, and supermodular games. On the other hand, Hopkins (2002) compares the properties of reinforcement learning and stochastic fictitious play. He shows that the expected motion of stochastic fictitious play and reinforcement learning with experimentation can both be written as a perturbed form of the evolutionary replicator dynamics. Consequently, the two models in many cases have the same asymptotic behavior.

Relatedly, Benaïm and Hirsch (1999b) proves convergence results for I player coordination games with two actions when the learning rate is constant. Benaïm and Weibull (2003) use an ODE approximation to study the finite-time and long-run behavior of evolutionary game dynamics. Arieli and Young (2016) study the time taken to come close to Nash equilibrium and provide explicit bounds on the speed of convergence. Mertikopoulos and Sandholm (2016) and Mertikopoulos et al. (2022) provide a range of connections between reinforcement learning algorithms and various dynamics.

Apart from convergence to Nash equilibria, there are several papers that study learning models that converge to correlated equilibria. Foster and Vohra (1997) study a calibrated learning model where players play a myopic best response to a calibrated forecast of the other's plays. Hart and Mas-Colell (2000, 2001) study a regret matching procedure. Lenzo and Sarver (2006) study the connection between correlated equilibrium and the multipopulation replicator dynamics, where each population is comprised of multiple subpopulations.

A common thread in these learning models is the strategies are limited over actions, so players are forced to be myopic. In contrast, our learning framework allows for bounded memory strategies, so players can be non-myopic.

Other Learning Models and ODEs Two other research communities have also made additional connections between learning models and appropriate ODEs. Sato and Crutchfield (2003) and Tuyls et al. (2003) independently derive identical equations that describe the dynamics of stateless Q-learning. Physicists are generally more interested in deterministic

chaos from the learning model, whereas computer scientists generally focus on the design of new algorithms and finding connections between learning algorithms and dynamical systems.

The physics community adopt a statistical physics approach where the underlying assumption is that players interact infinitely many times before they adapt their behavior. Deterministic chaos is found in replicator dynamics (see Sato et al., 2002), and the dynamics of experience weighted attraction learning (see Galla and Farmer, 2013).

On the other hand, the computer science community have made numerous connections between reinforcement learning models and evolutionary game dynamics (see Bloembergen et al., 2015, for an overview). Gomes and Kowalczyk (2009), Babes et al. (2009) and Wunder et al. (2010) derive the dynamics of stateless Q-learning with an ϵ-greedy policy. Kleinberg et al. (2009) and Kasbekar and Proutiere (2010) show that the dynamics of the Hedge algorithm and the exponential-weight algorithm for exploration and exploitation (EXP3) recover the replicator dynamics, respectively. With this approach, the community developed several algorithms: frequency adjusted Q-learning (see Kaisers and Tuyls, 2010), lenient Q-learning (see Panait et al., 2008), and lenient frequency adjusted Q-learning (see Bloembergen et al., 2010).

All the aforementioned papers have been restricted to the case of stateless algorithms learning the optimal action. Several papers have attempted to address this shortfall. Vrancx et al. (2008) introduce the piecewise replicator dynamics, Hennes et al. (2009) introduce the state-coupled replicator dynamics, and Hennes et al. (2010) introduce the state-coupled replicator-mutation dynamics. These papers have two shortfalls: first, they consider bandit-type algorithms where the algorithms do not optimize a Markov Decision Process (MDP), but focus on selecting the optimal action given the context. Second, the dynamics are obtained heuristically with no theoretical guarantees.

Barfuss et al. (2019) address the first shortfall. The authors derive specific versions of our ALEs for three particular learning algorithms in stochastic games. However, the ODEs posed by the authors are derived heuristically, assuming each algorithm utilizes batch learning, scaling time, and taking the batch size to infinity.

Strategy Learning

The limitations of action learning has been well recognized in the literature. Erev and Roth (1998) note that learning behavior generally cannot be analyzed in terms of actions alone, while Camerer and Ho (1999) points out that actions are not always the most natural candidates for the strategies that players learn about. Consequently, there have been several approaches in the literature that tries to address this.

A strand of learning analyzes learning a repeated game strategy over a finite set of repeated game strategies through evolutionary dynamics in a population game. These learning models, assume that a repeated game is repeated infinite times, where each repetition of a repeated game allows one to evaluate the average payoffs when matching the repeated game strategies from the populations. In this setting, Nowak et al. (2004) studies the replicator dynamics, Imhof et al. (2005) studies the replicator-mutation dynamics, and Fudenberg and Imhof (2006, 2008) studies the imitation dynamics. In contrast, in our learning framework, players learn a m-memory strategy in the space of all m-memory strategies for a fixed value of m, and learning takes place over a single repeated game in which players are infinitely-lived.

Ioannou and Romero (2014) proposes an approach to study repeated game strategies through belief learning. They restrict the space of repeated game strategies to one where players' strategies are implemented by a Moore machine. They use a fitness function to evaluate how well each candidate strategy fits the observed action profile sequence. Finally, a player's strategy set is updated asynchronously at the completion of block of periods. In contrast, our learning framework follows action learning models where players' strategies are updated synchronously at the end of each period.

Fundamentally, our learning framework is underpinned by stochastic approximation techniques. Our learning framework builds upon the work on Ma et al. (1990) to provide simple sufficient conditions to apply the ODE method from stochastic approximation to study learning algorithms with bounded memory strategies in repeated games. Our work further simplifies the sufficient conditions so that we only need to verify basic characteristics of the learning model, and basic characteristics of the monitoring structure. Relatedly, Perkins and Leslie (2012) provide sufficient conditions to apply stochastic approximation with differential inclusions.

State-Dependent Smooth Fictitious Play Our learning model is bounded in rationality because it assumes the future evolution of play evolves according to the belief over the stationary m-memory strategy profile, and following fictitious play, the belief is determined by the empirical frequency of past play. Our learning model resembles conditional smooth fictitious play (see Fudenberg and Levine, 1999), where outcomes are classified into categories, and for each category, players play a myopic smoothed best response to the historical frequency of opponent's play in that category. In our setting, the categories correspond to all combinations of m-memory histories so that the historical frequency of play in all categories defines the belief over the m-memory strategy profile of the players. Players then

use the belief to play a smoothed best response with respect to the action values induced by the belief.

Closer to our learning model is the seminal work by Kalai and Lehrer (1993) who demonstrate that rational (Bayesian) learning based on past play will converge to a Nash equilibrium of the repeated game provided that the initial beliefs satisfy a "grain-of-truth" assumption, i.e., the initial beliefs must be compatible with the eventual play. The lack of subgame perfection means that the best one can achieve with their learning model is a Nash Folk theorem. In contrast, our subgame perfect Folk theorem encodes the reward-punishment mechanism to show that there is collusion. Additionally, our work does not require the grain-of-truth assumption, which is known to be restrictive (see for example Nachbar, 1997; Foster and Young, 2001). Finally, consistent with learning algorithms, players in our learning model are not fully rational.

Foster and Young (2003) propose a learning rule where players build and test hypotheses of other players' strategies, and occasionally deviate from the best response strategies. Their learning model selects a subgame perfect equilibrium of the repeated game and does not require the grain-of-truth assumption. Our approach is different because we restrict our learning model to one that is compatible with current machine learning models. Specifically, the learning rule is a deterministic function that takes in stochastic inputs, so it can be implemented as a simple algorithm. This property also allows us to use stochastic approximation techniques to characterize the evolution of the bounded memory strategies as a system of ODEs. The ODEs allow us to describe the type of bounded memory strategies learned, which also makes it possible to characterize the payoffs achieved in the continuation game once the learning model converges. Moreover, the ODEs also allow us to address equilibrium selection.

Recently, Jindani (2022) builds upon the work of Foster and Young (2003) and proposes a learning rule that selects efficient subgame perfect equilibria of the repeated game. Jindani's learning model is designed to select efficient equilibria. In contrast, our learning model is not designed to select a particular equilibrium. However, in our setting, players can influence the outcome by encoding a strong initial belief. If they coordinate on the initial belief, then they can drive the learning process towards a particular equilibrium (collusive or otherwise). Alternatively, if players do not try to influence the outcome through the initial belief, then the nature of the collusion learned is best characterized as inadvertent tacit collusion because players in our learning model can inadvertently learn a collusive equilibrium even though they do not set out to learn to collude. The characterization of whether learning a collusive equilibrium is inadvertent depends on whether or not players try to influence the outcome through the initial belief.

By design, our learning model is similar to the decentralized learning algorithms proposed by computer scientists to achieve convergence in stochastic games (e.g., Sayin et al., 2021, 2022; Leonardos et al., 2022; Leslie et al., 2020; Baudin and Laraki, 2022a,b). The learning model of Maheshwari et al. (2023) can be seen as a model-free version of our learning model where the action values are estimated through the path of play with a Bellman equation. Beyond the similarity in the learning model, their approach is different from ours. They use the two-timescale differential inclusion approach from Perkins and Leslie (2012) to prove convergence to the ϵ-Markov perfect equilibrium that is the global maximizer of the perturbed potential function in a Markov potential game. In doing so, they implicitly assume the existence of a unique equilibrium, so their analysis does not address our question of interest. In contrast, our key contribution is the analysis of which m-memory subgame perfect equilibrium of the repeated game can be learned in the presence of multiple equilibria. To do this, we show that the state-dependent smoothed best response dynamics has finitely many rest points (which we show by proving a purification-type result), and we characterize the local stability of relevant rest points. This analysis is necessary to obtain a Folk theorem from learning, and to show that algorithms can learn to collude.

2.2 Collusion

Folk Theorems

In a static one-shot non-cooperative game, the Pareto outcome is not always an equilibrium because there is no way to ensure that the opponent will not undercut you. However, through repeated interactions, we can devise a strategy to ensure that the Pareto outcome is achieved as an equilibrium outcome. The classical paper by Stigler (1964) introduced the notion of self-policing to enforce monopolistic conduct, and has informed our understanding of how a collusive outcome can be achieved. Consequently, the standard approach to study collusion is through repeated games, and the main result is the Folk theorem.

Since then, various Folk theorems have been established for different contingent strategies with different information sets. That is, a strategy that maps a sequence of past signals to an action. There are three main types of information sets: perfect monitoring, public monitoring, and private monitoring. In perfect monitoring, the strategies of players are contingent on the full or partial history of past play (see Abreu, 1988). This setting is often used to study Bertrand oligopolies where players can monitor the prices. In public monitoring, the strategies of players are contingent on a perfect or possibly noisy public signal and the latter is often referred to as imperfect public monitoring (see Green and Porter, 1984). This setting is often used to study Cournot oligopolies where players observe

a common market price. In private monitoring, the strategies of players are contingent on a noisy private signal, which is often referred to as imperfect private monitoring (see Kandori and Matsushima, 1998).

Folk theorems have been established in a plethora of settings. In repeated games with perfect monitoring (e.g., Fudenberg and Maskin, 1986; Abreu, 1988), imperfect public monitoring (e.g., Abreu et al., 1990; Fudenberg et al., 1994), private monitoring (e.g., Sugaya, 2021), with bounded memory strategies (e.g., Barlo et al., 2009, 2016), overlapping generations (e.g., Kandori, 1992; Ellison, 1994; Clark et al., 2021), or when players aim to minimize the complexity of their strategies (e.g., Abreu and Rubinstein, 1988; Piccione, 1992); see Mailath and Samuelson (2006) for a comprehensive overview of various Folk theorems. These studies focus on when cooperative behavior can be sustained by a self-enforcing agreement. Our focus is similar, but we consider less than fully rational players that learn the strategies behind the Folk theorem.

Our work is also related to the strand of literature that studies Folk theorems in repeated games with unknown payoff distributions. In this literature, players learn the state of the world which corresponds to a different payoff matrix of a stage game chosen by Nature at the start of the repeated game (the state is fixed throughout the game and it is not observable to players, see Wiseman, 2005, 2012; Sugaya and Yamamoto, 2020). In contrast, we focus on learning bounded memory strategies to play the repeated game.

Tacit Collusion

The distinguishing factor between explicit collusion and tacit collusion is the communication mechanism to suppress rivalry. In explicit collusion, there is an agreement among players that relies on communication between players. In tacit collusion, a collusive agreement is achieved without communication. In antitrust law, explicit collusion is illegal, while tacit collusion is not.

There are two broad problems to solve to achieve a collusive arrangement: how to initiate the collusive arrangement and how to implement that arrangement. The problem of initiating collusion involves coming to agreement on what the collusive structures required to deter secret deviations will be. The problem of implementation involves managing the ongoing operation of the collusive arrangement, including the implementation of the collusive structures (see Green et al., 2014). The economics literature focuses only on the implementation stage of a collusive arrangement, hence, Folk theorems have nothing to say about initiating a collusive arrangement.

Our work provides a resolution to how players initiate a collusive arrangement without communication. Specifically, our results show that a collusive arrangement can be initiated through learning without communication.

Algorithmic Collusion

There are three broad fields of algorithmic collusion. The first strand studies action learning with a focus on learning non-equilibrium cooperative (Pareto dominant) outcomes. The second strand studies strategy learning to understand if algorithms can learn collusive equilibria. The final strand studies algorithmic pricing with limited commitment, where pricing algorithms are periodically revised.

Cooperative Outcomes The fact that action learning can lead to cooperative outcomes is not new. Indeed, Karandikar et al. (1998) and Cho and Matsui (2005) both prove that action learning can learn non-equilibrium cooperative outcomes. Recent interest from algorithmic collusion has lead researchers to further understand if these outcomes occur for other algorithms, and to further understand mechanisms that lead to cooperative outcomes.

Waltman and Kaymak (2008) study Q-learning in a Cournot oligopoly and find that action learning leads to supracompetitive prices. Hansen et al. (2021) prove that UCB-type algorithms will learn to play the cooperative action for symmetric 2×2 games when there is no stochasticity in the rewards received.

Asker et al. (2022) show the importance of the information that the algorithms leverage when they learn. They show that by conducting counterfactuals to assist learning, algorithms learn to price competitively, whereas without counterfactuals, the algorithms learn to price supracompetitive prices. Banchio and Skrzypacz (2022) show the importance of auction design and the type of outcomes the algorithms learn. They show that first-price auctions with no additional feedback lead to supracompetitive outcomes, while second-price auctions do not. Finally, Colliard et al. (2022) study Q-learning in a repeated Bertrand oligopoly, where the stage game is underpinned by a Glosten–Milgrom setup (see Glosten and Milgrom, 1985). They find that an increase in the variance of the payoffs leads to less competitive outcomes.

Collusive Strategies Action learning cannot lead to collusion where supracompetitive prices are sustained with a reward-punishment mechanism. However, an interesting question is if algorithms can learn collusive strategies.

Calvano et al. (2020) studies Q-learning in a repeated Bertrand oligopoly with differentiated goods, while Calvano et al. (2021) studies Q-learning in a repeated Cournot oligopoly

with stochastic demand. In both cases, they find that Q-learning can learn strategies to sustain supracompetitive outcomes through a behavior that resembles a reward-punishment mechanism. They check this with an impulse response to see how the algorithms behave after a deviation.

Klein (2021) studies Q-learning in a repeated sequential move pricing duopoly environment of Maskin and Tirole (1988). Using an impulse response, he also finds that Q-learning can learn strategies to sustain supracompetitive outcomes through a behavior that resembles a reward-punishment mechanism. Asker et al. (2023) extends Asker et al. (2022) to one-memory strategy learning. They find that learning with and without counterfactuals both lead to supracompetitive prices. However, the case with counterfactuals leads to significantly lower prices than in the case without counterfactuals. Finally, Dou et al. (2024) studies a Cournot oligopoly underpinned by the environment of Kyle (1985).

Recently, several researchers have questioned the viability of using an impulse response to check for collusion. Lambin (2023) demonstrates that high prices and the apparent punishment schemes result directly from simultaneous experimentation. Abada et al. (2024) argue that seemingly-collusive outcomes arise from insufficient exploration during the learning process. They further show that allowing for more thorough exploration leads to more competitive outcomes. Further related to a lack of experimentation is the importance of where the action grid lies. Epivent and Lambin (2024) demonstrate that shifting the action space can lead to convergence below the competitive equilibrium. Additionally, price wars also occur in these cases when there is a deviation. Neither are equilibrium behavior.

Our work provides the theory behind algorithmic collusion by analyzing how players learn the strategies behind Folk theorems.

Adaptive Pricing Algorithms Algorithms are not limited to ones that learn to play the game. Algorithms can also encode a strategy, and firms change strategies periodically. This literature focuses on pricing algorithms with limited commitment, where prices are set more frequently than algorithms are revised.

Salcedo (2015) studies a dynamic model where firms commit to pricing algorithms in the short run, and over time, their algorithms are inferred or decoded by their competitors who can revise their algorithms in response. Salcedo shows that if firms compete with algorithms that are fixed in the short run and can be revised over time, then collusion is inevitable. Lamba and Zhuk (2023) build upon Salcedo's work and show that, in a simple repeated duopoly with two possible prices, monopoly pricing is the unique equilibrium

outcome. Levine (2024) generalizes these results and shows that, in a simple repeated game environment, observable commitments leads to play that is efficient.

Brown and MacKay (2023) study a model that allows for asymmetric technology among firms. They show that asymmetry in pricing technology shifts the equilibrium behavior. Specifically, if one firm adopts superior technology, then all firms can obtain higher prices. If all firms adopt automated high-frequency algorithms, then collusive prices can be supported without the use of traditional collusive strategies.

Finally, Cho and Williams (2024) study a model of algorithmic pricing where firms use pricing algorithms from a parameterized family of model specification. The firms update both the parameters and the weights on models to adapt endogenously to market outcomes. Their model shuts down every channel for explicit or implicit collusion, but they show that the market experiences recurrent episodes where both firms set prices at collusive levels.

Detecting Collusion

Methods for discovering cartels can be broadly divided into those which are structural and those which are behavioral. A structural approach identifies markets with traits thought to be conducive to collusion, while a behavioral approach involves (i) observing the way in which firms coordinate, or (ii) observing the end result of the coordination.

The process of detecting collusion can be broken into three stages (see Harrington, 2005): screening, verification, and prosecution. Screening identifies markets where collusion is suspected. Verification systematically tries to exclude competition as an explanation for observed behavior and to provide evidence in support of collusion. Finally, prosecution develops economic evidence to persuade the court or some other administrative body that there has been a violation of the law.

Generally speaking, screening methods often rely on models of economic behavior. These models are simple enough to either explicitly characterize equilibria or statistical facts about equilibrium behavior. The statistical features of the model are then reformulated into a statistical test, where the null hypothesis is competition and the empirical task is to accept or reject that hypothesis. Examples of this approach include Porter and Zona (1993, 1999), Chassang and Ortner (2019), Kawai and Nakabayashi (2022), Ortner et al. (2022), and Kawai et al. (2023). For a more comprehensive review, see Harrington (2005), Porter (2005), and Chassang and Ortner (2023).

Examples

Financial Markets The case study we examine in Chapter 6 is related to Christie and Schultz (1994) who document an absence of odd-eighth quotes for Nasdaq stocks and offer implicit collusion as the most likely explanation.[1] The case study in Cartea et al. (2025) documents a phenomenon where HFTs break the anonymity of limit orders by signaling themselves to each other and offer collusion as a possible explanation. Dutta and Madhavan (1997) provide a model of dealer markets and rationalize the results of Christie and Schultz (1994) through a collusive arrangement. Chapter 6 provides a model of the limit order book and rationalizes signaling through a collusive arrangement. Tangentially, Bryzgalova et al. (2025) present a model where arbitrageurs choose to specialize in some markets, which leads to the highest combined profits. They further present evidence consistent with their theory from the options market.

Signaling The case study we examine also shares similarities with the bid signaling that occurred in the Federal Communications Commission (FCC) spectrum auctions, where rivals used "code bidding" to send messages to their rivals to coordinate on which licenses to bid and which to avoid (see Cramton and Schwartz, 2000, 2001).[2] Our model shows that HFTs signal their limit orders with large volumes to distinguish their limit orders from the limit orders sent by patient investors. This enables the HFTs to share the profitable benign flow of patient investors, and as a byproduct, enables the HFTs to receive additional benign flow from impatient investors who would have otherwise matched with a patient investor's limit order.

2.3 Market Microstructure

Anonymity and Transparency

Fundamental to our result in Chapter 6 is that signaling breaks the anonymity of limit orders. This artificial form of pre-trade transparency creates a different playing field for a subset of market participants. Therefore, our work is related to the literature that analyzes the effect of transparency on market dynamics (see De Jong and Rindi, 2009). Transparency comes in various dimensions, from post-trade transparency (see e.g., Madhavan, 1995; Friederich and Payne, 2014; Meling, 2021) to pre-trade quotation transparency (see e.g., Biais, 1993;

[1] This coordinated behavior was later identified as explicitly collusive and successfully litigated, which resulted in a collective $1.027 billion fine imposed on those who participated.

[2] See also Klemperer (2002) for other examples of signaling in auction markets.

Madhavan et al., 2005). Our work is closer to the literature on the pre-trade anonymity of limit orders.

Simaan et al. (2003) argue that the anonymity of limit orders reduces the probability of collusion among quote setters because it limits their ability to monitor and punish deviations from cooperation. Empirically, they show that the spread of Nasdaq dealer quotes posted through anonymous electronic communication networks are tighter than those of Nasdaq dealer quotes posted through the transparent dealer quotation system. In our model, monitoring is imperfect because signaling is voluntary. However, this does not affect the existence of a collusive equilibrium because our problem is similar to one of "secret price cutting" (see Green and Porter, 1984).

Foucault et al. (2007) show that limit order anonymity can result in tighter bid-ask spreads because a non-anonymous environment can lead to free riding, to which market makers respond by quoting a wider spread. Our model has a similar flavor. When HFTs signal themselves to each other, they partially reveal their private information. A collusive equilibrium exists when the gains outweigh the costs from revealing yourself.

Comerton-Forde et al. (2005) and Comerton-Forde and Tang (2009) use natural experiments to study the impact of pre-trade anonymity and find that limit order anonymity improves liquidity through tighter spreads. On the other hand, Comerton-Forde et al. (2011) study how brokers strategically disclose their identity to reduce their execution costs.

The main difference between our work and all the aforementioned papers on transparency is that the degree of transparency is imposed by the design of the market. In contrast, our setting is an artificial form of pre-trade transparency that is achieved only if HFTs decide to break their anonymity.

High-Frequency Trading

Our model contributes to the theoretical literature on HFTs (see Menkveld, 2016, for a review).[3,4] Our model generalizes the two-period trading game in Budish et al. (2024), and it retains the main elements of latency arbitrage (see e.g., Budish et al., 2015) and adverse selection (see e.g., Baldauf and Mollner, 2020). More similar to Li et al. (2021), our model introduces investors who send limit orders inside the quoted spread, but the main difference

[3] See for example Biais et al. (2011), Cartea and Penalva (2012), Aït-Sahalia and Saglam (2013), Foucault et al. (2013), Hoffmann (2014), Jovanovic and Menkveld (2016), Menkveld and Zoican (2017), Foucault et al. (2017), Bernales (2017).

[4] See for example Bergault et al. (2022), Drissi (2022, 2023), Cartea et al. (2015, 2023b,c,d,e, 2024a,b); Bergault and Sánchez-Betancourt (2025); Cartea et al. (2022e); Cartea and Sánchez-Betancourt (2023, 2025), Aqsha et al. (2024) for the type of algorithms used in high-frequency trading.

is that by allowing HFTs to be informed, we introduce an element of ambiguity as to *who* sent the limit order inside the quoted spread.

The empirical literature on HFTs is diverse. From understanding how HFTs affect market quality (see e.g., Brogaard, 2010; Hendershott et al., 2011; Hagströmer and Nordén, 2013; Brogaard et al., 2015; Brogaard and Garriott, 2019) to how HFTs increase price discovery (see e.g., Brogaard et al., 2014, 2019). Our work is closer to the literature that documents the behavior of HFTs (see e.g., Hendershott and Riordan, 2009; Hagströmer et al., 2014; Kirilenko et al., 2017; Aquilina et al., 2021). We rationalize a phenomenon where HFTs knowingly or unknowingly signal themselves to each other.

Our model also connects to the early literature on market microstructure that studies liquidity and asymmetric information (e.g., Glosten and Milgrom, 1985; Kyle, 1985; Glosten, 1994).[5] Similar to our work is Easley and O'Hara (1987), where the authors show that order flow signals the investor's type (i.e., informed or uninformed). Our model looks at the size of limit orders as a method to signal your type. In our model, the competitive equilibrium is a pooling equilibrium where one cannot differentiate between limit orders sent by informed HFTs or limit orders sent by patient investors, while the collusive equilibrium is a separating equilibrium where HFTs voluntarily distinguish themselves from patient investors.

[5]For more recent work on decentralized markets, see Capponi et al. (2025).

Chapter 3

Algorithmic Learning Equations

3.1 The Framework

3.1.1 Setup

Consider an infinitely repeated game $\mathcal{G}^\infty = \langle (\mathcal{A}_i)_{0<i\leq I}, (u^i)_{0<i\leq I}, \delta \rangle$ played by $I \geq 2$ players. Let $\mathcal{A}_i$ denote the finite set of actions for player i, let $\mathcal{A}_{-i} = \bigtimes_{j\neq i} \mathcal{A}_j$ denote the set of action profiles excluding player i, and let $\mathcal{A} = \bigtimes_{0<i\leq I} \mathcal{A}_i$ denote the set of action profiles. For simplicity and without loss of generality, we assume that $\mathcal{A}_i = \mathcal{A}_j$ for all $0 < i, j \leq I$. After an action profile $\mathbf{a} = (a^1, ..., a^I) \in \mathcal{A}$ is played at each period $n = 0, 1, 2, 3...$ of the repeated game, each player i receives a payoff according to a utility function $u^i : \mathcal{A} \to \mathbb{R}$, and $\delta \in [0, 1)$ is the common parameter used to discount the future stream of payoffs. At the end of each period, players observe a public signal y from the finite signal space Y, and each player observes an idiosyncratic signal y^i from the finite signal space Y_i.[1] The realization of the signals y and y^i depend on the action profile from that period.

The n-period history of public signals is a sequence of n public signals that identify the signal realized in periods 0 through $n-1$ given by $h_n^p = (y_0, y_1, ..., y_{n-1})$. The set of n-period public histories is given by $\mathcal{H}_n^p = Y^n$, where Y^n is the n-fold product of Y. The set of public histories is defined as $\mathcal{H}^p = \bigcup_{n\geq 0} \mathcal{H}_n^p$, where $\mathcal{H}_0^p = \{\emptyset\}$. The n-period history for player i includes both the public history and the history of idiosyncratic signals given by $h_n^i = (y_0, y_0^i; y_1, y_1^i; ...; y_{n-1}, y_{n-1}^i)$. The set of n-period histories for player i is given by $\mathcal{H}_n^i = (Y_i \times Y)^n$, so the set of histories for player i is $\mathcal{H}^i = \bigcup_{n\geq 0} \mathcal{H}_n^i$, where $\mathcal{H}_0^i = \{\emptyset\}$.

To analyze learning bounded m-memory strategies, it is useful to focus on continuation games from periods $n \geq m$ onwards. For periods $n \geq m$, a stationary m-memory strategy for player i is a mapping $\sigma^i : \mathcal{H}_m^i \to \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ is the set of probability measures

[1] Here, signals are not to be confused with signals from a Bayesian framework; rather, signals refer to observable variables following the terminology of Mailath and Samuelson (2006).

on $\mathcal{A}_i$. We define the state of the game $\mathbf{s} = (s^1, ..., s^I, s) \in \mathcal{S} := \mathcal{S}_1 \times ... \times \mathcal{S}_I \times \tilde{\mathcal{S}}$ as all the information sets required by the players' strategies, i.e., the information players condition their action on. For example, if all players use m-memory strategies, then $\tilde{\mathcal{S}} = Y^m$ and $\mathcal{S}_i = Y_i^m$ for all i. Let $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ denote the transition function that describes the transition between the information sets of the players' strategies based on their actions. Specifically, $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ is the probability that the subsequent state is $\mathbf{s}'$ given that the current state is $\mathbf{s}$ and the current action profile is $\mathbf{a}$.

Example 1 (m-memory perfect monitoring)**.** Consider players who use m-memory strategies that depends only on the public component h_n^p of their history $\mathcal{H}^i$. Further, if the public signal Y is the action profile $\mathcal{A}$, then the set of states $\mathcal{S} = \mathcal{A}^m$ is such that each state of the game is the action profile of the previous m stage games. The transition function is such that we have a sliding window of the most recent m-memory history by dropping the action profile from the oldest stage game and appending the new action profile. That is, $\mathbf{s}_n = (\mathbf{a}_{n-m}, \mathbf{a}_{n-m+1}, ..., \mathbf{a}_{n-2}, \mathbf{a}_{n-1})$ transitions to $\mathbf{s}_{n+1} = (\mathbf{a}_{n-m+1}, \mathbf{a}_{n-m+2}, ..., \mathbf{a}_{n-1}, \mathbf{a}_n)$ with probability one if $\mathbf{a}_n$ is played at period n.

This setup allows us to study learning algorithms with behavioral strategies that condition on the (recent) history. Specifically, we have an automaton-like representation of the space of bounded memory strategies used by the players so that we can analyze the evolution of their strategies through learning.

3.1.2 Learning Model

We consider a model of learning where players use algorithms to learn bounded memory strategies through repeated interactions. In our model, each player i has an algorithm that tracks a tuple of parameters $\boldsymbol{\theta}^i = \left(\theta^i(a \mid s^i, s)\right)_{a \in \mathcal{A}_i, s^i \in \mathcal{S}^i, s \in \tilde{\mathcal{S}}}$ through time as it learns. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}^i)_{i \leq I} \in G$ denote the parameter profile of the players, where $G \subset \mathbb{R}^K$ is bounded with $K \in \mathbb{N}$. Along with the parameters, each player has a choice rule that maps the tuple of parameters $\boldsymbol{\theta}^i$ to a bounded memory strategy σ_θ^i. Let $\sigma_\theta = (\sigma_\theta^i)_{i \leq I}$ denote the bounded memory strategy profile parameterized by $\boldsymbol{\theta}$.

At each period n, players play an action based on the state of the game $\mathbf{s}_n$, with a strategy profile σ_{θ_n} that is parameterized by the parameter profile $\boldsymbol{\theta}_n$. Based on the action profile $\mathbf{a}_n$, players receive new public and idiosyncratic signals y_n and y_n^i for all i, which leads to a new state $\mathbf{s}_{n+1}$ given by the transition function $p(\mathbf{s}_{n+1} \mid \mathbf{s}_n, \mathbf{a}_n)$. At the end of each period n, each player updates the parameters of their algorithms according to

$$\theta_{n+1}^i(a \mid s^i, s) = \theta_n^i(a \mid s^i, s) + \gamma_{n+1} f_{a \mid s^i, s}^i(\boldsymbol{\theta}_n, \mathbf{s}_n, \mathbf{a}_n, \mathbf{s}_{n+1}), \tag{ALG}$$

for all i, where $f^i_{a|s^i,s} : G \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the learning rule and $\gamma_n \in \mathbb{R}^+$ is the learning rate. The above is repeated ad infinitum.

Our analysis ignores the initial periods $n < m$ before we have a "valid" state of the game $\mathbf{s}_n$. For these periods, we assume each algorithm has a method to prescribe actions. For example, they can sample random actions until they have a valid state to condition their action on.

Our analysis centers on periods $n \geq m$, where the action of each player depends on the recent history, and the current action affects the history which future actions depend on. The difficulty here is that the transition dynamics of the states $\mathbf{s}$ are not stationary and change from period to period because the strategy profile σ_θ changes at each period as the parameter profile $\boldsymbol{\theta}$ changes according to (ALG). In effect, the state process is a controlled Markov chain where the control is changing through time.

3.1.3 Algorithmic Learning Equations

We use stochastic approximation techniques to analyze the evolution of (ALG). We show that when the value of the learning rate γ_n is small, the trajectories of (ALG) are approximated by solutions $\bar{\boldsymbol{\theta}}$ to the ODE

$$\dot{\bar{\boldsymbol{\theta}}} = F(\bar{\boldsymbol{\theta}}),$$

for an appropriately defined deterministic function F, where the dot denotes the time derivative. The usual choice of F is the expected value of the learning rule f so that F is the expected evolution of the learning rule. Intuitively, if we treat the learning rate γ_n as a time step, then as the value of γ_n decreases, the number of realizations of the stage game increases per unit time, and hence the number of updates to the parameters increases per unit time. Furthermore, if the value of the learning rate is small, the updates of the parameters will also be small; in which case, by the law of large numbers, each stochastic update is close to its theoretical mean.

In our setting, taking the expected value of the learning rule f is not straightforward. Here, the evolution of the learning rule f depends explicitly on the state of the game $\mathbf{s}_n$ because the action choice depends on the state of the game $\mathbf{s}_n$. Therefore, computing the expectation of (ALG) also requires us to compute the expectation with respect to $\mathbf{s}_n$. To take this expectation, we need $\mathbb{P}(\mathbf{s}_n = \mathbf{s})$, which is difficult to compute because the state process is a controlled Markov chain where the control is changing through time.

To overcome this difficulty, we approximate $\mathbb{P}(\mathbf{s}_n = \mathbf{s})$ with $\Gamma_{\sigma_{\theta_n}}(\mathbf{s})$, which represents the long-run average frequency of visiting state $\mathbf{s}$ assuming all players use a fixed strategy profile parameterized by a fixed parameter profile $\boldsymbol{\theta}_n$. The idea is that if this approximation

does not introduce 'too much' error, then the trajectories of the system of ODEs driven by $\Gamma_{\sigma_{\theta_n}}(\mathbf{s})$ will not deviate far from the system driven by $\mathbb{P}(\mathbf{s}_n = \mathbf{s})$. Hence, we retain the ability to approximate trajectories of (ALG).

We derive F as follows. First, take the conditional expectation of the learning rule f with respect to all information realized by period n to obtain a function $\bar{f} = \bar{f}(\boldsymbol{\theta}_n, \mathbf{s}_n)$. Next, to avoid averaging with respect to $\mathbb{P}(\mathbf{s}_n = \mathbf{s})$, we exploit the fact that for fixed $\boldsymbol{\theta}$, the state process is a Markov chain, which, under certain conditions, has a unique stationary distribution denoted by Γ_{σ_θ}. Finally, take the expectation of $\bar{f}$ with respect to Γ_{σ_θ} to obtain $F = F(\boldsymbol{\theta}_n)$.

Formally, we define the filtration $\mathcal{F}_n$ as the σ-algebra generated by $\{\mathbf{s}_k, \mathbf{a}_{k-1}, \boldsymbol{\theta}_k : k \leq n\}$, which describes the aggregate information accumulated throughout the game up to, and including, period n. We stress that no player has access to all of the information in this filtration, as it describes the evolution of the *entire* system. Define $\bar{f}^i_{a|s^i,s} : \mathbb{R}^K \times \mathcal{S} \to \mathbb{R}$ as

$$\begin{aligned} \bar{f}^i_{a|s^i,s}(\boldsymbol{\theta}_n, \mathbf{s}_n) &:= \mathbb{E}\left[f^i_{a|s^i,s}(\boldsymbol{\theta}_n, \mathbf{s}_n, \mathbf{a}_n, \mathbf{s}_{n+1}) \,\middle|\, \mathcal{F}_n \right] \\ &= \sum_{\mathbf{s}', \mathbf{a}} p(\mathbf{s}' \mid \mathbf{s}_n, \mathbf{a}) \prod_{j=1}^{I} \sigma^j_{\theta_n}(a^j \mid s^j_n, s_n)\, f^i_{a|s^i,s}(\boldsymbol{\theta}_n, \mathbf{s}_n, \mathbf{a}, \mathbf{s}') \,. \end{aligned} \tag{3.1}$$

Second, fix $\boldsymbol{\theta} \in G$, and consider the hypothetical evolution of the game when players use a fixed strategy profile σ_θ. Define $\mathbf{s}^{\sigma_\theta}$ as the corresponding sequence of states of the game. The process $\mathbf{s}^{\sigma_\theta}$ is a Markov chain with explicit transition dynamics $P_{\sigma_\theta}(\cdot \mid \cdot)$ given by

$$P_{\sigma_\theta}(\mathbf{s}' \mid \mathbf{s}) := \mathbb{P}(\mathbf{s}^{\sigma_\theta}_{n+1} = \mathbf{s}' \mid \mathbf{s}^{\sigma_\theta}_n = \mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) \prod_{j=1}^{I} \sigma^j_\theta(a^j \mid s^j, s) \,.$$

If the process $\mathbf{s}^{\sigma_\theta}$ is an aperiodic Markov chain with a single recurrent class, then there exists a stationary distribution $\Gamma_{\sigma_\theta} \in \Delta(\mathcal{S})$, where $\Gamma_{\sigma_\theta}(\mathbf{s})$ represents the long-run average frequency of visiting state $\mathbf{s}$ assuming all players use a fixed strategy profile parameterized by a fixed parameter profile $\boldsymbol{\theta}$.

Finally, take the expectation of $\bar{f}$ with respect to $\Gamma_{\sigma_{\theta_n}}$ to obtain the deterministic mapping

$$\begin{aligned} F^i_{a|s^i,s}(\boldsymbol{\theta}_n) &= \sum_{\mathbf{s}} \Gamma_{\sigma_{\theta_n}}(\mathbf{s})\, \bar{f}^i_{a|s^i,s}(\boldsymbol{\theta}_n, \mathbf{s}) \\ &= \sum_{\mathbf{s}} \Gamma_{\sigma_{\theta_n}}(\mathbf{s}) \sum_{\mathbf{s}', \mathbf{a}} p(\mathbf{s}' \mid \mathbf{s}_n, \mathbf{a}) \prod_{j=1}^{I} \sigma^j_{\theta_n}(a^j \mid s^j_n, s_n)\, f^i_{a|s^i,s}(\boldsymbol{\theta}_n, \mathbf{s}_n, \mathbf{a}, \mathbf{s}') \,. \end{aligned} \tag{3.2}$$

The indices ${}^{i}_{a|s^i,s}$ denote the components of the mappings. To streamline the notation, write

$$f^i = (f^i_{a|s^i,s})_{a \in \mathcal{A}^i, s^i \in \mathcal{S}^i, s \in \tilde{\mathcal{S}}}, \quad \bar{f}^i = (\bar{f}^i_{a|s^i,s})_{a \in \mathcal{A}^i, s^i \in \mathcal{S}^i, s \in \tilde{\mathcal{S}}}, \quad F^i = (F^i_{a|s^i,s})_{a \in \mathcal{A}^i, s^i \in \mathcal{S}^i, s \in \tilde{\mathcal{S}}},$$

$$f = (f^1, \ldots, f^I), \qquad \bar{f} = (\bar{f}^1, \ldots, \bar{f}^I), \qquad F = (F^1, \ldots, F^I).$$

Thus, the *algorithmic learning equations* are a deterministic system of ODEs defined as

$$\dot{\bar{\boldsymbol{\theta}}}(t) = \sum_{\mathbf{s}} \Gamma_{\sigma_{\bar{\theta}(t)}}(\mathbf{s}) \sum_{\mathbf{s}',\mathbf{a}} p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) \prod_{j=1}^{I} \sigma^j_{\bar{\theta}(t)}(a^j \mid s^j, s)\, f\left(\bar{\boldsymbol{\theta}}(t), \mathbf{s}, \mathbf{a}, \mathbf{s}'\right). \tag{ALE}$$

In what follows, we show that solutions to (ALE) approximate the trajectories of (ALG) under appropriate conditions.

3.2 Main Results

We present sufficient conditions that ensure that the trajectories of (ALE) approximate the trajectories of (ALG) in finite time. We also show that when we sharpen these conditions (in particular Assumption A.1), we obtain stronger convergence of the long-run behavior of (ALG).[2]

Assumption A.

A.1 *The learning rate* $(\gamma_n)_{n\in\mathbb{N}}$ *is non-increasing and positive, with* $\sum_n \gamma_n = \infty$.

A.2 $\boldsymbol{\theta}_n \in G \,\forall n$ *a.s., where* G *is a compact convex subset of a Euclidean space, i.e.,* $\sup_n |\boldsymbol{\theta}_n| < \infty$ *a.s.*.

A.3 *For each* $\boldsymbol{\theta} \in G$, $\mathbf{s}^{\sigma_\theta}$ *is an aperiodic Markov chain with a single recurrent class.*

A.4 *For each* $0 < i \leq I$, *the bounded memory strategy* σ^i_θ *is Lipschitz in* $\boldsymbol{\theta}$ *for all* $\boldsymbol{\theta} \in G$.

A.5 *For each* $0 < i \leq I$, $\mathbf{s} \in \mathcal{S}$, $\mathbf{a} \in \mathcal{A}$, *and* $\mathbf{s}' \in \mathcal{S}$, *the learning rule* $f^i(\boldsymbol{\theta}, \mathbf{s}, \mathbf{a}, \mathbf{s}')$ *is Lipschitz in* $\boldsymbol{\theta}$ *for all* $\boldsymbol{\theta} \in G$.

First, the proposition below shows that (ALE) has a unique solution.

Proposition 1. *Assume A.2, A.3, A.4, A.5 hold. Then, for any* $\boldsymbol{\theta}_0 \in G$, (ALE) *has a unique, global solution* $\left(\bar{\boldsymbol{\theta}}(t;\boldsymbol{\theta}_0)\right)_{t\in\mathbb{R}^+}$, *such that* $\bar{\boldsymbol{\theta}}(0;\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0$.

Proposition 1 is an immediate consequence of Lemma 5 in Appendix A. The dependence of $\bar{\boldsymbol{\theta}}(t;\boldsymbol{\theta}_0)$ on its initial condition is crucial. For simplicity, we write $\bar{\boldsymbol{\theta}}(t) = \bar{\boldsymbol{\theta}}(t;\boldsymbol{\theta}_0)$ when there is no chance of confusion.

Intuitive interpretations of Assumption A are as follows. In Assumption A.1, the learning rate may be constant or it may decrease, and the latter part of Assumption A.1

[2]Henceforth, when dealing with random objects, a.s. is implied as needed.

places a lower bound on the rate of decrease. On the other hand, later in Theorem 2, we obtain stronger convergence results when we place an upper bound on the rate of decrease. The condition $\sum_n \gamma_n = \infty$ is crucial for both Theorems 1 and 2 because it ensures that the algorithm will always continue to learn.

Assumption A.2 is a technical condition required in the proofs.[3] Assumption A.3 is a condition on the dynamics of the information that players receive as signals from the game. This assumption depends on the strategies used by the players because their strategies determine which signals form part of the state of the game $\mathbf{s}$ and their strategies also determine the evolution of the state. Assumption A.3 also ensures that for all possible parameters $\boldsymbol{\theta}$, the state of the game has a well-defined stationary distribution Γ_{σ_θ}, that is, the long-run average frequency of observed states is non-degenerate.

Assumptions A.4 and A.5 are conditions on the learning algorithm used by each player. These assumptions characterize the type of learning algorithms for which the approximation hold. Furthermore, these assumptions only need to hold for each player's learning algorithm. Hence, our framework allows each player to use a different learning algorithm.[4]

Assumption A.3 when paired with Assumptions A.4 and A.5 ensure that the evolution of Γ_{σ_θ} is well-behaved and does not introduce too much error when $\boldsymbol{\theta}$ evolves according to (ALG). Hence, we can bound and control the errors between (ALE) and (ALG). Intuitively, these conditions ensure that the evolution of Γ_{σ_θ} does not change drastically as $\boldsymbol{\theta}$ evolves according to (ALG), so as the value of γ_n decreases, $\Gamma_{\sigma_{\theta_n}}$ becomes an accurate approximation of $\mathbb{P}(\mathbf{s}_n = \mathbf{s})$.

3.2.1 Finite Time

To compare the trajectory of the discrete-time stochastic process $\boldsymbol{\theta}_n$ with the trajectory of the deterministic, continuous-time system of ODEs in (ALE), both trajectories must be set to a common time scale. To this end, we set $t_n = \sum_{i=1}^{n} \gamma_i$ with $t_0 = 0$. Furthermore, to compare the trajectories $\boldsymbol{\theta}_n$ and $\bar{\boldsymbol{\theta}}(t; \boldsymbol{\theta}_{t_n})$ between times t_n and $t_n + T$, we study $\boldsymbol{\theta}_k$ for integers k between n and $m(n,T)$, where $m(n,T) := \max\{k > n : t_n + T \geq t_k\}$ for all $n \in \mathbb{N}$. For simplicity, we write $m(T) = m(0,T)$.

Our first result shows that for any fixed and finite $T > 0$, solutions to (ALE) approximate $\boldsymbol{\theta}_n$ on $[0,T]$ provided the values of the learning rate γ_n are small enough.

[3] This assumption is not restrictive when the utility functions are bounded.

[4] If Assumption A.4 fails to hold, then the results in Perkins and Leslie (2012) may help to address this.

Theorem 1. *Let Assumption A hold. Then there exists $C(T,\gamma_1) > 0$ such that for all $T > 0$ and $\delta > 0$, we have*

$$\mathbb{P}\left(\sup_{n \leq m(T)} \left| \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}(t_n) \right| \geq \delta \right) \leq \frac{C(T,\gamma_1)}{\delta^2}$$

with the property that $C(T,\gamma_1) \to 0$ as $\gamma_1 \to 0$.

Theorem 1 shows that for any fixed and finite time $T > 0$, the behavior of $\boldsymbol{\theta}_n$ up to $m(T)$ is characterized by the trajectories of $\bar{\boldsymbol{\theta}}(t)$ for $t \in [0,T]$ with high probability for a small enough value of γ_1. This result allows us to obtain uniform convergence in probability between the trajectories of (ALE) and the trajectories of the discrete-time learning algorithms governed by (ALG) for any fixed and finite time horizon $T > 0$. However, this result does not guarantee much about $\boldsymbol{\theta}_n$ beyond $m(T)$, i.e., little can be said about the convergence or stability of the long-term behavior of $\boldsymbol{\theta}_n$.

3.2.2 Asymptotics

In Theorem 2, we show that if we sharpen Assumption A.1, then the trajectories of (ALE) characterize the behavior of (ALG) beyond T, and we show that $\boldsymbol{\theta}_n$ can converge to an asymptotically stable rest point $\boldsymbol{\theta}^*$ of (ALE) as $n \to \infty$.[5]

Before characterizing the long-term behavior of (ALG) with solutions of (ALE), we introduce a suitable notion of asymptotic approximation. We extend the discrete-time processes $\theta^i_n(a \,|\, s^i, s)$ to continuous-time processes $\hat{\theta}^i_t(a \,|\, s^i, s)$ for $t \in \mathbb{R}^+$ by interpolating the trajectories of $\theta^i_n(a \,|\, s^i, s)$ to define

$$\hat{\theta}^i_{t_n + r}(a \,|\, s^i, s) = \theta_n + r \, \frac{\theta^i_{n+1}(a \,|\, s^i, s) - \theta^i_n(a \,|\, s^i, s)}{t_{n+1} - t_n},$$

for all $n \in \mathbb{N}$ and $0 \leq r < \gamma_{n+1}$. We refer to $\hat{\boldsymbol{\theta}}_t$ as the *real-time interpolation* of $\boldsymbol{\theta}_n$.

Definition 1 (Asymptotic Pseudo-trajectory)**.** *Let $\tilde{\boldsymbol{\theta}} : \mathbb{R}^+ \to \mathbb{R}^K$ be a continuous function. If*

$$\lim_{t \to \infty} \sup_{0 \leq h \leq T} \left| \tilde{\boldsymbol{\theta}}(t+h) - \bar{\boldsymbol{\theta}}\left(h; \tilde{\boldsymbol{\theta}}(t)\right) \right| = 0, \tag{APT}$$

*for any finite $T > 0$, then $\tilde{\boldsymbol{\theta}}$ is said to be an **asymptotic pseudo-trajectory** of $\bar{\boldsymbol{\theta}}$. If $\tilde{\boldsymbol{\theta}}$ is a stochastic process and* (APT) *holds with probability one, then $\tilde{\boldsymbol{\theta}}$ is an asymptotic pseudo-trajectory of $\bar{\boldsymbol{\theta}}$.*

[5] Recall that a rest point $\boldsymbol{\theta}^*$ is (Lyapunov) stable if for every neighbourhood U of $\boldsymbol{\theta}^*$, there exists a neighbourhood V of $\boldsymbol{\theta}^*$ such that if $\bar{\boldsymbol{\theta}}(t) \in V$ then $\bar{\boldsymbol{\theta}}(t) \in U$ for all $t \geq 0$. A rest point $\boldsymbol{\theta}^*$ is (Lyapunov) asymptotically stable if it is stable and has a neighbourhood V such that $\bar{\boldsymbol{\theta}}(t) \to \boldsymbol{\theta}^*$ for $t \to \infty$.

With a precise notion of the approximation, Theorem 2 shows the real-time interpolation $\hat{\boldsymbol{\theta}}$ of (ALG) is an asymptotic pseudo-trajectory of solutions $\bar{\boldsymbol{\theta}}$ to (ALE). Intuitively, for $\boldsymbol{\theta}$ to be an asymptotic pseudo-trajectory of $\bar{\boldsymbol{\theta}}$, it must be the case that for fixed and finite $T > 0$ and for $n \in \mathbb{N}$, the solution of (ALE) originating at $\boldsymbol{\theta}_n$ approximates the evolution of (ALG) up to $\boldsymbol{\theta}_{m(n,T)}$. As the value of n grows, the approximation sharpens, until in the limit $n \to \infty$ the solutions to (ALE) approximate $\boldsymbol{\theta}_n$ uniformly. We also note that the limiting behavior of an asymptotic pseudo-trajectory is related to the notion of chain recurrence, which describes limit points of a solution to a system of ODEs when subjected to small shocks occurring at isolated moments in time (see Benaïm, 1999; Hofbauer and Sandholm, 2002, for a more detailed discussion). Furthermore, under certain conditions, the result also shows that the discrete-time stochastic process $\boldsymbol{\theta}_n$ converges to an asymptotically stable rest point as $n \to \infty$.

Theorem 2. *Let Assumption A hold, and further assume that for some $q \geq 2$,*

$$\sum_n \gamma_n^{1+q/2} < \infty\,. \tag{DLR}$$

2.1 Then the real-time interpolated process $\hat{\boldsymbol{\theta}}$ is almost surely an asymptotic pseudo-trajectory of $\bar{\boldsymbol{\theta}}$.

2.2 Furthermore, let $\boldsymbol{\theta}^*$ be a locally (Lyapunov) asymptotic stable solution to (ALE) *with a domain of attraction $D(\boldsymbol{\theta}^*)$. If there is a compact set $A \subset D(\boldsymbol{\theta}^*)$ such that $\boldsymbol{\theta}_n \in A$ infinitely often, then $\lim_{n\to\infty} \boldsymbol{\theta}_n = \boldsymbol{\theta}^*$ a.s..*

When we further bound the rate of decrease of the value of the learning rate in (DLR), we can analyze the asymptotic behavior of algorithms. The second part of the theorem shows that learning converges under suitable conditions. However, for cases when learning does not converge, we can at least describe the dynamics of (ALG) through the notion of an asymptotic pseudo-trajectory from the first part of the theorem.

Proof Sketch 1. To prove Theorems 1 and 2, we analyze the errors from approximating the learning rule of the discrete-time stochastic process with the deterministic mapping F from (ALG). Rewrite (ALG) as

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \gamma_{n+1} F(\boldsymbol{\theta}_n) + \gamma_{n+1} \varepsilon_{n+1}\,, \tag{3.3}$$

where F is defined in (3.2) and the error term is $\varepsilon_{n+1} := f(\boldsymbol{\theta}_n, \mathbf{s}_n, \mathbf{a}_n, \mathbf{s}_{n+1}) - F(\boldsymbol{\theta}_n)$.

For Theorem 1, we show that the accumulation of errors is bounded for a finite horizon. For Theorem 2, to show that $\hat{\boldsymbol{\theta}}$ is an asymptotic pseudo-trajectory of solutions $\bar{\boldsymbol{\theta}}$ of (ALE),

we show that the accumulation of errors ε_k from $k = n$ to $m(n,T)$ decays as $n \to \infty$ for any finite $T > 0$. Specifically, we show the *asymptotic rate of convergence* of ε_k tends to zero, that is,

$$\lim_{n\to\infty} \left(\sup_{n<\ell\leq m(n,T)} \left| \sum_{k=n+1}^{\ell} \gamma_k \, \varepsilon_k \right| \right) = 0 \,, \quad \text{a.s.}. \tag{ARC}$$

We analyze the accumulation of errors ε_k by decomposing ε_k into a martingale noise component, denoted M_k, and a remainder component, denoted U_k. Specifically, we write

$$\varepsilon_{n+1} = M_{n+1} + U_{n+1} \,, \tag{3.4a}$$

$$M_{n+1} := f(\boldsymbol{\theta}_n, \mathbf{s}_n, \mathbf{a}_n, \mathbf{s}_{n+1}) - \bar{f}(\boldsymbol{\theta}_n, \mathbf{s}_n) \,, \tag{3.4b}$$

$$U_{n+1} := \bar{f}(\boldsymbol{\theta}_n, \mathbf{s}_n) - F(\boldsymbol{\theta}_n) \,, \tag{3.4c}$$

where $\bar{f}$ is defined in (3.1). After bounding and controlling the error terms in (3.4a), apply Proposition 4.1 of Benaïm (1999) to obtain the first part of Theorem 2. Next, apply the Kushner-Clark Lemma in Kushner and Clark (1978) to obtain the second part of Theorem 2. □

Corollary 1. *Suppose that $\{\gamma_n\}$ is a sequence of random variables such that γ_{n+1} is $\mathcal{F}_n$ measurable. If $\sum_n \gamma_n = \infty$ and* (DLR) *holds almost surely, then the conclusions from Theorem 2.2 continue to hold.*

Corollary 1 follows as an immediate result of Remark 4.3 in Benaïm (1999). This result is useful to analyze asynchronous learning algorithms. These are algorithms where the learning rate depends on the state and the number of times the state has been visited.

3.3 Discussion

Our framework allows us to analyze learning in a variety of repeated games. First, if the public signal Y and the idiosyncratic signal for each player are both empty, then our framework allows us to analyze action learning. This includes the classical learning algorithms such as Cross learning (see Cross, 1973), Erev and Roth's learning model (see Erev and Roth, 1998), and multi-arm bandit algorithms such as the EXP3 algorithm (see Auer et al., 2002). From our notation, it is also easy to see that our framework allows us to analyze learning in stochastic games. This allows us to analyze learning models designed for Markov decision processes such as Q-learning, SARSA, and Actor Critic algorithms (see Sutton and Barto, 2018).

Importantly, our framework allows us to study bounded memory strategy learning under various monitoring structures. If $Y = \mathcal{A}$ and $Y^i = \{\emptyset\}$ for all i, then we have perfect monitoring; if Y is a noisy public signal and $Y^i = \{\emptyset\}$ for all i, then we have imperfect public monitoring; and if $Y = \{\emptyset\}$ and Y^i are idiosyncratic signals for all i, then we have private monitoring. Of course, our framework easily accommodates any combination of perfect, public, and private monitoring.

Additionally, our framework allows each player to use different learning rules provided they all satisfy Assumptions A.4 and A.5. Our framework allows for different lengths of memory used by each player so long as each player uses a bounded memory strategy. Although our framework allows for these asymmetries, the usefulness of this beyond numerical experiments is limited because proving convergence results under these asymmetries remain non-trivial.

Key to our result is that the state space is finite. This assumption is hardly restrictive through the numerous examples provided. Additionally, in the next Chapter, we show that a finite state space is sufficient to achieve a Folk theorem. Nonetheless, one can obtain results for the case of an infinite state space through the assumptions in Métivier and Priouret (1984). However, these assumptions are hard to verify because it requires one to directly verify properties of the Poisson equations associated with the Markov chain.

Finally, our approach offers numerous benefits over simulation studies because the ODEs allow us to (i) efficiently analyze full the parameter space, (ii) visualize the asymptotic behavior of the learning dynamics through visual inspection of the basins of attractions, and (iii) use numerous numerical packages that already exist for ODE solvers.

Chapter 4

A Folk Theorem from Learning

4.1 Learning Model

In this Chapter, we focus on a dynamic generalization of smooth fictitious play when monitoring is perfect and all players use m-memory strategies.

4.1.1 Setup

Consider an infinitely repeated game $\mathcal{G}^\infty = \langle (\mathcal{A}_i)_{0<i\leq I}, (u^i)_{0<i\leq I}, \delta \rangle$ played by $I \geq 2$ players. Let $\mathcal{A}_i$ denote the finite set of actions for player i, let $\mathcal{A}_{-i} = \bigtimes_{j\neq i} \mathcal{A}_j$ denote the set of action profiles excluding player i, and let $\mathcal{A} = \bigtimes_{0<i\leq I} \mathcal{A}_i$ denote the set of action profiles. For simplicity and without loss of generality, we assume that $\mathcal{A}_i = \mathcal{A}_j$ for all $0 < i, j \leq I$. After an action profile $\mathbf{a} = (a^1, ..., a^I) \in \mathcal{A}$ is played at each period $n = 0, 1, 2, 3...$ of the repeated game, each player i receives a payoff according to a utility function $u^i : \mathcal{A} \to \mathbb{R}$, and $\delta \in [0, 1)$ is the common parameter used to discount the future stream of payoffs.

The set of period $n \geq 0$ histories is given by $\mathcal{H}_n = \mathcal{A}^n$, where $\mathcal{A}^n$ is the n-fold product of $\mathcal{A}$. An n-stage history $h \in \mathcal{H}_n$ is a sequence of n action profiles that identify the actions played in periods 0 through $n-1$. The set of all histories is defined as $\mathcal{H} = \bigcup_{n\geq 0} \mathcal{H}_n$, where $\mathcal{H}_0 = \{\emptyset\}$. Let $\ell(h)$ denote the length of any history $h \in \mathcal{H}$. For any finite integer m and any history $h \in \mathcal{H}$ with $\ell(h) \geq m$, denote $T^m(h) = (\mathbf{a}_{\ell(h)-m}, ..., \mathbf{a}_{\ell(h)-1})$ as the most recent m action profiles of h.

To simplify notation, it is useful to develop a specialized notation for any continuation game from periods $n \geq m$ onwards. For any history $h \in \mathcal{H}$ with $\ell(h) \geq m$, let $T^m(h) = \mathbf{s} \in \mathcal{S}$ denote the state of the game, where $\mathcal{S} = \mathcal{A}^m$ is the set of m-memory histories. Let $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ denote the transition function that describes the transition between m-memory histories. Specifically, $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ is the probability that the subsequent state is $\mathbf{s}'$ given that the current state is $\mathbf{s}$ and the current action profile is $\mathbf{a}$. We further denote $p_{a^i}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}^{-i})$ as the conditional

transition function given that player i plays a^i, and given the action profile $\mathbf{a}^{-i}$ of all other players.

A stationary m-memory strategy profile is characterized as follows. Let $\theta^i(a\,|\,\mathbf{s})$, or more compactly $\theta^i_{a|\mathbf{s}}$, denote the probability that player i plays action a in state $\mathbf{s}$. The collection $\boldsymbol{\theta}^i = (\theta^i_{a|\mathbf{s}})_{a\in\mathcal{A}_i,\mathbf{s}\in\mathcal{S}} \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$ characterizes a stationary m-memory strategy for player i, where $\boldsymbol{\theta}^i_{\mathbf{s}} \in \Delta(\mathcal{A}_i)$ denotes the mixed strategy of player i in state $\mathbf{s}$, and $\Delta(\mathcal{A}_i)$ is the set of probability measures on $\mathcal{A}_i$. The collection $\boldsymbol{\theta} = (\boldsymbol{\theta}^i)_{0<i\leq I}$ characterizes a stationary m-memory strategy profile for all players where $\boldsymbol{\theta} \in \Delta(\mathcal{A}_i)^{I\times|\mathcal{S}|} := G$, and the collection $\boldsymbol{\theta}^{-i} = (\boldsymbol{\theta}^j)_{j\neq i}$ characterizes a stationary m-memory strategy profile excluding player i. Finally, we often refer to $\boldsymbol{\theta}$ as a strategy profile for simplicity (the bounded memory and stationarity are always implied).[1]

To complete the specialized notation, we denote $\mathcal{G}^\infty_m$ as any continuation game of $\mathcal{G}^\infty$ that begins from period $n \geq m$ onwards, and with the restriction that the strategies are limited to stationary m-memory strategies. Observe that we use the recursive structure of the repeated game to obtain a Markov, automaton-like representation of the space of m-memory strategy profiles so that we can analyze the evolution of m-memory strategies through learning. This setup allows us to study learning algorithms with behavioral strategies that can condition on the (recent) history.

4.1.2 Preliminaries

Consider a perturbed version of the game $\mathcal{G}^\infty(\tau)$, whose one-stage expected utility for player i, when the stage game strategy profile given state $\mathbf{s}$ is $\boldsymbol{\theta}_{\mathbf{s}} \in \Delta(\mathcal{A}_i)^I$, is given by

$$\tilde{u}^i(\boldsymbol{\theta}_{\mathbf{s}}) = \sum_{\mathbf{a}\in\mathcal{A}} \left[\prod_{j=1}^{I} \theta^j_{a^j|\mathbf{s}}\right] u^i(\mathbf{a}) - C^i(\tau;\boldsymbol{\theta}^i_s) := u^i(\boldsymbol{\theta}_{\mathbf{s}}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}),$$

where a^j are components of the action profile $\mathbf{a}$ and $\tau > 0$ is the perturbation parameter. The term $C^i(\tau;\boldsymbol{\theta}^i_s)$ is a deterministic perturbation function that satisfies three conditions:

1. $C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})$ is strictly convex and bounded in $\boldsymbol{\theta}^i_s$ for each $\tau > 0$,
2. $C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}) \to 0$ as $\tau \to 0$ and $C^i(0;\boldsymbol{\theta}^i_{\mathbf{s}}) = 0$ for all $\boldsymbol{\theta}^i_{\mathbf{s}} \in \Delta(\mathcal{A}_i)$, and
3. $|\nabla_{\boldsymbol{\theta}^i_s} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})| \to \infty$ as $\boldsymbol{\theta}^i_{\mathbf{s}}$ approaches the boundary of $\Delta(\mathcal{A}_i)$ for all $\tau > 0$.

[1] In our construction of a stationary m-memory strategy profile, we omit defining the strategy profile for periods $n < m$. Nonetheless, the strategy profiles we study are well defined for continuation games where $n \geq m$, and this is not a problem because $n \gg m$ through the process of learning.

These are standard admissibility conditions for deterministic perturbations (see Fudenberg and Levine, 1998). When $\tau = 0$, the perturbed game $\mathcal{G}^\infty(\tau)$ reduces to the unperturbed game $\mathcal{G}^\infty(0)$ because of the second condition.

For a given strategy profile $\boldsymbol{\theta}$, the *continuation payoff* of state $\mathbf{s} \in \mathcal{S}$ for player i of the perturbed game $\mathcal{G}_m^\infty(\tau)$ is given by

$$V_{\mathbf{s}}^i(\tau;\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u^i(\boldsymbol{\theta}_{\mathbf{s}_k^\theta}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}_k^\theta})\right) \middle| \mathbf{s}_0^\theta = \mathbf{s}\right], \tag{4.1}$$

which is the expected discounted (perturbed) payoff that player i achieves if state $\mathbf{s}$ is reached and everyone (including player i) continues to play according to $\boldsymbol{\theta}$. When necessary, we write the continuation payoff with respect to an arbitrary distribution over states $\mu \in \Delta(\mathcal{S})$ as $V_\mu^i(\tau;\boldsymbol{\theta}) := \sum_{\mathbf{s}\in\mathcal{S}} \mu(\mathbf{s})\, V_{\mathbf{s}}^i(\tau;\boldsymbol{\theta})$, where $\Delta(\mathcal{S})$ is the set of probability measures on $\mathcal{S}$.

In (4.1), the expectation is taken with respect to the hypothetical evolution of the game where actions are sampled from the fixed strategy profile $\boldsymbol{\theta}$, and the process $\mathbf{s}^\theta$ is the corresponding sequence of states of the game. The process $\mathbf{s}^\theta$ is a Markov chain with transition dynamics

$$P_\theta(\mathbf{s}' \,|\, \mathbf{s}) := \mathbb{P}(\mathbf{s}_{k+1}^\theta = \mathbf{s}' \,|\, \mathbf{s}_k^\theta = \mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}} p(\mathbf{s}' \,|\, \mathbf{s}, \mathbf{a}) \prod_{j=1}^{I} \theta^j_{a^j|\mathbf{s}}\,.$$

For a given strategy profile $\boldsymbol{\theta}$, action $a \in \mathcal{A}_i$, and state $\mathbf{s} \in \mathcal{S}$, the *action value* of a given $\mathbf{s}$ for player i is computed as

$$V_{a|\mathbf{s}}^i(\tau;\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u^i(\boldsymbol{\theta}_{\mathbf{s}_k^\theta}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}_k^\theta})\right) \middle| \mathbf{s}_0^\theta = \mathbf{s}, a_0^i = a\right], \tag{4.2}$$

which is the expected discounted (perturbed) payoff associated with the one-shot deviation principle that player i achieves if state $\mathbf{s}$ is reached and player i deviates with action a. Here, all players (including player i) continue to play according to $\boldsymbol{\theta}$ with the exception that player i deviates with action a at the first instance. For a fixed strategy profile $\boldsymbol{\theta}$ and perturbation parameter τ, the continuation payoffs in (4.1) and the action values in (4.2) are bounded for $\delta < 1$ because $\tilde{u}^i(\boldsymbol{\theta}_{\mathbf{s}})$ is bounded for all $0 < i \leq I$ and $\boldsymbol{\theta}_{\mathbf{s}} \in \Delta(\mathcal{A}_i)^I$. Finally, the continuation payoff of state $\mathbf{s} \in \mathcal{S}$ is $V_{\mathbf{s}}^i(\tau;\boldsymbol{\theta}) = \sum_{a\in\mathcal{A}_i} \theta^i_{a|\mathbf{s}} V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta})$.

4.1.3 State-Dependent Smooth Fictitious Play

In classical smooth fictitious play, each player has full knowledge of her own payoffs and assumes that her opponents play according to stationary stage game strategies. In each period, each player plays a smoothed best response to her expected payoff vector, which is

the expected payoff of the player's pure actions given the belief about the stationary stage game strategy profile of the opponents. The belief is given by the empirical frequency of past play. The smoothed best response is the best response subject to a deterministic perturbation or cost. Thus, players are myopic and optimize the immediate payoff (see Fudenberg and Kreps, 1993). Classical smooth fictitious play is an algorithm that learns both pure and mixed Nash equilibria of static games, and it is *stateless* in the sense that the strategies used for the belief and smoothed best response do not depend on the state of the game.

We introduce a dynamic generalization of smooth fictitious play with bounded m-memory strategies, whereby each player conditions her actions on the state of the game, i.e., the recent m-memory history. In our learning model, each player has full knowledge of the transition function $p(\mathbf{s}' \,|\, \mathbf{s}, \mathbf{a})$ and their own payoffs, and assumes that all players (themselves included) play according to a stationary m-memory strategy, i.e., a set of stage game mixed strategies for all combinations of m-memory histories. We assume the beliefs of all players are common throughout the entirety of the repeated game.[2] In each period, each player plays a smoothed best response to their perturbed action values, where the action values are computed based on the belief over everyone's stationary m-memory strategy. Thus, players are non-myopic and optimize the tradeoff between immediate and future payoffs.

Remark 1. In classical smooth fictitious play, players require only the belief over their opponents' stage game strategy to calculate their expected payoff vector. In our learning model, players also have a belief about how they will behave in the future to calculate the action values. In general, what one believes one would play in the future does not necessarily line up with one's behavioral response except at convergence. We think of this as a form of bounded rationality imposed by computational constraints. To reduce computational costs, players use their own empirical behavior thus far, as a proxy for their own future behavior, to calculate the continuation payoffs associated with a one-shot deviation.

Beliefs about the stationary m-memory strategies are based on the empirical frequency of play. Player i's empirical frequency of an action $a \in \mathcal{A}_i$ in state $\mathbf{s} \in \mathcal{S}$ up to period n is

$$\theta_n^i(a \,|\, \mathbf{s}) = \frac{1}{\mathbf{s}_n^{\#}} \sum_{k=0}^{n-1} \mathbb{1}_{\{\mathbf{s}_k = \mathbf{s}\}} \mathbb{1}_{\{a_k^i = a\}},$$

where $\mathbf{s}_n^{\#} = \sum_{k=0}^{n-1} \mathbb{1}_{\{\mathbf{s}_k = \mathbf{s}\}}$ counts the number of times state $\mathbf{s}$ was reached, and the indicator function on actions counts the number of times a particular action was played in state $\mathbf{s}$. To avoid excessive notation, we also use $\boldsymbol{\theta}$ to refer to the (common) belief.

[2] This assumption is not restrictive. If players draw random actions until they form a valid belief where each state has been visited at least once, then players will have a common belief throughout the game.

For a given belief $\boldsymbol{\theta}$, each player i samples an action from the smoothed best response function given by

$$\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = \underset{\mathbf{y}\in\operatorname{int}(\Delta(\mathcal{A}_i))}{\arg\max}\ J^i_{\mathbf{s}}(\mathbf{y},\boldsymbol{\theta};\tau) \text{ with } J^i_{\mathbf{s}}(\mathbf{y},\boldsymbol{\theta};\tau) = \mathbf{y}\cdot V^i_{\cdot|\mathbf{s}}(\tau;\boldsymbol{\theta}) - C^i(\tau;\mathbf{y}), \tag{4.3}$$

for all $\mathbf{s}\in\mathcal{S}$, where $V^i_{\cdot|\mathbf{s}}(\tau;\boldsymbol{\theta}) = (V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}))_{a\in\mathcal{A}_i} \in \mathbb{R}^{|\mathcal{A}_i|}$ is the vector of action values in state $\mathbf{s}$, the operator $\cdot$ denotes the dot product, $C^i(\tau;\mathbf{y})$ is the deterministic perturbation in (4.1) and (4.2) with the same perturbation parameter τ, and $\operatorname{int}(\Delta(\mathcal{A}_i))$ denotes the interior of the simplex $\Delta(\mathcal{A}_i)$.[3] Therefore, given a belief $\boldsymbol{\theta}$, the smoothed best response $\tilde{B}(\tau;\boldsymbol{\theta}) = (\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}))_{0<i\leq I,\mathbf{s}\in\mathcal{S}}$ forms a stationary m-memory strategy profile with $\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) \in \operatorname{int}(\Delta(\mathcal{A}_i))$ and the component $\tilde{B}^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta})$ that corresponds to an action a is the probability that player i plays a in state $\mathbf{s}$.

The evolution of the beliefs for state-dependent smooth fictitious play can be written as a learning algorithm given by the following discrete-time stochastic system:

$$\theta^i_{n+1}(a^i\,|\,\mathbf{s}) = \theta^i_n(a^i\,|\,\mathbf{s}) + \frac{\mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}}}{\mathbf{s}^{\#}_n + \mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}}}\left(\mathbb{1}_{\{a^i_n=a^i\}} - \theta^i_n(a^i\,|\,\mathbf{s})\right) \tag{4.4}$$

for all $0 < i \leq I$, $a\in\mathcal{A}_i$, and $\mathbf{s}\in\mathcal{S}$. The action a^i_n at period n is sampled based on the current state $\mathbf{s}_n$ from the smoothed best response function $\tilde{B}^i_{\mathbf{s}_n}(\tau;\boldsymbol{\theta}_n)$ associated with the belief at period n. The *learning rate* for each state $\mathbf{s}\in\mathcal{S}$ is defined as

$$\gamma^{\mathbf{s}}_{n+1} = \frac{\mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}}}{\mathbf{s}^{\#}_n + \mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}}}.$$

Thus, we write (4.4) more compactly in vector notation as

$$\boldsymbol{\theta}_{n+1}(\mathbf{s}) = \boldsymbol{\theta}_n(\mathbf{s}) + \gamma^{\mathbf{s}}_{n+1}\left(\mathbb{1}_{\{\mathbf{a}_n=\mathbf{a}\}} - \boldsymbol{\theta}_n(\mathbf{s})\right), \tag{SFP}$$

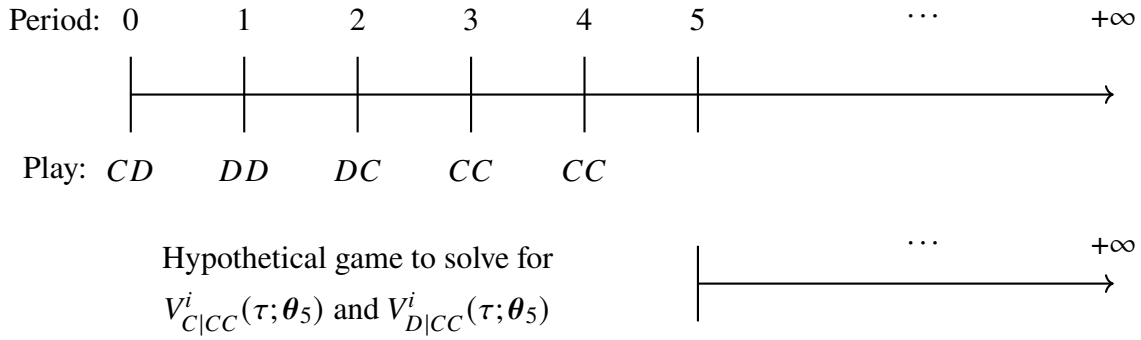
for all $\mathbf{s}\in\mathcal{S}$.

In short, state-dependent smooth fictitious play is summarized as follows. Given a belief $\boldsymbol{\theta}_n$ at period n, players compute the perturbed action values in (4.2) by taking the expectation with respect to the hypothetical evolution of the game where the actions are sampled from the fixed belief $\boldsymbol{\theta}_n$ at period n. Next, each player uses the perturbed action values to obtain the smoothed best response function $\tilde{B}^i(\tau;\boldsymbol{\theta}_n)$ in (4.3). Then, each player uses the smoothed best response function $\tilde{B}^i_{\mathbf{s}_n}(\tau;\boldsymbol{\theta}_n)$ from state $\mathbf{s}_n$ at period n to sample and play their action a^i_n, which generates the action profile $\mathbf{a}_n$. Finally, the action profile $\mathbf{a}_n$ feeds into the belief through (4.4), and the state of the game evolves according to $p(\mathbf{s}_{n+1}\,|\,\mathbf{s}_n,\mathbf{a}_n)$. The above is repeated ad infinitum.

[3] The payoff of the underlying game and the action choice are perturbed. A priori, the perturbations can be different provided they satisfy the standard admissibility conditions. However, using the same strictly convex perturbation function is essential to prove convergence of (SFP).

Remark 2. In our construction of (SFP), we assume that players randomize their behaviors independently in both the calculation of the perturbed action values in (4.2) and when sampling their action from the smoothed best response function $\tilde{B}(\tau;\boldsymbol{\theta}_n)$ in (4.3). In our setup, it is easy to capture additional correlation between the players actions by introducing a public correlating device, where the outcome of the public correlating device ω is sampled from a finite set of signals $\mathcal{W}$. In the case of public correlation, an n-stage history $h \in \mathcal{H}_n$ is a sequence of n action profiles and n realizations of the public correlating device, so the set of m-memory histories $\mathcal{S} = (\mathcal{A} \times \mathcal{W})^m$.

Example 2. Consider a repeated game with two players $I = 2$ and two actions $\mathcal{A}_i = \{C, D\}$ for each player. The players use one-memory strategies, so the set of states is given by $\mathcal{S} = \{CC, CD, DC, DD\}$.



At period $n = 5$, the state of the game is $\mathbf{s}_5 = CC$, and the belief is given by $\theta^1_5(C|CC) = 1$, $\theta^1_5(C|CD) = 0$, $\theta^1_5(C|DC) = 1$, $\theta^1_5(C|DD) = 0$ for player 1, and $\theta^2_5(C|CC) = 1$, $\theta^2_5(C|CD) = 0$, $\theta^2_5(C|DC) = 1$, $\theta^2_5(C|DD) = 1$ for player 2. Using the one-memory belief $\boldsymbol{\theta}_5$, each player plays a hypothetical game to solve for the action values in state CC. Each player then uses their action values to play a smoothed best response from state $\mathbf{s}_5 = CC$ given by $\tilde{B}_{CC}(\tau;\boldsymbol{\theta}_5)$. Suppose the players respond with an action profile $\mathbf{a}_5 = DD$, then the state of the game becomes $\mathbf{s}_6 = DD$, and the belief is updated to become $\theta^1_6(C|CC) = 0.5$, $\theta^2_6(C|CC) = 0.5$, and $\theta^2_6(C|\mathbf{s}) = \theta^2_5(C|\mathbf{s})$ for all $\mathbf{s} \in \{CD, DC, DD\}$. This entire process is repeated ad infinitum.

The remainder of the paper focuses on the case where the deterministic perturbation is given by the entropy function

$$C^i(\tau;\boldsymbol{\theta}^i_s) = \tau \sum_{a \in \mathcal{A}_i} \theta^i_{a|\mathbf{s}} \ln \theta^i_{a|\mathbf{s}} \tag{4.5}$$

for $\tau \geq 0$, so that the smoothed best response function in (4.3) reduces to the logit function

$$\tilde{B}^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}) = \frac{\exp\left(\tau^{-1} V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta})\right)}{\sum_{a' \in \mathcal{A}_i} \exp\left(\tau^{-1} V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta})\right)} \tag{4.6}$$

for all $\mathbf{s} \in \mathcal{S}$ and $a \in \mathcal{A}_i$; see Hofbauer and Sandholm (2002). The logit function is such that as $\tau \to 0$, the smoothed best response function converges to the best response function, and as $\tau \to \infty$, every action is played with equal probability.

In the memoryless case of $m = 0$, where $\mathcal{S} = \{\emptyset\}$ and $|\mathcal{S}| = 1$, (SFP) reduces to classical smooth fictitious play because the vector of action values for player i reduces to their expected payoff vector plus a constant that is the same across all actions. Thus, the smoothed best response is the same with or without the payoff perturbation because (4.6) is invariant under translations by a common constant. On the other hand, if $\delta = 0$, then (SFP) reduces to conditional smooth fictitious play from Fudenberg and Levine (1999) with a fixed categorization rule that corresponds to m-memory histories.

4.1.4 State-Dependent Smoothed Best Response Dynamics

Following Chapter 3, we use stochastic approximation techniques to show that as $n \to \infty$, the trajectories of (SFP) are approximated by solutions $\bar{\boldsymbol{\theta}}$ to the ODE

$$\dot{\bar{\boldsymbol{\theta}}} = d\bar{\boldsymbol{\theta}}/dt = F(\bar{\boldsymbol{\theta}}),$$

for an appropriately defined deterministic function $F : \mathbb{R}^{I \times |\mathcal{S}| \times |\mathcal{A}_i|} \to \mathbb{R}^{I \times |\mathcal{S}| \times |\mathcal{A}_i|}$. In standard smooth fictitious play without states, one obtains $\dot{\bar{\boldsymbol{\theta}}} = \tilde{B}(\tau; \bar{\boldsymbol{\theta}}) - \bar{\boldsymbol{\theta}}$, which is the expected increment of the belief $\boldsymbol{\theta}$.

In this setting, the belief $\boldsymbol{\theta}$ is the tuple of parameters that the algorithm tracks. The smoothed best response $\tilde{B}(\tau; \boldsymbol{\theta}_n)$ is the choice rule that maps the tuple of parameters to a bounded memory strategy σ^i_θ for each player i. As with Chapter 3, the distribution of $\Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s})$ is computed as follows. Fix the belief $\boldsymbol{\theta} \in G$ which also fixes the smoothed best response functions $\tilde{B}(\tau; \boldsymbol{\theta}) = (\tilde{B}^i_{\mathbf{s}}(\tau; \boldsymbol{\theta}))_{0<i\leq I, \mathbf{s}\in\mathcal{S}}$. Next, consider the hypothetical evolution of the game that evolves according to the fixed smoothed best response functions $\tilde{B}(\tau; \boldsymbol{\theta})$ and define $\mathbf{s}_k^{\tilde{B}(\tau;\theta)}$ as the corresponding sequence of states of the game. Then the process $\mathbf{s}_k^{\tilde{B}(\tau;\theta)}$ is a Markov chain with transition dynamics

$$P_{\tilde{B}(\tau;\theta)}(\mathbf{s}' \,|\, \mathbf{s}) := \mathbb{P}\left(\mathbf{s}_{k+1}^{\tilde{B}(\tau;\theta)} = \mathbf{s}' \,|\, \mathbf{s}_k^{\tilde{B}(\tau;\theta)} = \mathbf{s}\right) = \sum_{\mathbf{a}\in\mathcal{A}} p(\mathbf{s}' \,|\, \mathbf{s}, \mathbf{a}) \prod_{j=1}^{I} \tilde{B}^j_{a^j|\mathbf{s}}(\tau; \boldsymbol{\theta}).$$

If the process $\mathbf{s}^{\tilde{B}(\tau;\theta)}$ is an aperiodic Markov chain whose one recurrent class is $\mathcal{S}$, then there exists a unique probability distribution $\Gamma_{\tilde{B}(\tau;\theta)} \in \Delta(\mathcal{S})$, and by the Ergodic Theorem for Markov Chains, we have

$$\Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) = \lim_{K\to\infty} \frac{1}{K+1} \mathbb{E}\left[\sum_{k=0}^{K} \mathbb{1}_{\left\{\mathbf{s}_k^{\tilde{B}(\tau;\theta)} = \mathbf{s}\right\}}\right]. \tag{4.7}$$

Thus, the algorithmic learning equations for state-dependent smooth fictitious play is given by

$$\dot{\bar{\boldsymbol{\theta}}}^i_{\mathbf{s}}(t) = \Gamma_{\tilde{B}(\tau;\bar{\theta}(t))}(\mathbf{s}) \left(\tilde{B}^i_{\mathbf{s}}(\tau; \bar{\boldsymbol{\theta}}(t)) - \bar{\boldsymbol{\theta}}^i_{\mathbf{s}}(t) \right), \tag{SBRD}$$

for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$, where $\bar{\boldsymbol{\theta}}^i_{\mathbf{s}}(t) \in \Delta(\mathcal{A}_i)$ is the belief of player i's mixed strategy in state $\mathbf{s}$ at time t. We call these system of equations the state-dependent smoothed best response dynamics. Intuitively, $\Gamma_{\tilde{B}(\tau;\bar{\theta}(t))}(\mathbf{s})$ can be thought of as the relative 'instantaneous' rate with which beliefs of the mixed strategy in state $\mathbf{s}$ are updated. If $|\mathcal{S}| = 1$, then $\Gamma_{\tilde{B}(\tau;\bar{\theta}(t))} \equiv 1$, and (SBRD) reduces to the standard smoothed best response dynamics for smooth fictitious play without states. In what follows, we show that the trajectories of (SBRD) approximate the trajectories of (SFP) in the sense of an asymptotic pseudo-trajectory; see Definition 1.

4.2 Convergence Results

4.2.1 Convergence to State-Dependent Smoothed Best Response

We use the properties of (SBRD) to show that (SFP) converges to an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}^\infty_m(0)$. Therefore, we first show that trajectories of (SBRD) approximate the asymptotic behavior of (SFP).

First, observe that for each $\boldsymbol{\theta} \in G$, the state process $\mathbf{s}^{\tilde{B}(\tau;\theta)}$ is an aperiodic Markov chain with a single recurrent class when the perturbation parameter $\tau > 0$. This observation follows because the smoothed best response $\tilde{B}(\tau;\boldsymbol{\theta})$ places a positive probability of playing any action for any $\boldsymbol{\theta} \in G$ when $\tau > 0$. This observation ensures that for all possible beliefs $\boldsymbol{\theta}$, the distribution defined in (4.7) is well defined and unique, and by Perkins and Leslie (2012), there is $\eta > 0$ such that for all $\boldsymbol{\theta} \in G$, $\Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) \geq \eta$ for all $\mathbf{s} \in \mathcal{S}$. Hence, each state $\mathbf{s} \in \mathcal{S}$ will be visited infinitely many times at a non-negligible relative frequency; specifically

$$\liminf_{n\to\infty} \frac{\mathbf{s}^\#_n}{n} > 0 \quad \text{and} \quad \lim_{n\to\infty} \mathbf{s}^\#_n = \infty, \quad a.s.. \tag{4.8}$$

Next, we use the learning rate as a connection to compare the discrete-time trajectories of (SFP) with the continuous-time solutions of (SBRD). The difficulty here is that here the learning rate $\gamma^{\mathbf{s}}_n$ is specific to each state and updates the beliefs with respect to its own local clock $\mathbf{s}^\#_n$ and not the global clock n.[4] This asynchronicity is addressed by following Borkar

[4]This asynchrony is different from that in Asker et al. (2022), where the learning rate is the same for each state, but each state is updated asynchronously. Here, each state is updated asynchronously, but the learning rate is also unique for each state.

(2008). Rewrite (SFP) so that there is a common stochastic learning rate for all $\mathbf{s} \in \mathcal{S}$, and define

$$\bar{\gamma}_n = \max_{\mathbf{s}\in\mathcal{S}} \gamma_n^{\mathbf{s}} > 0$$

for all n. By (4.8), we have that $\sum_n \gamma_n^{\mathbf{s}} = \infty$ and $\sum_n \left(\gamma_n^{\mathbf{s}}\right)^2 < \infty$ almost surely, so

$$\sum_n \bar{\gamma}_n = \infty \quad \text{and} \quad \sum_n \bar{\gamma}_n^2 < \infty \quad a.s.$$

because $\mathcal{S}$ is finite. Therefore, we rewrite (SFP) as

$$\begin{aligned}
\boldsymbol{\theta}_{n+1}(\mathbf{s}) &= \boldsymbol{\theta}_n(\mathbf{s}) + \bar{\gamma}_{n+1}\, \hat{\gamma}_{n+1}^{\mathbf{s}} \left(\mathbb{1}_{\{\mathbf{a}_n=\mathbf{a}\}} - \boldsymbol{\theta}_n(\mathbf{s})\right) \\
&= \boldsymbol{\theta}_n(\mathbf{s}) + \bar{\gamma}_{n+1}\, \mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}} \left(\mathbb{1}_{\{\mathbf{a}_n=\mathbf{a}\}} - \boldsymbol{\theta}_n(\mathbf{s})\right) \\
&= \boldsymbol{\theta}_n(\mathbf{s}) + \bar{\gamma}_{n+1}\, f_{\mathbf{s}}(\boldsymbol{\theta}_n, \mathbf{a}_n)\,,
\end{aligned}$$

with a synchronized stochastic learning rate because the random variable $\hat{\gamma}_{n+1}^{\mathbf{s}} = \gamma_{n+1}^{\mathbf{s}} / \bar{\gamma}_{n+1} = \mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}}$ for all n (observe that for each n, $\gamma_n^{\mathbf{s}}$ is only non-zero for one $\mathbf{s} \in \mathcal{S}$), and the learning rule $f_{\mathbf{s}} : G \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{I \times |\mathcal{A}_i|}$ in state $\mathbf{s}$ is given by

$$f_{\mathbf{s}}(\boldsymbol{\theta}_n, \mathbf{a}_n) = \mathbb{1}_{\{\mathbf{s}_n=\mathbf{s}\}} \left(\mathbb{1}_{\{\mathbf{a}_n=\mathbf{a}\}} - \boldsymbol{\theta}_n(\mathbf{s})\right). \tag{4.9}$$

Following the same steps as Chapter 3, to compare the discrete-time trajectories of (SFP) with the continuous-time solutions of (SBRD), we scale time by treating the stochastic learning rate $\bar{\gamma}_n$ as an increment of the real-time t with respect to the global clock n and define an affine interpolation of (SFP). To this end, set $t_n = \sum_{i=1}^{n} \bar{\gamma}_i$ with $t_0 = 0$. We write the solution of the ODEs with initial condition $\boldsymbol{\theta}_{t_n}$ as $\bar{\boldsymbol{\theta}}(t; \boldsymbol{\theta}_{t_n})$. To compare the trajectories $\boldsymbol{\theta}_n$ and $\bar{\boldsymbol{\theta}}(t; \boldsymbol{\theta}_{t_n})$ between times t_n and $t_n + T$, we study $\boldsymbol{\theta}_k$ for integers k between n and $m(n,T)$, where $m(n,T) := \max\{k > n : t_n + T \geq t_k\}$ for all $n \in \mathbb{N}$, and interpolate the discrete-time processes $\theta_n^i(a \,|\, \mathbf{s})$ to a continuous-time processes $\hat{\theta}_t^i(a \,|\, \mathbf{s})$ for $t \in \mathbb{R}^+$ to define

$$\hat{\theta}_{t_n+h}^i(a \,|\, \mathbf{s}) = \theta_n^i(a \,|\, \mathbf{s}) + h\, \frac{\theta_{n+1}^i(a \,|\, \mathbf{s}) - \theta_n^i(a \,|\, \mathbf{s})}{t_{n+1} - t_n}$$

for all $n \in \mathbb{N}$ and $0 \leq h < \bar{\gamma}_{n+1}$. We refer to $\hat{\boldsymbol{\theta}}_t$ as the real-time interpolation of $\boldsymbol{\theta}_n$.

Theorem 3. *Let the perturbation parameter $\tau > 0$, then the real-time interpolated process $\hat{\boldsymbol{\theta}}$ of* (SFP) *is almost surely an asymptotic pseudo-trajectory to the solution of* (SBRD).

The theorem follows once we verify that the conditions in Corollary 1 are satisfied. This theorem is a building block for Theorem 4 because it allows us to appeal to asymptotic convergence results to describe the limiting behavior of (SFP). Specifically, if (SBRD) admits a smooth and strict Lyapunov function, then (SFP) will converge to the rest points of (SBRD) with probability one; see for example Chapters 5 and 6 in Benaïm (1999).

4.2.2 Convergence to ϵ-subgame Perfect Equilibria

To prove that play from (SFP) converges to an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$, we focus on Markov potential games to obtain a smooth and strict Lyapunov function for (SBRD), and for this class of games, we prove that (SFP) converges to a rest point of (SBRD). We then show that the rest points of (SBRD) are m-memory ϵ-subgame perfect equilibria of the unperturbed game $\mathcal{G}_m^\infty(0)$ to complete the result.

We focus on Markov potential games because they admit a suitable potential function which is key to obtain a strict Lyapunov function for (SBRD).

Definition 2 (Markov Potential Game)**.** *A game $\mathcal{G}_m^\infty(0)$ is a Markov potential game if there exists a (state-dependent) potential function $\Phi_{\mathbf{s}} : \boldsymbol{\theta} \to \mathbb{R}$ such that*

$$\Phi_{\mathbf{s}}(\boldsymbol{\theta}^i, \boldsymbol{\theta}^{-i}) - \Phi_{\mathbf{s}}(\boldsymbol{\theta}^{i\prime}, \boldsymbol{\theta}^{-i}) = V_{\mathbf{s}}^i(0; \boldsymbol{\theta}^i, \boldsymbol{\theta}^{-i}) - V_{\mathbf{s}}^i(0; \boldsymbol{\theta}^{i\prime}, \boldsymbol{\theta}^{-i}),$$

for all $0 < i \leq I$, $\mathbf{s} \in \mathcal{S}$, and $\boldsymbol{\theta}^i, \boldsymbol{\theta}^{i\prime} \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$, $\boldsymbol{\theta}^{-i} \in \Delta(\mathcal{A}_i)^{(I-1)\times|\mathcal{S}|}$.

A Markov potential game is closely related to a potential game. In fact, we have the following characterization.

Lemma 1. *A repeated potential game with bounded memory strategies is a Markov potential game.*

In (SFP), the players are playing according to the perturbed game for $\tau > 0$. Therefore, the following lemma guarantees that the perturbed game $\mathcal{G}_m^\infty(\tau)$ is a Markov potential game when the underlying unperturbed game $\mathcal{G}_m^\infty(0)$ is a Markov potential game.

Lemma 2. *If the game $\mathcal{G}_m^\infty(0)$ is a Markov potential game, then the perturbed game $\mathcal{G}_m^\infty(\tau)$ for $\tau \geq 0$ is a Markov potential game with potential function*

$$\Phi_{\mathbf{s}}(\tau; \boldsymbol{\theta}) = \Phi_{\mathbf{s}}(\boldsymbol{\theta}) - \mathbb{E}\left[\sum_{i=1}^{I}\sum_{k=0}^{\infty} \delta^k C^i(\tau; \boldsymbol{\theta}^i_{\mathbf{s}_k^\theta}) \,\middle|\, \mathbf{s}_0^\theta = \mathbf{s}\right] \text{ for all } \mathbf{s} \in \mathcal{S}.$$

Notice that $\Phi_{\mathbf{s}}(\tau; \boldsymbol{\theta}) = \Phi_{\mathbf{s}}(\boldsymbol{\theta})$ for $\tau = 0$.

The following proposition shows that the perturbed potential function of a Markov potential game is a strict Lyapunov function for (SBRD).

Proposition 2. *Consider a Markov potential game and let the perturbation parameter $\tau > 0$. Then the perturbed potential function $\Phi_\mu(\tau; \boldsymbol{\theta}) = \sum_{\mathbf{s}\in\mathcal{S}} \mu(\mathbf{s})\, \Phi_{\mathbf{s}}(\tau; \boldsymbol{\theta})$ is a strict Lyapunov function for* (SBRD) *for any state distribution $\mu \in \Delta(\mathcal{S})$ such that $\mu(\mathbf{s}) > 0$ for all $\mathbf{s} \in \mathcal{S}$.*

The existence of a strict Lyapunov function ensures that the dynamics of (SBRD) are sufficiently well behaved. Together with the fact that the interpolated trajectories of (SFP) are an asymptotic pseudo-trajectory of (SBRD), we can show that (SFP) converges to rest points of (SBRD) as $n \to \infty$. Therefore, it is useful to first characterize the rest points of (SBRD).

The following defines an m-memory subgame perfect equilibrium of the game $\mathcal{G}_m^\infty(\tau)$.

Definition 3 (Subgame Perfect Equilibrium)**.** *An m-memory strategy profile $\boldsymbol{\theta}$ is a subgame perfect equilibrium of $\mathcal{G}_m^\infty(\tau)$ if for all $\boldsymbol{\theta}^{i\prime} \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$, we have*

$$V_{\mathbf{s}}^i(\tau; \boldsymbol{\theta}^{i\prime}, \boldsymbol{\theta}^{-i}) - V_{\mathbf{s}}^i(\tau; \boldsymbol{\theta}^i, \boldsymbol{\theta}^{-i}) \leq 0 \tag{4.10}$$

for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$. Moreover, $\boldsymbol{\theta}$ is an m-memory ϵ-subgame perfect equilibrium of $\mathcal{G}_m^\infty(\tau)$ if the right-hand side of (4.10) *is replaced with ϵ.*

The subgame perfect equilibrium we consider is more demanding because of the restriction to m-memory strategies. The following proposition shows that the rest points of (SBRD) are m-memory ϵ-subgame perfect equilibria of the unperturbed game $\mathcal{G}_m^\infty(0)$.

Proposition 3.

3.1 A strategy profile $\boldsymbol{\theta}^*$ is an m-memory subgame perfect equilibrium of the perturbed game $\mathcal{G}_m^\infty(\tau)$ if and only if $\boldsymbol{\theta}^*$ is a rest point of (SBRD)*.*

3.2 For any $\epsilon > 0$, there exists $\tilde{\tau} > 0$ such that for any $\tau \in (0, \tilde{\tau})$, a rest point of (SBRD) *is an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$.*

A property of the rest points of (SBRD) is that the m-memory behavioral response coincides with the m-memory belief because (SBRD) equals zero if and only if $\tilde{B}(\tau; \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$. Moreover, the perturbed action value converges to the unperturbed action value as $\tau \to 0$. Thus, by construction, the rest points of (SBRD) satisfy the one-shot deviation principle with an ϵ error, and are therefore an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$. Therefore, with Propositions 2 and 3, the following result proves that play from (SFP) converges to an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$. This is a necessary first step to obtain a Folk theorem from learning and show that algorithms can learn to collude.

Theorem 4. *In a repeated potential game, for any $\epsilon > 0$, there exists $\tilde{\tau} > 0$ such that for any $\tau \in (0, \tilde{\tau})$, play from* (SFP) *converges to a connected subset of ϵ-subgame perfect equilibria of the unperturbed game $\mathcal{G}_m^\infty(0)$ with probability one.*

The theorem shows that, in a repeated potential game, play from (SFP) will converge to a subset of m-memory ϵ-subgame perfect equilibria of the unperturbed game $\mathcal{G}_m^\infty(0)$ with probability one. However, by the connectedness of this subset, there is a continuum of potential equilibria to which (SFP) could converge. In our next result, we impose a generic regularity condition on the unperturbed game $\mathcal{G}_m^\infty(0)$ to sharpen this result and to characterize a set of ϵ-subgame perfect equilibria that can be learned.

4.2.3 Learnable Strategies

While Theorem 4 establishes that (SFP) learns to play an m-memory ϵ-subgame perfect equilibrium, the result is insufficient to establish a Folk theorem and insufficient to prove that algorithms can learn to collude because it does not characterize the equilibria that can be learned. A well-known result from stochastic approximation is that (SFP) cannot converge to an unstable rest point of (SBRD), see for example Pemantle (1990) or Brandière (1998). Indeed, it could be the case that the only learnable m-memory subgame perfect equilibrium is to perpetually play a stage game Nash equilibrium, which makes the result vacuous to prove a Folk theorem and vacuous to prove that algorithms can learn to collude.

To characterize a set of equilibria that can be learned, several definitions are required. Consider a collection of actions $a_{\mathbf{s}}^i$ for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$. Define $M : G_\epsilon \to \mathbb{R}^{I \times |\mathcal{S}| \times |\mathcal{A}_i|}$ whose components are given by

$$M_{a^i|\mathbf{s}}^i(\boldsymbol{\theta}) = \begin{cases} \sum_{a^i \in \mathcal{A}_i} \theta_{a^i|\mathbf{s}}^i - 1 & \text{if } a^i = a_{\mathbf{s}}^i, \\ \theta_{a^i|\mathbf{s}}^i \left(V_{a^i|\mathbf{s}}^i(0;\boldsymbol{\theta}) - V_{a_{\mathbf{s}}^i|\mathbf{s}}^i(0;\boldsymbol{\theta}) \right) & \text{if } a^i \in \mathcal{A}_i \setminus \{a_{\mathbf{s}}^i\}, \end{cases} \tag{4.11}$$

for all $0 < i \leq I$, $a^i \in \mathcal{A}_i$, and $\mathbf{s} \in \mathcal{S}$. The set G_ϵ is an open set that strictly contains G, and its construction is detailed in Appendix A. Here, $a_{\mathbf{s}}^i$ is a reference action for player i in state $\mathbf{s}$ so that if $\boldsymbol{\theta}$ is an m-memory subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$, such that $\theta_{a_{\mathbf{s}}^i|\mathbf{s}}^i > 0$ for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$, then $M(\boldsymbol{\theta}) = \mathbf{0}$.

Definition 4 (Regular Equilibrium)**.** *An m-memory subgame perfect equilibrium $\boldsymbol{\theta}$ of the unperturbed game $\mathcal{G}_m^\infty(0)$ is regular if the Jacobian of M with respect to $\boldsymbol{\theta}$ has full rank for some selection of actions $a_{\mathbf{s}}^i \in \mathcal{A}_i$ such that $\theta_{a_{\mathbf{s}}^i|\mathbf{s}}^i > 0$ for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$.*

This notion of regularity is from Doraszelski and Escobar (2010) and it is closely related to that introduced by Harsanyi (1973a,b).

Definition 5 (Regular Game)**.** *A game $\mathcal{G}_m^\infty(0)$ is regular if all equilibria are regular.*

A regular game may seem like a strict requirement, but Theorem 1 of Doraszelski and Escobar (2010) proves that for almost all games $\mathcal{G}_m^\infty(0)$, all equilibria are regular.[5] The first step to refine Theorem 4 is to have a regular game because the number of equilibria is finite, see Lemma 12. We also prove a purification-type result to show that there is a rest point of (SBRD) near each regular equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$. This ensures that we have finitely many rest points of (SBRD) so that play from (SFP) converges to a rest point instead of a continuum of rest points.

The next definition describes the type of equilibria that can be learned by (SFP).

Definition 6 (Pure Equilibrium)**.** *An m-memory subgame perfect equilibrium $\boldsymbol{\theta}$ of the unperturbed game $\mathcal{G}_m^\infty(0)$ is pure if $\boldsymbol{\theta}$ is a pure strategy profile.*

Along with the purification result, we use the Gershgorin circle theorem to prove that a rest point of (SBRD) near a pure equilibrium is locally asymptotically stable. From this it follows that these equilibria have a non-zero probability of attracting the limiting trajectory of (SFP).

Theorem 5. *Consider a regular repeated potential game. Fix $\epsilon > 0$ and let $\boldsymbol{\theta}_{pure}$ denote a strategy profile that is a pure (m-memory subgame perfect) equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$ for a $\delta \in [0,1)$. Then, there exists $\bar{\tau} > 0$ such that for all $\tau \in (0,\bar{\tau})$, play from* (SFP) *with a fixed (δ,τ) pair has a non-zero probability of converging to a strategy profile $\tilde{B}(\tau;\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$ such that the continuation payoffs of the strategy profile $\boldsymbol{\theta}^*$ are within ϵ of the continuation payoffs of $\boldsymbol{\theta}_{pure}$, i.e., $|\mathbf{V}^i(0;\boldsymbol{\theta}^*) - \mathbf{V}^i(0;\boldsymbol{\theta}_{pure})| \leq \epsilon$ for all i, where $\mathbf{V}^i$ is a vector over states $\mathbf{s} \in \mathcal{S}$. Moreover, $\boldsymbol{\theta}^*$ is an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$, and $|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{pure}| \leq \epsilon$.*

Theorem 5 shows that play from (SFP) has a non-zero probability of learning an approximate pure equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$. The non-zero probability of learning the equilibrium is a result of the stochastic nature of (SFP). This characterization is necessary to obtain a Folk theorem from learning because it describes the equilibria that can be learned. However, this result does not say if mixed equilibria can be learned in a regular repeated potential game. We conjecture that mixed equilibria cannot be learned in a regular repeated potential game because mixed equilibria cannot be learned in a potential game with classical smooth fictitious play (see Hofbauer and Hopkins, 2005). However, this remains an open question beyond the scope of this paper. Nonetheless, our characterization

[5] Doraszelski and Escobar (2010) prove this for stochastic games, but the result readily extends to $\mathcal{G}_m^\infty(0)$ because of the recursive structure of the repeated game and our restriction to stationary m-memory strategy profiles.

in Theorem 5 is sufficient to obtain a Folk theorem from learning because all that remains is to establish a Folk theorem using m-memory strategy profiles that are pure equilibria of the unperturbed regular repeated potential game $\mathcal{G}_m^\infty(0)$. Similarly, the result is sufficient to prove that algorithms can learn to collude once we find collusive equilibria that are pure equilibria of the unperturbed regular repeated potential game $\mathcal{G}_m^\infty(0)$.

4.3 A Folk Theorem and Equilibrium Selection

In this section, we apply Theorem 5 to prove a Folk theorem from learning. To provide additional insights into the equilibrium selection process we also refine Theorem 5 and provide a lower bound on the probability of converging to a particular equilibrium.

4.3.1 A Folk Theorem from Learning

Before presenting the complete Folk theorem, we build an understanding of our results with a partial result that uses one-memory strategies in repeated potential games.

First, we define the relevant payoff spaces. Let $\mathbf{v} = (v^1, ..., v^I)$ denote a stage game payoff profile. The set of stage game payoffs generated by the pure action profiles in $\mathcal{A}$ is $\mathcal{U}^a = \{\mathbf{v} \in \mathbb{R}^I : \exists \mathbf{a} \in \mathcal{A} \text{ s.t. } v^i = u^i(a^i, a^{-i}) \text{ for all } i\}$. The set of stage game Nash equilibrium payoffs generated by the pure action profiles in $\mathcal{A}$ is $\mathcal{U}^e = \{\mathbf{v} \in \mathcal{U}^a : \mathbf{e} \in \mathcal{A} \text{ is a stage game Nash equilibrium}\}$.[6] The set of feasible payoffs $\mathcal{U} = \text{co}(\mathcal{U}^a)$ is the convex hull of the set of payoffs $\mathcal{U}^a$. Let $\underline{v}^i = \min_{\mathbf{a}^{-i} \in \mathcal{A}_{-i}} \max_{a^i \in \mathcal{A}_i} u^i(a^i, \mathbf{a}^{-i})$ denote the (pure action) minmax payoff for player i, then $\mathcal{U}^\dagger = \{\mathbf{v} \in \text{co}(\mathcal{U}^a) : v^i \geq \underline{v}^i \text{ for all } i\}$ denotes the set of feasible and individually rational payoff profiles. Finally, suppose that players play according to a fixed strategy profile $\boldsymbol{\theta}^*$ in the repeated game. The normalized continuation payoff $(1-\delta)V_{\mathbf{s}}^i(\tau; \boldsymbol{\theta}^*)$ is the average payoff for player i in the continuation game when starting from state $\mathbf{s}$ with the strategy profile $\boldsymbol{\theta}^*$.

One-Memory Perfect Monitoring

Suppose that players use one-memory strategies so that $\mathcal{S} = \mathcal{A}$. Let $\mathbf{v_a} \in \mathcal{U}^a$ be a payoff profile from an action profile $\mathbf{a}$ that Pareto dominates any payoff profile $\mathbf{v_e} \in \mathcal{U}^e$ from a stage game Nash equilibrium $\mathbf{e}$. A one-memory Nash-reversion strategy profile is given by $\theta^i_{a^i|\mathbf{s}} = 1$ if $\mathbf{s} = \mathbf{a}$ and $\theta^i_{e^i|\mathbf{s}} = 1$ whenever $\mathbf{s} \neq \mathbf{a}$ for all $0 < i \leq I$. It is clear that there exists a

[6] In potential games, the set $\mathcal{U}^e$ is non-empty, see Monderer and Shapley (1996).

value of the discount factor $\underline{\delta}$ such that for all $\delta \in (\underline{\delta}, 1)$, the one-memory Nash-reversion strategy profile is a pure equilibrium of the unperturbed game $\mathcal{G}_1^\infty(0)$ that is collusive.[7]

Corollary 2. *Consider a regular repeated potential game and suppose that players use one-memory strategies. Fix $\epsilon > 0$ and let $\mathbf{v_a} \in \mathcal{U}^a$ be a payoff profile that Pareto dominates any $\mathbf{v_e} \in \mathcal{U}^e$. Denote $\boldsymbol{\theta}_{NR}$ as the one-memory Nash-reversion strategy profile for this payoff profile. Then, there exists $\underline{\delta} \in (0,1)$ and $\bar{\tau} > 0$ such that for all $\delta \in (\underline{\delta}, 1)$ and $\tau \in (0, \bar{\tau})$, play from* (SFP) *with any fixed (δ, τ) pair has a non-zero probability of converging to a one-memory strategy profile $\tilde{B}(\tau; \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$ such that the average payoff of the strategy profile $\boldsymbol{\theta}^*$ is within ϵ of $\mathbf{v_a}$ for an appropriate continuation game. Moreover, $\boldsymbol{\theta}^*$ is a one-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_1^\infty(0)$, and $|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{NR}| \leq \epsilon$.*

The "appropriate" continuation game is used to capture certain subtleties when evaluating the expected discounted payoffs from the strategies learned in our Folk theorem. Specifically, there are two reasons why the payoff profile $\mathbf{v_a}$ will only be achieved for an appropriate continuation game. First, (SFP) has to learn or converge to the appropriate one-memory Nash-reversion strategy profile. Second, when (SFP) converges, the appropriate state (i.e., the state with a normalized continuation payoff that approximately recovers the payoff profile $\mathbf{v_a}$) may not be realized at the time of convergence. However, the appropriate state will eventually be realized at some point in time after convergence (in fact, it will be realized infinitely often) because of the occasional errors in the smoothed best response. Therefore, the appropriate continuation game is the first time the appropriate state is realized after convergence.[8] In the context of Corollary 2, the appropriate state is $\mathbf{s} = \mathbf{a}$ because $(1-\delta) V_{\mathbf{s}}^i(\tau; \boldsymbol{\theta}^*) \approx v_{\mathbf{a}}^i$ for all $0 < i \leq I$ when $\mathbf{s} = \mathbf{a}$ and $(1-\delta) V_{\mathbf{s}}^i(\tau; \boldsymbol{\theta}^*) \approx v_{\mathbf{e}}^i$ for all $0 < i \leq I$ when $\mathbf{s} \neq \mathbf{a}$.

Key to our Folk theorem from learning is to compute the average payoff by discounting the future stream of payoffs to the first time the appropriate state is realized after convergence. To see the importance of this, consider a two player Prisoners' dilemma. Let $\mathbf{v_a} = \mathbf{v}_{CC}$ and $\mathbf{v_e} = \mathbf{v}_{DD}$. Suppose that play converges to an approximate one-memory Nash-reversion strategy of (almost always) cooperate (C) when players cooperated in the previous stage game, and (almost always) defect (D) whenever either player did not cooperate in the previous stage game. The occasional errors in the action choice from the smoothed best

[7] It is easy to extend this to any feasible payoff profile $\mathbf{v} \in \mathcal{U}$ that Pareto dominates any $\mathbf{v_e} \in \mathcal{U}^e$ from a stage game Nash equilibrium $\mathbf{e}$ by introducing a correlation device.

[8] For illustration purposes, we explicitly define the appropriate continuation game as the first time the appropriate state is realized after convergence. However, an appropriate continuation game is whenever the appropriate state is realized after convergence because the average payoff from discounting holds in expectation.

response is essential to ensure that every state will be visited infinitely often so that players can learn an optimal response for every state $\mathbf{s}$.

A consequence of these occasional errors in the action choice is that after convergence, a large portion of time will be spent in the paths of play corresponding to the punishment phase. Using the example above, any occasional error in the action choice from the cooperation phase will instigate a punishment phase, and returning to cooperation will require a simultaneous error in the action choice from both players during the punishment phase (which will eventually happen with probability one). Therefore, if one computes the average payoff (after convergence) as the time average payoff without discounting, then the strategy profile produces an average payoff near $\mathbf{v}_{DD}$. On the other hand, if one computes the average payoff (after convergence) by discounting the future stream of payoffs to the first time the appropriate state ($\mathbf{s} = CC$) is realized after convergence, then one recovers an average payoff near $\mathbf{v}_{CC}$ because $(1-\delta)\, V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}^*) \approx v^i_{CC}$ for all $0 < i \leq I$ when $\mathbf{s} = CC$. The subtlety here is that the occasional errors in the action choice during the cooperation phase are infrequent, so when an error does occur, the discounting subdues the impact of the future stream of payoffs from the punishment phase.

One-memory Nash-reversion strategies are not the only pure equilibria with one-memory perfect monitoring. We use these strategies for illustration purposes due to their simplicity. Using the same example above, win-stay, lose-shift (also known as the Pavlov strategy) is also a symmetric pure equilibrium of the unperturbed game $\mathcal{G}^\infty_1(0)$. With this strategy profile, returning to cooperation does not require a simultaneous error in the action choice from both players, so the appropriate state will appear soon after convergence.[9] The takeaways are that the target payoff profile will be achieved, that it will be achieved after play reaches an appropriate continuation game, and that when the appropriate continuation game is realized depends on the strategies learned.

Bounded-Memory Perfect Monitoring

Our previous result is limited because the result relies on simple one-memory strategies, which makes it difficult to provide the necessary incentives for players to carry out the punishment.[10] In the following theorem, we provide a complete result with m-memory strategies that does not require public randomization. The result shows that learning with

[9] Here, forgiveness in the win-stay, lose-shift strategy is encoded into the decision rule, which is different to the "forgiveness" in the approximate one-memory Nash-reversion strategy profile where forgiveness is a result of simultaneous errors in the choice rule.

[10] Although difficult, it is not impossible. For example, Barlo et al. (2009) prove a subgame perfect Folk theorem with one-memory, provided that the set of actions is sufficiently rich.

bounded rationality in a repeated potential game, which we capture through (SFP), leads to a Folk theorem.

Theorem 6. *Consider a regular repeated potential game that satisfies the nonequivalent utilities (NEU) condition.*[11] *Fix* $\epsilon > 0$ *and let* $\mathbf{v} \in \mathcal{U}^\dagger$ *be a feasible and individually rational payoff profile. Then, there exists* $\underline{\delta} \in (0,1)$, $\bar{\tau} > 0$, *and* $m \in \mathbb{N}$ *such that for all* $\delta \in (\underline{\delta}, 1)$ *and* $\tau \in (0, \bar{\tau})$, *play from* (SFP) *with any fixed* (δ, τ) *pair has a non-zero probability of converging to an* m*-memory strategy profile* $\tilde{B}(\tau; \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$ *such that the average payoff is within* ϵ *of* $\mathbf{v}$ *for an appropriate continuation game. Moreover,* $\boldsymbol{\theta}^*$ *is an* m*-memory* ϵ*-subgame perfect equilibrium of the unperturbed game* $\mathcal{G}_m^\infty(0)$.

This result is a consequence of Theorem 5 and Theorem 1 of Barlo et al. (2016) where they show that for any $\mathbf{v} \in \mathcal{U}^\dagger$, there exists an m-memory pure strategy profile that approximately achieves the payoff profile $\mathbf{v}$ and is a subgame perfect equilibrium of the repeated game. We refer the interested reader to their paper for a full exposition of the strategy profile.

In Theorem 6, we highlight that the average payoff profile achieved with $\boldsymbol{\theta}^*$ has two sources of errors when compared with the payoff profile $\mathbf{v}$. The first source of error is from the m-memory pure strategy profile from Barlo et al. (2016) that approximately achieves the payoff profile $\mathbf{v}$. The length of memory m controls the degree of approximation from this first source of error, and a larger value of the memory m reduces the error. The second source of error is from Theorem 5 because (SFP) learns a strategy profile that is close to the required m-memory pure strategy profile. The value of the perturbation parameter τ controls the degree of approximation from this second source of error, and a smaller value of the perturbation parameter τ reduces the error. This differs from Corollary 2, where the only source of error is the second source from the perturbation parameter τ.

4.3.2 Equilibrium Selection

Until now, our results showed that there is a non-zero probability of learning an approximate pure equilibrium that supports the payoff profile. Below, we provide a sharper characterization of the equilibrium selection process by providing a lower bound on the probability of converging to a particular equilibrium.

Our following result uses the notion of a strong belief, which refers to a belief such that a new empirical observation has little impact when updating said belief, i.e., the learning rate

[11] Given two players i and j, the NEU condition states that one cannot obtain the utility function of player i as a linear combination of the utility function of player j, i.e., there are no constants $c, d > 0$ such that $u^i(\mathbf{a}) = c + d\,u^j(\mathbf{a})$ for all $\mathbf{a} \in \mathcal{A}$; see Abreu et al. (1994).

$\gamma_n^{\mathbf{s}}$ has a small value for all $\mathbf{s} \in \mathcal{S}$. There are two ways to achieve a strong belief, and they correspond to two different approaches to view the equilibrium selection process. First, the progression of play strengthens the belief because $\mathbf{s}_n^{\#}$ increases as n increases for all $\mathbf{s} \in \mathcal{S}$. Therefore, for any $N \in \mathbb{N}$, the progression of play will ensure that $\mathbf{s}_n^{\#} \geq N$ for all $\mathbf{s} \in \mathcal{S}$ for some period n onwards. Second, a strong belief is achieved by encoding a strong prior. This is achieved by modifying $\mathbf{s}_n^{\#} = N + \sum_{k=0}^{n-1} \mathbb{1}_{\{\mathbf{s}_k = \mathbf{s}\}}$. In both cases, a sufficiently strong belief means that N is sufficiently large so that the evolution of the belief is sufficiently slow.

Theorem 7. *Consider a regular repeated potential game. Let $\boldsymbol{\theta}^*$ denote a rest point of* (SBRD) *near a pure (m-memory subgame perfect) equilibrium of the unperturbed game $\mathcal{G}_m^{\infty}(0)$ for a value of $\delta \in [0,1)$ and a small value of $\tau > 0$. Let the belief lie within a neighborhood of $\boldsymbol{\theta}^*$. For any $\varepsilon > 0$, there exists a large value of $N \in \mathbb{N}$ such that play from* (SFP) *converges to $\tilde{B}(\tau; \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$ with probability greater than or equal to $1 - \varepsilon$.*

The neighborhood of $\boldsymbol{\theta}^*$ refers to a compact subset of the domain of attraction of $\boldsymbol{\theta}^*$. The result is intuitive because a strong belief ensures that random shocks are less likely to take the belief out of the domain of attraction of the equilibrium $\boldsymbol{\theta}^*$ so that play is more likely to converge to the equilibrium $\boldsymbol{\theta}^*$. In the first approach where the progression of play strengthens the belief, the players have no influence on the equilibrium they reach because equilibrium selection is effectively random due to the high degree of stochasticity in the early stages of play. In the second approach where a strong belief is achieved by encoding a strong prior, the players can influence the equilibrium they reach by encoding a strong initial prior.

When players try to influence the equilibrium they reach, their strong prior must lie within the basin of attraction of the equilibrium $\boldsymbol{\theta}^*$. The radii of the basins of attraction depend on the value of the discount factor δ. A full characterization is beyond the scope of the paper because these radii are specific to each game. In Chapter 5, we use a toy example of a duopoly with one-memory perfect monitoring to illustrate this dependency, and we find that the basin of attraction of the collusive equilibrium from the one-memory Nash-reversion strategy profile increases as the value of the common discount factor δ increases.

We highlight that the probability of (SFP) not converging to $\boldsymbol{\theta}^*$ when the belief lies within a neighborhood of $\boldsymbol{\theta}^*$ is precisely what ensures that (SFP) has a non-zero probability of converging to a strategy profile near any pure equilibrium. Therefore, depending on the strength of the belief and where the belief lies, the non-zero probability of converging to the particular equilibrium in Theorem 5 can be negligible.

4.4 Discussion

This chapter provides an approach to model human behavior in repeated games where stage game strategies are not enough (see Erev and Roth, 1998). A limitation of our learning model is that it assumes a common m-memory bound. Although such an assumption is standard (see e.g., Barlo et al., 2016), it does mean that the length of the m-memory is fixed and cannot be changed during play. Nonetheless, our result demonstrates that by endowing standard learning models, such as smooth fictitious play, with memory, we gain a lot of mileage in terms of convergence results. Specifically, we go from convergence to a stage game Nash equilibrium to convergence to m-memory subgame perfect equilibrium. Moreover, with sufficient memory, we also recover a Folk theorem from learning in repeated potential games.

More generally, our result provides a partial resolution to the question posed in Green et al. (2014). Specifically, how do players initiate a collusive arrangement without communication? Our results show that a collusive arrangement can be initiated through learning without communication. Our resolution is only partial because we assume that players use a common m-memory bound. The question is then how do players arrive at a common m-memory bound? Is there a focal point for the strategy specification? In the experimental literature on tacit algorithmic collusion, there is a focal point on one-memory strategies because of the long convergence times from learning (see for example Calvano et al., 2020, 2021). In practice, a common m-memory bound might arise from technological constraints, but this question remains an open problem.

On the risk of tacit algorithmic collusion, we emphasize that our results are highly asymptotic. Convergence from learning is an asymptotic result that does not guarantee convergence to a collusive equilibrium, only that it can happen with positive probability. Moreover, as we highlighted in the discussion of Corollary 2, waiting for the appropriate state to be realized after convergence can itself be an asymptotic result (depending on the strategies learned). Nonetheless, the risk of tacit algorithmic collusion is greatest when there is a clear focal point for the strategy specification, when the value of the discount factor δ is sufficiently large, and when there is a clear focal point for where to set the initial belief. Overall, our results theoretically prove that tacit algorithmic collusion is possible, but also that the Folk theorem continues to hold when less than fully rational players learn as they play a repeated potential game.

Finally, our result on equilibrium selection may help to uncover intent on colluding. If a firm delegates pricing or decision-making to an algorithm, it must choose its initial parameters. When these parameters are set close to a collusive equilibrium, rather than

drawn randomly or using standard practices like optimistic initialization, this may suggest a form of intent to induce collusion. Therefore, our equilibrium selection results may help regulators further study unnatural backtesting, hyperparameter choices, or proximity to collusive basins as methods to uncover intention to collude.

Chapter 5

Numerical Experiments

We consider a toy example of a duopoly with a restricted action space and signal space to study algorithmic collusion. By limiting the action space and signal space, we visualize the space of one-memory strategies and its evolution through learning.

5.1 Setting

Consider a duopoly ($I = 2$), where each firm i has a choice between a monopolistic (M) action or a competitive (C) action, i.e., $\mathcal{A}_i = \{M, C\}$. Based on the action profile, the firms receive a public signal $Y = \{G, B\}$ (good and bad). We consider a Bertrand duopoly to study the case of perfect monitoring, and a Cournot duopoly with stochastic demand to study the case of imperfect public monitoring. In the case where players use one-memory strategies, the set of states of the game $\mathcal{S}$ is the set of public signal Y.

Perfect Monitoring

In a Bertrand duopoly, firms perfectly monitor the prices set by their rivals. To simplify the signal space, we consider a public signal that monitors whether or not both firms set the monopolistic price. Specifically,

$$\mathbb{P}(G \mid \mathbf{a}) = \begin{cases} 1 & \text{if } \mathbf{a} = MM \\ 0 & \text{if } \mathbf{a} = \{MC, CM, CC\}, \end{cases}$$

so the firms perfectly monitor any deviations from cooperating at the monopolistic price. The payoff matrix for each firm is given by

	M		C	
M	1	1	$-g$	$1+g$
C	$1+g$	$-g$	0	0

(5.1)

where $g > 0$. This simple example captures the key strategical elements present in a Bertrand duopoly. Specifically, the competitive outcome is the unique Nash equilibrium in the static setting, and the monopolistic outcome Pareto dominates the competitive outcome. However, when interacting repeatedly, the monopolistic outcome can be sustained as an equilibrium outcome if the firms are sufficiently patient. For example, if $\delta > \frac{g}{1+g}$, then a one-memory Nash reversion strategy profile, given by $\sigma^i(M \mid G) = 1$ and $\sigma^i(C \mid B) = 1$ for all i, supports the monopolistic outcome as a one-memory subgame perfect equilibrium.

Imperfect Public Monitoring

In a Cournot duopoly with stochastic demand, firms cannot monitor the output of their rivals. However, they observe a market price that depends on the firms' outputs and the stochastic demand. With the simplified signal space, we interpret a good signal as a high market price and a bad signal as a low market price. The resulting market price conditional on the output profile is given by

$$\mathbb{P}(G \mid \mathbf{a}) = \begin{cases} p & \text{if } \mathbf{a} = MM \\ q & \text{if } \mathbf{a} = \{MC, CM\} \\ r & \text{if } \mathbf{a} = CC, \end{cases}$$

where $p > q > r$ and $p - q > q - r$. The payoff to each firm is given by

$$u^i(a^i, y) = \begin{cases} 1 + \frac{(1+g)(1-p)}{p-q} & \text{if } a^i = H, y = G \\ 1 - \frac{(1+g)p}{p-q} & \text{if } a^i = H, y = B \\ \frac{(1+g)(1-r)}{q-r} & \text{if } a^i = L, y = G \\ -\frac{(1+g)r}{q-r} & \text{if } a^i = L, y = B, \end{cases}$$

and depends on the market price. The payoff is designed so that the expected utility $u^i(\mathbf{a}) = \sum_{y \in Y} \mathbb{P}(y \mid \mathbf{a})\, u^i(a^i, y)$ matches the payoff in (5.1). This simple example captures similar strategical elements as the Bertrand duopoly example, but with the added inability to perfectly infer their rival's output. Nonetheless, if $\frac{1}{(p-r)+(q-r)} > \delta > \frac{1}{(p-q)+(p-r)}$, then a one-memory trigger price strategy profile, given by $\sigma^i(M \mid G) = 1$ and $\sigma^i(C \mid B) = 1$ for all i, also supports the monopolistic outcome as a one-memory subgame perfect equilibrium.

5.2 Learning Algorithms

We present a model of belief learning and a model of reinforcement learning that learns one-memory strategies as examples we use for the remainder of the chapter.

Belief Learning

We use the dynamic generalization of smooth fictitious play, presented in Chapter 4, as a model of belief learning. Recall the model works as follows: given a belief $\boldsymbol{\theta}_n$ at period n, players compute the perturbed action values in (4.2) by taking the expectation with respect to the hypothetical evolution of the game where the actions are sampled from the fixed belief $\boldsymbol{\theta}_n$ at period n. Next, each player uses the perturbed action values to obtain the smoothed best response function $\tilde{B}^i(\tau;\boldsymbol{\theta}_n)$ in (4.3). Then, each player uses the smoothed best response function $\tilde{B}^i_{\mathbf{s}_n}(\tau;\boldsymbol{\theta}_n)$ from state $\mathbf{s}_n$ at period n to sample and play their action a^i_n, which generates the action profile $\mathbf{a}_n$. Finally, the action profile $\mathbf{a}_n$ feeds into the belief through (4.4), and the state of the game evolves according to $p(\mathbf{s}_{n+1} \,|\, \mathbf{s}_n, \mathbf{a}_n)$. The above is repeated ad infinitum.

This model of belief learning is not compatible with the case of imperfect public monitoring because the learning model uses the action to update the belief. However, a key feature of the case of imperfect public monitoring is that opponents' actions cannot be observed nor inferred. Nonetheless, we ignore this inconsistency as an exercise to see how the model learns when behavior is conditioned on a noisy public signal instead of a perfect public signal.

Reinforcement Learning

We use Q-learning as a model of reinforcement learning. The algorithm stems from the machine learning literature and its popularity has grown in the algorithmic collusion literature. However, little is known about its economic interpretation. The exception is Beggs (2022) who provides an economic interpretation of temporal difference algorithms through reference points and recursive preferences.

Q-learning is a model-free algorithm that is part of the family of temporal difference algorithms. The algorithm dynamically estimates the action values, and it uses the estimate of the action values in its choice rule. The action values, also known as Q-values, are estimated through a Bellman-like learning rule given by

$$Q^i_{n+1}(a \,|\, \mathbf{s}) = Q^i_n(a \,|\, \mathbf{s}) + \gamma_{n+1} \left(u^i(\mathbf{a}_n) + \delta \max_{a'} Q^i_n(a' \,|\, \mathbf{s}_{n+1}) - Q^i_n(a \,|\, \mathbf{s}) \right), \tag{5.2}$$

when $a^i_n = a, \mathbf{s}_n = \mathbf{s}$, and $Q^i_{n+1}(a \,|\, \mathbf{s}) = Q^i_n(a \,|\, \mathbf{s})$ otherwise.

The choice rule is a logit probability function, so the strategy parameterized by Q-values

is given by

$$\sigma_Q^i(a\,|\,\mathbf{s}) = \frac{\exp\left(\tau^{-1}\,Q^i_{a|\mathbf{s}}\right)}{\sum_{a'\in\mathcal{A}_i}\exp\left(\tau^{-1}\,Q^i_{a'|\mathbf{s}}\right)}, \tag{5.3}$$

for all $a \in \mathcal{A}_i$ and $\mathbf{s} \in \mathcal{S}$.[1]

Q-learning is an off-policy algorithm, where the behavioral strategy is different from the behavior used to update (5.2). The behavioral strategy in (5.3) is a fully mixed strategy, whereas the estimate from the optimal future value $\max_{a'} Q^i_n(a'\,|\,\mathbf{s}_{n+1})$ follows a pure strategy. Nonetheless, in the single agent setting, Q-learning is known to converge to the optimal strategy (see for example Singh et al., 2000).

Next, we confirm that Assumption A holds for Q-learning.

Proposition 4. *Let* $\hat{u} = \sup_{\mathbf{a},i}\left|u^i(\mathbf{a})\right|$. *If* $\max \boldsymbol{Q}_0 \le \hat{u}/(1-\delta)$, *then Assumptions A.2, A.4, and A.5 hold.*

Proof. By assumption, initialize $\boldsymbol{Q}_0$ so that $\max \boldsymbol{Q}_0 \le \hat{u}/(1-\delta)$. Without loss of generality, assume that $Q^i(a\,|\,\mathbf{s})$ updates in every iteration. We verify Assumption A.2 by induction. We show that $\max \boldsymbol{Q}_n \le \hat{u}/(1-\delta)$ implies that $\max \boldsymbol{Q}_{n+1} \le \hat{u}/(1-\delta)$. Use the learning rule to write

$$\begin{aligned}\left|Q^i_{n+1}(a\,|\,\mathbf{s})\right| &= \left|(1-\gamma_{n+1})Q^i_n(a\,|\,\mathbf{s}) + \gamma_{n+1}\,u^i(\mathbf{a}_n) + \gamma_{n+1}\,\delta\,\max_{a'} Q^i_n(a'\,|\,\mathbf{s}')\right| \\ &\le (1-\gamma_{n+1})\,\frac{\hat{u}}{1-\delta} + \gamma_{n+1}\,\hat{u} + \gamma_{n+1}\,\delta\,\frac{\hat{u}}{1-\delta} = \hat{u}/(1-\delta)\,,\end{aligned}$$

which verifies Assumption A.2. Assumption A.4 is satisfied because the logit probability function is Lipschitz in Q. Finally, f^i is Lipschitz because the maximum operator and the coordinate mapping $Q \mapsto Q^i(a\,|\,\mathbf{s})$ are both Lipschitz. Hence, Assumption A.5 is satisfied. □

With Assumption A verified, the algorithmic learning equations for Q-learning is given by

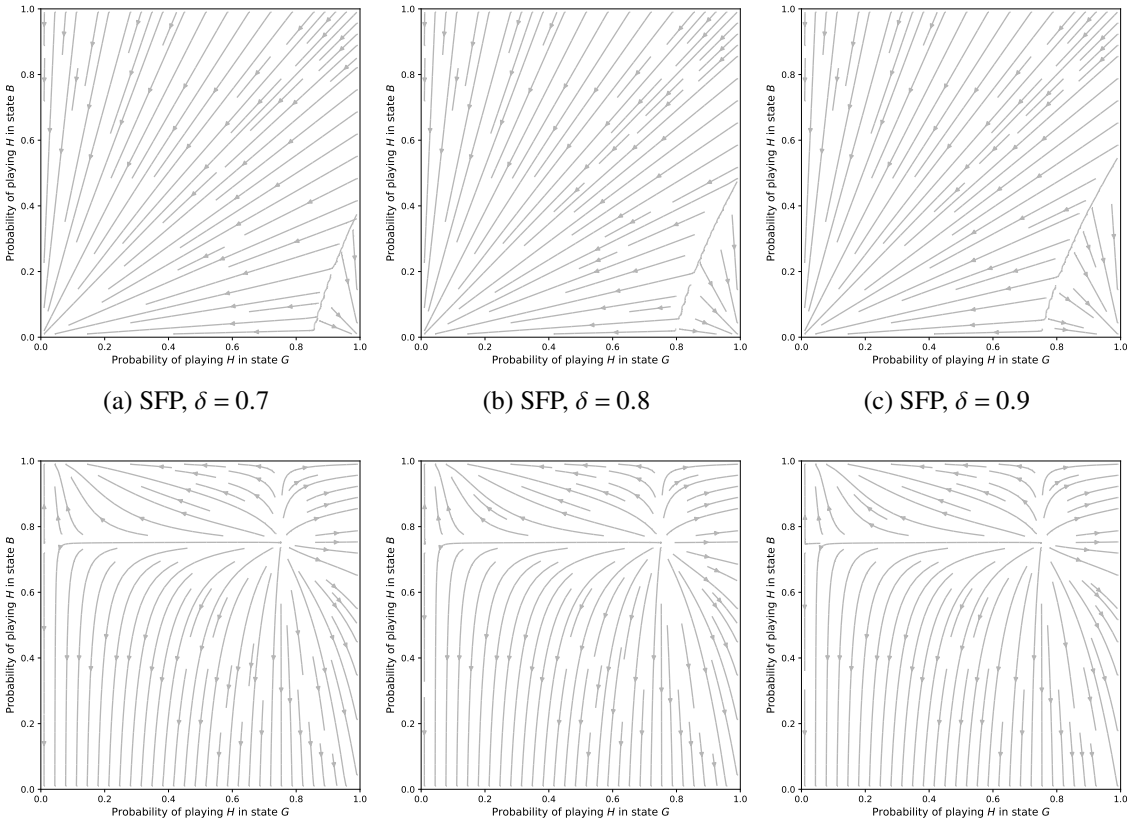
$$\begin{aligned}\dot{\bar{Q}}^i_{a|\mathbf{s}}(t) = \sigma^i_{\bar{Q}(t)}(a\,|\,\mathbf{s})\,\Gamma_{\bar{Q}(t)}(\mathbf{s}) \sum_{\substack{\mathbf{s}'\in\mathcal{S}\\ a^{-i}\in\mathcal{A}_{-i}}} p(\mathbf{s}'\,|\,\mathbf{s},(a,a^{-i}))\prod_{j\neq i}\sigma^j_{\bar{Q}(t)}(a^j\,|\,\mathbf{s}) \\ \times\left(u^i(a,a^{-i}) + \delta\,\max_{a'}\bar{Q}^i_{a|\mathbf{s}'}(t) - \bar{Q}^i_{a|\mathbf{s}}(t)\right),\end{aligned} \tag{5.4}$$

for each $0 < i \le I$, $a \in \mathcal{A}_i$, and $\mathbf{s} \in \mathcal{S}$.

[1] Another popular choice rule is the ϵ-greedy rule: with probability $1-\epsilon$ play the action with the highest action value, with probability ϵ all actions are equally likely to be played. We use the logit choice rule because it satisfies Assumption A.4.

5.3 Results

We numerically solve the algorithmic learning equations for both learning models to analyze the learning outcomes. For the dynamic generalization of smooth fictitious play, the field plots illustrate the evolution of the empirical frequency of play. On the other hand, for Q-learning, we solve the Q-values according to (5.4), but we visualize the corresponding evolution of the parameterized strategy. For both learning models, we impose symmetry in the parameter values of the players to visualize the dynamics in two-dimensions.[2] In effect, we study the evolution of symmetric strategies.



(a) SFP, $\delta = 0.7$ (b) SFP, $\delta = 0.8$ (c) SFP, $\delta = 0.9$

(d) Q-learning, $\delta = 0.7$ (e) Q-learning, $\delta = 0.8$ (f) Q-learning, $\delta = 0.9$

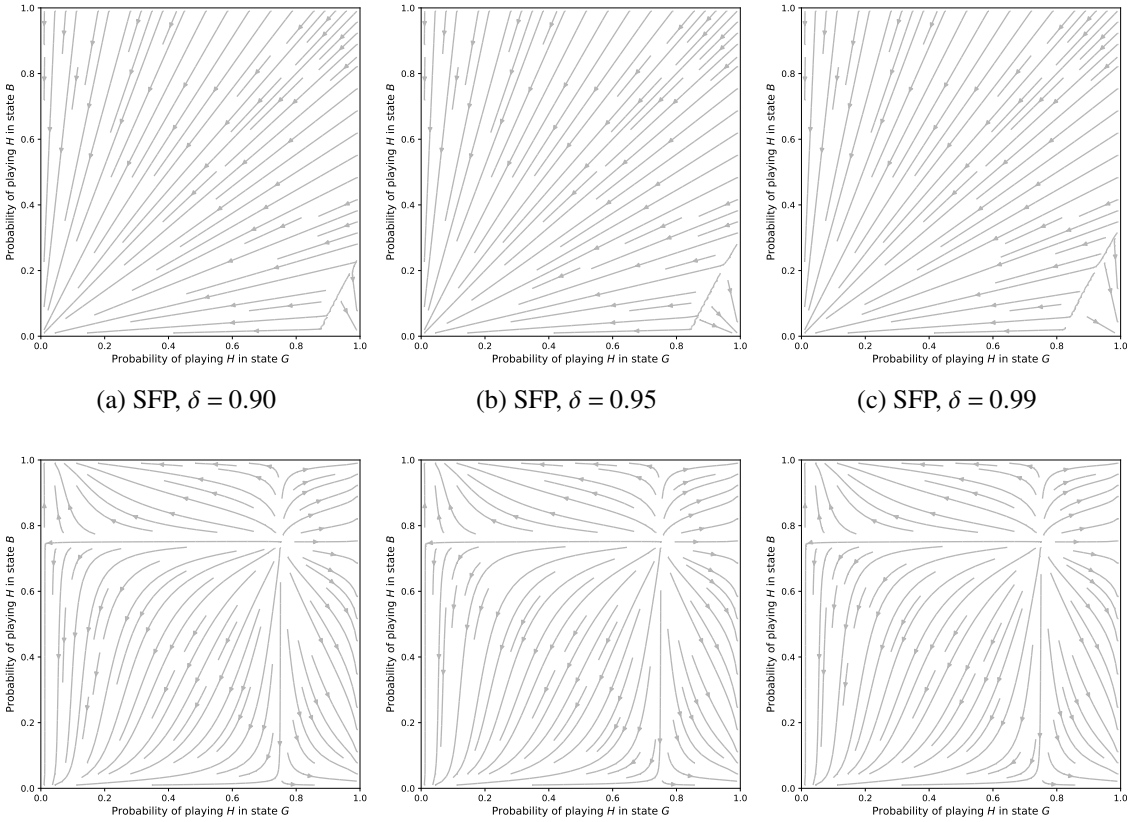
Figure 5.1: Perfect monitoring.

Figure 5.1 visualizes the dynamics for the case of perfect monitoring. We see that (SFP) only converges to one of two equilibrium strategies: the static Nash equilibrium (bottom left corner), or the one-memory Nash-reversion strategy (bottom right corner). On the other

[2]For Q-learning, the initial conditions are set in terms of the strategy, but the trajectories of the ODEs are solved in terms of Q-values. Thus, for a given initial strategy, we use (5.3) without the denominator to obtain the initial Q-values. For the initial conditions we consider, this procedure satisfies the assumption of the initial Q-values in Proposition 4.

hand, we see that Q-learning converges to any one of the pure strategies: two equilibrium strategies (bottom corners), and two non-equilibrium strategies (top corners).[3]

Similarly, Figure 5.2 visualizes the dynamics for the case of imperfect public monitoring. Again, (SFP) only converges to one of two equilibrium strategies: the static Nash equilibrium (bottom left corner), or the one-memory trigger price strategy (bottom right corner); whereas Q-learning converges to any one of the pure strategies: two equilibrium strategies (bottom corners), and two non-equilibrium strategies (top corners).



(a) SFP, $\delta = 0.90$ (b) SFP, $\delta = 0.95$ (c) SFP, $\delta = 0.99$

(d) Q-learning, $\delta = 0.90$ (e) Q-learning, $\delta = 0.95$ (f) Q-learning, $\delta = 0.99$

Figure 5.2: Imperfect public monitoring.

Convergence of (SFP) to an equilibrium strategy under both monitoring structures is expected because the underlying game is a potential game, and from Theorem 4, we know that (SFP) will converge to a ϵ-subgame perfect equilibria in potential games. Interestingly, we see that the basin of attraction of the collusive equilibrium increases as the value of the common discount factor δ increases. Combining this observation together with Theorem 7, we conclude that (SFP) is more likely to learn to collude when players become more patient.

[3]The non-equilibrium strategies here are related to experience based equilibria in Asker et al. (2023).

This result follows because the trajectory is more likely to stumble into the basin of attraction of the collusive equilibrium.

On the other hand, the basins of attraction for the dynamics of Q-learning appear to be invariant to the discount factor and the monitoring structure. This is likely due to the reinforcement nature of Q-learning because the expected utility is the same under both monitoring structures and the discount factor simply acts as a scaling factor when estimating Q-values.

Despite the idiosyncrasies of Q-learning, there are two interesting observations we obtain when comparing Q-learning to (SFP). First, the basin of attraction of the collusive equilibrium from Q-learning is significantly larger than that of (SFP). Thus, Q-learning is more likely to learn to collude when compared to (SFP). Second, Q-learning has large basins of attraction to non-equilibrium strategies. Therefore, some of the learning outcomes reported for Q-learning in the literature may be from non-equilibrium behavior. Indeed, this is consistent with recent numerical simulations of Lambin (2023), Abada et al. (2024), and Epivent and Lambin (2024), who find various non-equilibrium behavior from Q-learning.

Chapter 6

Anonymity and Signaling

6.1 Background

Anonymity is important for strategic interactions in financial markets. In limit order books, anonymity enables informed investors to blend in with uninformed investors (see Kyle, 1985), it facilitates investors to hide their trading intentions to prevent back-running and predatory trading (see Yang and Zhu, 2019; Brunnermeier and Pedersen, 2005), and it enables market makers to conceal their inventory positions (see Biais, 1993). Importantly, anonymity improves market quality (see Comerton-Forde et al., 2005). Therefore, why would HFTs reveal themselves to each other? What happens when HFTs break the pre-trade anonymity of limit orders and reveal themselves to each other?

In Cartea et al. (2025), we use proprietary data from Euronext Amsterdam to study the anonymity of the limit order book in the ETF market. We find that HFTs knowingly or unknowingly signal themselves to each other. HFTs signal themselves through

- limit orders with volumes that are very large compared with the size of limit orders sent by other market participants and compared with the size of transactions, and
- by adding trailing digits to their limit orders.

The consequence is that signaling breaks the anonymity of limit orders and it introduces an artificial form of pre-trade transparency that creates a different playing field for a subset of market participants. Cartea et al. (2025) show that the trading behavior of the HFTs depends on the identity of the counterparty. Specifically, spread-improving limit orders that are close to crossing the quoted spread are very likely to be sniped (62.58% probability) if the limit orders originate from a retail trader, and very unlikely to be sniped (0.08% probability) if the limit orders originate from an HFT.[1]

[1] A spread-improving limit order is sniped if it enters inside the quoted spread, and all or part of it is taken

Industry Response

In response to the first draft of our article, industry practitioners have publicly confirmed our hypothesis that the limit order book (designed to be anonymous) is no longer anonymous (see Clancy and Cesa, 2025):

> Traders adding trailing digits to their limit orders is a well-established and legitimate practice in the European ETF market.
>
> *Matthijs Pars – Association of Proprietary Traders*

> Market makers use [trailing digits] to track their own limit orders.
>
> *Matthijs Pars – Association of Proprietary Traders*

> In some markets it is almost seen as a "professional courtesy" for large traders to identify themselves by their signature quantities.
>
> *Head of surveillance at a large US bank*

> Confirms that liquidity providers used distinct order quantities to recognise their own orders – at least "in the early days". This allowed them to "see in the blink of an eye how you are ordering".
>
> Then, they began to recognise the handles of rival firms and began factoring this information into their own strategies – for instance, adjusting their pricing, or not, in response to what competitors were doing.
>
> Market-makers have since found other uses for these signals, such as filtering out retail orders when fitting their volatility curves to the market.
>
> The practice of signalling continues [...] because it does not breach any rules and provides valuable information to market makers.
>
> *A trader that has worked at two large market making firms*

In an anonymous limit order book, revealing yourself to others is a voluntary and costly decision. Since rational agents do not give information for free, the question is then: what are the offsetting benefits for market makers?

out by an aggressive order within 1 millisecond.

Overview

In this Chapter, we propose a model of the limit order book that considers competitive and collusive equilibria. Our model rationalizes the behavior observed in Cartea et al. (2025) and highlights the economic forces that underlie the incentives of market makers, i.e., HFTs, to reveal themselves to each other. Our model also answers the following questions: under what conditions would an HFT reveal herself to others? How do HFTs mutually benefit from revealing themselves, and how can they collectively enforce this behavior?

The model consists of impatient investors who take liquidity, patient investors who provide liquidity, and strategic HFTs who play a repeated trading game. The trading game is a generalization of the trading game in Budish et al. (2024), and it retains the main elements of latency arbitrage and adverse selection. In our model, one of the HFTs may receive short-lived private information in each trading game (see Foucault et al., 2016, for a model where HFTs trade on short-lived information), which differs from the usual high-frequency trading models with informed traders (see Baldauf and Mollner, 2020; Budish et al., 2024). The primary generalization in our model is the arrival of patient investors who send limit orders that improve the quoted spread.[2] This introduces strategic ambiguity because market participants do not know *who* sent the spread-improving limit order. A limit order posted inside the spread could be benign flow from patient investors, or it could be toxic flow from a privately informed HFT who is pretending to be a patient investor.

6.2 Setup

Fundamental Value Consider security x whose fundamental value is given by y, and at the end of each trading game, x can be liquidated at this fundamental value with zero cost, i.e., no frictions or fees. The value y evolves across trading games as a discrete-time jump process. In each trading game, if there is an innovation in y, then y jumps up or down with equal probability, and J is the distribution of the size of the jumps. The price grid is continuous to abstract from the queuing dynamics when non-zero tick sizes are a binding constraint.

Impatient Investor In each trading game, an impatient investor arrives stochastically with an inelastic need to buy or to sell one unit of x. Impatient investors arrive with probability p_{LT}, and are equally likely to buy or to sell a unit of x. On arrival, impatient

[2] Our patient investors are different from the slow trading firms in Budish et al. (2024) that have no intrinsic need to buy or to sell. Our patient investors, like impatient investors, have a need to buy or to sell one unit of the asset, but are unwilling to cross the half-spread.

investors transact immediately with a marketable limit order.[3] Impatient investors are the usual "liquidity traders" in Glosten and Milgrom (1985) or "noise traders" in Kyle (1985). These participants include mutual funds, pension funds, hedge funds, retail traders, etc. In the Euronext dataset, the majority of impatient investors are retail traders.

Patient Investor In each trading game, a patient investor arrives stochastically with an elastic need to buy or to sell one unit of x. Patient investors arrive with probability p_{LO}, and are equally likely to buy or to sell one unit of x. Patient investors care only about buying or selling x at their reservation price, and they do so by submitting limit orders inside the quoted spread. Patient investors are similar to the execution algorithms in Li et al. (2021). However, the key differences are that execution algorithms have an inelastic demand and are strategic, while patient investors have an elastic demand and are not strategic.[4] Patient investors, as seen in the dataset, are the retail traders who send limit orders that improve, but do not cross, the quoted spread. For analytical tractability, patient investors cancel their unexecuted limit orders at the end of the trading game.

If a patient investor arrives to buy one unit of the security, then he sends a buy limit order with a limit price of $y+\delta$, so he is content with buying at any price below or equal to $y+\delta$. Similarly, if a patient investor arrives to sell one unit of the security, then he sends a sell limit order with a limit price of $y-\delta$, so he is content with selling at any price above or equal to $y-\delta$. The value $\delta \in (0, s/2)$ is the improvement in the quoted spread by a patient investor. We focus on the case where δ is positive, so the reservation price is mispriced because the limit order improves the best bid or best offer beyond the fundamental value of the security. The improvement value δ is bounded above by the half-spread $s/2$ set by the HFTs, otherwise the limit order gets executed as a marketable limit order, in which case patient investors are subsumed as impatient investors.

HFTs There are $N \geq 3$ HFTs who make and take liquidity, and who are present throughout all iterations of the trading game. HFTs are risk neutral and they have no intrinsic need to buy or to sell x. They buy x at prices lower than y and sell x at prices higher than y to maximize profits. Furthermore, they discount future payoffs with the common discount factor $\rho \in [0, 1)$.

[3] Marketable limit orders are limit orders with a bid price weakly greater than the best ask (if buying), or an ask price weakly lower than the best bid (if selling).

[4] Although patient investors are not strategic, Section 6.5.3 analyzes how the behavior of patient investors affects the equilibrium outcomes.

Public and Private Information The probability that there is a jump in y that is public information and seen by all market participants at the same time is p_{public}, while the probability that there is a jump in y seen only by HFT i is p^i_{private}. HFTs are equally likely to receive private information, so $p^i_{\text{private}} = p_{\text{private}}/N$, where p_{private} is the probability that there is private information.

If an HFT observes private information, she can act on that information in the current trading game. Regardless of her actions, any private information becomes public at the end of the trading game. Given that ETFs are indices, it is unlikely that informed traders can acquire private information about all the constituents of an ETF. Therefore, the usual assumption that informed traders possess long-lived private information is unlikely to hold in an ETF market. Here, information is "news" as in Foucault et al. (2016), where every quote update or trade in any exchange is a source of information for the fundamental value y. Therefore, short-lived private information of HFTs is a consequence of market fragmentation. HFTs are present in multiple, but not all, exchanges and integrate information across markets (see Baron et al., 2019; Brogaard et al., 2019), but importantly, HFTs are not always connected to the same set of exchanges (we verified this with MiFID II data). If an HFT is not connected to a venue, then she can infer that there is private information from the behavior of an HFT connected to that venue, or wait until the information becomes public once she receives the data from a third-party data vendor who processes and disseminates the data.

Latency HFTs operate with no latency. There is no delay in sending or receiving updates from the exchange. When multiple messages reach the exchange at the same time, there is a random tie-break to decide which message is processed first. In contrast, investors are slow, so when they race against HFTs to send a message to the exchange, the HFTs always win.

6.3 Timing of the Trading Game

Our trading game consists of four periods, and the four-period trading game is repeated infinitely many times. At the start of each trading game, there is a publicly observed state (y, ω), which consists of the fundamental value y, and of the outstanding bids and asks in the limit order book ω. At the first instance of the trading game, the initial fundamental value is y_0, and the order book is empty. In all subsequent instances of the trading game, the state is determined at the conclusion of the previous trading game, and ω consists of all outstanding limit orders.

The four-period trading game proceeds sequentially as follows:

1. **Period 1:** HFTs observe the state (y, ω) and send instructions to the exchange. These instructions are messages to submit limit orders, cancel existing limit orders, submit marketable limit orders, and submit immediate-or-cancel orders. Formally, o^i is the set of messages sent by HFT i and $o^i \in O$, where O is the set of all possible combinations of messages. Messages are processed by the exchange and ω is updated.

2. **Period 2:** Nature moves and selects one of two possibilities:

   (i) With probability p_{LO}, a patient investor, who is equally likely to buy or to sell a unit of x, arrives and sends a limit order inside the quoted spread at his reservation price.

   (ii) With probability $p_{\text{private}} = 1 - p_{LO}$, nature releases private information to one HFT. The informed HFT has an opportunity to send instructions to the exchange.

3. **Period 3:** If a limit order arrived in the second period, then the HFTs have an opportunity to snipe the limit order. Afterwards, HFTs have an opportunity to send instructions to the exchange.

4. **Period 4:** Depending on nature's draws in period two, nature either does nothing or moves again. If nature selected an informed HFT in the second period, then nature does nothing. Otherwise, nature moves and selects one of three possibilities:

   (i) With probability p_{LT}, an impatient investor arrives and he is equally likely to buy or to sell a unit of x.

   (ii) With probability p_{public}, there is a publicly observable jump in y. HFTs participate in latency arbitrage as in Budish et al. (2015).

   (iii) With probability $p_N = 1 - p_{LT} - p_{\text{public}}$, there is no event.

   Finally, at the end of each trading game, any outstanding limit order from a patient investor is canceled.

Figure 6.1 illustrates nature's actions and their associated probabilities. Within a single trading game, nature can call upon both a patient investor and an impatient investor. This captures the incentives of HFTs to snipe limit orders from patient investors who improve the quoted spread. If HFTs do not snipe limit orders from patient investors, then these limit orders preclude HFTs from earning the half-spread on one side of the book in the event an impatient investor arrives in period four.

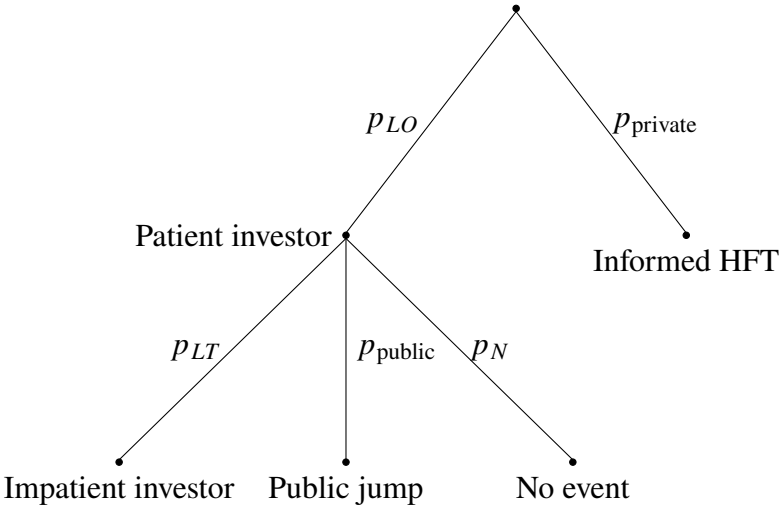


Figure 6.1: Nature's actions in each trading game.

6.4 Competitive Equilibrium

For the competitive equilibrium in our infinitely repeated trading game, we restrict our attention to pure strategies where market participants condition their actions on the state (y, ω) and update their beliefs based on Bayes' rule whenever possible. The relevant beliefs are about *who* arrives in period two. We analyze each trading game in isolation, and then show that repeated play of the equilibrium for a single trading game remains an equilibrium for the infinitely repeated trading game.

6.4.1 Solution Concept

With the restriction to pure strategies, one solution concept would be a pure strategy weak perfect Bayesian equilibrium (WPBE). However, a pure strategy WPBE does not exist. As in Budish et al. (2024), excess liquidity (i.e., liquidity that supplies a quantity greater than the one unit required by the impatient investor) exposes the liquidity provider to adverse selection and latency arbitrage, but without the benefits of liquidity provision. Thus, there are no other liquidity providers willing to provide excess liquidity to constrain the spread, so the liquidity provider has a profitable deviation by widening the spread.

We extend the order book equilibrium from Budish et al. (2024) to restore the existence of equilibrium. As in Budish et al. (2024), a Bayesian order book equilibrium (BOBE) strictly weakens a WPBE by allowing for profitable unilateral deviations to exist, provided that these deviations are rendered unprofitable by one of two specific reactions by rivals: withdrawal of liquidity (canceling limit orders), or safe profitable price improvements (of

limit orders). Withdrawals are message sets that strictly reduce the amount of liquidity provided relative to a particular candidate equilibrium message set $\mathbf{o} = \{o^i\}_{0<i\leq N}$. Price improvements are message sets that, relative to $\mathbf{o}$, reduce the cost to trade. A safe profitable price improvement is a price improvement that is strictly profitable and it remains profitable even if some other HFT withdraws liquidity in response. The definition of an BOBE is given below.

Definition 7. *A Bayesian order book equilibrium of our trading game is a message set* $\mathbf{o} = \{o^i\}_{0<i\leq N}$ *submitted by all HFTs in period 3 given state* (y, ω) *such that:*

1. *No HFT i has a safe profitable price improvement.*

2. *No HFT i has any other strictly profitable deviation (that is not a price improvement) that remains strictly profitable if, in response to HFT i's deviation, some other HFT engages in a profitable reaction that is either a withdrawal of liquidity or a safe profitable price improvement.*

The BOBE captures the spirit of competitive liquidity provision as discussed and used in Glosten and Milgrom (1985). The solution concept ensures that even if excess liquidity is not provided in equilibrium, the presence of other potential liquidity providers will discipline equilibrium price levels.[5] For a more detailed discussion, see Budish et al. (2024).

6.4.2 Equilibrium Behavior of HFTs

In the first period, the weakly dominant strategy is for one of the HFTs to maintain two-sided quotes. One quote to buy one unit of x at price $y - s/2 - \varepsilon/2$, and the other quote to sell one unit of x at price $y + s/2 + \varepsilon/2$, where $s \geq 0$ is the HFT's bid-ask spread in period three that is solved through an equilibrium indifference condition. The small and positive constant $\varepsilon \to 0$ ensures that there are incentives for someone to undercut the quotes in period three, so the HFT can withdraw liquidity in period three (without raising suspicion) in the event she becomes informed. Individually, each HFT strictly prefers to become an informed trader upon receiving private information in the second period, so none of the HFTs set quotes with $\varepsilon = 0$ in period one.

[5] Baldauf and Mollner (2020) refer to these potential liquidity providers as enforcers.

Informed HFT In the second period, if an HFT receives private information, then the weakly dominant strategy is to pretend to be a patient investor (when profitable) by submitting a toxic limit order for one unit of x in the same direction as the jump. Specifically, if $y' > y + \delta$, then the informed HFT sends a buy limit order with a price of $y + \delta$; similarly, if $y' < y - \delta$, then the informed HFT sends a sell limit order with a price of $y - \delta$. On the other hand, if the jump size does not exceed the improvement level δ, then the informed HFT does not submit a limit order because it is not profitable to do so.[6]

In the fourth period, if the value y' is not profitable to trade (i.e., $|y' - y| \leq s/2$), then the informed HFT does nothing. However, if the value y' is profitable to trade (i.e., $|y' - y| > s/2$), then the informed HFT adversely selects the market maker and earns $|y' - y| - s/2$.[7]

Uninformed HFTs In the second period, if HFT i does not receive private information, then Nature either called upon a patient investor with probability p_{LO}, or released private information to another HFT with probability $p^{-i}_{\text{private}} = \sum_{j \neq i} p^{j}_{\text{private}} = \frac{N-1}{N} p_{\text{private}}$. There are two cases to consider as an uninformed HFT in period two: (i) no limit order arrives, or (ii) a limit order arrives.

In the first case where no limit order arrives, it means that Nature released private information to another HFT, but the jump size did not exceed the improvement level δ. If the jump size does not exceed the improvement level δ, then the jump size cannot exceed $s/2$. In this case, all HFTs do nothing and earn nothing.

In the second case where a limit order arrives, the probability that an uninformed HFT observes a limit order arrive in period two is $\alpha = p_{LO} + p^{-i}_{\text{private}} \mathbb{P}(J > \delta)$. Applying Bayes' rule to this information set, where a limit order arrives, the probability of the limit order originating from a patient investor is p_{LO}/α, while the probability of the limit order originating from an informed HFT is $p^{-i}_{\text{private}} \mathbb{P}(J > \delta)/\alpha$.

If a limit order arrives in the second period, then, in the third period, the uninformed HFTs have an opportunity to snipe the limit order, or they can wait until the fourth period

[6] In equilibrium, the uninformed HFTs know that if the order originates from an informed HFT, then the jump size exceeds the improvement level δ. However, for the uninformed HFT who provides liquidity in period three, the cost of adverse selection remains the same even if the informed HFTs do not send toxic orders. Therefore, the weakly dominant strategy of the informed HFT is always to send a toxic limit order whenever it is profitable to do so.

[7] Although the informed HFT receives private information in the second period, the dominant strategy in equilibrium is to wait until the fourth period to adversely select the market maker. Adversely selecting the market maker in the second period will make other HFTs aware of an informed HFT, which will inhibit the informed HFT's ability to profit from providing liquidity with a toxic limit order by pretending to be a patient investor. Additionally, as the quotes will tighten by ε at the end of period three, it is more profitable to adversely select the market maker in the fourth period.

when there is a public innovation in y to ensure that the limit order is from a patient investor so that it is safe and profitable to trade against. Afterwards, at the end of period three, the HFT who provided liquidity in period one cancels her quotes, and the uninformed HFTs endogenously sort themselves into one of two roles. As in Budish et al. (2024), one uninformed HFT (who may or may not be the same HFT providing liquidity in period one) takes the role of a market maker and the remaining uninformed HFTs take the role of a latency arbitrageur.

First, consider the case where at least one uninformed HFT chooses to snipe in the third period whenever a limit order arrives in the second period. If the limit order is from a patient investor, then the uninformed HFTs (who choose to snipe) have an equal probability of executing against the mispriced limit order to earn an expected profit of δ. However, if the limit order is from an informed HFT, then the informed HFT earns a profit of $|y' - y| - \delta$, while the uninformed HFT who sniped the informed HFT's toxic limit order will incur a loss of $|y' - y| - \delta$.

With the limit order from period two cleared out, the trading game becomes the two-period trading game in Budish et al. (2024), but with event probabilities given by the updated beliefs. Specifically, at the end of period three, one uninformed HFT will make two-sided quotes. One quote to buy one unit of x at price $y - s/2$, and one quote to sell one unit of x at price $y + s/2$. In period four, if the public jump to the value y' is not profitable to trade (i.e., $|y' - y| \leq s/2$), then nothing happens. However, if the value y' is profitable to trade (i.e., $|y' - y| > s/2$), then a latency arbitrage race occurs.

In the latency arbitrage race in period four, the market maker races to cancel her unprofitable stale quote. The market maker will successfully cancel her quote with probability $1/N$ and lose nothing, and she will unsuccessfully cancel her quote with probability $(N-1)/N$ and lose $|y' - y| - s/2$. Simultaneously, the latency arbitrageur races to pick up the market maker's stale quote. The latency arbitrageur will win the race with probability $1/N$ and earn $|y' - y| - s/2$, and she will lose the race with probability $(N-1)/N$ and earn nothing.

Next, consider the case when all HFTs wait until there is a public innovation in y in the fourth period to decide whether to trade against the limit order posted in period two. In this case where the limit order from period two is not cleared out in period three, the uninformed HFT will only provide liquidity on one side of the book. Providing liquidity on both sides of the book means that one of her orders does not receive the benefits of liquidity provision, but it is exposed risk of latency arbitrage. Therefore, if a buy limit order arrives in period two and remains unexecuted by the end of period three, then the market maker will provide a quote to sell one unit of x at price $y + s/2$; similarly, if a sell limit order arrives in period

two and remains unexecuted by the end of period three, then the market maker will provide a quote to buy one unit of x at price $y - s/2$.

In period four, if the public jump to the value y' is such that the market maker's quote is stale, then a latency arbitrage race occurs between the HFTs. On the other hand, if the value y' is such that the limit order from period two is stale, then all the HFTs attempt to latency arbitrage the patient investor. All HFTs have an equal probability of trading against the limit order from the patient investor to earn $|y' - y| + \delta$. Finally, if the value y' is such that no quotes are stale, then nothing happens and the limit order from the patient investor is canceled.

In what follows, we characterize the HFT's bid-ask spread in period three, and provide the conditions to determine whether an HFT snipes the limit order in period three, or waits until there is an innovation in y in period four.

6.4.3 Equilibrium Analysis

To analyze the equilibrium, we first solve the equilibrium bid-ask spread set by uninformed HFTs at the end of period three, and then we solve the equilibrium of whether uninformed HFTs snipe in the third period or not. First, we introduce some notation. Let

$$L(x) = \mathbb{P}(J > x)\, \mathbb{E}[J - x \mid J > x] \quad \text{and} \quad \tilde{L}(x) = \mathbb{P}(J < x)\, \mathbb{E}[x - J \mid J < x]$$

denote the expected payoffs associated with latency arbitrage. Furthermore, let

$$L_\delta(x) = \mathbb{P}(J_\delta > x)\, \mathbb{E}[J_\delta - x \mid J_\delta > x]$$

denote the expected loss from adverse selection, where $J_\delta \sim J | J > \delta$ is the distribution of the size of the jump given that the size of the jump is greater than δ. Finally, let d^i denote the binary decision variable of HFT i sniping in the third period. If HFT i snipes in the third period, then $d^i = 1$, otherwise $d^i = 0$.

In equilibrium, the bid-ask spread set by uninformed HFTs at the end of period three leaves them indifferent between the role of a market maker or that of a latency arbitrageur given their updated beliefs.

Uninformed HFTs do not snipe in period three The market maker has fewer opportunities to provide liquidity (the market maker only provides liquidity on one side of the book). Therefore, latency arbitrageurs have fewer opportunities to latency arbitrage the market maker. However, because the uninformed HFTs do not snipe in period three, then, conditional on a public innovation of y in the appropriate direction, all the uninformed HFTs (including the market maker) have the opportunity to latency arbitrage the outstanding limit order from the patient investor in period four for a profit of $L(-\delta) + \tilde{L}(\delta)$.

At least one uninformed HFT snipes in period three The market maker provides liquidity on both sides of the book, so latency arbitrageurs can latency arbitrage the market maker on both sides of the book. If an uninformed HFT decides to snipe in period three, then she takes a share of the limit orders from patient investors at an expected payoff of δ, but simultaneously, she is also exposed to trading with toxic limit orders from informed HFTs, who pretend to be a patient investor, at a loss of $L(\delta)$.

Therefore, conditional on HFT i being uninformed and observing a limit order arrive in period two, her expected payoff acting as a market maker in period three is

$$\begin{aligned}\frac{1}{\alpha}\Bigg[& \underbrace{\mathbb{1}_{\left\{\sum_{j=1}^{N} d^j=0\right\}} p_{LO}\,\frac{p_{\text{public}}}{2N}\left(L(-\delta)+\tilde{L}(\delta)\right)}_{\text{latency arbitraging patient investors in period four}} + \underbrace{d^i\left(\frac{p_{LO}}{\sum_j d^j}\,\delta - \sum_{j\neq i}\frac{p^j_{\text{private}}}{\sum_{k\neq j} d^k}\,L(\delta)\right)}_{\text{sniping in period three}} \\ & + \underbrace{\mathbb{1}_{\left\{\sum_{j=1}^{N} d^j=0\right\}}\frac{1}{2}\,p_{LO}\left(p_{LT}\,\frac{s}{2} - p_{\text{public}}\,\frac{N-1}{N}\,L\left(s/2\right)\right)}_{\text{liquidity provision on one side of the book}} \\ & + \underbrace{\mathbb{1}_{\left\{\sum_{j=1}^{N} d^j>0\right\}} p_{LO}\left(p_{LT}\,\frac{s}{2} - p_{\text{public}}\,\frac{N-1}{N}\,L\left(s/2\right)\right)}_{\text{liquidity provision on both sides of the book}} \\ & - \underbrace{\sum_{j\neq i} p^j_{\text{private}}\,\mathbb{P}\left(J>\delta\right)\,L_\delta\left(s/2\right)}_{\text{adverse selection}}\Bigg].\end{aligned} \tag{6.1}$$

All the terms are pre-multiplied by $1/\alpha$ to account for the updated beliefs about who arrived in period two. The second term (sniping in period three) includes a double sum because when the informed HFT pretends to be a patient investor, the cost is shared equally among all uninformed HFTs who choose to snipe in period three. However, this cost excludes the informed HFT because she does not snipe her own order. The last term (adverse selection cost) accounts for the fact that when informed HFTs send toxic orders, it means that the size of the jump exceeds the improvement level δ. The last term can be rewritten as $\sum_{j\neq i} p^j_{\text{private}}\,L\left(s/2\right)$ because $\mathbb{P}\left(J>\delta\right)\,L_\delta\left(s/2\right) = L\left(s/2\right)$, which demonstrates that the cost of adverse selection remains the same even if the informed HFTs do not send toxic orders.

Next, conditional on HFT i being uninformed and observing a limit order arrive in

period two, her expected payoff acting as a latency arbitrageur in period three is

$$\frac{1}{\alpha}\Bigg[\underbrace{\mathbb{1}_{\left\{\sum_{j=1}^{N} d^j=0\right\}}\, p_{LO}\,\frac{p_{\text{public}}}{2N}\left(L(-\delta)+\tilde{L}(\delta)\right)}_{\text{latency arbitraging patient investors in period four}}+\underbrace{d^i\left(\frac{p_{LO}}{\sum_j d^j}\,\delta-\sum_{j\neq i}\frac{p^j_{\text{private}}}{\sum_{k\neq j} d^k}\,L(\delta)\right)}_{\text{sniping in period three}}$$

$$+\quad\underbrace{\mathbb{1}_{\left\{\sum_{j=1}^{N} d^j=0\right\}}\,\frac{1}{2}\,p_{LO}\,p_{\text{public}}\,\frac{1}{N}\,L\left(s/2\right)}_{\text{latency arbitraging the market maker on one side of the book}} \tag{6.2}$$

$$+\quad\underbrace{\mathbb{1}_{\left\{\sum_{j=1}^{N} d^j>0\right\}}\,p_{LO}\,p_{\text{public}}\,\frac{1}{N}\,L\left(s/2\right)}_{\text{latency arbitraging the market maker on both sides of the book}}\Bigg],$$

where $1/\alpha$ accounts for the updated beliefs.

When (6.1) and (6.2) are equal, uninformed HFTs are indifferent between taking the role of a latency arbitrageur or that of a market maker. Thus, equating (6.1) and (6.2) leads to the following two equilibrium indifference conditions

$$\frac{1}{2}\,p_{LO}\,p_{LT}\,s/2=\frac{1}{2}\,p_{LO}\,p_{\text{public}}\,L\left(s/2\right)+\frac{N-1}{N}\,p_{\text{private}}\,L\left(s/2\right), \tag{6.3a}$$

$$p_{LO}\,p_{LT}\,s/2=p_{LO}\,p_{\text{public}}\,L\left(s/2\right)+\frac{N-1}{N}\,p_{\text{private}}\,L\left(s/2\right). \tag{6.3b}$$

If uninformed HFTs do not snipe in period three, then (6.3a) uniquely pins down the equilibrium bid-ask spread s_0^* set by uninformed HFTs in period three, whereas if at least one uninformed HFT snipes in period three, then (6.3b) uniquely pins down the equilibrium bid-ask spread s_1^* set by uninformed HFTs in period three. Both s_0^* and s_1^* are positive and unique because in both equations the left-hand sides are strictly increasing in s and equal to zero when $s=0$, while the right-hand sides are strictly decreasing in s and are positive for $s=0$.

The difference between these indifference conditions is that if uninformed HFTs do not snipe in period three, then the market maker has fewer opportunities to provide liquidity (and latency arbitrageurs have fewer opportunities to latency arbitrage the market maker), but the market maker still faces the same level of adverse selection from informed HFTs. To compensate for this difference, market makers require a wider spread, i.e., $s_0^*\geq s_1^*$, with equality when $p_{\text{private}}=0$.

Next, we show when uninformed HFTs snipe in the third period. This requires an equilibrium profile $\mathbf{d}=(d^1,\ldots,d^N)$ such that there are no profitable deviations in d^i. We restrict our attention to symmetric equilibria when either $\mathbf{d}=\mathbf{1}=(1,\ldots,1)$, all HFTs snipe in

period three, or $\mathbf{d} = \mathbf{0} = (0, ..., 0)$, HFTs do not snipe in period three. The following lemma provides the existence conditions for each equilibrium.

Lemma 3. *Let s_0^* and s_1^* be the unique solutions to* (6.3a) *and* (6.3b)*, respectively. If*

$$\begin{aligned} & p_{LO} \frac{p_{public}}{2N} L\left(s_0^*/2\right) + p_{LO} \frac{p_{public}}{2N} \left(L(-\delta) + \tilde{L}(\delta)\right) \\ & > \\ & p_{LO} \frac{p_{public}}{N} L\left(s_1^*/2\right) + p_{LO}\,\delta - \frac{N-1}{N} p_{private} L(\delta) \end{aligned} \tag{D0}$$

holds, then HFTs do not snipe in period three, so $\mathbf{d} = \mathbf{0}$*. On the other hand, if*

$$p_{LO}\,\delta/N > p_{private} L(\delta)/N \tag{D1}$$

holds, then HFTs snipe in period three, so $\mathbf{d} = \mathbf{1}$*.*

To show the result, consider a deviation from the symmetric profile, and write the inequalities to ensure that the deviation is not profitable. If the expected cost of being fooled by informed HFTs is too high, then condition (D0) holds and in equilibrium it is optimal for all HFTs not to snipe in period three. On the other hand, when the expected profit from sniping patient investors outweighs the expected loss from being fooled by informed HFTs, then condition (D1) holds and in equilibrium it is optimal for all HFTs to snipe in period three.

The following proposition summarizes the equilibrium behavior.

Proposition 5. *If the no-snipe condition* (D0) *or the snipe condition* (D1) *holds, then there exists a symmetric equilibrium (with respect to* $\mathbf{d}$*) where the following occurs in every iteration of the trading game given state* (y, ω)*:*

- *At the end of period 1:*
  - *If* (D0) *holds, then there is one unit of liquidity provided at $y - s_0^*/2 - \varepsilon/2$ and one unit of liquidity provided at $y + s_0^*/2 + \varepsilon/2$, where s_0^* is the unique solution to* (6.3a)*.*
  - *If* (D1) *holds, then there is one unit of liquidity provided at $y - s_1^*/2 - \varepsilon/2$ and one unit of liquidity provided at $y + s_1^*/2 + \varepsilon/2$, where s_1^* is the unique solution to* (6.3b)*.*
- *In period 2: a patient investor, upon arrival, submits a limit order for a unit of security x at his reservation price; or an informed HFT, when profitable, submits a limit order at the patient investor's reservation price according to the direction of her privately observed jump.*

- *In period 3:*
  - *If* (D0) *holds, then uninformed HFTs do not snipe the limit order from period 2. If a buy limit order arrived in period 2, then at the end of period 3, an uninformed HFT provides one unit of liquidity to sell at* $y + s_0^*/2$*. If a sell limit order arrived in period 2, then at the end of period 3, an uninformed HFT provides one unit of liquidity to buy at* $y - s_0^*/2$*. The quoted spread following period 3 is* $s_0^*/2 - \delta$*.*
  - *If* (D1) *holds, then uninformed HFTs race to snipe the limit order from period 2. At the end of period 3, there is one unit of liquidity provided at* $y - s_1^*/2$ *and one unit of liquidity provided at* $y + s_1^*/2$*, so the quoted spread following period 3 is* s_1^**.*
- *In period 4: an impatient investor, upon arrival, purchases or sells one unit of x at the best bid or best offer; an informed HFT, when profitable, purchases or sells one unit of x at the best bid or best offer; HFTs race to latency arbitrage stale quotes when a publicly observed jump is profitable, if the stale quote belongs to the market maker, then she attempts to cancel her stale quote.*

Repeated play of the equilibrium for a single trading game is an equilibrium for the infinitely repeated trading game. This follows because there is no queuing motive to rest orders that carry over to subsequent games, and because information resets at the end of each trading game.

6.5 Collusive Equilibrium

The collusive strategy we consider is a reversion strategy, one in which observed deviations or suspected deviations from cooperation lead to a reversion to competitive play that lasts for T iterations of the trading game. The collusive strategy of interest recovers the key features we observe from the behavior of HFTs in the data. In our collusive strategy, we do not consider the possibility of cooperating to widen the spread (see Dutta and Madhavan, 1997) so that we focus on the benefits that arise from cooperating by revealing yourself to others.[8] In the collusive equilibrium, HFTs signal to each other to avoid trading with each other. This enables the HFTs to share the benign flow of patient investors, and as a byproduct, enables the HFTs to receive additional benign flow from impatient investors

[8]This restriction also rules out infinitely many collusive equilibria that arise as a consequence of the Folk theorem. In practice, the spread is constrained by other exchanges. Table 3 in Cartea et al. (2025) shows that the spread in Euronext is similar to the spread in other European exchanges. If HFTs cooperate to widen the spread, then brokers may divert benign flow away to other exchanges with tighter spreads.

who would have otherwise matched with a patient investor's limit order at better prices than those quoted by HFTs.

A prerequisite for cooperating through signaling is that there is a need to signal. This need arises only when uninformed HFTs do not snipe in period three in the competitive equilibrium (i.e., condition (D0) holds) because the costs of sniping toxic limit orders from privately informed HFTs is too high. This is a consequence of restricting the collusive strategy to one that does not widen the spread. If there is no need to signal (i.e., condition (D1) holds), then a collusive equilibrium can still exist if HFTs cooperate to widen the spread, but we do not consider this case.

6.5.1 Equilibrium Behavior of HFTs

There are two phases in the collusive strategy: a cooperation phase and a punishment phase. HFTs cooperate until they observe a deviation or suspect a deviation from cooperation, both of which trigger a punishment phase that lasts for T iterations of the trading game, after which, play returns to the cooperation phase. The collusive strategy is as follows.

In the cooperation phase,

- HFTs signal all their limit orders with a quantity $Q \gg 1$ so that their limit orders are easily differentiated from a patient investor's limit order for one unit of the security. HFTs take turns to provide liquidity. Each HFT i takes the role of the market maker for a proportion $1/N$ of the iterations played. HFTs do not latency arbitrage or adversely select market makers, and a market maker is compliant if she respects the agreement to take turns to provide liquidity at $y - s_0^*/2$ and at $y + s_0^*/2$, where s_0^* is the unique solution to (6.3a).

- HFTs do not pretend to be a patient investor, and in period 3 HFTs race to snipe limit orders that arrived in period 2 without a signal, i.e., they race to snipe patient investors.

In the punishment phase, HFTs revert to play from the competitive equilibrium. HFTs play according to the behavior described in Section 6.4.2 where HFTs do not snipe in period three.

Coordinating Play In the collusive equilibrium, there is a queuing motive for resting orders that carry over to subsequent iterations of the game, because when cooperating, HFTs forgo latency arbitraging and adversely selecting market makers, and instead share the role of the market maker. Therefore, HFTs need to coordinate their proportion of time as

the market maker. This coordination uses public and private innovations in y as a correlation device.

At the start of the infinitely repeated trading game, HFTs race to provide quotes around y_0, and the probability of each HFT winning the race is $1/N$. The winner of the race remains as the incumbent market maker until the next innovation in y.[9] Whenever there is an innovation in y (public or private), HFTs race to become the new market maker, and the incumbent market maker providing stale quotes complies by canceling her stale quotes. Losers of the race cancel their limit orders because the volume Q of the limit order is too large for losers of the race to receive flow from impatient investors. When the jump in y is public information, HFTs have an equal probability of winning the race to become the new market maker, and when the jump in y is private information, the privately informed HFT becomes the new market maker.

Consider a publicly observed jump from y to y'. If the jump is such that $|y' - y| < s_0^*$, then HFTs race to provide new quotes around y', and the incumbent market maker providing stale quotes around y complies by canceling her outstanding orders. On the other hand, if the jump is such that $|y' - y| \geq s_0^*$, then the incumbent market maker complies by canceling her stale orders around y, and HFTs race to provide new quotes around y'. For large public innovations in y, the incumbent market maker is provided with an opportunity to first cancel her stale quotes around y so that new quotes from the race around y' do not trade with the stale quotes of the incumbent market maker. This ensures that HFTs do not latency arbitrage market makers.

Next, consider a privately observed jump from y to y'. The informed HFT always wins the race because only she knows the value y' until the end of the trading game. If the jump is such that $|y' - y| < s_0^*$, then the informed HFT provides new quotes around y', and the incumbent market maker providing stale quotes around y complies by canceling her outstanding orders. On the other hand, if the jump is such that $|y' - y| \geq s_0^*$, then the informed HFT submits a signaled limit order at the patient investor's reservation price according to the direction of her privately observed jump. The incumbent market maker providing stale quotes around y complies by canceling her outstanding orders, and the informed HFT cancels her signaled limit order and provides new quotes around y'. For large private innovations in y, the signaled limit order at the patient investor's reservation price serves as a warning to the incumbent market maker that there is a privately informed

[9] This is different to the competitive equilibrium where the HFT providing liquidity can be different in period one and period three. Here, the same HFT continues to provide liquidity until there there is an innovation in y.

HFT. This provides the incumbent market maker with an opportunity to comply and cancel her stale quotes so that HFTs do not adversely select market makers.

By coordinating the role of the market maker through innovations in y, each HFT i becomes the incumbent market maker for a proportion $1/N$ of the iterations played.

Monitoring To sustain cooperation, HFTs must be able to monitor deviations from cooperative play so that profitable deviations from cooperation can be deterred with the threat of punishment, i.e., reversion to competitive play. If HFTs cannot reliably monitor deviations, then the threat of punishment is not credible.

In the collusive strategy, there are three punishment triggers:

1. a market maker is latency arbitraged or adversely selected,

2. a market maker is not compliant, and

3. HFTs snipe a non-signaled limit order from period two and lose money.

In the model, monitoring is perfect for the first two triggers. If a market maker is traded against for a quantity greater than one unit, then it is easy to infer that a market maker was latency arbitraged or adversely selected by another HFT because patient investors only demand one unit of x. On the other hand, if a market maker is traded against for a quantity of one unit of x, then HFTs cannot infer who (HFT or impatient investor) traded with the market maker from the quantity alone. However, in equilibrium, HFTs can perfectly infer that a market maker was latency arbitraged or adversely selected by another HFT because these trades are associated with an innovation in y, while trades initiated by an impatient investor are not associated with an innovation in y, see Figure 6.1. Similarly, if a market maker is not compliant because they do not respect the agreement to take turns to provide liquidity, or because they undercut the agreed upon spread s_0^*, then the non-compliance is immediately observed by all HFTs.

On the other hand, monitoring is imperfect for the third trigger. There are two possibilities that lead to an HFT losing money after sniping a non-signaled limit order from period two. First, the limit order is from a patient investor and the public innovation in y went in the wrong direction (the sniper was unlucky), or there was an informed HFT pretending to be a patient investor. When private information becomes public at the end of the trading game, this looks like public information from the perspective of uninformed HFTs, so uninformed HFTs cannot differentiate if the jump is public or it was a private jump that became public.[10]

[10] If one assumes HFTs can determine if a jump was public or private, then the problem becomes one with perfect monitoring.

Hence, the problem is one related to "secret price cutting" (see Green and Porter, 1984), because HFTs cannot perfectly infer whether the trigger was a result of bad luck or because there was a cheater.

In what follows, we analyze and characterize conditions for the collusive strategy to be an equilibrium of the infinitely repeated trading game.

6.5.2 Equilibrium Analysis

In the collusive strategy, if play is in the punishment phase, then the expected payoff of HFT i in the competitive equilibrium, per trading game, is

$$\frac{1}{2N} p_{LO}\, p_{\text{public}}\, L\left(s_0^*/2\right) + \frac{1}{N} p_{\text{private}}\, L\left(s_0^*/2\right) + p_{LO}\, \frac{p_{\text{public}}}{2N} \left(L(-\delta) + \tilde{L}(\delta)\right), \tag{6.4}$$

where s_0^* is the unique solution to (6.3a). The first term is the expected payoff from latency arbitraging the market maker. The second term is the expected payoff from becoming informed, while the final term is the expected payoff from latency arbitraging limit orders posted by patient investors. On the other hand, if play is in the cooperative phase, then the expected payoff of HFT i when cooperating, per trading game, is

$$\frac{1}{N} p_{LO}\, p_{LT}\, s_0^*/2 + p_{LO}\, \frac{1}{N} \delta, \tag{6.5}$$

which consists of the expected payoff from sharing the role of the market maker, and the expected payoff from sniping limit orders posted by patient investors in period two. A necessary condition for collusion is that the difference between the expected payoff from cooperation in (6.4) is greater than that from competitive play in (6.5). This difference is given by

$$\begin{aligned} c = \underbrace{\frac{1}{N} p_{LO}\, p_{LT}\, s_0^*/2 - \frac{1}{2N} p_{LO}\, p_{\text{public}}\, L\left(s_0^*/2\right) - \frac{1}{N} p_{\text{private}}\, L\left(s_0^*/2\right)}_{\text{gains from benign flow of impatient investors}} \\ + \underbrace{\frac{p_{LO}}{N}\delta - p_{LO}\, \frac{p_{\text{public}}}{2N} \left(L(-\delta) + \tilde{L}(\delta)\right)}_{\text{gains from benign flow of patient investors}}. \end{aligned} \tag{6.6}$$

The tradeoff is clear. The first term is the benign flow from impatient investors, the second term is the opportunity cost from no longer latency arbitraging the market maker, and the third term is the opportunity cost from no longer adversely selecting the market maker. The sum of these three terms is positive, and the additional profits arise from the additional benign flow from impatient investors that would have otherwise matched with a patient investor's limit order. The fourth term is the profitable benign flow from sniping patient

investors in period three, and the final term is the opportunity cost from no longer adversely selecting patient investors in period four when there is a public innovation in y because patient investors are sniped in period three. When $c > 0$, the profits from cooperation are higher than the profits from competitive play, which is a necessary condition to collude. However, when cooperating, HFTs have an incentive to cheat.

There are three profitable deviations from cooperation. One, HFTs latency arbitrage or adversely select market makers. Two, HFTs are not compliant and do not respect the agreement to take turns to provide liquidity. Three, HFTs pretend to be a patient investor to fool other HFTs into sniping their informed toxic limit order.

With the first two deviations, the upper bound on the gain from deviation that one can achieve in expectation, per trading game, is

$$k = Q\, L(s_0^*/2)\,. \tag{6.7}$$

Even though the incentive to deviate is large, the incentive is easily deterred because these deviations are perfectly observed. If HFTs are sufficiently patient (i.e., the discount factor ρ is sufficiently close to one) and if the punishment phase is sufficiently long (i.e., T is large enough), then the loss in future payoffs from entering into a punishment phase outweighs the immediate gain from deviation. Therefore, in equilibrium, these deviations never occur and punishment phases are never triggered by these deviations.

Next, the expected gain, per trading game, from the third deviation is

$$g = \frac{p_{\text{private}}}{N}\, L(\delta)\,. \tag{6.8}$$

This deviation is harder to disincentivize because HFTs cannot perfectly infer if the punishment phase was triggered as a result of bad luck or because there was a cheater. To disincentivize informed HFTs pretending to be a patient investor, the trigger condition must be sufficiently informative. An informative trigger condition allows the HFTs to monitor each other with sufficient accuracy so that cheating can be appropriately punished by reverting to competitive play. If the trigger condition is sufficiently informative, if HFTs are sufficiently patient, and if the punishment phase T is sufficiently long, then cheating by pretending to be a patient investor can be deterred with intertemporal incentives.

In the collusive strategy, if play is in the cooperative phase, then the probability of triggering a punishment phase in the next trading game is the probability of sniping a patient investor and losing money, which is given by

$$1 - q_C = p_{LO}\, p_{\text{public}}\, \frac{\mathbb{P}(J > \delta)}{2}\,.$$

On the other hand, if play is in the cooperative phase and one HFT deviates by pretending to be a patient investor, then the probability of triggering a punishment phase in the next trading game is the probability of sniping a patient investor and losing money plus the probability of the deviator pretending to be a patient investor, which is given by

$$1 - q_D = p_{\text{private}} \frac{\mathbb{P}(J > \delta)}{N} + p_{LO}\, p_{\text{public}} \frac{\mathbb{P}(J > \delta)}{2}.$$

Therefore, the informativeness of the trigger condition is given by

$$q_C - q_D = p_{\text{private}} \frac{\mathbb{P}(J > \delta)}{N}.$$

To deter deviations by pretending to be a patient investor, the trigger condition must be sufficiently informative (i.e., $q_C - q_D$ must be sufficiently large) so that cheating is noticed and punished.

A key property of the collusive strategy is that, in equilibrium, reverting to competitive play is necessary even though no HFT will pretend to be a patient investor. If the collusive strategy did not specify such a negative repercussion, then HFTs would have an incentive to cheat and the strategy profile would no longer be an equilibrium.

The following proposition provides the conditions for the collusive equilibrium to exist and summarizes the equilibrium behavior.

Proposition 6. *Let the competitive equilibrium be one where HFTs do not snipe in period three, i.e., condition* (D0) *holds. If the gains c in* (6.6) *from cooperation are positive, and if the two inequalities*

$$\begin{aligned} \rho\left[c\,(q_C - q_D)\left(1-\rho^T\right) + g\,q_C + \rho^T\,(1-q_C)\,g\right] &\geq g \quad \textit{and} \\ \rho\left[c\,q_C\left(1-\rho^T\right) + k\,q_C + \rho^T\,(1-q_C)\,k\right] &\geq k \end{aligned} \tag{ICC}$$

hold, where the gains from deviation k and g are in (6.7) *and* (6.8)*, respectively, then the following reversion strategy profile is a collusive equilibrium:*

- *in punishment phases, HFTs play according to the competitive behavior described in Proposition 5, and*
- *in cooperation phases:*
  - *HFTs signal all their limit orders with a quantity $Q \gg 1$. HFTs take turns to provide liquidity. Each HFT i takes the role of the market maker for a proportion $1/N$ of the iterations played. HFTs do not latency arbitrage or adversely select market makers, and a market maker is compliant if she respects the agreement to take turns to provide liquidity at $y - s_0^*/2$ and at $y + s_0^*/2$, where s_0^* is the unique solution to* (6.3a).

– *HFTs do not pretend to be a patient investor, and in period 3 HFTs race to snipe limit orders that arrived in period 2 without a signal.*

Play begins and remains in the cooperation phase until any one of the following is observed:

- *a market maker is latency arbitraged or adversely selected,*
- *a market maker is not compliant, or*
- *an HFT snipes a limit order from period two without a signal and loses money.*

Upon observing any of the triggers above, a punishment phase proceeds for T iterations of the trading game, after which, play returns to the cooperation phase.

The result follows from calculating the continuation values to check that there are no profitable deviations with the one-shot deviation principle. When play is in the punishment phase, there are no profitable deviations because myopic play is an equilibrium and no deviations can influence the continuation values. On the other hand, when play is in the cooperation phase, we check that the value obtained from each myopic deviation is less than the continuation value of continued cooperation. If condition (ICC) holds, then the myopic deviations are not profitable.

The (ICC) conditions hold if HFTs are sufficiently patient, if the punishment phase is sufficiently long, if the gains c from cooperation are large enough, and if the trigger condition is sufficiently informative (i.e., $q_C - q_D$ must be sufficiently large). The conditions ensure that deviations from cooperation are deterred through intertemporal incentives.

In the collusive equilibrium, supracompetitive profits from cooperation arise from the ability to identify benign limit orders from patient investors and the ability to identify toxic limit orders from informed HFTs. This allows HFTs to safely profit by sniping limit orders from patient investors; a source of excess profits that would otherwise be unavailable in the competitive equilibrium where HFTs do not snipe in period three. The supracompetitive profits from cooperation, in turn, provide the necessary incentives for the HFTs to willingly reveal themselves to each other, while the incentive to cheat by pretending to be a patient investor is deterred through intertemporal incentives with the threat of reverting to competitive play. This leads to an equilibrium where: (i) quoted spreads are on average wider than that in the competitive equilibrium where the type of limit order (toxic or benign) cannot be identified, and (ii) the trading costs for impatient investors are higher because they are forced to trade at the spread set by HFTs (they cannot trade with limit orders from a patient investor unless play is in the punishment phase).

6.5.3 Parameter Analysis

To understand the robustness of the collusive equilibrium, we analyze how the values of the model parameters affect the existence of the collusive equilibrium. We provide both theoretical and numerical results to gain insights into what breaks the collusive equilibrium.

Theoretical Analysis We provide a theoretical result to show how the number N of HFTs affects the existence of the collusive equilibrium, and a result to highlight how private information affects the existence of the collusive equilibrium.

Corollary 3. *Consider the set of model parameterizations such that the conditions in Proposition 6 hold. The size of this set decreases as the number N of HFTs increases.*

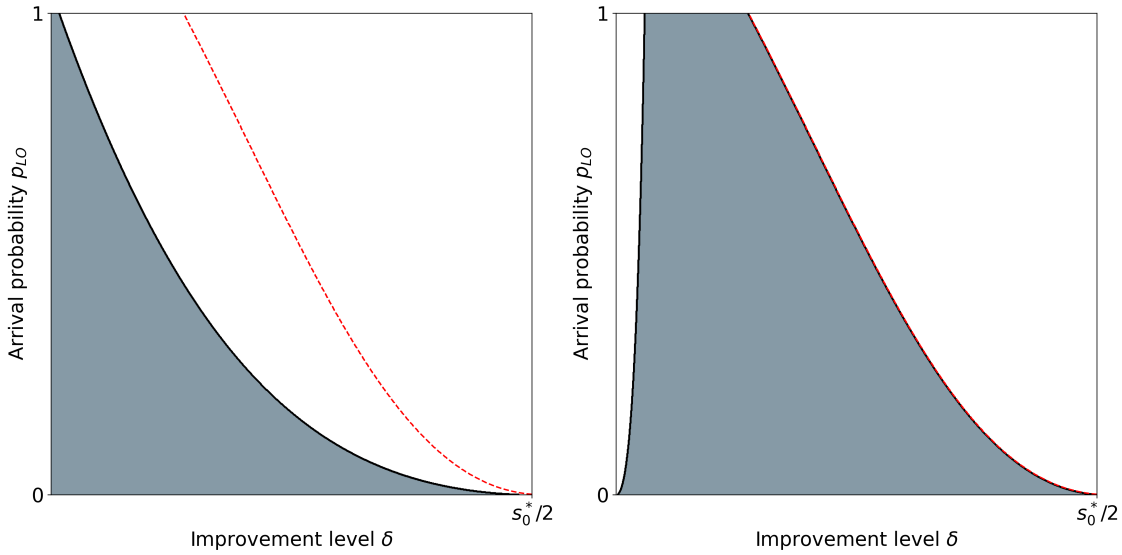
Put simply, as the number N of HFTs increases, the possibility of colluding with play described in Proposition 6 decreases. There are two forces driving this result. First, the gains c from cooperation decrease as the number of HFTs increases. Therefore, each HFT gets a smaller share of the pie, which decreases their incentives to cooperate. Second, the informativeness of the trigger condition $q_C - q_D$ decreases as N increases. Therefore, deviations are harder to detect, so it is harder to deter cheaters, which makes it more difficult to sustain a collusive arrangement.

Corollary 4. *In the absence of private information, the collusive equilibrium described in Proposition 6 does not exist.*

This result is straightforward. The motivation behind signaling is that benign limit orders from patient investors are distinguished from toxic limit orders sent by informed HFTs. If there are no informed HFTs, then there is no need to signal. Although the result is obvious, it is worth stating it because of a potential policy response: eliminate short-lived private information that arises from market fragmentation. For example, this can potentially be addressed through frequent batch auctions (see Budish et al., 2015).

Numerical Analysis We use a toy example to analyze numerically the driver behind each of the conditions in Proposition 6 as a function of the improvement value δ of limit orders from patient investors and the arrival probability of a patient investor p_{LO}. If all of the conditions hold, then a collusive equilibrium exists.

In Figure 6.2, regions of (δ, p_{LO}) that satisfy each of the conditions in Proposition 6 are in gray. In Figure 6.2a, condition (D0) only holds when the expected payoff from sniping a patient investor in period 3, i.e., $\delta\, p_{LO}$, is not too large. In Figures 6.2b and 6.2c, conditions (ICC) and $c > 0$ hold when the limit orders of patient investors are sufficiently mispriced



(a) Condition (D0), i.e., competitive equilibrium where HFTs do not snipe in period three

(b) Condition $c > 0$, i.e., positive gains from cooperation

(c) Condition (ICC)

(d) All conditions

Figure 6.2: Improvement value δ of limit orders from patient investors and the arrival probability of a patient investor p_{LO} that satisfy each of the conditions in Proposition 6. If all of the conditions hold, then collusive equilibrium exists. Shaded regions are where the conditions are satisfied. Dashed red line is the solution to (6.3a) as a function of the arrival probability of a patient investor p_{LO}, which is the upper bound on the improvement level δ. Model parameters: $N = 3$, $p_{LT} = p_{\text{public}} = 0.4$, $Q = 3$, and $J \sim U(0, 10)$.

and when the arrival probability of patient investors is not too high.[11] Figure 6.2d outlines

[11] We say condition (ICC) holds whenever there exists a combination of (ρ, T) such that the inequality is

the regions of (δ, p_{LO}) that satisfy all conditions in Proposition 6. We see that the regions where the collusive equilibrium exists are relatively small. The collusive equilibrium only exists when the limit orders posted by patient investors are sufficiently mispriced, and when the arrival probability of a patient investor is low. The requirement for sufficiently mispriced limit orders posted by patient investors is consistent with Table 6 in Cartea et al. (2025), which shows that for limit orders sent by retail investors, the sniping probability increases as the level of mispricing increases.

The condition for a low arrival probability of a patient investor is unexpected because one would expect that more patient investors means that there is more to gain from cooperation. However, the reason why collusion breaks down is intuitive. The limit orders of patient investors must be sufficiently mispriced so that conditions (ICC) and $c > 0$ hold. Now, if there are too many patient investors, then the expected payoff from sniping a patient investor's limit order covers the cost of being fooled by a toxic limit order from an informed HFT. Therefore, the need to signal disappears and collusion breaks down. It is worth highlighting that the main driver behind this result is our restriction to collusive strategies that do not widen the spread, so supracompetitive profits arise from revealing yourself to others.

6.6 Discussion

In the cooperation phase of the collusive equilibrium, the way in which HFTs can coordinate play is not unique. Similar to Folk theorems, our goal is not to specify the exact strategies used by HFTs, but rather to highlight the economic forces that underlie the incentives of HFTs to reveal themselves and how they can collectively enforce a collusive arrangement. Therefore, our model does not explain why HFTs would choose one collusive strategy profile over another.

Similarly, our model cannot explain how such a collusive arrangement might be initiated, whether it is tacit or explicit, because our model only analyzes the incentive structures in the collusive arrangement. However, Chapter 4 shows that a collusive arrangement can be initiated through learning without communication.

Finally, we highlight that in the model, HFTs are invariant to the signaling mechanism. The only requirements are that limit orders posted by HFTs are easily differentiated from limit orders posted by a patient investor, and that play in the cooperation phase is coordinated to accommodate for the signaling mechanism.

satisfied.

Appendix A

Proofs

A.1 Chapter 3

Following the proof sketch, to analyze the error terms in (3.4a), we first establish several useful implications of Assumption A.

First, an immediate consequence of Assumptions A.2 and A.5 is that there exists a constant L_1 such that $|f(\boldsymbol{\theta}, \mathbf{s}, \mathbf{a}, \mathbf{s}')| \leq L_1$ for all $\boldsymbol{\theta} \in G$, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, and $\mathbf{a} \in \mathcal{A}$. By construction, M_n, defined in (3.4b) is an $\mathcal{F}_n$-martingale, which is therefore uniformly bounded with $|M_n| \leq 2\, L_1$ almost surely.

Second, an immediate consequence of Assumption A.4 is that the transition probability $P_{\sigma_\theta}(\mathbf{s}' \,|\, \mathbf{s})$ is Lipschitz in $\boldsymbol{\theta}$ for all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, while Assumptions A.2, A.4, and A.5 ensure that $\bar{f}(\boldsymbol{\theta}, \mathbf{s})$ is Lipschitz in $\boldsymbol{\theta}$ for all $\mathbf{s} \in \mathcal{S}$.

Under Assumption A.3, for each $\boldsymbol{\theta} \in G$, the Markov chain $\mathbf{s}^{\sigma_\theta}$ induces a unique stationary distribution $\Gamma_{\sigma_\theta}(\mathbf{s})$ on the states $\mathbf{s} \in \mathcal{S}$ that satisfies the following system of equations

$$\Gamma_{\sigma_\theta}(\mathbf{s}) = \sum_{\mathbf{s}'} \Gamma_{\sigma_\theta}(\mathbf{s}')\, P_\theta(\mathbf{s} \,|\, \mathbf{s}')\,, \quad \sum_{\mathbf{s}} \Gamma_{\sigma_\theta}(\mathbf{s}) = 1$$

for all $\mathbf{s} \in \mathcal{S}$. The following lemma shows that the regularity of the transition probabilities $P_{\sigma_\theta}(\mathbf{s}' \,|\, \mathbf{s})$ establishes regularity results for the components of Γ_{σ_θ}.

Lemma 4. *Assume Assumptions A.3–A.4 hold, then the stationary distribution* $\Gamma_{\sigma_\theta}(\mathbf{s})$ *is Lipschitz in* $\boldsymbol{\theta}$ *for all* $\mathbf{s} \in \mathcal{S}$.

The proof of Lemma 4 follows from Lemma 4.2 in Ma et al. (1990). The key to bounding the errors in (3.4c) is to establish the regularity of the Poisson Equations. The collection of functions $\boldsymbol{\upsilon} = (\upsilon^i_{a|s,s'})_{0<i\leq I, a\in\mathcal{A}^i, \mathbf{s}\in\mathcal{S}}$, where $\upsilon^i_{a|s^i,s} : G \times \mathcal{S} \to \mathbb{R}$, is said to solve the Poisson Equations if the following holds for every $\boldsymbol{\theta} \in G$ and for every $\mathbf{s} \in \mathcal{S}$:

$$\boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s}) - \sum_{\mathbf{s}'} P_{\sigma_\theta}(\mathbf{s}' \,|\, \mathbf{s})\, \boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s}') = \bar{f}(\boldsymbol{\theta}, \mathbf{s}) - F(\boldsymbol{\theta})\,. \tag{A.1}$$

The following lemma establishes the regularity of F and $\boldsymbol{\upsilon}$ with respect to $\boldsymbol{\theta}$.

Lemma 5. *Assume Assumptions A.2 to A.5 hold, then the solution pair $(F, \boldsymbol{\upsilon})$ to the Poisson equation is Lipschitz in $\boldsymbol{\theta}$, i.e.,*

5.1 $|F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}')| \leq L_2 |\boldsymbol{\theta} - \boldsymbol{\theta}'|$ *for all* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in G$.

5.2 $|\boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s}) - \boldsymbol{\upsilon}(\boldsymbol{\theta}', \mathbf{s})| \leq L_2 |\boldsymbol{\theta} - \boldsymbol{\theta}'|$ *for all* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in G$ *and* $\mathbf{s} \in \mathcal{S}$. *In particular,* $\boldsymbol{\upsilon}$ *is bounded, i.e.,* $L_3 = \sup_{\boldsymbol{\theta}} \sup_{\mathbf{s}} |\boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s})| < \infty$.

The proof of Lemma 5 follows from Theorem 4.3 in Ma et al. (1990). As F is Lipschitz with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n \in G$, Proposition 1 is proved immediately by appealing to the Cauchy–Lipschitz Theorem. Lemma 5 admits the following corollary.

Corollary 5. *There exists* $L_4 > 0$ *such that* $\left| \mathbb{E}^{\theta} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s}_n) \,|\, \mathcal{F}_{n-1} \right] - \mathbb{E}^{\theta'} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}', \mathbf{s}_n) \,|\, \mathcal{F}_{n-1} \right] \right| \leq L_4 |\boldsymbol{\theta} - \boldsymbol{\theta}'|$ *for all* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in G$ *and* $\mathbf{s}_n \in \mathcal{S}$.

Proof. For any function $f : \mathcal{S} \to \mathbb{R}$, define $\mathbb{E}^{\mathbf{s}} \left[f(\mathbf{s}_k) \right] := \mathbb{E} \left[f(\mathbf{s}_k^{\theta}) \right]$. Notice that

$$\sum_{\mathbf{s}'} P_{\sigma_{\theta_n}}(\mathbf{s}' \,|\, \mathbf{s}_n) \, \boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}') = \mathbb{E}^{\theta_n} \left[\upsilon(\boldsymbol{\theta}_n, \mathbf{s}_{n+1}) | \mathcal{F}_n \right] . \tag{A.2}$$

Use (A.2) and rearrange (A.1) to obtain the identity

$$\mathbb{E}^{\theta} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s}_n) | \mathcal{F}_{n-1} \right] = \boldsymbol{\upsilon}(\boldsymbol{\theta}, \mathbf{s}_{n-1}) + F(\boldsymbol{\theta}) - \bar{f}(\boldsymbol{\theta}, \mathbf{s}_{n-1}) . \tag{A.3}$$

By Lemma 5, each term on the right-hand side of (A.3) is Lipschitz with respect to $\boldsymbol{\theta}$ for fixed $\mathbf{s}_n \in \mathcal{S}$. □

To analyze the error terms U_n in (3.4c), use the expression in (A.3) to decompose U_n as follows:

$$\begin{aligned}
U_{n+1} &= \bar{f}(\boldsymbol{\theta}_n, \mathbf{s}_n) - F(\boldsymbol{\theta}_n) \\
&= \boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}_n) - \mathbb{E}^{\boldsymbol{\theta}_n} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}_{n+1}) | \mathcal{F}_n \right] \\
&= \boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}_n) - \mathbb{E}^{\boldsymbol{\theta}_n} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}_n) | \mathcal{F}_{n-1} \right] \\
&\quad + \mathbb{E}^{\boldsymbol{\theta}_n} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}_n) | \mathcal{F}_{n-1} \right] - \mathbb{E}^{\boldsymbol{\theta}_{n+1}} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_{n+1}, \mathbf{s}_{n+1}) | \mathcal{F}_n \right] \\
&\quad + \mathbb{E}^{\boldsymbol{\theta}_{n+1}} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_{n+1}, \mathbf{s}_{n+1}) | \mathcal{F}_n \right] - \mathbb{E}^{\boldsymbol{\theta}_n} \left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_n, \mathbf{s}_{n+1}) | \mathcal{F}_n \right] \\
&:= Z_{n+1}^{(1)} + Z_{n+1}^{(2)} + Z_{n+1}^{(3)} .
\end{aligned}$$

As a consequence of Lemma 5.2 and (A.2), $Z_n^{(1)}$ is a uniformly bounded $\mathcal{F}_n$-martingale.

Now, with the lemmas and corollary from above, we bound the M_n and $Z_n^{(j)}$ terms for $j = 1, 2, 3$. We first prove Theorem 2 because we only need to show that (ARC) holds under Assumption A with (DLR).

Proof of Theorem 2.

By Proposition 4.1 in Benaïm (1999), if (ARC) holds, then Theorem 2.1 holds. Verifying

$$\lim_{n\to\infty}\left(\sup_{n<\ell\le m(n,T)}\left|\sum_{k=n+1}^{\ell}\gamma_k M_k\right|\right)=0, \quad \text{a.s., and} \tag{A.4a}$$

$$\lim_{n\to\infty}\left(\sup_{n<\ell\le m(n,T)}\left|\sum_{k=n+1}^{\ell}\gamma_k Z_k^{(j)}\right|\right)=0, \quad \text{a.s.} \tag{A.4b}$$

for $j=1,2,3$ is sufficient to verify (ARC). Define an increasing sequence of integers to partition the positive real line in the following way: $n_0=0, n_1=m(n_0,T), n_2=m(n_1,T),\ldots,n_{r+1}=m(n_r,T)$. Finally, define $S_n^{(j)}=\sum_{k=1}^{n}\gamma_k Z_k^{(j)}$ for $j=1,2,3$ and $S_n=\sum_{k=1}^{n}\gamma_k M_k$.

First, we address the martingale terms M_n in a manner similar to Benaïm (1999). Because the M_n terms are a martingale, use the Burkholder–Davis–Gundy inequality to obtain

$$\mathbb{E}\left[\sup_{n_r<\ell\le n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k M_k\right|^q\right]\le C_q\,\mathbb{E}\left[\left(\sum_{k=1+n_r}^{n_{r+1}}\gamma_k^2|M_k|^2\right)^{q/2}\right]$$

for some constant C_q that depends on q. Use Hölder's Inequality to obtain

$$\begin{aligned}\tilde{S}_r:=\mathbb{E}\left[\sup_{n_r<\ell\le n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k M_k\right|^q\right]&\le C_q\,\mathbb{E}\left[\left(\sum_{k=1+n_r}^{n_{r+1}}\gamma_k\right)^{q/2-1}\sum_{k=1+n_r}^{n_{r+1}}\gamma_k^{1+q/2}\,|M_k|^q\right]\\&\le C_q\,T^{q/2-1}\,\mathbb{E}\left[\sum_{k=1+n_r}^{n_{r+1}}\gamma_k^{1+q/2}\,|M_k|^q\right]\\&\le C_{q,T}\,2^q\,L_1^q\sum_{k=1+n_r}^{n_{r+1}}\gamma_k^{1+q/2}.\end{aligned}$$

Thus, we have

$$\tilde{S}_r\le C_{q,T,L_1}\sum_{k=1+n_r}^{n_{r+1}}\gamma_k^{1+q/2},$$

where C_{q,T,L_1} is a constant that depends on q, T, and L_1. Therefore, by (DLR) we have

$$\sum_{r\ge0}\tilde{S}_r\le C_{q,T,L_1}\sum_{k=1}^{\infty}\gamma_k^{1+q/2}<\infty.$$

By the Markov inequality,

$$\sum_{r\ge0}\mathbb{P}\left(\sup_{n_r<\ell\le n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k M_k\right|^q\ge\delta\right)\le\frac{\sum_{r\ge0}\tilde{S}_r}{\delta}<\infty.$$

Thus, by the Borel-Cantelli Lemma, we have

$$\mathbb{P}\left(\sup_{n_r<\ell\leq n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k\,M_k\right|^q>0\quad i.o.\right)=0\,.$$

The same statement is true for $q=2$, so we obtain

$$\lim_{r\to\infty}\sup_{n_r<\ell\leq n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k\,M_k\right|=0\,.$$

Finally, noting that for n such that $n_r\leq n<n_{r+1}$, we have

$$\sup_{n<\ell\leq m(n,T)}\left|\sum_{k=1+n_r}^{\ell}\gamma_k\,M_k\right|\leq 2\sup_{n_r<\ell\leq n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k\,M_k\right|+\sup_{n_{r+1}<\ell\leq n_{r+2}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k\,M_k\right|,$$

so (A.4a) holds for the M_n terms.

Recall that $Z_n^{(1)}$ is a bounded martingale, so following the same procedure as for the M_n terms, one writes the inequality

$$\tilde{S}_r^{(1)}:=\mathbb{E}\left[\sup_{n_r<\ell\leq n_{r+1}}\left|\sum_{k=1+n_r}^{\ell}\gamma_k\,Z_k^{(1)}\right|^q\right]\leq C_{q,T,L_3}\sum_{k=1+n_r}^{n_{r+1}}\gamma_k^{1+q/2},$$

where C_{q,T,L_3} is a constant that depends on q, T, and L_3 which is the constant from Lemma 5.2. Following the arguments as before, we have that (A.4b) holds for $j=1$.

For the $Z_n^{(2)}$ terms, notice that for $0<n<\ell$, we have

$$\begin{aligned}S_\ell^{(2)}-S_n^{(2)}=&-\sum_{k=n}^{\ell-1}(\gamma_k-\gamma_{k+1})\,\mathbb{E}^{\theta_k}\left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_k,\mathbf{s}_k)|\mathcal{F}_{k-1}\right]\\&+\gamma_n\,\mathbb{E}^{\theta_n}\left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_n,\mathbf{s}_n)|\mathcal{F}_{n-1}\right]-\gamma_\ell\,\mathbb{E}^{\theta_\ell}\left[\boldsymbol{\upsilon}(\boldsymbol{\theta}_\ell,\mathbf{s}_\ell)|\mathcal{F}_{\ell-1}\right].\end{aligned}$$

By Lemma 5.2, we obtain

$$\begin{aligned}\left|S_\ell^{(2)}-S_n^{(2)}\right|&\leq L_3\sum_{k=n}^{\ell-1}(\gamma_k-\gamma_{k+1})+L_3\,(\gamma_n+\gamma_\ell)\\&=L_3\,\gamma_n-L_3\,\gamma_\ell+L_3\,(\gamma_n+\gamma_\ell)=2\,L_3\,\gamma_n\end{aligned}$$

for any $\ell>n$. Thus, (A.4b) holds for $j=2$ because $\lim_{n\to\infty}\gamma_n=0$ by (DLR).

To show (A.4b) for $j=3$, use the Lipschitz property from Corollary 5 and the boundedness of f to obtain

$$|Z_{n+1}^{(3)}|\leq L_4\,|\boldsymbol{\theta}_{n+1}-\boldsymbol{\theta}_n|\leq L_1\,L_4\,\gamma_{n+1}\,.$$

Therefore,

$$\sup_{n<\ell\le m(n,T)}\left|\sum_{k=n+1}^{\ell}\gamma_k Z_k^{(3)}\right| \le \sum_{k=n+1}^{m(n,T)}\gamma_k|Z_k^{(3)}| \le L_1 L_4 \sum_{k=n+1}^{m(n,T)}\gamma_k^2 \le L_1 L_4 T\,\gamma_{n+1}\,,$$

which tends to zero as $n\to\infty$, and yields (A.4b) for $j=3$.

Apply Proposition 4.1 of Benaïm (1999) to obtain part 2.1 of Theorem 2. Next, apply the Kushner-Clark Lemma in Kushner and Clark (1978) to obtain part 2.2 of Theorem 2. □

Proof of Theorem 1.

First, for any n, use Taylor's Theorem on the solution of (ALE) to obtain

$$\bar{\boldsymbol{\theta}}(t_{n+1})-\bar{\boldsymbol{\theta}}(t_n)=\int_{t_n}^{t_{n+1}} F\left(\bar{\boldsymbol{\theta}}(s)\right)ds=\gamma_{n+1}F\left(\bar{\boldsymbol{\theta}}(t_n)\right)+\kappa_{n+1}\,, \tag{A.5}$$

for some κ_{n+1} with $|\kappa_{n+1}|\le L_4\gamma_{n+1}^2$ where L_4 is the constant from Corollary 5.

Use (3.3) and (A.5) to write

$$\boldsymbol{\theta}_n-\bar{\boldsymbol{\theta}}(t_n)=\boldsymbol{\theta}_{n-1}-\bar{\boldsymbol{\theta}}(t_{n-1})+\gamma_n\left(F(\boldsymbol{\theta}_{n-1})-F(\bar{\boldsymbol{\theta}}(t_{n-1}))\right)+\kappa_n+\gamma_n\varepsilon_n\,.$$

Therefore,

$$\boldsymbol{\theta}_n-\bar{\boldsymbol{\theta}}(t_n)=\sum_{k=0}^{n-1}\gamma_{k+1}\left(F(\boldsymbol{\theta}_k)-F(\bar{\boldsymbol{\theta}}(t_k))\right)+\sum_{k=0}^{n-1}\kappa_{k+1}+\sum_{k=0}^{n-1}\gamma_{k+1}\varepsilon_{k+1}\,,$$

and

$$\left|\boldsymbol{\theta}_n-\bar{\boldsymbol{\theta}}(t_n)\right|\le L_4\sum_{k=0}^{n-1}\gamma_{k+1}|\boldsymbol{\theta}_k-\bar{\boldsymbol{\theta}}(t_k)|+L_4\sum_{k=0}^{n-1}\gamma_{k+1}^2+\sup_{\ell\le n}\left|\sum_{k=0}^{\ell-1}\gamma_{k+1}\varepsilon_{k+1}\right|.$$

For convenience, denote

$$Y_n^{(1)}:=L_4\sum_{k=0}^{n-1}\gamma_{k+1}^2\quad\text{and}\quad Y_n^{(2)}:=\sup_{\ell\le n}\left|\sum_{k=0}^{\ell-1}\gamma_{k+1}\varepsilon_{k+1}\right|.$$

Use the discrete version of the Gronwall Lemma (see Benveniste et al., 1990, Lemma 8 Chapter 1 of Part II), to obtain

$$\left|\boldsymbol{\theta}_n-\bar{\boldsymbol{\theta}}(t_n)\right|\le\exp\left(L_4\sum_{k=0}^{n-1}\gamma_{k+1}\right)\times\left(Y_n^{(1)}+Y_n^{(2)}\right),$$

and hence,

$$\mathbb{E}\left[\sup_{n\le m(T)}\left|\boldsymbol{\theta}_n-\bar{\boldsymbol{\theta}}(t_n)\right|^2\right]\le 2\exp\left(2L_4T\right)\times\left(\left(Y_{m(T)}^{(1)}\right)^2+\mathbb{E}\left[\left(Y_{m(T)}^{(2)}\right)^2\right]\right). \tag{A.6}$$

Bound $\left(Y^{(1)}_{m(T)}\right)^2$ by noticing that

$$\left(Y^{(1)}_{m(T)}\right)^2 = \left(L_4 \sum_{k=0}^{m(T)-1} \gamma_{k+1}^2\right)^2 = L_4^2 T^2 \gamma_1^2 .$$

To bound $\mathbb{E}\left[\left(Y^{(2)}_{m(T)}\right)^2\right]$, recall that M_k and $Z_k^{(1)}$ are bounded by L_1 and L_3, respectively, for any $q \geq 1$. By martingale inequalities, we have

$$\mathbb{E}\left[\sup_{n \leq m(T)} \left|\sum_{k=1}^{n} \gamma_k M_k\right|^2\right] \leq 16 L_1^2 \sum_{k=1}^{m(T)} \gamma_k^2 \leq 16 L_1^2 T \gamma_1 \quad \text{and,}$$
$$\mathbb{E}\left[\sup_{n \leq m(T)} \left|\sum_{k=1}^{n} \gamma_k Z_k^{(1)}\right|^2\right] \leq 16 L_3^2 \sum_{k=1}^{m(T)} \gamma_k^2 \leq 16 L_3^2 T \gamma_1 ,$$

respectively. For $Z_k^{(2)}$ and $Z_k^{(3)}$, use the inequalities from the proof of Theorem 2 to obtain

$$\mathbb{E}\left[\left|\sum_{k=1}^{m(T)} \gamma_k Z_k^{(2)}\right|\right]^2 \leq 4 L_3^2 \gamma_1^2 \quad \text{and,}$$
$$\mathbb{E}\left[\sup_{n \leq m(T)} \left|\sum_{k=1}^{n} \gamma_k Z_k^{(3)}\right|^2\right] \leq \mathbb{E}\left[\sum_{k=1}^{m(T)} \gamma_k \left|Z_k^{(3)}\right|\right]^2 \leq L_1^2 L_4^2 T^2 \gamma_1^2 ,$$

respectively. Combine the above inequalities together to obtain

$$\mathbb{E}\left[\left(Y^{(2)}_{m(T)}\right)^2\right] \leq 16 T \left(L_1^2 + L_3^2\right) \gamma_1 + \left(4 L_3^2 + L_1^2 L_4^2 T^2\right) \gamma_1^2 .$$

Combine the bounds for $Y^{(1)}_{m(T)}$ and $Y^{(2)}_{m(T)}$ to bound (A.6) with

$$\mathbb{E}\left[\sup_{n \leq m(T)} \left|\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}(t_n)\right|^2\right] \leq 2 \exp\left(2 L_4 T\right) \left(16 T \left(L_1^2 + L_3^2\right) \gamma_1 + \left(4 L_3^2 + L_1^2 L_4^2 T^2 + L_4^2 T^2\right) \gamma_1^2\right) .$$

Denote the constant on the right-hand side as $C(T, \gamma_1)$, and note that $C(T, \gamma_1) \to 0$ as $\gamma_1 \to 0$. Write

$$\mathbb{P}\left(\sup_{n \leq m(T)} \left|\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}(t_n)\right| \geq \delta\right) \leq \frac{\mathbb{E}\left[\sup_{n \leq m(T)} \left|\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}(t_n)\right|^2\right]}{\delta^2} \leq \frac{C(T, \gamma_1)}{\delta^2} ,$$

to complete the proof of Theorem 1 by appealing to Markov's inequality. □

A.2 Chapter 4

The continuation payoff in (4.1) can be rewritten as a system of linear equations

$$V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = \sum_{\mathbf{a}\in\mathcal{A}} \left[\prod_{j=1}^{I} \theta^j_{a^j|\mathbf{s}}\right] \left[u^i(\mathbf{a}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}) + \delta \sum_{\mathbf{s}'} p(\mathbf{s}'\,|\,\mathbf{s},\mathbf{a}) V^i_{\mathbf{s}'}(\tau;\boldsymbol{\theta})\right]$$
$$= u^i(\boldsymbol{\theta}_{\mathbf{s}}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}) + \delta \sum_{\mathbf{s}'} P_\theta(\mathbf{s}'\,|\,\mathbf{s}) V^i_{\mathbf{s}'}(\tau;\boldsymbol{\theta}) \tag{A.7}$$

for all $0 < i \leq I$, $\mathbf{s} \in \mathcal{S}$, and $\tau \geq 0$. Similarly, the action values can also be rewritten as a system of linear equations

$$V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}) = \sum_{\mathbf{a}^{-i}\in\mathcal{A}_{-i}} \left[\prod_{j\neq i} \theta^j_{a^j|\mathbf{s}}\right] \left[u^i(a^i,\mathbf{a}^{-i}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}) + \delta \sum_{\mathbf{s}'} p_{a^i}(\mathbf{s}'\,|\,\mathbf{s},\mathbf{a}^{-i})\, V^i_{\mathbf{s}'}(\tau;\boldsymbol{\theta})\right]$$
$$= u^i(a^i,\boldsymbol{\theta}^{-i}_{\mathbf{s}}) - C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}) + \delta \sum_{\mathbf{s}'} p_{a^i}(\mathbf{s}'\,|\,\mathbf{s},\boldsymbol{\theta}^{-i}_{\mathbf{s}})\, V^i_{\mathbf{s}'}(\tau;\boldsymbol{\theta})$$

for all $0 < i \leq I$, $a^i \in \mathcal{A}_i$, $\mathbf{s} \in \mathcal{S}$, and $\tau \geq 0$. Furthermore, notice the relation $V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) := \boldsymbol{\theta}^i_{\mathbf{s}} \cdot V^i_{\cdot|\mathbf{s}}(\tau;\boldsymbol{\theta})$.

Finally, recall that $G = \Delta(\mathcal{A}_i)^{I\times|\mathcal{S}|}$. Define $G_{\text{int}} = \text{int}(\Delta(\mathcal{A}_i))^{I\times|\mathcal{S}|}$ as the interior of G so that $C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})$ is differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{\text{int}}$ when $\tau > 0$. In addition, for any fixed $\tau > 0$, we define G_τ as a compact and convex subset of G_{int} such that the rest points of (SBRD) lie within the interior of G_τ.

A.2.1 Building Blocks for Theorems 3 and 4

Lemma 6. *In a Markov potential game $\mathcal{G}^\infty_m(0)$ with a potential $\Phi_{\mathbf{s}}(\boldsymbol{\theta})$ for $\mathbf{s} \in \mathcal{S}$, the following hold for the continuation payoffs $V^i_{\mathbf{s}}(0;\boldsymbol{\theta})$ for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$:*

6.1 There exists a function $\hat{U}^i_{\mathbf{s}} : \Delta(\mathcal{A}_i)^{(I-1)\times|\mathcal{S}|} \to \mathbb{R}$ such that for each $\boldsymbol{\theta} = (\boldsymbol{\theta}^i, \boldsymbol{\theta}^{-i})$, we have $V^i_{\mathbf{s}}(0;\boldsymbol{\theta}) = \Phi_{\mathbf{s}}(\boldsymbol{\theta}) + \hat{U}^i_{\mathbf{s}}(\boldsymbol{\theta}^{-i})$.

6.2 $\partial_{\theta^i_{a^i|\mathbf{s}}} \Phi_{\mathbf{s}}(\boldsymbol{\theta}) = \partial_{\theta^i_{a^i|\mathbf{s}}} V^i_{\mathbf{s}}(0;\boldsymbol{\theta})$ for all $0 < i \leq I$, $a^i \in \mathcal{A}_i$ and $\mathbf{s} \in \mathcal{S}$.

Proof. The result follows from Proposition B.1 in Leonardos et al. (2022). □

Lemma 7. *For any $\mu \in \Delta(\mathcal{S})$, $\mathbf{s} \in \mathcal{S}$, $\boldsymbol{\theta} \in G_{int}$, $0 < i \leq I$, $a \in \mathcal{A}_i$, and $\tau > 0$, we have*

$$\partial_{\theta^i_{a|\mathbf{s}}} V^i_\mu(\tau;\boldsymbol{\theta}) = d^\theta_\mu(\mathbf{s}) \left(V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \partial_{\theta^i_{a|\mathbf{s}}} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})\right),$$

where $d^\theta_\mu(\mathbf{s})$ is the discounted state visitation frequency given by

$$d^\theta_\mu(\mathbf{s}) = \sum_{\mathbf{s}^\theta_0\in\mathcal{S}} \mu(\mathbf{s}^\theta_0) \sum_{k=0}^{\infty} \delta^k\, \mathbb{P}(\mathbf{s}^\theta_k = \mathbf{s}\,|\,\mathbf{s}^\theta_0).$$

Proof. The proof follows a similar approach to that in Maheshwari et al. (2023). We claim that for any $\mu \in \Delta(\mathcal{S})$, $0 < i \leq I$, $\boldsymbol{\theta} \in G_{\text{int}}$, $\mathbf{s} \in \mathcal{S}$, $a \in \mathcal{A}_i$, and $\tau > 0$,

$$\partial_{\theta^i_{a|\mathbf{s}}} V^i_\mu(\tau;\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{k=0}^{K} \delta^k \, \mathbb{1}_{\{\mathbf{s}^\theta_k = \mathbf{s}\}}\right]\left(V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \partial_{\theta^i_{a|\mathbf{s}}} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})\right) + \delta^{K+1}\,\mathbb{E}\left[\partial_{\theta^i_{a|\mathbf{s}}} V^i_{\mathbf{s}^\theta_{K+1}}(\tau;\boldsymbol{\theta})\right]$$

holds for any integer $K \geq 0$. The claim follows from a straightforward proof by induction. Hence, take $K \to \infty$ and write

$$\begin{aligned}
\partial_{\theta^i_{a|\mathbf{s}}} V^i_\mu(\tau;\boldsymbol{\theta}) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \, \mathbb{1}_{\{\mathbf{s}^\theta_k = \mathbf{s}\}}\right]\left(V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \partial_{\theta^i_{a|\mathbf{s}}} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})\right) \\
&= \sum_{\mathbf{s}^\theta_0 \in \mathcal{S}} \mu(\mathbf{s}^\theta_0) \sum_{k=0}^{\infty} \delta^k \, \mathbb{P}\left(\mathbf{s}^\theta_k = \mathbf{s} \,|\, \mathbf{s}^\theta_0\right)\left(V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \partial_{\theta^i_{a|\mathbf{s}}} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})\right) \\
&= d^\theta_\mu(\mathbf{s})\left(V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \partial_{\theta^i_{a|\mathbf{s}}} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}})\right).
\end{aligned}$$

□

Lemma 8. *Let $\boldsymbol{\theta} \in G_{int}$ and let and $\tau > 0$, then*

$$d^\theta_\mu(\mathbf{s})\, \partial_1 J^i_{\mathbf{s}}(\mathbf{y}, \boldsymbol{\theta}; \tau) \cdot \Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s})\left(\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) - \boldsymbol{\theta}^i_{\mathbf{s}}\right) \geq 0$$

with equality only when $\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = \boldsymbol{\theta}^i_{\mathbf{s}} = \mathbf{y}$ for all $0 < i \leq I$ and $\mathbf{s} \in \mathcal{S}$.

Proof. Observe that

$$\frac{\partial^2 J^i_{\mathbf{s}}(\mathbf{y}, \boldsymbol{\theta}; \tau)}{\partial_{y_a} \partial_{y_{a'}}} = -\mathbb{1}_{\{a = a'\}} \frac{\tau}{y_a}$$

for all $y_a, y_{a'}$, which correspond to the actions of $\mathbf{y} \in \text{int}(\Delta(\mathcal{A}_i))$. Thus, $J^i_{\mathbf{s}}$ is strictly concave in $\mathbf{y}$ because the Hessian matrix of $J^i_{\mathbf{s}}$ is a diagonal matrix with negative entries. Furthermore, observe that $\partial_1 J^i_{\mathbf{s}}(\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}), \boldsymbol{\theta}; \tau) = 0$ by definition of $\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta})$. Therefore,

$$\left(\partial_1 J^i_{\mathbf{s}}(\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}), \boldsymbol{\theta}; \tau) - \partial_1 J^i_{\mathbf{s}}(\mathbf{y}, \boldsymbol{\theta}; \tau)\right) \cdot \left(\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) - \boldsymbol{\theta}^i_{\mathbf{s}}\right) \leq 0,$$

with equality only when $\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = \mathbf{y} = \boldsymbol{\theta}^i_{\mathbf{s}}$ because of the strict concavity and the definition of the smoothed best response. Thus,

$$\partial_1 J^i_{\mathbf{s}}(\mathbf{y}, \boldsymbol{\theta}; \tau) \cdot \left(\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) - \boldsymbol{\theta}^i_{\mathbf{s}}\right) \geq 0,$$

with equality only when $\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = \mathbf{y} = \boldsymbol{\theta}^i_{\mathbf{s}}$. Finally, the lemma follows because $\Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) > 0$ and $d^\theta_\mu(\mathbf{s}) > 0$ for all $\mathbf{s} \in \mathcal{S}$. □

Lemma 9. *For all $0 < i \leq I$, $\mathbf{s} \in \mathcal{S}$, $a^i \in \mathcal{A}_i$, and $\tau > 0$, the continuation payoff $V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ and the action values $V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta})$ are infinitely differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{int}$.*

Proof. Write (A.7) in vector notation over the states $\mathbf{s} \in \mathcal{S}$ as

$$\mathbf{V}^i(\tau;\boldsymbol{\theta}) = (\mathrm{Id} - \delta \mathbf{P}_\theta)^{-1} \left[\mathbf{u}^i(\boldsymbol{\theta}) - \mathbf{C}^i(\tau;\boldsymbol{\theta}^i)\right], \tag{A.8}$$

where Id is the identity matrix. The matrix $\mathbf{P}_\theta$ is stochastic, so its largest eigenvalue is 1. The matrix $\mathrm{Id} - \delta \mathbf{P}_\theta$ is invertible because its lowest eigenvalue is $1 - \delta > 0$. Recall the identity

$$(\mathrm{Id} - \delta \mathbf{P}_\theta)^{-1} = \frac{1}{\det(\mathrm{Id} - \delta \mathbf{P}_\theta)} \operatorname{adj}(\mathrm{Id} - \delta \mathbf{P}_\theta). \tag{A.9}$$

The determinant is a polynomial in the components of $\boldsymbol{\theta} \in G_{\text{int}}$. Each entry of the adjugate is the determinant of a cofactor matrix, and hence, each entry is a polynomial in the components of $\boldsymbol{\theta} \in G_{\text{int}}$. Therefore, $\mathbf{V}^i(\tau;\boldsymbol{\theta})$ is infinitely differentiable in $\boldsymbol{\theta}$ because $\mathbf{u}^i(\boldsymbol{\theta})$ and $\mathbf{C}^i(\tau;\boldsymbol{\theta}^i)$ are both infinitely differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{\text{int}}$. Moreover, $V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta})$ is also infinitely differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{\text{int}}$ because $V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta})$ is the sum of infinitely differentiable functions. □

Corollary 6. *For all $0 < i \leq I$, $\mathbf{s} \in \mathcal{S}$, $a^i \in \mathcal{A}_i$, and $\tau > 0$, the continuation payoff $V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ and the action values $V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta})$ are Lipschitz continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$.*

Proof. Observe that G_τ is a compact set. Thus $\nabla_\theta \mathbf{V}^i(\tau;\boldsymbol{\theta})$ is bounded on G_τ because it is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$. Therefore, $\mathbf{V}^i(\tau;\boldsymbol{\theta})$ is Lipschitz continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$. Moreover, $V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta})$ is a sum of Lipschitz functions, so it is also Lipschitz continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$. □

Corollary 7. *In a Markov potential game $\mathcal{G}^\infty_m(0)$, the perturbed potential function $\Phi_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ is infinitely differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{int}$, $\mathbf{s} \in \mathcal{S}$, and $\tau > 0$.*

Proof. The perturbed game $\mathcal{G}^\infty_m(\tau)$ is also a Markov potential game with potential function $\Phi_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ and corresponding continuation payoffs $V^i_s(\tau;\boldsymbol{\theta})$, by Lemma 2. Therefore, by Lemma 6, we have $\Phi_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) - \hat{U}^i_{\mathbf{s}}(\boldsymbol{\theta}^{-i})$. Hence, for all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, $0 < i \leq I$, $a \in \mathcal{A}_i$, we have $\partial_{\theta^i_{a|\mathbf{s}'}} \Phi_{\mathbf{s}}(\tau;\boldsymbol{\theta}) = \partial_{\theta^i_{a|\mathbf{s}'}} V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ for all $\tau > 0$. From Lemma 9, $V^i_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ is infinitely differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{\text{int}}$, so $\Phi_{\mathbf{s}}(\tau;\boldsymbol{\theta})$ is also infinitely differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_{\text{int}}$. □

A.2.2 Building Blocks for Theorems 5 and 7

We use a result from algebraic topology that Govindan et al. (2003) and Doraszelski and Escobar (2010) use to prove purification theorems.

Proposition 7. *Suppose that U is a bounded, open set in $\mathbb{R}^m$ and $H^\tau, M : \bar{U} \to \mathbb{R}^m$ are continuous, where $\bar{U}$ denotes the closure of U. Further, suppose that M is continuously differentiable on U, that x_0 is the only zero of M in U, and that the Jacobian of M at x_0 has full rank. If the function $\kappa H^\tau + (1-\kappa)M$ has no zero on the boundary of U for all $\kappa \in [0,1]$, then H^τ has a zero in U.*

Instead of using Proposition 7 to establish the existence of equilibrium of the perturbed game near the equilibrium of the unperturbed game, we use Proposition 7 to show that there exists a rest point of (SBRD) for $\tau > 0$ near a regular equilibrium $\boldsymbol{\theta}^*$ of the unperturbed game $\mathcal{G}_m^\infty(0)$. To do so, we extend the domain of G and the entropy function C^i to establish suitable functions H^τ and M to consider.

We construct G_ϵ according to Doraszelski and Escobar (2010). First, observe that (A.9) has strictly dominant diagonals. Therefore, for all $\boldsymbol{\theta}^* \in G$, one can find ϵ_{θ^*} such that (A.9) is invertible for all $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ satisfying $|\boldsymbol{\theta}^* - \boldsymbol{\theta}| < \epsilon_{\theta^*}$. Next, take a finite covering $\left(N_{\epsilon_{\theta_j^*}}(\boldsymbol{\theta}_j^*)\right)_{j\in J}$ of G because G is compact, and define G_ϵ to be open such that its closure $\bar{G}_\epsilon$ is contained in the open set $\bigcup_{j\in J} N_{\epsilon_{\theta_j^*}}(\boldsymbol{\theta}_j^*)$. Extend the entropy function (4.5) so that it is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\epsilon$.

The function $M : G_\epsilon \to \mathbb{R}^{I\times|\mathcal{S}|\times|\mathcal{A}_i|}$ was defined in Section 4.2.3 given by (4.11). Furthermore, M is continuously differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\epsilon$ as a consequence of (A.8), and the Jacobian of M at a regular equilibrium $\boldsymbol{\theta}^*$ has full rank by definition.

Next, for $\tau > 0$, define $H^\tau : G_\epsilon \to \mathbb{R}^{I\times|\mathcal{S}|\times|\mathcal{A}_i|}$ so that the components are given by

$$H^{i,\tau}_{a^i|\mathbf{s}}(\boldsymbol{\theta}) = \begin{cases} \sum_{a^i\in\mathcal{A}_i} \theta^i_{a^i|\mathbf{s}} - 1 & \text{if } a^i = a^i_{\mathbf{s}}, \\ \tilde{B}^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \theta^i_{a^i|\mathbf{s}} & \text{if } a^i \neq a^i_{\mathbf{s}}, \end{cases} \tag{A.10}$$

for all $0 < i \leq I$, $a^i \in \mathcal{A}_i$, and $\mathbf{s} \in \mathcal{S}$. The rest points of (SBRD) correspond to $\boldsymbol{\theta}^*_\tau$ such that $H^\tau(\boldsymbol{\theta}^*_\tau) = \mathbf{0}$ because for all $\boldsymbol{\theta} \in G$, $\Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) \geq \eta > 0$ for all $\mathbf{s} \in \mathcal{S}$. Observe that H^τ is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\epsilon$ as a consequence of (A.8) and because we extended the entropy function (4.5) so that it is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\epsilon$.

Finally, by the proof of Proposition 2 in Doraszelski and Escobar (2010), there exists an open set $U \subset G_\epsilon$ that satisfies the following conditions:

C1. $\boldsymbol{\theta}^* \in U$.

C2. For all $\boldsymbol{\theta} \in U$, $|\boldsymbol{\theta}^* - \boldsymbol{\theta}| < \epsilon_{\theta^*}$.

C3. $\boldsymbol{\theta}^*$ is the only zero of M in U.

C4. For all $0 < i \leq I$, $a^i \in \mathcal{A}_i$, and $\mathbf{s} \in \mathcal{S}$, if $\theta^{i,*}_{a^i|\mathbf{s}} > 0$, then $\theta^i_{a^i|\mathbf{s}} > 0$ for all $\boldsymbol{\theta} \in U$.

C5. For all $0 < i \leq I$, $a^i \in \mathcal{A}_i$, and $\mathbf{s} \in \mathcal{S}$, if $V^i_{a^i|\mathbf{s}}(0;\boldsymbol{\theta}^*) - V^i_{a^i_{\mathbf{s}}|\mathbf{s}}(0;\boldsymbol{\theta}^*) < 0$, then $V^i_{a^i|\mathbf{s}}(0;\boldsymbol{\theta}) - V^i_{a^i_{\mathbf{s}}|\mathbf{s}}(0;\boldsymbol{\theta}) < 0$ for all $\boldsymbol{\theta} \in U$.

To apply Proposition 7, the following lemma is required.

Lemma 10. *For a small enough value of $\tau > 0$ and all $\kappa \in [0,1]$, the function $\kappa H^\tau + (1-\kappa)M$ has no zero on the boundary of U.*

Proof. The result follows the same arguments as those in the proof of Theorem 2 in Doraszelski and Escobar (2010). □

Therefore, we have the following lemma.

Lemma 11. *Let $\boldsymbol{\theta}^* \in G$ be a regular equilibrium of the unperturbed game $\mathcal{G}^\infty_m(0)$. Then, for all $\epsilon > 0$, there exists $\hat{\tau} > 0$ such that for all $\tau \in (0, \hat{\tau})$, there exists $\boldsymbol{\theta}^*_\tau \in G_{int}$ such that $|\boldsymbol{\theta}^*_\tau - \boldsymbol{\theta}^*| < \epsilon$, where $\boldsymbol{\theta}^*_\tau \in G_{int}$ is a rest point of* (SBRD).

Proof. The result is an immediate consequence of applying Proposition 7 and the suitable construction of H^τ and M. Additionally, $\boldsymbol{\theta}^*_\tau \in G_{\text{int}}$ because $\tilde{B}^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}) > 0$ for $\tau > 0$, so a zero of H^τ is contained in G_{int}. □

Lemma 12. *If the unperturbed game $\mathcal{G}^\infty_m(0)$ is regular, then the number of equilibria is finite.*

Proof. Restrict the domain of M to a compact subset G. All equilibria of M are isolated because the Jacobian of M at a regular equilibrium has full rank. We proceed by contradiction. Suppose that the set of equilibria of M is an infinite set, then the equilibria of M will have an accumulation point $\boldsymbol{\theta}'$ because G is compact. Observe that M is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G \subset G_\epsilon$. Then, $\boldsymbol{\theta}'$ is also an equilibrium of M, which contradicts the fact that all equilibria of M are isolated. Therefore, the number of equilibria is finite. □

Corollary 8. *If the unperturbed game $\mathcal{G}^\infty_m(0)$ is regular, then* (SBRD) *has finitely many rest points.*

Proof. An immediate consequence of Lemmas 11 and 12, and Proposition 3. □

Lemma 13. *Let $\boldsymbol{\theta}^* \in G$ be a pure equilibrium of the regular unperturbed game $\mathcal{G}^\infty_m(0)$. Then, there exists $\hat{\tau} > 0$ such that all the eigenvalues of $\nabla_\theta H^\tau(\boldsymbol{\theta}^*_\tau)$ have a negative real part for all $\tau \in (0, \hat{\tau})$, where $\boldsymbol{\theta}^*_\tau \in G_{int}$ is a rest point of* (SBRD) *near the pure equilibrium $\boldsymbol{\theta}^*$.*

Proof. By Lemma 11, there is a rest point $\boldsymbol{\theta}^*_\tau$ of (SBRD) near the pure equilibrium $\boldsymbol{\theta}^*$. Rewrite $H^\tau : G \to \mathbb{R}^{I\times|\mathcal{S}|\times|\mathcal{A}_i|}$ so that the components are given by

$$H^{i,\tau}_{a^i|\mathbf{s}}(\boldsymbol{\theta}) = \tilde{B}^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \theta^i_{a^i|\mathbf{s}} \tag{A.11}$$

for all $0 < i \le I$, $a^i \in \mathcal{A}_i$, and $\mathbf{s} \in \mathcal{S}$. This reformulation is equivalent to (A.10) for $\boldsymbol{\theta}$ that is a zero of H^τ. Next, observe that $V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau) \to V^i_{a|\mathbf{s}}(0;\boldsymbol{\theta}^*_\tau)$ as $\tau \to 0$ for all $a \in \mathcal{A}_i$; together with C5 implies that there exists a dominant action $a^i_{\mathbf{s}}$ for each i and each state $\mathbf{s}$ such that $V^i_{a^i_{\mathbf{s}}|\mathbf{s}}(\tau,\boldsymbol{\theta}^*_\tau) > V^i_{a^i|\mathbf{s}}(\tau,\boldsymbol{\theta}^*_\tau)$ for a value of τ that is sufficiently small. Use the quotient rule to obtain

$$\begin{aligned}
&\partial_{\theta^i_{a|\mathbf{s}}} H^{i,\tau}_{a^i|\mathbf{s}}(\boldsymbol{\theta}^*_\tau) \\
&= \frac{\tau^{-1}e^{\tau^{-1}V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}\left(\partial_{\theta^i_{a|\mathbf{s}}} V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)\sum_{a'} e^{\tau^{-1}V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)} - \sum_{a'}\partial_{\theta^i_{a|\mathbf{s}}} V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)e^{\tau^{-1}V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}\right)}{\left(\sum_{a'} e^{\tau^{-1}V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}\right)^2} - \partial_{\theta^i_{a|\mathbf{s}}}\theta^{i,*}_{a^i|\mathbf{s},\tau} \\
&= \frac{-\tau^{-1}e^{\tau^{-1}V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}\sum_{a'\neq a^i}\partial_{\theta^i_{a|\mathbf{s}}} V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)e^{\tau^{-1}V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}}{\left(\sum_{a'} e^{\tau^{-1}V^i_{\mathbf{a}'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}\right)^2} - \partial_{\theta^i_{a|\mathbf{s}}}\theta^{i,*}_{a^i|\mathbf{s},\tau}\,.
\end{aligned}$$

Observe that

$$\partial_{\theta^i_{a|\mathbf{s}}} V^i_{a'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau) = c_1 + \tau\left(1+\log\theta^{i,*}_{a|\mathbf{s},\tau}\right) + c_2 + c_3\,,$$

where c_1, c_2, c_3, and

$$\log\theta^{i,*}_{a|\mathbf{s},\tau} = \tau^{-1}V^i_{a|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau) - \log\left(\sum_{a'} e^{\tau^{-1}V^i_{\mathbf{a}'|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)}\right)$$

are bounded by Lemma 9. Furthermore, for a small value of τ, the largest term in the denominator will be of the form $\exp\left(2\tau^{-1}V_{a^i_{\mathbf{s}}|\mathbf{s}}(\tau;\boldsymbol{\theta}^*_\tau)\right)$, which will strictly dominate all the terms in the numerator. Therefore, $\partial_{\theta^i_{a|\mathbf{s}}} H^{i,\tau}_{a^i|\mathbf{s}}(\boldsymbol{\theta}^*_\tau) \to -\partial_{\theta^i_{a|\mathbf{s}}}\theta^{i,*}_{a^i|\mathbf{s},\tau}$ as $\tau \to 0$. A similar argument shows that $\partial_{\theta^i_{a|\mathbf{s}'}} H^{j,\tau}_{a^j|\mathbf{s}}(\boldsymbol{\theta}^*_\tau) \to 0$ as $\tau \to 0$ for $\mathbf{s}' \neq \mathbf{s}$ and $i \neq j$.

Therefore, we have $\nabla_\theta H^\tau(\boldsymbol{\theta}^*_\tau) \to -\mathrm{Id}$ as $\tau \to 0$. Hence, for $\tau \in (0,\hat{\tau})$, $\nabla_\theta H^\tau(\boldsymbol{\theta}^*_\tau)$ is a diagonally dominant matrix. Pick $\hat{\tau} > 0$ so that the Gershgorin disc does not intersect 0 for all the rows of $\nabla_\theta H^\tau(\boldsymbol{\theta}^*_\tau)$. Then by the Gershgorin circle theorem, we have that all the eigenvalues of $\nabla_\theta H^\tau(\boldsymbol{\theta}^*_\tau)$ have a negative real part for all $\tau \in (0,\hat{\tau})$. □

A.2.3 Main Results

Proof of Theorem 3. The proof follows once we verify the conditions in Corollary 1. By Corollary 6, the action value $V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta})$ is Lipschitz continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$. Moreover, the logit function is Lipschitz continuous in the action value with Lipschitz constant τ^{-1}.

Therefore, the smoothed best response function in (4.6) is Lipschitz in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$ because function composition preserves Lipschitz continuity.

We consider a projected version of (SFP), where the projection operator projects $\boldsymbol{\theta}$ onto the nearest point in G_τ (see for example Kushner and Clark, 1978).[1] Assumption A.2 is trivially satisfied because $\boldsymbol{\theta} \in G_\tau$ for the projected version of (SFP). Assumption A.3 is satisfied because the state process $\mathbf{s}^{\tilde{B}(\tau;\theta)}$ is an aperiodic Markov chain with a single recurrent class for any $\boldsymbol{\theta} \in G$. Assumption A.5 is also satisfied because (4.9) is Lipschitz in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in G_\tau$. □

Proof of Lemma 1. A repeated potential game with bounded memory strategies is such that there exists $\phi_{\mathbf{s}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $u^i_{\mathbf{s}}(a^i, \mathbf{a}^{-i}) - u^i_{\mathbf{s}'}(a^i, \mathbf{a}^{-i}) = \phi_{\mathbf{s}}(a^i, \mathbf{a}^{-i}) - \phi_{\mathbf{s}'}(a^i, \mathbf{a}^{-i})$ for all $0 < i \leq I$, $(a^i, \mathbf{a}^{-i}) \in \mathcal{A}$, and $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$. This is because $u^i = u^i_{\mathbf{s}} = u^i_{\mathbf{s}'}$ for all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ in a repeated potential game. Therefore, the result follows from Proposition 4 in Mguni et al. (2021). □

Proof of Lemma 2. The result follows from Lemma 3.1 (a) in Maheshwari et al. (2023). □

Proof of Proposition 2. Fix a distribution on states $\mu \in \Delta(\mathcal{S})$ such that $\mu(\mathbf{s}) > 0$ for all $\mathbf{s} \in \mathcal{S}$. The perturbed game $\mathcal{G}^\infty_m(\tau)$ is a Markov potential game by Lemma 2, so by Lemma 6.2 and Lemma 7 we write

$$\begin{aligned}
\frac{d}{dt}\Phi_\mu(\tau;\boldsymbol{\theta}) = \dot{\Phi}_\mu(\tau;\boldsymbol{\theta}) &= \sum_{i=1}^{I} \sum_{\mathbf{s}\in\mathcal{S}} \sum_{a^i \in \mathcal{A}_i} \partial_{\theta^i_{a^i|\mathbf{s}}} \Phi^i_\mu(\tau;\boldsymbol{\theta})\, \dot{\theta}^i_{a^i|\mathbf{s}} = \sum_{i=1}^{I} \sum_{\mathbf{s}\in\mathcal{S}} \sum_{a^i \in \mathcal{A}_i} \partial_{\theta^i_{a^i|\mathbf{s}}} V^i_\mu(\tau;\boldsymbol{\theta})\, \dot{\theta}^i_{a^i|\mathbf{s}} \\
&= \sum_{i=1}^{I} \sum_{\mathbf{s}\in\mathcal{S}} \sum_{a^i \in \mathcal{A}_i} d^\theta_\mu(\mathbf{s}) \left(V^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \partial_{\theta^i_{a^i|\mathbf{s}}} C^i(\tau;\boldsymbol{\theta}^i_{\mathbf{s}}) \right) \Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) \left(\tilde{B}^i_{a^i|\mathbf{s}}(\tau;\boldsymbol{\theta}) - \theta^i_{a^i|\mathbf{s}} \right) \\
&= \sum_{i=1}^{I} \sum_{\mathbf{s}\in\mathcal{S}} d^\theta_\mu(\mathbf{s})\, \partial_1 J^i_{\mathbf{s}}(\boldsymbol{\theta}^i_{\mathbf{s}}, \boldsymbol{\theta}; \tau) \cdot \Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) \left(\tilde{B}^i_{\mathbf{s}}(\tau;\boldsymbol{\theta}) - \boldsymbol{\theta}^i_{\mathbf{s}} \right).
\end{aligned}$$

Thus, by Lemma 8, $\dot{\Phi}_\mu(\tau;\boldsymbol{\theta}) \geq 0$ with equality only when $\tilde{B}(\tau;\boldsymbol{\theta}) = \boldsymbol{\theta}$. □

Proof of Proposition 3. The result follows from Lemma 4.7 (d) and Lemma 3.1 (b) in Maheshwari et al. (2023). □

Proof of Theorem 4. From Proposition 3, rest points of (SBRD) are an ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}^\infty_m(0)$. Furthermore, by the proof of Proposition 2, critical points of the Lyapunov function are rest points of (SBRD). Moreover, the chain recurrent points of (SBRD) are the rest points of (SBRD) by Propositions 5.3 and 6.4 of

[1] This projection is required because Assumption A.4 is only satisfied for $\boldsymbol{\theta} \in G_\tau$. However, this technicality does not affect the proof or any of our subsequent analysis because G_τ is defined so that all the rest points of (SBRD) lie within the interior of G_τ. Thus, we do not have to consider the projected equivalent for (SBRD).

Benaïm (1999). Therefore, the limit trajectories of (SFP) converge to a connected subset of rest points of (SBRD) with probability one from Proposition 5.3 and Theorem 5.7 of Benaïm (1999).

Moreover, if the unperturbed game $\mathcal{G}_m^\infty(0)$ is regular, then (SBRD) has finitely many rest points by Corollary 8. Therefore, the limit trajectory of (SFP) converges to a rest point of (SBRD) with probability one as a result of Corollary 6.6 of Benaïm (1999). □

Proof of Theorem 5. By Proposition 2, Corollary 8, Lemmas 11 and 13, (SBRD) has an isolated and asymptotically stable rest point $\boldsymbol{\theta}_\tau^*$ within a neighborhood of $\boldsymbol{\theta}_{pure}$ for each $\tau \in (0, \hat{\tau})$.[2] Furthermore, $V_{\mathbf{s}}^i(\tau; \boldsymbol{\theta}_\tau^*) \to V_{\mathbf{s}}^i(0; \boldsymbol{\theta}_\tau^*)$ as $\tau \to 0$. Thus, for any $\epsilon > 0$, choose $\bar{\tau} < \hat{\tau}$ such that $|\mathbf{V}^i(0; \boldsymbol{\theta}_\tau^*) - \mathbf{V}^i(0; \boldsymbol{\theta}_{pure})| \leq \epsilon$ for all i and $|\boldsymbol{\theta}_\tau^* - \boldsymbol{\theta}_{pure}| \leq \epsilon$ simultaneously hold. Therefore, by Theorem 4 and Theorem 7.3 of Benaïm (1999), for a fixed (δ, τ) with $\tau \in (0, \bar{\tau})$, play from (SFP) has a non-zero probability of converging to $\boldsymbol{\theta}_\tau^*$. Finally, from Proposition 3, $\boldsymbol{\theta}_\tau^*$ is an m-memory ϵ-subgame perfect equilibrium of the unperturbed game $\mathcal{G}_m^\infty(0)$. □

Proof of Corollary 2 and Theorem 6. A consequence of Theorem 5 because the strategy profiles considered are pure equilibria of the unperturbed game $\mathcal{G}_m^\infty(0)$. □

Proof of Theorem 7. First, $\boldsymbol{\theta}^*$ is locally asymptotically stable by Lemma 13. Let Q be a compact subset of the domain of attraction of $\boldsymbol{\theta}^*$. Use the steps in the proof of Theorem 2 and follow the steps in the proof of Theorem 13 in Chapter 1 (Part II) of Benveniste et al. (1990) to show that there exists a constant c_4 such that

$$\mathbb{P}(\boldsymbol{\theta}_n \to \boldsymbol{\theta}^* | \boldsymbol{\theta}_n \in Q) \geq 1 - c_4 \sum_{k=n+1}^{\infty} \bar{\gamma}_k^2 \geq 1 - c_4 \sum_{k=n+1}^{\infty} \left(\frac{|\mathcal{S}|}{N+k} \right)^2 = 1 - \varepsilon,$$

where the value of N is determined from $\mathbf{s}_n^\# \geq N$ for all $\mathbf{s} \in \mathcal{S}$. Finally, observe that as the value of N increases, the value of ε decreases. □

A.3 Chapter 6

Proof of Lemma 3. Consider the symmetric profile where HFTs do not snipe in period three. Conditional on HFT i being uninformed and observing a limit order arrive in period two, her expected payoff acting as a latency arbitrageur (which is equivalent to the expected payoff acting as a market maker) in period three is

$$\frac{1}{\alpha} \left[p_{LO} \frac{p_{\text{public}}}{2N} L\left(s_0^*/2\right) + p_{LO} \frac{p_{\text{public}}}{2N} \left(L(-\delta) + \tilde{L}(\delta) \right) \right].$$

[2] In Lemma 13, we evaluate the local stability of H^τ given by (A.11). The stability properties of H^τ translates to (SBRD) as a consequence of Proposition 2 and because for all $\boldsymbol{\theta} \in G$, $\Gamma_{\tilde{B}(\tau;\theta)}(\mathbf{s}) \geq \eta > 0$ for all $\mathbf{s} \in \mathcal{S}$ (see for example Borkar, 2008, pp. 85-86).

Now, consider a deviation (by HFT i) from the symmetric profile where HFTs do not snipe in period three. Her expected payoff is

$$\frac{1}{\alpha}\left[p_{LO}\,\frac{p_{\text{public}}}{N}\,L\left(s_1^*/2\right)+p_{LO}\,\delta-\frac{N-1}{N}\,p_{\text{private}}\,L(\delta)\right].$$

For the symmetric profile to be an equilibrium, it must be that

$$\begin{gathered}p_{LO}\,\frac{p_{\text{public}}}{2N}\,L\left(s_0^*/2\right)+p_{LO}\,\frac{p_{\text{public}}}{2N}\,\left(L(-\delta)+\tilde{L}(\delta)\right)\\ >\\ p_{LO}\,\frac{p_{\text{public}}}{N}\,L\left(s_1^*/2\right)+p_{LO}\,\delta-\frac{N-1}{N}\,p_{\text{private}}\,L(\delta)\end{gathered}$$

holds. Hence, we obtain condition (D0).

Next, consider the symmetric profile where HFTs snipe in period three. Conditional on HFT i being uninformed and observing a limit order arrive in period two, her expected payoff acting as a latency arbitrageur in period three is

$$\frac{1}{\alpha}\left[p_{LO}\,\frac{p_{\text{public}}}{N}\,L\left(s_1^*/2\right)+p_{LO}\,\delta/N-p_{\text{private}}\,L(\delta)/N\right].$$

Consider a deviation (by HFT i) from the symmetric profile where HFTs snipe in period three. Her expected payoff is

$$\frac{1}{\alpha}\left[p_{LO}\,\frac{p_{\text{public}}}{N}\,L\left(s_1^*/2\right)\right].$$

For the symmetric profile to be an equilibrium, it must be that

$$\frac{1}{\alpha}\left[p_{LO}\,\frac{p_{\text{public}}}{N}\,L\left(s_1^*/2\right)+p_{LO}\,\delta/N-p_{\text{private}}\,L(\delta)/N\right]>\frac{1}{\alpha}\left[p_{LO}\,\frac{p_{\text{public}}}{N}\,L\left(s_1^*/2\right)\right].$$

Rearrange the equation to obtain condition (D1). □

Proof of Proposition 5. We solve the problem through backward induction with updated beliefs based on Bayes' rule whenever possible.

In period four, the strategy described is the weakly dominant strategy following the proof of Budish et al. (2024). The only difference is that the uninformed HFT who acts as a market maker in period three can also race to latency arbitrage the stale quote from a patient investor when a publicly observed jump is profitable, because this is the dominant strategy.

At the end of period three when uninformed HFTs sort themselves into the role of a market maker and latency arbitrageurs, the result follows the same arguments as those in

Budish et al. (2024), but with expected payoffs given by (6.1) and (6.2). Equating (6.1) and (6.2) leads to the indifference conditions in (6.3a) and (6.3b), which pins down the equilibrium bid-ask spread s_0^* and s_1^* that depends on if at least one uninformed HFT snipes in period three. Existence of the equilibrium is restored through the BOBE following the same arguments as Budish et al. (2024). Hence, the presence of other potential uninformed liquidity providers will discipline equilibrium price levels. At the start of period three when uninformed HFTs decide to snipe or not, the symmetric equilibrium follows from Lemma 3.

The equilibrium spread set by the uninformed market maker in period three does not tighten if no limit order arrived in period two since the cost of adverse selection does not change because $\mathbb{P}(J > \delta)\, L_\delta(s/2) = L(s/2)$. Hence, the informed HFT cannot profit by hiding information. Therefore, in period two, the weakly dominant strategy of the informed HFT is always to send a toxic limit order whenever it is profitable to do so (even if the toxic order will not be sniped in period three).

Finally, in period one, no HFT is willing to provide liquidity with $\varepsilon = 0$ because each HFT strictly prefers to become an informed trader upon receiving private information in the second period. If an HFT provides liquidity with $\varepsilon = 0$, then she cannot withdraw liquidity in period three (without raising suspicion) in the event she becomes informed, which will hinder her ability to become an informed trader. Since each HFT can become informed, they all face the same problem, so no HFT will provide liquidity with $\varepsilon = 0$. □

Proof of Proposition 6. We verify that the collusive strategy profile is an equilibrium through the one-deviation principle. First, we normalize the expected payoff per trading game by subtracting the expected payoff from the competitive equilibrium. Let V_c denote the average payoff under the cooperation phase, and V_d denote the average payoff upon entering the punishment phase. The average payoffs are given by the solution to the following system of equations:

$$\begin{aligned} V_c &= (1-\rho)\, c + \rho\left(q_c\, V_c + (1-q_C)\, V_d\right) \\ V_d &= \rho^T\, V_c\,. \end{aligned} \tag{A.12}$$

Rearrange to obtain

$$\begin{aligned} V_c &= \frac{1-\rho}{1-\rho\, q_c - \rho^{T+1}\,(1-q_C)}\, c \\ V_d &= \frac{(1-\rho)\,\rho^T}{1-\rho\, q_c - \rho^{T+1}\,(1-q_C)}\, c \end{aligned}$$

as the solution to (A.12).

Next, we verify that there are no profitable deviations. First, under the punishment phase, there are no profitable deviations because (i) play follows the competitive equilibrium, and

(ii) no deviation can alter the continuation values by reverting early to the cooperation phase. Second, under the cooperation phase, we solve for the incentive compatible conditions such that there are no profitable deviations.

First, consider the deviation where HFTs pretend to be a patient investor to fool other HFTs into sniping their informed toxic limit order. This deviation is not profitable if

$$V_c \geq (1-\rho)(c+g) + \rho\left(V_c\, q_D + (1-q_D)\, V_d\right).$$

Substitute (A.12) to write

$$\begin{aligned}(1-\rho)\, c + \rho\left(q_c V_c + (1-q_C)\, V_d\right) &\geq (1-\rho)(c+g) + \rho\left(V_c\, q_D + (1-q_D)\, V_d\right)\\ \rho\,(q_C - q_D)\,(V_c - V_d) &\geq (1-\rho)\, g\,.\end{aligned}$$

Use

$$V_c - V_d = \frac{1-\rho}{1-\rho\, q_c - \rho^{T+1}(1-q_C)}\, c\left(1-\rho^T\right)$$

to write

$$\rho\,(q_C - q_D)\,(1-\rho)\, c\left(1-\rho^T\right) \geq (1-\rho)\, g\left(1-\rho\, q_c - \rho^{T+1}(1-q_C)\right).$$

Simplify and rearrange to obtain the first condition in (ICC) given by

$$\rho\left[c\,(q_C - q_D)\left(1-\rho^T\right) + g\, q_C + \rho^T (1-q_C)\, g\right] \geq g\,.$$

Finally, consider the remaining deviations that are perfectly monitored. These deviations are not profitable if

$$V_c = (1-\rho)\, c + \rho\left(q_c V_c + (1-q_C)\, V_d\right) \geq (1-\rho)(c+k) + \rho\, V_d\,.$$

Follow the same steps as above to obtain the second condition in (ICC) given by

$$\rho\left[c\, q_C\left(1-\rho^T\right) + k\, q_C + \rho^T (1-q_C)\, k\right] \geq k\,.$$

□

Proof of Corollary 3. Consider the first condition of (ICC) given by

$$\rho\left[c\,(q_C - q_D)\left(1-\rho^T\right) + g\, q_C + \rho^T (1-q_C)\, g\right] \geq g\,.$$

Multiply both sides by N and take the derivative with respect to N. Observe the left-hand side is decreasing in N and the right-hand side is constant with respect to N.

Consider the set of model parameterizations that satisfy the inequality above for a fixed N. Take any element from the set. For that model parameterization, there exists $N' > N$ such that the inequality above does not hold for N'. Therefore, the size of the set decreases as N increases. □

Proof of Corollary 4. If $p_{\text{private}} = 0$, then $q_C - q_D = 0$. If $q_C - q_D = 0$, then

$$\rho \left[c \left(q_C - q_D \right) \left(1 - \rho^T \right) + g \, q_C + \rho^T \left(1 - q_C \right) g \right] \geq g$$

never holds. □

Bibliography

Abada, Ibrahim and Xavier Lambin (2023): "Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?" *Management Science*, 69, 5042–5065.

Abada, Ibrahim, Xavier Lambin, and Nikolay Tchakarov (2024): "Collusion by Mistake: Does Algorithmic Sophistication Drive Supra-Competitive Profits?" *European Journal of Operational Research*, 318, 927–953.

Abreu, Dilip (1988): "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica*, 56, 383–396.

Abreu, Dilip, Prajit K. Dutta, and Lones Smith (1994): "The Folk Theorem for Repeated Games: A Neu Condition," *Econometrica*, 62, 939–948.

Abreu, Dilip, David Pearce, and Ennio Stacchetti (1990): "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, 1041–1063.

Abreu, Dilip and Ariel Rubinstein (1988): "The Structure of Nash Equilibrium in Repeated Games with Finite Automata," *Econometrica*, 56, 1259–1281.

Aït-Sahalia, Yacine and Mehmet Saglam (2013): "High Frequency Traders: Taking Advantage of Speed," Tech. rep., National Bureau of Economic Research.

Aqsha, Alif, Fayçal Drissi, and Leandro Sánchez-Betancourt (2024): "Strategic Learning and Trading in Broker-Mediated Markets," *arXiv preprint arXiv:2412.20847*.

Aquilina, Matteo, Eric Budish, and Peter O'Neill (2021): "Quantifying the High-Frequency Trading "Arms Race"," *The Quarterly Journal of Economics*, 137, 493–564.

Arieli, Itai and H. Peyton Young (2016): "Stochastic Learning Dynamics and Speed of Convergence in Population Games," *Econometrica*, 84, 627–676.

Arthur, W. Brian (1991): "Designing Economic Agents that Act like Human Agents: A Behavioral Approach to Bounded Rationality," *The American Economic Review*, 81, 353–359.

——— (1993): "On designing economic agents that behave like human agents," *Journal of Evolutionary Economics*, 3, 1–22.

Asker, John, Chaim Fershtman, and Ariel Pakes (2022): "Artificial Intelligence, Algorithm Design, and Pricing," *AEA Papers and Proceedings*, 112, 452–56.

——— (2023): "The Impact of AI Design on Pricing," *Journal of Economics and Management Strategy*.

Assad, Stephanie, Emilio Calvano, Giacomo Calzolari, Robert Clark, Vincenzo Denicolò, Daniel Ershov, Justin Johnson, Sergio Pastorello, Andrew Rhodes, Lei Xu, and Matthijs Wildenbeest (2021): "Autonomous Algorithmic Collusion: Economic Research and Policy Implications," *Oxford Review of Economic Policy*, 37, 459–478.

Auer, Peter, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire (2002): "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, 32, 48–77.

Babes, Monica, Michael Wunder, and Michael L. Littman (2009): "Q-learning in Two-Player Two-Action Games," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '09.

Baldauf, Markus and Joshua Mollner (2020): "High-Frequency Trading and Market Performance," *The Journal of Finance*, 75, 1495–1526.

Banchio, Martino and Andrzej Skrzypacz (2022): "Artificial Intelligence and Auction Design," .

Barfuss, Wolfram, Jonathan F. Donges, and Jürgen Kurths (2019): "Deterministic Limit of Temporal Difference Reinforcement Learning for Stochastic Games," *Phys. Rev. E*, 99, 043305.

Barlo, Mehmet, Guilherme Carmona, and Hamid Sabourian (2009): "Repeated Games with One-Memory," *Journal of Economic Theory*, 144, 312–336.

——— (2016): "Bounded Memory Folk Theorem," *Journal of Economic Theory*, 163, 728–774.

Baron, Matthew, Jonathan Brogaard, Björn Hagströmer, and Andrei Kirilenko (2019): "Risk and Return in High-Frequency Trading," *Journal of Financial and Quantitative Analysis*, 54, 993–1024.

Baudin, Lucas and Rida Laraki (2022a): "Fictitious Play and Best-Response Dynamics in Identical Interest and Zero-Sum Stochastic Games," in *Proceedings of the 39th International Conference on Machine Learning*, ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, PMLR, vol. 162 of *Proceedings of Machine Learning Research*, 1664–1690.

——— (2022b): "Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions," in *Advances in Neural Information Processing Systems*, ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Curran Associates, Inc., vol. 35, 20243–20256.

Beggs, Alan (2005): "On the Convergence of Reinforcement Learning," *Journal of Economic Theory*, 122, 1–36.

——— (2022): "Reference Points and Learning," *Journal of Mathematical Economics*, 100, 102621.

Benaïm, Michel (1999): "Dynamics of Stochastic Approximation Algorithms," in *Séminaire de Probabilités XXXIII*, ed. by Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, Berlin, Heidelberg: Springer Berlin Heidelberg, 1–68.

Benaïm, Michel and Morris W. Hirsch (1999a): "Mixed Equilibria and Dynamical Systems Arising from Fictitious Play in Perturbed Games," *Games and Economic Behavior*, 29, 36–72.

——— (1999b): "Stochastic Approximation Algorithms with Constant Step Size whose Average is Cooperative," *The Annals of Applied Probability*, 9, 216 – 241.

Benaïm, Michel and Jörgen W Weibull (2003): "Deterministic Approximation of Stochastic Evolution in Games," *Econometrica*, 71, 873–903.

Benveniste, Albert, Michel Métivier, and Pierre Priouret (1990): *Adaptive Algorithms and Stochastic Approximations*, Heidelberg: Springer Berlin.

Bergault, Philippe, Fayçal Drissi, and Olivier Guéant (2022): "Multi-Asset Optimal Execution and Statistical Arbitrage Strategies under Ornstein–Uhlenbeck Dynamics," *SIAM Journal on Financial Mathematics*, 13, 353–390.

Bergault, Philippe and Leandro Sánchez-Betancourt (2025): "A Mean Field Game between Informed Traders and a Broker," *SIAM Journal on Financial Mathematics*, 16, 358–388.

Bernales, Alejandro (2017): "Algorithmic and High Frequency Trading in Dynamic Limit Order Markets," *Available at SSRN 2352409*.

Biais, Bruno (1993): "Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets," *The Journal of Finance*, 48, 157–185.

Biais, Bruno, Thierry Foucault, and Sophie Moinas (2011): "Equilibrium High Frequency Trading," in *Proceedings from the fifth annual Paul Woolley Centre conference, London School of Economics*.

Bloembergen, Daan, Michael Kaisers, and Karl Tuyls (2010): "Lenient frequency adjusted Q-learning," in *Proceedings of the 22nd Benelux Conference on Artificial Intelligence*, BNAIC '10, 19–26.

Bloembergen, Daan, Karl Tuyls, Daniel Hennes, and Michael Kaisers (2015): "Evolutionary Dynamics of Multi-Agent Learning: A Survey," *Journal of Artificial Intelligence Research*, 53, 659–697.

Borkar, Vivek S. (2008): *Stochastic Approximation: A Dynamical Systems Viewpoint*, Cambridge University Press.

Brandière, Odile (1998): "The Dynamic System Method and the Traps," *Advances in Applied Probability*, 30, 137–151.

Brogaard, Jonathan (2010): "High Frequency Trading and its Impact on Market Quality," *Northwestern University Kellogg School of Management Working Paper*, 66, 10.

Brogaard, Jonathan and Corey Garriott (2019): "High-Frequency Trading Competition," *Journal of Financial and Quantitative Analysis*, 54, 1469–1497.

Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan (2015): "Trading Fast and Slow: Colocation and Liquidity," *The Review of Financial Studies*, 28, 3407–3443.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan (2014): "High-Frequency Trading and Price Discovery," *The Review of Financial Studies*, 27, 2267–2306.

——— (2019): "Price Discovery without Trading: Evidence from Limit Orders," *The Journal of Finance*, 74, 1621–1658.

Brown, George W. (1951): "Iterative Solution of Games by Fictitious Play," in *Activity Analysis of Production and Allocation*, ed. by T. C. Koopmans, New York: Wiley.

Brown, Zach Y. and Alexander MacKay (2023): "Competition in Pricing Algorithms," *American Economic Journal: Microeconomics*, 15, 109–56.

Brunnermeier, Markus K. and Lasse Heje Pedersen (2005): "Predatory Trading," *The Journal of Finance*, 60, 1825–1863.

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorskaya (2025): "Strategic Arbitrage in Segmented Markets," *Journal of Financial Economics*, 166, 104008.

Budish, Eric, Peter Cramton, and John Shim (2015): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *The Quarterly Journal of Economics*, 130, 1547–1621.

Budish, Eric, Robin Lee, and John J. Shim (2024): "A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?" *Journal of Political Economy*, 132, 1209 – 1246.

Börgers, Tilman and Rajiv Sarin (1997): "Learning Through Reinforcement and Replicator Dynamics," *Journal of Economic Theory*, 77, 1–14.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicoló, and Sergio Pastorello (2020): "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, 110, 3267–97.

——— (2021): "Algorithmic Collusion with Imperfect Monitoring," *International Journal of Industrial Organization*, 79, 102712.

Camerer, Colin and Teck-Hua Ho (1999): "Experienced-Weighted Attraction Learning in Normal Form Games," *Econometrica*, 67, 827–874.

Capponi, Agostino, Álvaro Cartea, and Fayçal Drissi (2025): "Do Longer Block Times Impair Market Efficiency in Decentralized Markets?" *Available at SSRN 5290232*.

Cartea, Álvaro, Patrick Chang, and Gabriel García-Arenas (2023a): "Spoofing and Manipulating Order Books with Learning Algorithms," *Available at SSRN 4639959*.

Cartea, Álvaro, Patrick Chang, and Rob Graumans (2025): "Anonymity, Signaling, and Collusion in Limit Order Books," *Available at SSRN 5080700*.

Cartea, Álvaro, Patrick Chang, Mateusz Mroczka, and Roel Oomen (2022a): "AI-Driven Liquidity Provision in OTC Financial Markets," *Quantitative Finance*, 22, 2171–2204.

Cartea, Álvaro, Patrick Chang, and José Penalva (2022b): "Algorithmic Collusion in Electronic Markets: The Impact of Tick Size," *Available at SSRN 4105954*.

Cartea, Álvaro, Patrick Chang, José Penalva, and Harrison Waldon (2022c): "Algorithmic Collusion and a Folk Theorem from Learning with Bounded Rationality," *Available at SSRN*.

Cartea, Álvaro, Patrick Chang, José Penalva, and Harrison Waldon (2022d): "The Algorithmic Learning Equations: Evolving Strategies in Dynamic Games," *Available at SSRN 4175239*.

Cartea, Álvaro, Fayçal Drissi, and Marcello Monga (2023b): "Decentralised Finance and Automated Market Making: Execution and Speculation," *arXiv preprint arXiv:2307.03499*.

——— (2023c): "Execution and Statistical Arbitrage with Signals in Multiple Automated Market Makers," in *2023 IEEE 43rd International Conference on Distributed Computing Systems Workshops (ICDCSW)*, IEEE, 37–42.

——— (2023d): "Predictable Losses of Liquidity Provision in Constant Function Markets and Concentrated Liquidity Markets," *Applied Mathematical Finance*, 30, 69–93.

——— (2024a): "Decentralized Finance and Automated Market Making: Predictable Loss and Optimal Liquidity Provision," *SIAM Journal on Financial Mathematics*, 15, 931–959.

Cartea, Álvaro, Fayçal Drissi, and Pierre Osselin (2023e): "Bandits for Algorithmic Trading with Signals," *Available at SSRN 4484004*.

Cartea, Álvaro, Fayçal Drissi, Leandro Sánchez-Betancourt, David Siska, and Lukasz Szpruch (2024b): "Strategic Bonding Curves in Automated Market Makers," *Available at SSRN 5018420*.

Cartea, Álvaro, Sebastian Jaimungal, and José Penalva (2015): *Algorithmic and High-Frequency Trading*, Cambridge University Press.

Cartea, Álvaro and José Penalva (2012): "Where is the Value in High Frequency Trading?" *The Quarterly Journal of Finance*, 2, 1250014.

Cartea, Álvaro and Leandro Sánchez-Betancourt (2023): "Optimal Execution with Stochastic Delay," *Finance and Stochastics*, 27, 1–47.

——— (2025): "Brokers and Informed Traders: Dealing with Toxic Flow and Extracting Trading Signals," *SIAM Journal on Financial Mathematics*, 16, 243–270.

Cartea, Álvaro, Imanol Pérez Arribas, and Leandro Sánchez-Betancourt (2022e): "Double-Execution Strategies using Path Signatures," *SIAM Journal on Financial Mathematics*, 13, 1379–1417.

Chassang, Sylvain and Juan Ortner (2019): "Collusion in Auctions with Constrained Bids: Theory and Evidence from Public Procurement," *Journal of Political Economy*, 127, 2269–2300.

——— (2023): "Regulating Collusion," *Annual Review of Economics*, 15, 177–204.

Cho, In-Koo and Akihiko Matsui (2005): "Learning Aspiration in Repeated Games," *Journal of Economic Theory*, 124, 171–201.

Cho, In-Koo and Noah Williams (2024): "Collusive Outcomes Without Collusion: Algorithmic Pricing in a Duopoly Model," *Available at SSRN 4753617*.

Christie, William G and Paul H Schultz (1994): "Why do NASDAQ market makers avoid odd-eighth quotes?" *The Journal of Finance*, 49, 1813–1840.

Clancy, Luke and Mauro Cesa (2025): "Crossed Signals: Row Over Collusion Pits Scholars Against Traders," `https://www.risk.net/markets/7961037/crossed-signals-row-over-collusion-pits-scholars-against-traders`, accessed: 6 February 2025.

Clark, Daniel, Drew Fudenberg, and Alexander Wolitzky (2021): "Record-Keeping and Cooperation in Large Societies," *The Review of Economic Studies*, 88, 2179–2209.

Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo (2022): "Algorithmic Pricing and Liquidity in Securities Markets," *HEC Paris Research Paper No. FIN-2022-1459*.

Comerton-Forde, Carole, Alex Frino, and Vito Mollica (2005): "The Impact of Limit Order Anonymity on Liquidity: Evidence from Paris, Tokyo and Korea," *Journal of Economics and Business*, 57, 528–540.

Comerton-Forde, Carole, Tālis J Putniņš, and Kar Mei Tang (2011): "Why do Traders Choose to Trade Anonymously?" *Journal of Financial and Quantitative Analysis*, 46, 1025–1049.

Comerton-Forde, Carole and Kar Mei Tang (2009): "Anonymity, Liquidity and Fragmentation," *Journal of Financial Markets*, 12, 337–367.

Cramton, Peter and Jesse A Schwartz (2000): "Collusive bidding: Lessons from the FCC spectrum auctions," *Journal of Regulatory Economics*, 17, 229–252.

——— (2001): "Collusive Bidding in the FCC Spectrum Auctions," *Contributions in Economic Analysis & Policy*, 1, 1538–0645.1078.

Cross, John G. (1973): "A Stochastic Learning Model of Economic Behavior," *The Quarterly Journal of Economics*, 87, 239–266.

De Jong, Frank and Barbara Rindi (2009): *The Microstructure of Financial Markets*, Cambridge University Press.

Doraszelski, Ulrich and Juan F. Escobar (2010): "A Theory of Regular Markov Perfect Equilibria in Dynamic Stochastic Games: Genericity, Stability, and Purification," *Theoretical Economics*, 5, 369–402.

Dou, Winston Wei, Itay Goldstein, and Yan Ji (2024): "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency," *Available at SSRN 4452704*.

Drissi, Fayçal (2022): "Solvability of Differential Riccati Equations and Applications to Algorithmic Trading with Signals," *Applied Mathematical Finance*, 29, 457–493.

——— (2023): "Models of Market Liquidity: Applications to Traditional Markets and Automated Market Makers," *Available at SSRN 4424010*.

Duffy, John and Ed Hopkins (2005): "Learning, Information, and Sorting in Market Entry Games: Theory and Evidence," *Games and Economic Behavior*, 51, 31–62.

Dutta, Prajit K. and Ananth Madhavan (1997): "Competition and Collusion in Dealer Markets," *The Journal of Finance*, 52, 245–276.

Easley, David and Maureen O'Hara (1987): "Price, Trade Size, and Information in Securities Markets," *Journal of Financial Economics*, 19, 69–90.

Ellison, Glenn (1994): "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching," *The Review of Economic Studies*, 61, 567–588.

Epivent, Andréa and Xavier Lambin (2024): "On Algorithmic Collusion and Reward–Punishment Schemes," *Economics Letters*, 237, 111661.

Erev, Ido and Alvin E. Roth (1998): "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *The American Economic Review*, 88, 848–881.

Foster, Dean P. and Rakesh V. Vohra (1997): "Calibrated Learning and Correlated Equilibrium," *Games and Economic Behavior*, 21, 40–55.

Foster, Dean P. and H. Peyton Young (2001): "On the Impossibility of Predicting the Behavior of Rational Agents," *Proceedings of the National Academy of Sciences*, 98, 12848–12853.

——— (2003): "Learning, Hypothesis Testing, and Nash Equilibrium," *Games and Economic Behavior*, 45, 73–96.

Foucault, Thierry, Johan Hombert, and Ioanid Roşu (2016): "News Trading and Speed," *The Journal of Finance*, 71, 335–381.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel (2013): "Liquidity Cycles and Make/Take Fees in Electronic Markets," *The Journal of Finance*, 68, 299–341.

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham (2017): "Toxic Arbitrage," *The Review of Financial Studies*, 30, 1053–1094.

Foucault, Thierry, Sophie Moinas, and Erik Theissen (2007): "Does Anonymity Matter in Electronic Limit Order Markets?" *The Review of Financial Studies*, 20, 1707–1747.

Friederich, Sylvain and Richard Payne (2014): "Trading Anonymity and Order Anticipation," *Journal of Financial Markets*, 21, 1–24.

Fudenberg, Drew and Lorens A. Imhof (2006): "Imitation Processes with Small Mutations," *Journal of Economic Theory*, 131, 251–262.

——— (2008): "Monotone Imitation Dynamics in Large Populations," *Journal of Economic Theory*, 140, 229–245.

Fudenberg, Drew and David M. Kreps (1993): "Learning Mixed Equilibria," *Games and Economic Behavior*, 5, 320–367.

Fudenberg, Drew, David Levine, and Eric Maskin (1994): "The Folk Theorem with Imperfect Public Information," *Econometrica*, 62, 997–1039.

Fudenberg, Drew and David K. Levine (1998): *The Theory of Learning in Games*, MIT Press, Cambridge, MA.

——— (1999): "Conditional Universal Consistency," *Games and Economic Behavior*, 29, 104–130.

——— (2009): "Learning and Equilibrium," *Annual Review of Economics*, 1, 385–420.

Fudenberg, Drew and Eric Maskin (1986): "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, 54, 533–554.

Galla, Tobias and J. Doyne Farmer (2013): "Complex Dynamics in Learning Complicated Games," *Proceedings of the National Academy of Sciences*, 110, 1232–1236.

Glosten, Lawrence R. (1994): "Is the Electronic Open Limit Order Book Inevitable?" *The Journal of Finance*, 49, 1127–1161.

Glosten, Lawrence R. and Paul R. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14, 71–100.

Gomes, Eduardo Rodrigues and Ryszard Kowalczyk (2009): "Dynamic Analysis of Multiagent Q-Learning with ϵ-Greedy Exploration," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 369–376.

Govindan, Srihari, Philip J. Reny, and Arthur J. Robson (2003): "A short proof of Harsanyi's purification theorem," *Games and Economic Behavior*, 45, 369–374, special Issue in Honor of Robert W. Rosenthal.

Green, Edward J., Robert C. Marshall, and Leslie M. Marx (2014): "Tacit Collusion in Oligopoly," in *The Oxford Handbook of International Antitrust Economics, Volume 2*, ed. by Roger D. Blair and D. Daniel Sokol, Oxford University Press, 464–497.

Green, Edward J. and Robert H. Porter (1984): "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, 52, 87–100.

Hagströmer, Björn and Lars Nordén (2013): "The Diversity of High-Frequency Traders," *Journal of Financial Markets*, 16, 741–770.

Hagströmer, Björn, Lars Nordén, and Dong Zhang (2014): "How Aggressive Are High-Frequency Traders?" *Financial Review*, 49, 395–419.

Hansen, Karsten T, Kanishka Misra, and Mallesh M Pai (2021): "Frontiers: Algorithmic Collusion: Supra-Competitive Prices via Independent Algorithms," *Marketing Science*, 40, 1–12.

Harrington, Joseph E. (2005): "Detecting Cartels," Tech. rep., Working paper.

——— (2018): "Developing Competition Law for Collusion by Autonomous Artificial Agents," *Journal of Competition Law & Economics*, 14, 331 – 363.

Harsanyi, John C. (1973a): "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points," *International Journal of Game Theory*, 2, 1–23.

——— (1973b): "Oddness of the number of equilibrium points: A new proof," *International Journal of Game Theory*, 2, 235–250.

Hart, Sergiu and Andreu Mas-Colell (2000): "A Simple Adaptive Procedure Leading to Correlated Equilibrium," *Econometrica*, 68, 1127–1150.

——— (2001): *A Reinforcement Procedure Leading to Correlated Equilibrium*, Berlin, Heidelberg: Springer Berlin Heidelberg, 181–200.

——— (2003): "Uncoupled Dynamics Do Not Lead to Nash Equilibrium," *American Economic Review*, 93, 1830–1836.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld (2011): "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance*, 66, 1–33.

Hendershott, Terrence and Ryan Riordan (2009): "Algorithmic Trading and Information," *Manuscript, University of California, Berkeley*.

Hennes, Daniel, Michael Kaisers, and Karl Tuyls (2010): "RESQ-Learning in Stochastic Games," in *Adaptive and Learning Agents (ALA 2010) Workshop*, AAMAS '10, 8–15.

Hennes, Daniel, Karl Tuyls, and Matthias Rauterberg (2009): "State-Coupled Replicator Dynamics," in *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*, ed. by Carles Sierra, Cristiano Castelfranchi, Keith S. Decker, and Jaime Simão Sichman, AAMAS '09, 789–796.

Hofbauer, Josef and Ed Hopkins (2005): "Learning in Perturbed Asymmetric Games," *Games and Economic Behavior*, 52, 133–152.

Hofbauer, Josef and William H. Sandholm (2002): "On the Global Convergence of Stochastic Fictitious Play," *Econometrica*, 70, 2265–2294.

Hoffmann, Peter (2014): "A Dynamic Limit Order Market with Fast and Slow Traders," *Journal of Financial Economics*, 113, 156–169.

Hopkins, Ed (2002): "Two Competing Models of How People Learn in Games," *Econometrica*, 70, 2141–2166.

Hopkins, Ed and Martin Posch (2005): "Attainability of Boundary Points under Reinforcement Learning," *Games and Economic Behavior*, 53, 110–125.

Imhof, Lorens A., Drew Fudenberg, and Martin A. Nowak (2005): "Evolutionary Cycles of Cooperation and Defection," *Proceedings of the National Academy of Sciences*, 102, 10797–10800.

Ioannou, Christos A. and Julian Romero (2014): "A Generalized Approach to Belief Learning in Repeated Games," *Games and Economic Behavior*, 87, 178–203.

Jindani, Sam (2022): "Learning Efficient Equilibria in Repeated Games," *Journal of Economic Theory*, 205, 105551.

Jovanovic, Boyan and Albert J Menkveld (2016): "Middlemen in Limit Order Markets," *Available at SSRN 1624329*.

Kaisers, Michael and Karl Tuyls (2010): "Frequency Adjusted Multi-Agent Q-Learning," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, AAMAS '10, 309–316.

Kalai, Ehud and Ehud Lehrer (1993): "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 61, 1019–1045.

Kandori, Michihiro (1992): "Repeated Games Played by Overlapping Generations of Players," *The Review of Economic Studies*, 59, 81–92.

Kandori, Michihiro and Hitoshi Matsushima (1998): "Private Observation, Communication and Collusion," *Econometrica*, 66, 627–652.

Karandikar, Rajeeva, Dilip Mookherjee, Debraj Ray, and Fernando Vega-Redondo (1998): "Evolving Aspirations and Cooperation," *Journal of Economic Theory*, 80, 292–331.

Kasbekar, Gaurav and Alexandre Proutiere (2010): "Opportunistic Medium Access in Multi-Channel Wireless Systems: A Learning Approach," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1288–1294.

Kawai, Kei and Jun Nakabayashi (2022): "Detecting Large-Scale Collusion in Procurement Auctions," *Journal of Political Economy*, 130, 1364–1411.

Kawai, Kei, Jun Nakabayashi, Juan Ortner, and Sylvain Chassang (2023): "Using Bid Rotation and Incumbency to Detect Collusion: A Regression Discontinuity Approach," *The Review of Economic Studies*, 90, 376–403.

Kirilenko, Andrei, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun (2017): "The Flash Crash: High-Frequency Trading in an Electronic Market," *The Journal of Finance*, 72, 967–998.

Klein, Timo (2021): "Autonomous Algorithmic Collusion: Q-learning under Sequential Pricing," *The RAND Journal of Economics*, 52, 538–558.

Kleinberg, Robert, Georgios Piliouras, and Eva Tardos (2009): "Multiplicative Updates Outperform Generic No-Regret Learning in Congestion Games," in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, New York, NY, USA: Association for Computing Machinery, STOC '09, 533–542.

Klemperer, Paul (2002): "What Really Matters in Auction Design," *The Journal of Economic Perspectives*, 16, 169–189.

Kushner, Harold Joseph and Dean S. Clark (1978): *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, New York: Springer.

Kyle, Albert S (1985): "Continuous Auctions and Insider Trading," *Econometrica*, 1315–1335.

Lamba, Rohit and Sergey Zhuk (2023): "Pricing with Algorithms," *Available at SSRN 4085069*.

Lambin, Xavier (2023): "Less Than Meets the Eye: Simultaneous Experiments as a Source of Algorithmic Seeming Collusion," *Available at SSRN 4498926*.

Lenzo, Justin and Todd Sarver (2006): "Correlated Equilibrium in Evolutionary Models with Subpopulations," *Games and Economic Behavior*, 56, 271–284.

Leonardos, Stefanos, Will Overman, Ioannis Panageas, and Georgios Piliouras (2022): "Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games," in *International Conference on Learning Representations*.

Leslie, David S., Steven Perkins, and Zibo Xu (2020): "Best-Response Dynamics in Zero-Sum Stochastic Games," *Journal of Economic Theory*, 189, 105095.

Levine, David K (2024): "Efficiently Breaking the Folk Theorem by Reliably Communicating Long Term Commitments," *Working Paper*.

Li, Sida, Xin Wang, and Mao Ye (2021): "Who Provides Liquidity, and When?" *Journal of Financial Economics*, 141, 968–980.

Ma, Dye-Jyan, Armand M Makowski, and Adam Shwartz (1990): "Stochastic Approximations for Finite-State Markov Chains," *Stochastic Processes and Their Applications*, 35, 27–45.

Madhavan, Ananth (1995): "Consolidation, Fragmentation, and the Disclosure of Trading Information," *The Review of Financial Studies*, 8, 579–603.

Madhavan, Ananth, David Porter, and Daniel Weaver (2005): "Should Securities Markets be Transparent?" *Journal of Financial Markets*, 8, 265–287.

Maheshwari, Chinmay, Manxi Wu, Druv Pai, and S. Shankar Sastry (2023): "Independent and Decentralized Learning in Markov Potential Games," Available at arXiv:2205.14590v4.

Mailath, George J and Larry Samuelson (2006): *Repeated Games and Reputations: Long-Run Relationships*, Oxford university press.

Maskin, Eric and Jean Tirole (1988): "A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles," *Econometrica*, 56, 571–599.

Meling, Tom Grimstvedt (2021): "Anonymous Trading in Equities," *The Journal of Finance*, 76, 707–754.

Menkveld, Albert J (2016): "The Economics of High-Frequency Trading: Taking Stock," *Annual Review of Financial Economics*, 8, 1–24.

Menkveld, Albert J and Marius A Zoican (2017): "Need for speed? Exchange Latency and liquidity," *The Review of Financial Studies*, 30, 1188–1228.

Mertikopoulos, Panayotis, Ya-Ping Hsieh, and Volkan Cevher (2022): "Learning in Games from a Stochastic Approximation Viewpoint," .

Mertikopoulos, Panayotis and William H. Sandholm (2016): "Learning in Games via Reinforcement and Regularization," *Mathematics of Operations Research*, 41, 1297–1324.

Métivier, M. and P. Priouret (1984): "Applications of a Kushner and Clark Lemma to General Classes of Stochastic Algorithms," *IEEE Transactions on Information Theory*, 30, 140–151.

Mguni, David, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang (2021): "Learning in Nonzero-Sum Stochastic Games with Potentials," in *International Conference on Machine Learning (ICML)*, 7688–7699.

Monderer, Dov and Lloyd S. Shapley (1996): "Potential Games," *Games and Economic Behavior*, 14, 124–143.

Nachbar, John H. (1997): "Prediction, Optimization, and Learning in Repeated Games," *Econometrica*, 65, 275–309.

Nowak, Martin A., Akira Sasaki, Christine Taylor, and Drew Fudenberg (2004): "Emergence of Cooperation and Evolutionary Stability in Finite Populations," *Nature*, 428, 646–650.

OECD (2017): "Algorithms and Collusion: Competition Policy in the Digital Age," Available at `https://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm`.

Ortner, Juan M, Sylvain Chassang, Kei Kawai, and Jun Nakabayashi (2022): "Screening Adaptive Cartels," Tech. rep., National Bureau of Economic Research.

Panait, Liviu, Karl Tuyls, and Sean Luke (2008): "Theoretical Advantages of Lenient Learners: An Evolutionary Game Theoretic Perspective," *Journal of Machine Learning Research*, 9, 423–457.

Pemantle, Robin (1990): "Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations," *The Annals of Probability*, 18, 698–712.

Perkins, Steven and David S. Leslie (2012): "Asynchronous Stochastic Approximation with Differential Inclusions," *Stochastic Systems*, 2, 409–446.

Piccione, Michele (1992): "Finite Automata Equilibria with Discounting," *Journal of Economic Theory*, 56, 180–193.

Porter, Robert H (2005): "Detecting Collusion," *Review of Industrial Organization*, 26, 147–167.

Porter, Robert H and J Douglas Zona (1993): "Detection of Bid Rigging in Procurement Auctions," *Journal of Political Economy*, 101, 518–538.

Porter, Robert H. and J. Douglas Zona (1999): "Ohio School Milk Markets: An Analysis of Bidding," *The RAND Journal of Economics*, 30, 263–288.

Salcedo, Bruno (2015): "Pricing Algorithms and Tacit Collusion," .

Sato, Yuzuru, Eizo Akiyama, and J. Doyne Farmer (2002): "Chaos in Learning a Simple Two-Person Game," *Proceedings of the National Academy of Sciences*, 99, 4748–4751.

Sato, Yuzuru and James P. Crutchfield (2003): "Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems," *Phys. Rev. E*, 67, 015206.

Sayin, Muhammed O., Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar (2021): "Decentralized Q-learning in Zero-sum Markov Games," *Advances in neural information processing systems*, 35.

Sayin, Muhammed O., Francesca Parise, and Asuman Ozdaglar (2022): "Fictitious Play in Zero-Sum Stochastic Games," *SIAM Journal on Control and Optimization*, 60, 2095–2114.

Simaan, Yusif, Daniel G. Weaver, and David K. Whitcomb (2003): "Market Maker Quotation Behavior and Pretrade Transparency," *The Journal of Finance*, 58, 1247–1267.

Singh, Satinder, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári (2000): "Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms," *Machine Learning*, 38, 287–308.

Stigler, George J. (1964): "A Theory of Oligopoly," *Journal of Political Economy*, 72, 44–61.

Sugaya, Takuo (2021): "Folk Theorem in Repeated Games with Private Monitoring," *The Review of Economic Studies*, 89, 2201–2256.

Sugaya, Takuo and Yuichi Yamamoto (2020): "Common Learning and Cooperation in Repeated Games," *Theoretical Economics*, 15, 1175–1219.

Sutton, Richard S. and Andrew G. Barto (2018): *Reinforcement Learning: An Introduction*, The MIT Press, second ed.

Tuyls, Karl, Katja Verbeeck, and Tom Lenaerts (2003): "A Selection-Mutation Model for Q-Learning in Multi-Agent Systems," in *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, NY, USA: Association for Computing Machinery, AAMAS '03, 693–700.

Vrancx, Peter, Karl Tuyls, and Ronald Westra (2008): "Switching Dynamics of Multi-Agent Learning," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '08, 307–313.

Waltman, Ludo and Uzay Kaymak (2008): "Q-learning Agents in a Cournot Oligopoly Model," *Journal of Economic Dynamics and Control*, 32, 3275–3293.

Wiseman, Thomas (2005): "A Partial Folk Theorem for Games with Unknown Payoff Distributions," *Econometrica*, 73, 629–645.

——— (2012): "A Partial Folk Theorem for Games with Private Learning," *Theoretical Economics*, 7, 217–239.

Wunder, Michael, Michael Littman, and Monica Babes (2010): "Classes of Multiagent Q-learning Dynamics with ϵ-greedy Exploration," in *Proceedings of the 27th Annual International Conference on Machine Learning*, ICML '10, 1167–1174.

Yang, Liyan and Haoxiang Zhu (2019): "Back-Running: Seeking and Hiding Fundamental Information in Order Flows," *The Review of Financial Studies*, 33, 1484–1533.