REINFORCEMENT LEARNING AND COLLUSION

CLEMENS POSSNIG

Abstract. This paper develops an analytical framework to characterize the long-run policies learned by repeatedly interacting algorithms. Algorithms observe a state variable and update policies to maximize long-term payoffs; their long-run policies correspond to the stable equilibria of a tractable differential equation. In a repeated Bertrand game, I derive necessary and sufficient conditions under which Nash equilibria are learned. This reveals how the interplay between monitoring technology (state variables) and market conditions determines whether competitive or collusive outcomes emerge. I apply these insights to evaluate two key regulatory policies: limiting algorithmic data inputs and imposing competition in the software provider market.

JEL classification. C62, C73, D43, D83.

Keywords. Multi-Agent Reinforcement Learning, Repeated Games, Collusion, Learning in Games.

More and more companies are using artificial intelligence-based tools in their pursuit of profit maximization. Such algorithms take market data to determine current price levels, updating in real-time. Algorithms can help firms adapt to rapidly changing market environments, and potentially better serve their markets. However, these tools appear to have an inherent ability to collude. Studying the German gasoline retail market, Assad et al. 2024 observe that after a critical mass of firms deployed pricing algorithms, profit margins rose by

Date: July 28, 2025.

Clemens Possnig: University of Waterloo, cpossnig@uwaterloo.ca.

I thank my committee members Li Hao, Vitor Farinha Luz, and Michael Peters for years of guidance and conversations. I am grateful to Rohit Lamba, Alexander Frankel, Kevin Leyton-Brown, Wei Li, Vadim Marmer, Jesse Perla, Chris Ryan, and Kevin Song for many helpful discussions. I thank the participants at SEA '23, SAET '23, INFORMS '23, EC '22, GTA '22, CORS/INFORMS '22, and CETC '22 for insightful comments. I also thank participants of the theory lunches at VSE for their extensive feedback. I gratefully acknowledge support through a University of Waterloo SSHRC Institutional Grant (SIG) .

28%. This finding, alongside numerous simulation-based studies[1] , raises an urgent question: Which algorithms, and which markets, are likely to support such outcomes?

I first introduce a common family of RL algorithms that repeatedly play a game. While the results are more general, the leading application considered is Bertrand price competition. The algorithms observe a common state variable without knowing their payoff function or state transition likelihoods, and adapt by repeatedly experimenting with price choices and estimating a value function. I show that to pin down the long run behavior of the algorithmic learners, it is enough to find the stable rest points of a differential equation.

Next, I use this characterization to study whether the algorithms can learn to repeat the static Nash equilibrium, which we can think of as the non-collusive benchmark. It turns out that the answer depends on properties of the state variable and market conditions. Learning to repeat the static Nash equilibrium may be impossible, even within classical, well-studied[2] stage games.

Allowing for arbitrary, finite, state variables, the possibilities for collusive behavior are ample. I therefore focus on the payoff-optimal collusive equilibrium as a bound to the possibilities of cartelization in this setting. Through an approximation exercise, I show that optimal collusive equilibria are closely related to optimal imperfect monitoring equilibria of the bang-bang kind, as characterized in Abreu, Pearce, and Stacchetti 1986. Conditions for such equilibria to be learned also come down to properties of the state variable and the market competed in.

The algorithms' state variable serves as a model of the data input chosen by the algorithm designer.[3] Even without the designer's intent, this data acts as a monitoring technology through the correlation between data outcomes and price choices of the algorithms. Whether static Nash can be learned by algorithms comes down to an interplay between the sensitivity of the monitoring technology, and market conditions such as own- and opponents-price elasticities.

[1]Klein 2021, Calvano, Calzolari, Denicoló, et al. 2021 show that algorithms may learn to play repeated game strategies akin to typical carrot-and-stick type strategies studied in the economic theory literature.
[2]Linear or logit demand, and constant marginal cost.
[3]E.g., aggregate market conditions, time of day, sales volume, consumer data, etc.

One can think of monitoring sensitivity a a measure of the ability to identify deviations in the continuous space, related to Fudenberg, D. Levine, and Maskin 1994's pairwise identifiability condition. Fixing market conditions, higher monitoring sensitivity makes learning static Nash less likely. Fixing the monitoring technology, this equilibrium is more likely to be learned if there is weaker competition and higher own-price elasticity. Turning to collusion, conditions for such equilibria to be learned are, again, related to the sensitivity of the monitoring technology, and market conditions. A less sensitive monitoring technology may make it more likely for this optimal equilibrium to be learned, but at the same time may decrease the payoffs to that equilibrium.

The inuition for these insights comes down to the stochastic nature of the algorithmic learning process. Any real-world application of algorithmic learning involves estimation, e.g. of payoff functions, and the possibility of making mistakes. Mistakes induce perturbations in currently played policy profiles. For perturbations to have any impact down the line, they must be detected, i.e. monitoring must be sensitive enough. On the other side, given that perturbations are detected, they must matter; here market conditions come into play. Perturbations matter less if competitor's prices don't have much of an impact (weak competition), or if small adjustments are enough to correct for perturbations (large own-price elasticity).

I then consider two policy options: a restriction to the data algorithms may be allowed to employ as their inputs; and imposing competition in the algorithmic software provider market. I show that the former has potential, as it can be used to decrease the sensitivity of the monitoring technology, enabling the learning of the static Nash equilibrium. On the other hand, the latter comes only with ambiguous predictions.

I finish by extending the Bertrand application to one allowing for market dynamics. This setting is especially interesting as it directly applies as a model of Assad et al. 2024's study on algorithmic collusion in the German gasoline retail market. I conclude that also in this market, a restriction on data inputs can facilitate the learning of the most competitive option, static Nash.

To the best of my knowledge, this is the first theoretical work on algorithmic collusion that uncovers the relationship between standard economic concepts such as price elasticity and monitoring technologies, and the learning of collusive behavior.

Related Literature

This project speaks to results in the fast-growing literature on algorithmic collusion, the theory of learning in games, as well as the study of asymptotic behavior of algorithms in the computer science literature. A more detailed discussion can be found in the online appendix[4].

Firstly, the literature on algorithmic collusion has received increasing attention in recent years. Assad et al. 2024 provide an empirical study supporting the hypothesis that algorithms may learn to play collusively, while there are many simulation studies suggesting the same, of which Calvano, Calzolari, Denicolo, et al. 2020, Calvano, Calzolari, Denicoló, et al. 2021, and Klein 2021 are important examples. A paper close in spirit to this study is Banchio and Mantegazza 2022. They consider a fluid approximation technique related to the stochastic approximation approach applied here, and recover novel phenomena regarding the learning of cooperation for finite-action class of RL algorithms without memory, with a focus on Q-learning. The family of algorithms studied here concerns learning in continuous-action games with repeated game strategies, such as repeated Bertrand or Cournot competition. While a main example is an extension of Q-learning (ACQ, see discussion below), it cannot accommodate Q-learning as a special case, nor is it a special case of Q-learning . Cartea et al. 2022 show how stochastic approximation can be applied in the analysis of finite-action reinforcement learners such as Q-learners as well. Meylahn and V. den Boer 2022, Loots and V. den Boer 2023 use ODE methods related to the ones applied in this paper to prove that specific algorithms can learn to collude in a pricing game. The framework developed here is more general, shifting the focus from the properties of a particular algorithm and its determininig parameters to the interplay between market structure and the data available to a family of algorithms. Further important recent work in the area of algorithmic collusion includes Lamba and Zhuk 2022, Brown and MacKay 2021, Johnson, Rhodes, and Wildenbeest 2020, and Salcedo 2015. These papers feature stylized models of algorithmic

[4] https://cjmpossnig.github.io/papers/RLColl_onlineApp_0.pdf

competition, abstracting away from issues of learning and estimation, which are an important aspect of my analysis.

Secondly, this paper connects to the theory of learning in games. Classically, this literature has been concerned with the ability of agents to learn a Nash equilibrium of the stage game when following a given learning rule (e.g. Milgrom and Roberts 1991, Fudenberg and Kreps 1993). More recent results concern learning in stochastic games (e.g. Leslie, Perkins, and Xu 2020), where the state variable is taken as an exogenous object. The class of algorithms studied here has the ability to learn *repeated game strategies*, i.e. strategies that condition on summaries of the history of the game, implemented as automaton strategies. The games that can be studied here therefore contain stochastic games as a special case, but also allow for the case where the state that agents observe represents a history of the repeated interaction.

My class contains algorithms that impose little informational assumptions as a special case, known as "model free". Such algorithms do not carry a model of opponent behavior, and also no model of their environment and own payoffs. Thus, this class falls into the family of adaptive uncoupled learning rules as defined in Hart and Mas-Colell 2003. Further foundational papers in this literature include Milgrom and Roberts 1990 , Fudenberg and D. K. Levine 2009, Gaunersdorfer and Hofbauer 1995, and many more.

Thirdly, this paper makes use of and generalizes an extensive body of research related to stochastic approximation theory (see e.g. Borkar 2009) and hyperbolic theory (Palis Jr, Melo, et al. 1982). The generalizations are relegated to the online appendix. There is a growing strand of the computer science literature devoted to establishing convergence proofs in multi-agent algorithmic environments. The paper in that area closest to this one is Mazumdar, Ratliff, and Sastry 2020.

I. Multi-Agent Learning

This section introduces the updating rule (algorithm) and main assumptions used as running example in this paper. The algorithm is known as actor-critic Q-learning (ACQ). These algorithms keep track of an estimated performance criterion (the "critic", or Q-function, essentially a value function) and a policy function (the "actor") that is updated towards the maximizer of the performance criterion. The policy is a mapping from observables (states), such as past prices or other market data, to actions, in this case prices. A main advantage of ACQ over the simpler and more commonly known Q-learning (Watkins 1989) is that it directly applies to continuous-action problems, which are the focus of this paper.

In standard Q-learning, a Q-matrix is the algorithmic updating object. As these matrices can only take finitely many elements, Q-learning is ideally suited for finite-action problems, requiring discretization in the continuous-action world. The matrix represents what are called "state-action values", which one can think of as a value function before maximiziation over one-shot deviations is applied. This learning rule then uses an estimation-based Bellman iteration to estimate this value function, known as the Q-function. Importantly, the policy found by such an algorithm can be directly 'read-off' of the Q-function. This implies that Q-policies move more erratically than actor-critic methods, as such methods dampen the relationship between policy and critic. Advantages of actor-critic methods therefore include variance reduction of updates, among others.

The results presented in this paper are more general than the case of a Q-function used as the critic. A broader characterisation of algorithms for which the results stated here hold can be found in Appendix A. The online appendix further extends results to allow for non-vanishing bias in the critic estimation. As critic estimation generally involves function approximation (especially in the continuous action case), non-vanishing estimation bias is an important concern.

Actor-critic reinforcement learning, which is a superset of the class studied here, has become popular in the reinforcement learning commmunity, due to the variance reduction property mentioned before and higher flexibility than pure critic- or actor-based methods (Q-learning being a critic-based method). See e.g. the substantial popularity of PPO (Schulman et al. 2017). Actor-critic learning has been studied in stage game settings before, see e.g. Leslie and Collins 2003.

I.A. **Payoffs and Strategies**

There are N algorithms indexed by i, each having as action space an interval $Y_i \subseteq \mathbb{R}$, with profile space $Y = \times_i Y_i$. A state variable S taking values in space O_S with $|O_S| = K < \infty$ comes with a transition probability function, twice differentiable in Y, $T : O_S^2 \times Y \to [0, 1]$.

Each algorithm has a stage game payoff function $\pi^i : Y \times O_S \to \mathbb{R}$, $\mathcal{C}^2$ in Y, and common discount factor $\delta \in (0,1)$.[5]

Throughout, it is important to keep in mind that I define an environment competed on not by rational agents, but by algorithms constrained to play policies based on a fixed domain: O_S. I will take S as an exogenous object chosen by whoever initialized the algorithm. I will assume throughout that the state variable and current state s is a common observable to all algorithms.

Algorithms update a policy function $\rho_{i,t} : O_S \to Y_i$. Since states are finite, policy profile $\boldsymbol{\rho} \in \overline{Y} = Y^{NL}$ can be represented as a vector in $\mathbb{R}^{NL}$.

Assumption 1. *For all $\rho \in \overline{Y}$, the Markov chain induced by $T_{ss'}[\rho(s)]$ is irreducible and aperiodic.*[6]

In fact, one can view any such policy as a stationary Markov strategy given state variable S. Next, define $\overline{Y}_i = Y_i^L$, and $\overline{Y}_{\text{-}i} = \times_{j \neq i} \overline{Y}_j$.

Expected future discounted payoffs $W^i(\boldsymbol{\rho}_i, \rho_{\text{-}i}, s_0)$ are defined given stationary policy profiles $(\boldsymbol{\rho}_i, \boldsymbol{\rho}_{\text{-}i}) \in \overline{Y}$, and any initial state $s_0 \in O_S$:

$$W^i(\boldsymbol{\rho}_i, \boldsymbol{\rho}_{\text{-}i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t \pi^i \left(\boldsymbol{\rho}(s_t), s_t \right), \tag{1}$$

where the expectation is taken over the randomness in the stage game payoffs and state transitions. Define $B_S^i(\boldsymbol{\rho}_{\text{-}i})$ as the optimal policy for i given a profile $\boldsymbol{\rho}_{\text{-}i} \in \overline{Y}_{\text{-}i}$, chosen from the constraint set of stationary, S-state policies:

$$B_S^i(\boldsymbol{\rho}_{\text{-}i}) = \arg\max_{\boldsymbol{\rho} \in \overline{Y}_i} W^i(\boldsymbol{\rho}, \rho_{\text{-}i}, s_0), \tag{2}$$

where due to our assumption on irreducibility of the state space the optimal policy does not depend on the initial state s_0. The optimal policy is indeed optimal over all possible history-dependend policies since given a Markov stationary opponent profile $\boldsymbol{\rho}_{\text{-}i}$ there must be a Markov stationary best response. In what follows, write $\bar{B}_S(\boldsymbol{\rho})$ as the stacked best response correspondence over i.

[5]Let $\mathcal{C}^i[Y, C]$ be the set of functions that are i times continuously differentiable, with domain Y and range C. When domain and range are clear, I write $\mathcal{C}^i$.

[6]For definitions, see e.g. Appendix A in Puterman 2014

Definition 1. *Define*

(1) $E_S \subset \overline{Y}$ *to be the set of Nash equilibria in policy profiles based on payoff functions* W^i*. In other words,* E_S *is the set of profiles* $\boldsymbol{\rho}^*$ *s.t.* $\boldsymbol{\rho}^* \in \bar{B}_S(\boldsymbol{\rho}^*)$*.*

(2) $\rho^* \in E_S$ *as 'differential Nash equilibrium' if* $\boldsymbol{\rho}^*$ *is interior, first order conditions hold for each agent at* $\boldsymbol{\rho}^*$*, and the Hessian of each agent's optimization problem at* $\boldsymbol{\rho}^*$ *is negative definite.*

Given these definitions regarding the underlying payoff environment, assume:

Assumption 2 (Equilibrium existence and differentiability)**.**

(1) *Given state variable* S*, stationary equilibrium profiles* $\rho^* \in \overline{Y}$ *exist.*

(2) *There exist* $\boldsymbol{\rho}^* \in E_S$ *that are differential Nash equilibria.*

A sufficient condition for both points in Assumption 2 to hold is the existence of a differential static Nash equilibrium, given $\pi(a, s)$ for all $s \in O_S$. As our analysis of limiting strategies will depend on a smoothness condition of an underlying differential equation at some equilibria, the second point will prove crucial.

I.B. **Algorithmic Learning in Reduced Form**

The purpose of the following analysis is to understand long run outcomes of machine learning algorithms that satisfy consistency in their critic estimation step, in a manner introduced below. This can be seen as a benchmark for ideal individual behavior of a family of algorithms. For each algorithm i, let $F_S^i : \overline{Y} \to \mathbb{R}^K$ refer to the performance criterion, i.e. critic term representing the function approximation target the learning algorithm intends to estimate.

For ACQ learning, one would have $F_S^i(\boldsymbol{\rho}) = B_S^i(\boldsymbol{\rho}_{\text{-}i}) - \boldsymbol{\rho}_i$[7], which is an extension of best response dynamics (Gilboa and Matsui 1991) to a setting of repeated-games payoffs. A common alternative are gradient updating schemes, in which case one would take $F_S^i(\boldsymbol{\rho}) = \frac{\partial}{\partial \boldsymbol{\rho}_i} W^i(\boldsymbol{\rho}_i, \boldsymbol{\rho}_{\text{-}i}, s_0)$, the vector of payoff gradients of the learner. General results in the next section apply to both[8]. Some results in Section III apply only partly to both, which will be

[7]The results presented in the following can be generalized to a setting where $B_S^i(\cdot)$ is not single-valued on a positive-measure set of opponent strategies, as long as focus remains on differential Nash equilibria.

[8]And their combination, as discussed in the policy subsection.

noted. When specificity is required, I will refer to the former as $F^i_{S,B}$ and to the latter as $F^i_{S,G}$.

This paper remains agnostic about the specificities of the critic estimation part of the algorithms. The goal is to gain insights about what can be learned as long as this function approximation step is reasonably well behaved, a property to be defined below. Note that well-behavedness in the critic estimation does not imply convergence of the algorithms to a Nash equilibrium, or convergence to some specific ideal strategy.

For each i, policies $\rho_{i,t}$ update according to

$$\boldsymbol{\rho}_{i,t+1} = \boldsymbol{\rho}_{i,t} + \alpha^i_t \left[F^i_S \left(\boldsymbol{\rho}_t \right) + \boldsymbol{d}^i_{t+1} + \boldsymbol{M}^i_{t+1} \right], \tag{3}$$

where $\alpha^i_t > 0$ is a sequence of stepsizes converging to $\underline{\alpha} \geq 0$, $\boldsymbol{d}^i_{t+1}$ is a bias term converging to zero, and $\boldsymbol{M}^i_{t+1}$ is an error term of bounded variance. Bias and error term represent the estimation error involved in the estimation of F^i_S. For some estimator $\hat{F}^i_{S,t}$, one can write $\boldsymbol{d}^i_{t+1} + \boldsymbol{M}^i_{t+1} = \hat{F}^i_{S,t}(\boldsymbol{\rho}_t) - F^i_S(\boldsymbol{\rho}_t)$. I assume throughout that stepsizes of all agents lie within an order of magnitude of each other. Detailed sufficient conditions on stepsizes, bias, and error terms are relegated to Appendix A. The conditions ensure that (3) can be interpreted as a Robbins-Monro scheme (Robbins and Monro 1951), to which an extensive machinery for asymptotic results has been developed (c.f. Borkar 2009).

II. Long Run Behavior: Main Results

This section presents the main results regarding characterisation of long run behavior of the algorithms. For a set A, let $cl\,(A)$ be its closure.

Definition 2. *Take the algorithm defined in* (3). *The limit set is defined as*

$$L_S = \bigcap_{t \geq 0} cl \left(\{ \boldsymbol{\rho}_\ell \,|\, \ell \geq t \} \right),$$

the set of limits of convergent subsequences $\boldsymbol{\rho}_{t_k}$.

The limit set represent what I refer to as the long run policies learned by the algorithms. This definition allows for the possibility of non-convergence. I write S as subscript to underline the dependence of the limiting set on the state variable S. As the characterizations introduced here will require properties of a differential equation, I present next some useful definitions:

Definition 3. *Given some ODE $\dot{\boldsymbol{\rho}} = f(\boldsymbol{\rho})$, let $\boldsymbol{\rho}^*$ be a rest point of $f(\boldsymbol{\rho})$. Let $\Lambda = eig[Df(\boldsymbol{\rho}^*)]$ the set of eigenvalues of the linearization of f at $\boldsymbol{\rho}^*$. For a complex number z, let $Re[z] \in \mathbb{R}$ be the real part. $\boldsymbol{\rho}^*$ is*

- *Asymptotically stable if $Re[\lambda] < 0$ holds for all $\lambda \in \Lambda$.*
- *Linearly unstable if $Re[\lambda] > 0$ holds for at least one $\lambda \in \Lambda$.*

One can think of $Re[\lambda] < 0$ as a contraction property of the dynamical system around the rest point. Asymptotically stable rest points are *attractors* of the ODE. In other words, if the dynamical system were to start close to such a rest point, it will converge to it. On the other side, linearly unstable rest points don't come with a contraction property. There is at least some repelling direction of the ODE around the rest point.

To save notation, write $F_S(\boldsymbol{\rho})$ as stacked version of critic terms $F_S^i(\boldsymbol{\rho})$ for $\boldsymbol{\rho} \in \overline{Y}$.

Theorem 1. *Let $\boldsymbol{\rho}^* \in E_S$ be asymptotically stable for F_S. Then*

$$\mathbb{P}[L_S = \{\boldsymbol{\rho}^*\}] > 0.$$

Proof Sketch of Theorem 1

The full proof for this and the following Theorems can be found in the online appendix.

Firstly, I make a connection between the recursion in (3) and the differential equation induced by F_S. One can relate a time-interpolated version of the recursion $\boldsymbol{\rho}_t$ to solutions to the ordinary differential equation

$$\dot{\boldsymbol{\rho}} = \boldsymbol{A} F_S(\boldsymbol{\rho}(t)),$$

where $\boldsymbol{A}$ is a diagonal matrix of strictly positive weights, representing the limiting relative stepsizes of different algorithms i. When considering that the updating rates α_t^i converge to zero, one may convince oneself that the recursion looks similar to a discrete approximation to a time-derivative. The idea is to show that the time-interpolated version of $\boldsymbol{\rho}_t$ must stay close, almost surely, to solutions of $\boldsymbol{A} F_S(\boldsymbol{\rho})$. Attracting points of the differential system are then natural candidates to also attract $\boldsymbol{\rho}_t$.

On the other hand, learning to play unstable rest points is an issue:

Theorem 2. *Let $\boldsymbol{\rho}^* \in E_S$ be linearly unstable for $\boldsymbol{A}F_S$. Then there exists an open neighborhood U of $\boldsymbol{\rho}^*$ such that*

$$\mathbb{P}[L_S \in U] = 0.$$

Proof Sketch of Theorem 2

$\boldsymbol{\rho}^*$ being unstable implies that there exists an unstable manifold that $\boldsymbol{\rho}^*$ lies on, which acts as a repeller to the differential equation based on F_S. I go on to show that due to the instability of $\boldsymbol{\rho}^*$ and nonvanishing variance of noise term $\boldsymbol{M}_{t+1}$, no matter how close the algorithmic process gets to $\boldsymbol{\rho}^*$, and no matter how large t is, there is always a nonzero probability that $\boldsymbol{\rho}_t$ lands on the unstable manifold and therefore must move away from $\boldsymbol{\rho}^*$.

Hence, asymptotically stable equilibria are equilibria that can be limiting points of the RL learning procedure, while unstable equilibria are not. The intuition is related to how RL learn to play: since such agents make errors due to estimation and also to explore their action space, opponent's strategy profiles are constantly perturbed. In other words, out of the view of a fixed agent i, the other agents are frequently deviating to policies nearby in the policy space. Now suppose the current profile $\boldsymbol{\rho}_t$ is close to an equilibrium $\boldsymbol{\rho}^*$. Since i's updating rule tracks F_S, their policy will only stay close to $\boldsymbol{\rho}^*$ if the dynamics of F_S are somehow robust to deviations. This robustness is implied by asymptotic stability, and broken by unstable equilibria.

There is a caveat here, however: Theorem 1 does not state that all limiting points in L_S will be equilibria of the underlying repeated game as played by rational players. Depending on details of the stage game and state variable, one may or may not be able to rule out the case where algorithm updates get trapped in a cycle, or other more complex behavior not involving rest points (see Papadimitriou and Piliouras 2018). I do not include cycles in the above definition, however it is straightforward to extend Theorem 1 to the case of attracting cycles as in Faure and Roth 2010, and there exist results considering linearly unstable cycles (Michel Benaïm and Faure 2012) that suggest one may extend Theorem 2 to such linearly unstable cycles also.[9] Notice that this observation implies that the Folk theorem is neither necessary nor sufficient in describing the possible payoffs achieveable by learning algorithms.

III. Bertrand Application

With the theoretical framework of long run learning in place, we can move to an application of market competition-a differentiated goods Bertrand model (Bertrand model for short).

There are two firms, $i \in \{1,2\}$. Firms choose prices $p_i \in Y \subseteq \mathbb{R}_+$. After prices are posted, consumers choose whether to consume, and at which firm with some idiosyncratic noise; this leads to a random aggregate demand shock observable by all. Let $\tilde{A} \in \Omega \subset \mathbb{R}_+$ be the random variable representing the aggregate demand shock. Assume Ω is compact. It is natural that consumer's choices will be affected by posted prices. Hence, conditional on the price vector $\boldsymbol{p}$, $\tilde{A}$ has an absolutely continuous distribution with a density $g(a;\boldsymbol{p})$. I assume that this density only depends on aggregate prices, what one can think of as a price index level. We can write $g(a;\boldsymbol{p}) = g(a;P)$ for $P = p_1 + p_2$. Demand for each individual firm is then a random variable, denoted $\tilde{X}_i$. $\tilde{X}_i$ depends on other firm's actions only through $\tilde{A}$. The expectation of $\tilde{X}_i$ given $\boldsymbol{p}$ is then written as $X_i(\boldsymbol{p}) = \mathbb{E}[\tilde{X}_i \mid \boldsymbol{p}]$.

Expected profits given $\boldsymbol{p}$ are then given by $\pi_i(\boldsymbol{p}) = (p_i - c)X_i(\boldsymbol{p})$, for some common marginal cost $c \geq 0$. For ease of exposition, we assume that $\pi_i(\boldsymbol{p})$ are symmetric. The results extend in quality to the asymmetric setting. The next assumption ensures that the games considered here adhere to standard intuitions about Bertrand games.

Assumption 3. *For all $\boldsymbol{p} \in \mathbb{R}^2_+$, $p_{\text{-}i} \geq 0$:*

(1) $\frac{\partial}{\partial p_i} X_i(\boldsymbol{p}) \leq 0$, $\frac{\partial}{\partial p_{\text{-}i}} X_i(\boldsymbol{p}) \geq 0$.
(2) $\frac{\partial}{\partial p_i} X_i(0, \boldsymbol{p}_{\text{-}i}) < 0$.
(3) $\pi_i(\boldsymbol{p})$ *is quasiconcave in* p_i.
(4) $\left|\frac{\partial}{\partial p_i} X_i(\boldsymbol{p})\right| \geq \left|\frac{\partial}{\partial p_{\text{-}i}} X_i(\boldsymbol{p})\right|$
(5) $\left|\frac{\partial^2}{(\partial p_i)^2} X_i(\boldsymbol{p})\right| \geq \left|\frac{\partial^2}{\partial p_i \partial p_{\text{-}i}} X_i(\boldsymbol{p})\right|$.
(6) For all $a \in \Omega$, $g(a;\boldsymbol{p})$ *is twice differentiable in* p.

[9] The inclusion of an analysis of limit cycles is an interesting avenue of further research, but would be beyond the scope of this paper.

(1) is a natural assumption in the Bertrand game. (2) ensures that positive prices will optimally be played. (3) implies that first order conditions are sufficient for best responses in the stage game. (4) and (5) ensure that effects of own price decisions dominate opponent's decision's impact on own demand. (6) is a regularity condition that will become useful for the results on stability of equilibria.

Example 1. *Consider the linear demand differentiated Bertrand model, with $X_i(\boldsymbol{p}) = A - bp_i + \gamma p_{\text{-}i}$, $b > \gamma > 0$. If one takes $\tilde{A}$ to feature linearly in the individual demand $\tilde{X}_i$, and has $\mathbb{E}[\tilde{A} \mid \boldsymbol{p}] = Z + p_1 + p_2$ for some $Z \geq 0$, this model accomodates our motivation above, and satisfies Assumption 3.*

Example 2. *A slight extension to the logit-demand model also accomodates this setting. Suppose a mass of consumers $Z > 0$, upon observing a price index $P = p_1 + p_2$, stochastically decide to participate in the market. The expected mass of consumers given P is $M(P) = Z - P \geq 0$. Conditional on participation, the standard logit-demand model is carried out. Then we have $X_i(\boldsymbol{p}) = M(P)\exp(\mu_i - \beta_i p_i) / \left(\sum_{j=1}^{2} \exp(\mu_j - \beta_j p_j)\right)$, where $\mu_i, \beta_i > 0$.*[10]

Under the trivial state variable S_0 which takes only one value (i.e. only stage game learning is possible), refer to $F_{S_0,B}, F_{S_0,G}$ as static best response dynamics (where payoffs are stage game payoffs, not repeated game payoffs as in payoff function W), and static gradient dynamics, respectively.

Lemma 1. *Under Assumption 3, there is a unique interior Nash equilibrium p_N, which is symmetric. Given state S_0, this Nash equilibrium will be learned with probability* 1 *by ACQ learners, and with positive probability by gradient learners.*

Hence, if algorithms were not able to learn based on state variables correlated to past actions, learning to play Nash is a likely outcome. The introduction of a state variable can be seen as the introduction of a monitoring technology, as policies can now condition on random variables correlated to each agent's past actions. This interpretation is useful

[10]Assumption 3 needs to be weakened for this example. The remaining claims in this section carry over here as well for bounded p_i, and large enough β_i, which ensures that signs in the assumption aren't flipped for a large enough set of prices.

also in that the results in this section directly relate to the effectiveness of the monitoring technology available to the algorithms.

The results in this section are stated and proved under the assumption that $\frac{\alpha_t^1}{\alpha_t^2} \to 1$ as $t \to \infty$, i.e. stepsizes of the two agents are asymptotically equal. The more general setting is discussed later on.

III.A. **Learning Nash**

Recall that we defined $T_{ss'}(p_1, p_2)$ as the transition probability of moving to s' given current state s and choices p_i in state s. Assume that $T_{ss'}(p_1, p_2) = \mathbb{P}\left[s' \middle| s; p_1 + p_2\right]$ for all s, p_i, i.e. as is true for aggregate shock $\tilde{A}$, transition probabilities only depend on aggregate prices, what one can think of as a price index level. I will therefore commonly write $T_{ss'}(p_1, p_2) = T_{ss'}(P)$ with $P = p_1 + p_2$.

Lemma 1 implies that there exist trivial state variables (taking a constant value) under which static Nash can be expected to be learned by ACQ learners and gradient learners. I show next that even though that is true, a larger family of state variables exist so that when they are used, ACQ learners will not converge to this Nash equilibrium.

For the remainder of this section, if not further specified, results are stated regarding ACQ learners ($F_{S,B}$) only. To ease intuitions, consider a state variable S the transitions of which depend on realizations of the aggregate shock $\tilde{A}$. Since $\tilde{A}$ is the only payoff-relevant variable correlated to all agent's choices, we call such state variables PR-states, or just PR (payoff-relevant). For such state variables, transitions are pinned down by a deterministic mapping $f_S(s, \tilde{A}) \in O_S$ for all $s \in O_S$. Define own-and-opponent's price elasiticty of demand as

$$\xi_o(\boldsymbol{p}) = \left(\frac{\partial}{\partial p_1} X_1(\boldsymbol{p})\right) \frac{p_1}{X_1(\boldsymbol{p})}, \qquad \xi_c(\boldsymbol{p}) = \left(\frac{\partial}{\partial p_2} X_1(\boldsymbol{p})\right) \frac{p_2}{X_1(\boldsymbol{p})},$$

and let $L(p) = \frac{p-c}{c}$, with $L_N = L(p_N)$ being the Lerner index at Nash. Finally, define the growth rate of the Lerner index as $G_N = \left(\frac{\partial}{\partial p_1} L(p)\right)\Big|_{p=p_N} / L_N$. Write

$$d\xi_o(\boldsymbol{p}) = \frac{\partial}{\partial p_1} \xi_o(\boldsymbol{p}) + \frac{\partial}{\partial p_2} \xi_o(\boldsymbol{p}).$$

$d\xi_o(\boldsymbol{p}) < 0$ implies that price hikes by both agents lead to an overall increase in magnitude of elasticity for agent 1, as $\xi_o(\boldsymbol{p}) \leq 0$ by Assumption 3. Call such markets *balanced*.[11] Finally, note that π_i depends on conditional density $g(a;p)$ only through expected demand $X_i(\boldsymbol{p})$. Say two densities g, g' *induce* π if both densities lead to the same expected demand function.

Theorem 3. *Consider a* PR-*state variable* S_{PR} *with* $|O_{S_{PR}}| = K \geq 1$ *states, and let* $\rho_N : O_{S_{PR}} \to \mathbb{R}_+$ *be the policy that plays* p_N *in every state. Given* π*, there exists* $0 < C < \infty$ *such that for any* g *inducing* π*,*

(1) ρ_N *will be learned with positive probability if* $\sup_{a\in\Omega} \left| \frac{\partial}{\partial P} g(a;P)\big|_{P=P_N} \right| < C$.

(2) ρ_N *will not be learned if* $\inf_{a\in\Omega} \left| \frac{\partial}{\partial P} g(a;P)\big|_{P=P_N} \right| > C$.

(3) C *is proportional to*

$$\frac{1}{\xi_c(p_N)} \left(|\xi_o(p_N)| G_N - d\xi_o(p_N) \right).$$

Whether ρ_N can be approached by learning algorithms comes down to sensitivity of transition probabilities to deviations in prices from p_N, in conjuction with market properties. Crucially, note that derivatives of $g(a;p)$ that are small in magnitude indicate that deviations in price would be harder to detect, i.e. result in an ineffective monitoring technology.

This finding has immediate policy relevance: if regulators can limit the sensitivity of the data algorithms use, they might be able to promote more competitive outcomes. I explore this possibility in detail in Section III.C.

Point (3) indicates how such market properties may lead to an increase in the set of transition probabilities that may allow for the learning of ρ_N. Overall, this point indicates an interesting dichotomy: on the one hand, weaker competition (low ξ_c, balanced market) facilitates learning, while on the other hand, higher own-price elasticity $|\xi_o|$ and more sensitive Lerner indices (G_N large) also facilitate learning. Note that this results holds for *any* PR-state variable.[12]

The conclusion is concerning, as achieving favorable market conditions for the learning of ρ_N may be difficult, weak competition being commonly associated also with low own-price

[11] Examples 1, 2 satisfy this when increases in own price have large enough effects on own demand relative to increases in opponent price.

[12] The bound readily generalizes to any state variable with differentiable transition function by replacing the bound in (1), (2) by one that depends on transition probabilities directly.

elasticity. An example of markets favorable to the learning of ρ_N according to point (3) would be luxury goods markets with high brand recognition. Strong branding leads to weak competition, while luxury goods may be avoided by some consumers if prices are too high, which constitutes high ξ_o.

In terms of intuitions, recall that attractiveness of ρ_N with respect to F_S determines whether this equilibrium may be learned by algorithms. Attractiveness is a property of robustness to perturbations in the policy space; being able to detect likely deviations (i.e. having an effective monitoring technology) is necessary for perturbations to have any bite at all. Furthermore, the properties in (3) all come down to robustness should deviations be detected, by firstly ensuring that opponent's deviations don't have too large of an effect (see 'small enough' ξ_c, 'large enough' $d\xi_o$), while secondly own deviations must readily correct for perturbations, requiring a 'large enough' $|\xi_o|$.

It is clear that state variables of arbitrary $K > 1$ have the potential to support complex collusive schemes featuring a variety of different prices over time. The bounds to producer and consumer surplus due to such equilibria are characterised by the profit-maximizing collusive scheme. This, in turn, can be pinned down by simple binary strategies, as is known from Abreu, Pearce, and Stacchetti 1990, henceforth APS. Under additional conditions, binary schemes will be the only way to support optimal collusive schemes. The stability of such optimal collusive schemes then serves as a bound to the extend of possible collusion among algorithms.

III.B. **Relationship to the best Equilibrium**

APS provide a result stating that the best strongly symmetric sequential equilibrium (SSE) of the repeated game can be supported by a bang-bang solution, under their setting.

Such a bang-bang strategy, by definition, is constructed using subsets of Ω, intended as punishment and reward regions. Translated into this paper's setup, there exists a state variable S^* with $O^*_S = \{A, B\}$ and sets Ω_A, Ω_B such that $s = A$ and $\tilde{A} \in \Omega_A$ implies the next period's state is A, and $s = B$ and $\tilde{A} \in \Omega_B$ implies next period's state is B. The reverse holds for $\Omega \setminus \Omega_s$.

Notice that any binary partition of Ω affects payoffs of players only by pinning down their transition probability functions $T_{ss'}(P)$. As $g(a; p)$ is twice differentiable in p by Assumption 3, we can restrict attention to twice continuously differenetiable $T_{ss'}$. Call the space of such transition probability functions $\mathcal{T}$. For any $T_{ss'} \in \mathcal{T}$, let $E^*(T_{ss'})$ be the set of symmetric Nash equilibria given expected discounted payoffs $W(\sigma)$ when $T_{ss'}$ governs state transitions. This set is nonempty for all $T_{ss'} \in \mathcal{T}$ due to the repetition of the static Nash equilibrium p_N.

As APS' result was shown only for finite strategy sets, I introduce an approximation result to the continuous action case. Let Y_L be a discretization of cardinality $0 < L < \infty$ of price-set Y, where for any discretization I impose that $p_N \in Y_L$.

Define $W(\boldsymbol{\rho}, T)$ for symmetric profiles $\boldsymbol{\rho} \in \overline{Y}$, transition probabilities $T \in \mathcal{T}$ as long run expected payoffs. Define $E_L(T)$ as the set of symmetric equilibria given discretization Y_L. Here, APS's bang-bang result directly applies. By the observation above, we can alternatively characterise the maximal SSE as

$$V_L = \sup_{\substack{\boldsymbol{\rho} \in E_L(T) \\ T \in \mathcal{T}}} W(\boldsymbol{\rho}, T).$$

Analogously, define

$$V = \sup_{\substack{\boldsymbol{\rho} \in E^*(T) \\ T \in \mathcal{T}}} W(\boldsymbol{\rho}, T). \tag{4}$$

Define V^* to be the best SSE payoff among all SSE of Γ^∞.

Proposition 1. *Given additional regularity conditions on W, $\mathcal{T}$*[13]*, there exists an SSE $\boldsymbol{\rho}$ of Γ^∞ supported by a binary-state policy, under some $T^* \in \mathcal{T}$ such that $V = W(\boldsymbol{\sigma}, T^*)$. It holds that*

(1) $V \leq V^*$.
(2) For any ε there exists $\bar{L}$ such that for all $L \geq \bar{L}$, $|V - V_L| < \varepsilon$.

Proposition 1 tells us that there exist binary state variables such that if used by algorithms, they may learn to play strategies that achieve the best SSE payoff for any discretization of their game. Whether such strategies may be learned comes down to sensitivity of transition

[13]These stronger conditions ensure continuity of the equilibrium correspondence, see Assumption 6

probabilities in a similar fashion as in Theorem 3. To state this formally, a few more variables require defining.

Let $T^*_{ss'}$ be the transition probability function supporting an SSE that achieves V. Call the payoff-maximizing equilibrium policy based on this state variable $\boldsymbol{\rho}^*$, then $\boldsymbol{\rho}^*(A) = p_A \geq p_N \geq p_B = \boldsymbol{\rho}^*(B)$. Call a state variable *generic* if its associated transition matrix $T_{ss'}(\rho^*(s))$ is generic in the space of matrices.

Given PR-state variable S with $K > 1$ states, say policy $\bar{\boldsymbol{\rho}} : O_S \to Y$ supports $\boldsymbol{\rho}^*$ if all the prices played under $\bar{\boldsymbol{\rho}}$ equal either p_A or p_B. To save notation, we also say $\bar{\boldsymbol{\rho}}$ is generic if there is a generic state variable generating $\bar{\boldsymbol{\rho}}$ supporting $\boldsymbol{\rho}^*$.

Theorem 4. *There exist $0 < C_{\pi,1}, C_{\pi,2} < \infty$, such that for any generic $\bar{\boldsymbol{\rho}}$ supporting $\boldsymbol{\rho}^*$,*

(1) $\bar{\boldsymbol{\rho}}$ will be learned with positive probability if

$$\max\left\{\max_{s\in\{A,B\}}\sup_{a\in\Omega}\left|\frac{\partial}{\partial P}g(a;P)\big|_{P=P_s}\right|, \max_{s\in\{A,B\}}\sup_{a\in\Omega}\left|\frac{\partial^2}{(\partial P)^2}g(a;P)\big|_{P=P_s}\right|\right\} < C_{\pi,1}.$$

(2) $\bar{\boldsymbol{\rho}}$ will not be learned if

$$\min_{s\in\{A,B\}}\inf_{a\in\Omega}\left|\frac{\partial}{\partial P}g(a;P)\big|_{P=P_s}\right| > C_{\pi,2}.$$

Hence, a similar insight to Theorem 3 applies here as well: If state transitions (i.e. derivatives of g in the case of PR-states) are insensitive enough to deviations in prices, then collusive equilibria such as $\boldsymbol{\rho}^*$ may be learned by algorithms. Note however that less sensitive transitions also affect the equilibrium set E_S, associated equilibrium payoffs, and consumer welfare down the line. This is discussed in the next section.

III.C. **Analysis of Policy Options**

Covariate Restriction. It follows from Theorem 3 that if one were able to affect the distribution over observed states, one may be able to ensure ρ_N would be learned. This channel represents a feasible, and realistic policy instrument. Indeed, restrictions to the inputs (i.e. state variables) of algorithms have been implemented in the United States after succesful lawsuits (see e.g. the Supreme Court decision in *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*).

It is important to note, however, that according to Theorem 4, de-sensitizing the density g of aggregate shock $\tilde{A}$ may also introduce the possibility of collusive outcomes being learned. Nevertheless, less sensitive g can be interpreted as less accurate monitoring technology available to players; this in turn can only lead to (weakly) less concerning collusive possibilities.

Formally, suppose initially that algorithms follow some PR-state variable given commonly observable $\tilde{A}$. I model restrictions to state variables of the algorithm as a garbling of observable $\tilde{A}$. Introduce $\beta \in (0, 1]$, and let $\tilde{U} \sim U([0, 1])$ be a uniformly distributed random variable, independent of $\tilde{A}$. Algorithms can only condition actions on $\tilde{C} \sim (1-\beta) \circ \tilde{A} + \beta \circ \tilde{U}$, i.e. $\tilde{C}$ is distributed as a convex combination of $\tilde{A}$ and $\tilde{U}$. The support of $\tilde{C}$ is still compact and positive, just as for $\tilde{A}$, so that all previous results go through for PR-state variables S given commonly observed $\tilde{C}$. For $\beta = 0$, we recover the unrestricted case, and $\beta = 1$ leads to E_S consisting of only stage game Nash equilibria as the state would carry no information about the past. With some abuse of notation, let $V(\beta)$ be the sup bang-bang payoff as defined in (4), when $\tilde{C}$ given β is used as common observable.

Proposition 2. *Take any stage game and $\tilde{A}$ satisfying Assumption 3. When* PR*-state variables given $\tilde{C}$ are used by algorithms,*

(1) there exists $\beta \in [0, 1)$ s.t. ρ_N will be learned with positive probability,

(2) $V(\beta)$ weakly decreases in β.

It is important to note that a restriction on the observability of $\tilde{A}$ also hampers the firms' ability to respond efficiently to real demand shocks. Forcing firms to set prices based on a noisier signal may introduce welfare loss to producers.

A regulator using this instrument, therefore, faces a trade-off: allow firms access to perfect information and risk facilitating tacit collusion, or force them to use noisy information and accept a degree of allocative inefficiency. Determining the optimal level of data restriction, i.e. β, would involve a complex welfare analysis that balances these two competing effects, which is beyond the scope of this paper, but interesting to consider in future work.

Upstream Competition. Here I consider an extension where the two agents may use algorithms that differ either via their stepsizes α_t^i, or their critic function F_S. This analysis can be interpreted as one of competition among algorithmic software providers selling learning algorithms to firms; in such a case it is more likely to find asymmetric learning rates, or critic functions.

Recall that I assume throughout that stepsizes of all agents lie within an order of magnitude from each other (see Assumption 5 in the appendix). It turns out that, in the case of symmetric equilibria, all insights stated for equal stepsizes carry over to the more general case. Suppose that $\frac{\alpha_t^1}{\alpha_t^2} \to \bar{\alpha}$ as $t \to \infty$. So far in this section, results have been stated under $\bar{\alpha} = 1$. Also define the system asymmetric in critic functions

$$\dot{\boldsymbol{\rho}} = F_{A,S} \equiv \begin{bmatrix} B_S^1(\boldsymbol{\rho}_2) - \boldsymbol{\rho}_1 \\ W_{1,S}^2(\boldsymbol{\rho}_2, \boldsymbol{\rho}_1) \end{bmatrix},$$

where $W_{1,S}^2$ refers to player 2's derivative vector of long run payoffs with respect to $\boldsymbol{\rho}_2$, given the same state variable S as used by player 1.

Proposition 3. *Take any state variable S, and any symmetric $\rho^* \in E_S$. Consider best response dynamics $F_{S,B}$. Under a regularity condition on $W(\boldsymbol{\rho}^*)$,*[14]

(1) $\boldsymbol{\rho}^*$ will be approached with positive probability given $\bar{\alpha} \in (0,1)$ if and only if this is true given $\bar{\alpha} = 1$.

(2) When $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$, (i) holds also when replacing $F_{B,S}$ with gradient dynamics $F_{G,S}$.

(3) When $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$, $\boldsymbol{\rho}^*$ will be approached with positive probability given $F_{A,S}$ if and only if $\boldsymbol{\rho}^*$ will be approached with positive probability given $\bar{\alpha} = 1$ under $F_{B,S}$.

This result indicates that upstream competition among software providers, selling algorithmic pricing solutions to firms, may not affect likely learning outcomes of those firms, as long as stepsizes remain within an order of magnitude of each other and firms' incentives are sufficiently symmetric. When restricting attention to the benchmark $\boldsymbol{\rho}_N$, the result extends to gradient dynamics, and also to the case of asymmetric critics $F_{A,S}$.

If it were true that, e.g., $\bar{\alpha} = 0$, the limiting behavior of algorithms could be quite different. The machinery applied in this paper extends to this case, with the important change that the interaction among algorithms can be interpreted as sequential: if 2's updates are an order

[14] All eigenvalues of the jacobian of the best response function at $\boldsymbol{\rho}^*$ are real.

of magnitude faster than 1's updates to their policy, in the limit, the behavior observed will be as if 1 commits to a (Markov-) policy, 2 observes this, and best responds (in the case of ACQ learning). Clearly, the equilibrium set of such a game would be inherently different. Whether such a setting would lead to less collusive outcomes is beyond the scope of this paper.

III.D. **Extension to Market Dynamics**

Practical applications of algorithmic pricing commonly involve the need to adjust prices to fluctuating market conditions. Examples include gasoline retail pricing, which involves changing demand over day and night, and weekdays to weekend as discussed in the introduction (Assad et al. 2024). To accomodate this setting, consider the following extension:

There is a random variable S_D representing aggregate market conditions $s_D \in O_{S_D}$, which takes finitely many values $K_D \geq 1$. S_D evolves as an irreducible Markov chain with transition matrix T_D. At every period t, before price choices are made, $s_D \in O_{S_D}$ is revealed to all firms. As before, $\tilde{A} \in \Omega$ is a random variable representing market conditions that are affected by current price choices, and is revealed at the end of each period. Assume that $\tilde{A}$ is independent of S_D.[15] Stage game payoffs are then written as conditional on S_D, in expectation over $\tilde{A}$. Define $X_i(p, s_D)$ as the expected demand given s_D and price vector p, where expectation is taken over $\tilde{A}$.

$$\pi_i(\boldsymbol{p}, s_D) = X_i(\boldsymbol{p}, s_D)(p_i - c).$$

We can extend Assumption 3 to this setting in a straightforward manner:

Assumption 4. *For all $s_D \in O_{S_D}$, all $\boldsymbol{p} \in \mathbb{R}^2_+$, $p_{\text{-}i} \geq 0$:*

(1) $\frac{\partial}{\partial p_i} X_i(\boldsymbol{p}, s_D) \leq 0$, $\frac{\partial}{\partial p_{\text{-}i}} X_i(\boldsymbol{p}, s_D) \geq 0$.

(2) $\frac{\partial}{\partial p_i} X_i(0, \boldsymbol{p}_{\text{-}i}, s_D) < 0$.

(3) $\pi_i(\boldsymbol{p}, s_D)$ *is quasiconcave in* p_i.

(4) $\left|\frac{\partial}{\partial p_i} X_i(\boldsymbol{p}, s_D)\right| \geq \left|\frac{\partial}{\partial p_{\text{-}i}} X_i(\boldsymbol{p}, s_D)\right|$

(5) $\left|\frac{\partial^2}{(\partial p_i)^2} X_i(\boldsymbol{p}, s_D)\right| \geq \left|\frac{\partial^2}{\partial p_i \partial p_{\text{-}i}} X_i(\boldsymbol{p}, s_D)\right|$.

[15] This can be weakened to $\tilde{A}$ having a distribution that depends on the realization s_D, while not affecting transitions of S_D in the future.

(6) For all $a \in \Omega$, $g(a;p)$ is twice differentiable in p.

An argument analogous to that of Lemma 1 gives that for all $s_D \in O_{S_D}$, there is a unique symmetric (static) Nash equilibrium $p_N(s_D)$. Given some K-sized PR-state variable $O_{S_{PR}}$, policies of algorithms would map as $\rho : O_{S_{PR}} \times O_{S_D} \to \mathbb{R}_+$. As before, let $\boldsymbol{\rho}_N \in \mathbb{R}_+^{K_D K}$ be the repetition of the static Nash equilibria, so that with some abuse of notation, we write $\boldsymbol{\rho}_N(s, s_D) = p_N(s_D)$ for all $s \in O_{S_{PR}}, s_D \in O_{S_D}$. Analogously to Theorem 3 we define here

$$\begin{aligned}
\gamma_1^m &= X^m \left(\xi_{o,1}^m L_N^m + \xi_o^m \frac{\partial}{\partial p_1} L_N^m \right) \\
\gamma_2^m &= X^m \xi_{o,2}^m L_N^m \\
\gamma_3^m &= X^m \xi_c^m L_N^m,
\end{aligned}$$

where superscript m denotes payoff terms given state s_D^m, all evaluated at $p_N(s_D^m)$, which is dropped for ease of notation. Then let

$$\lambda_1 = \max_m \left| \frac{\gamma_2^m}{\gamma_1^m} \right|; \qquad \lambda_2 = \max_m \left| \frac{1}{\gamma_1^m} \right|; \qquad \lambda_3 = \max_m |\gamma_3^m|,$$

where notably $-\frac{\gamma_2^m}{\gamma_1^m}$ is the slope of the static best response function under s_d^m, evaluated at $p_N(s_D^M)$. Hence, under Assumption 4 we have $\lambda_1 < 1$.

Theorem 5. *For any S_D state variable satisfying this setting, given π, there exists $0 < C < \infty$ such that for any g inducing π, $\boldsymbol{\rho}_N$ will be learned with positive probability if*

$$\sup_{a \in \Omega} \left| \frac{\partial}{\partial P} g(a; P)\big|_{P = P_N} \right| < C.$$

Furthermore, given any S_{PR} state variable,

(1) there exist λ_1 and $\lambda_2 \lambda_3$ small enough so that ρ_N will be learned with positive probability.

(2) there exist m_1, m_2, m_3 and $|\gamma_1^{m_1}|, |\gamma_2^{m_2}|, |\gamma_3^{m_3}|$ large enough so that $\boldsymbol{\rho}_N$ will not be learned.

Hence, inuitions extend from the $K_D = 1$ setting discussed in Theorem 3 to this more general setting: whether competitive outcomes can be learned depends on the senstivity of the monitoring technology and the magnitudes of elasticities, lerner indices and their growth rates. Again, low enough monitoring sensitivity ensures that static Nash can be learned with positive probability.

IV. Conclusion

This paper considers the long run behavior of a class of RL algorithms and shows how it can intepreted via the stability of repeated game equilibria according to an underlying differential equation. The application of collusion in repeated games is employed to show the usefulness of this framework. An important insight from my analysis is the dependence of the attractability of a given equilibrium of the repeated game on state variables observed by algorithms, i.e. their implied monitoring technology. This insight, as discussed, may serve as a tool to curb algorithmic collusion. Interesting future research directions include more detailed considerations of asymmetric learning settings, as touched upon in the discussion on upstream competition. Furthermore, the characterization of long run behaviors serves as a methodology that can allow for a variety of interesting economic application. The method enables researches to pick a given interaction of interest, e.g. an auction, a stock market, or multilateral platform, then pick a class of algorithms, and evaluate long run outcomes in the chosen setting.

Appendix A. The Reduced Form Algorithm

The following assumptions are sufficient for the results stated in Section II to go through, upon minor extensions to known results from stochastic approximation theory, to be found in Benaïm 1999, Borkar 2009, Michel Benaïm and Faure 2012. A thorough argument generalizing results further to the non-vanishing bias case can be found in the online appendix[16].

For notational ease, write $F(\boldsymbol{\rho}) = F_S(\boldsymbol{\rho})$, as the stacking over i of $F_S^i(\boldsymbol{\rho})$. For all results to follow, state variables will be fixed. The algorithm (3) can be written as

$$\boldsymbol{\rho}_{n+1} = \boldsymbol{\rho}_n + \alpha_n \left[F(\boldsymbol{\rho}_n) + \boldsymbol{\delta}_n + \boldsymbol{M}_{n+1} \right], \tag{5}$$

where $\boldsymbol{\delta}_n$ is a vector of $\boldsymbol{\delta}_n^i$ stacked over i. We switch to an identification of time periods by n in order to distinguish the continuous timescale t used in the associated continuous time systems.

Assumption 5. *Let $\mathcal{F}_n$ be the σ-field generated by $\{\boldsymbol{\rho}_n, \boldsymbol{\delta}_n, M_n, \boldsymbol{\rho}_{n-1}, \boldsymbol{\delta}_{n-1}, \boldsymbol{M}_{n-1} \ldots, \boldsymbol{\rho}_0, \boldsymbol{\delta}_0, \boldsymbol{M}_0\}$, i.e. all the information available to the updating rule at a given period n.*

(1) Stepsizes α_n^i satisfy, for all i, to be square-summable, but not summable.

(2) For all i, j, $\lim_{n\to\infty} \frac{\alpha_n^i}{\alpha_n^j}$ exists and lies in (c, ∞), for some $c > 0$.

(3) F is Liptschitz continuous and grows sublinearly, i.e.

$$\limsup_{\|\boldsymbol{\rho}\|\to\infty} \frac{\|F(\boldsymbol{\rho})\|}{\|\boldsymbol{\rho}\|} < \infty.$$

(4) $\boldsymbol{M}_{n+1}$ is a Martingale-difference noise. There is $0 < \bar{M} < \infty$ and $q \geq 2$ such that for all n

$$\mathbb{E}[\boldsymbol{M}_{n+1} | \mathcal{F}_n] = 0; \quad \mathbb{E}\left[\|\boldsymbol{M}_{n+1}\|^q \mid \mathcal{F}_n \right] < \bar{M} \;\; \mathcal{F}_0 - \textit{almost surely.}$$

(5) There exists a continuous function

$$\Omega : \overline{Y} \mapsto \mathcal{J}(\overline{Y}),$$

[16] https://cjmpossnig.github.io/papers/RLColl_onlineApp_0.pdf

where $\mathcal{J}(\overline{Y})$ is the space of positive definite matrices given vectors in $\overline{Y}$, such that for all n

$$\mathbb{E}[\boldsymbol{M}_{n+1}\boldsymbol{M}'_{n+1}\,|\mathcal{F}_n] = \Omega(\boldsymbol{\rho}_n),$$

whenever $\boldsymbol{\rho}_n \in \mathcal{U}$.

(6)

$$E\left[\|\boldsymbol{\delta}_n\|\right] = o(b_n),$$

where $b_n \to 0$ satisfies $\max_i \lim_{n\to\infty} \frac{\alpha_n^i}{b_n} = 0$, α_n^i being i's stepsize.

(7)

$$\sup_{n\geq 0} \mathbb{E}\left[\|\boldsymbol{\delta}_n\|^2\right] < \infty,$$

(8) For all $n' < n''$, $\boldsymbol{\delta}_{n'}, \boldsymbol{\delta}_{n''}$ are uncorrelated conditional on $\mathcal{F}_{n'}$.

(9) Iterates stay bounded almost surely:

$$\sup_n \|\boldsymbol{\rho}_n\| < \infty, \text{ a.s.}.$$

Point (1) is known as the Robbins-Monro condition (Robbins and Monro 1951) on stepsizes. It ensures that stepsizes converge slowly enough so that the whole real line can be mapped (as a continuous-time interval), while converging not too slowly in order for error terms to be averaged out. (2) ensures that all stepsizes lie within the same order of magnitude. Point (3) ensures global integrability and uniqueness of solutions to $\dot{\boldsymbol{\rho}} = F(\boldsymbol{\rho})$. In the example of ACQ, it is an assumption on payoffs W^i, and that best responses can't grow too quickly. Point (4) implies that given current information in period t, new errors due to $t+1$'s estimator of F are well-behaved. It is a common assumption in stochastic approximation theory. Point (5) ensures that some variance in error terms remains for all n; this is satisfied e.g. if the estimation of F involves exploratory noise, or stochasticity during the estimation as is true under randomized Bellman-iteration schemes. This assumption will be the main driver that pushes iterations away from unstable equilibria. Point (6) ensures that the bias term vanishes faster than stepsizes. Points (7), (8) are further regularity conditions on the bias term. Even though commonly made, point (9) is often difficult to verify. It is common for results to be stated conditioning on the event that (9) holds, see for example Michel Benaïm and Faure 2012. For a more general discussion of sufficient conditions for bounded iterates, see Borkar 2009, Chapter 2.

Appendix B. Proofs

B.A. Proof of Theorems 1, 2

These results are straightforward applications of known results in stochastic approximation theory, for Theorem 1 see e.g. Borkar 2009, Theorem 2, which pins down the connection between limit set L_S and the continuous time system

$$\dot{\boldsymbol{\rho}} = \boldsymbol{A} F_S(\boldsymbol{\rho}(t)).$$

Then, to conclude convergence with positive probability to an attractor, one can apply Faure and Roth 2010, Theorem 2.15. For Theorem 2, see Michel Benaïm and Faure 2012, Theorem 3.12 to conclude the result about zero-probability convergence to a linearly unstable equilibrium. Detailed arguments together with an extension to non-vanishing bias are relegated to the online appendix.

For the following proofs, it will be useful to recall a fact about block symmetric matrices:

Remark 1. *Suppose $\boldsymbol{A}, \boldsymbol{B}$ are square matrices of the same dimension. Let*

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B} & \boldsymbol{A} \end{bmatrix}.$$

Then one can show

$$det(\boldsymbol{T}) = det(\boldsymbol{A} - \boldsymbol{B}) det(\boldsymbol{A} + \boldsymbol{B}).$$

Given a square matrix $\boldsymbol{A}$, define Λ as the set of eigenvalues of the $\boldsymbol{A}$. Then define

$$\kappa(\boldsymbol{A}) = \max\{|\lambda| : \lambda \in \Lambda\},$$

as the spectral radius of $\boldsymbol{A}$. For some policy profile $\boldsymbol{\sigma}^* \in \overline{Y}$, Define $J^i(\sigma^*)$ as the Jacobian of $B^i_S(\boldsymbol{\sigma}^*)$, which is the matrix of best response derivatives of a given player. For symmetric $\boldsymbol{\sigma}^*$, we drop the i superscript to save notation.

Lemma 2. *Suppose $\boldsymbol{\sigma}^* \in E_S$ is a differential, symmetric Nash equilibrium. Let $\bar{\kappa}$ be the real part of the spectral radius of $J(\boldsymbol{\sigma}^*)$. Then $\boldsymbol{\sigma}^*$ is asymptotically stable under $F_{S,B}$ if $\bar{\kappa} < 1$, and unstable if $\bar{\kappa} > 1$.*

Proof. Using Remark 1, we get that

$$ch(\lambda) = det(J(\boldsymbol{\sigma}^*) - (1+\lambda)\boldsymbol{I}_2)det(J(\boldsymbol{\sigma}^*) + (1+\lambda)\boldsymbol{I}_2),$$

where $\boldsymbol{I}_k$ is the $k-$dimensional unit matrix. Thus, if μ is an eigenvalue of $J(\boldsymbol{\sigma}^*)$, then $\pm|\mu - 1|$ is an eigenvalue of $X(\boldsymbol{\sigma}^*)$, the Jacobian of $F_S(\sigma^*)$. The conclusion follows, since asymptotic stability requires that all eigenvalues of $X(\boldsymbol{\sigma}^*)$ have negative real parts. $\square$

Hence, it is enough to be concerned with the eigenvalues of individual jacobians of F_S^i for player i, when considering symmetric Nash equilibria.

The following Lemma will help with bounding eigenvalues in later proofs:

Lemma 3. *Consider a block-matrix $\boldsymbol{M}$ with*

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

For m, k s.t. $m + k = n$, $\boldsymbol{A}$ is an invertible $m \times m$ matrix, $\boldsymbol{D}$ is an invertible $k \times k$ matrix, and $\boldsymbol{C}$, $\boldsymbol{B}$ are $m \times k, k \times m$ matrices respectively. All blocks are assumed generic. Let $\eta_1 = \max(\|\boldsymbol{A}\|_\infty, \|\boldsymbol{D}\|_\infty)$, $\eta_2 = \|\boldsymbol{B}\|_\infty \|\boldsymbol{C}\|_\infty$. Then, for all $\lambda \in eig(\boldsymbol{M})$:

$$|\lambda| \leq \frac{\eta_1}{2} + \sqrt{\left(\frac{\eta_1}{2}\right)^2 + \eta_2}.$$

Proof. Using the Schur complement and genericity of the block matrices, we can write the characteristic equation of $\boldsymbol{M}$ as

$$char(\lambda) = \det(\boldsymbol{D} - \lambda \boldsymbol{I}_k) \det\left(\boldsymbol{A} - \lambda \boldsymbol{I}_m - \boldsymbol{B}\left((\boldsymbol{D} - \lambda \boldsymbol{I}_k)^{-1} \boldsymbol{C}\right)\right.$$

Now, any eigenvalue $\lambda \in eig(\boldsymbol{M})$ with $\lambda \notin eig(\boldsymbol{D})$ must satisfy

$$\lambda \in eig\left(\boldsymbol{A} - \boldsymbol{B}(\boldsymbol{D} - \lambda \boldsymbol{I}_k)^{-1} \boldsymbol{C}\right) \equiv eig(\boldsymbol{M}(\lambda)).$$

Letting $\rho(\boldsymbol{A})$ be the spectral radius of some matrix $\boldsymbol{A}$, recall that for any (sub-multiplicative) matrix norm $\|\|$, we have $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$. Consider $\|\boldsymbol{A}\|_\infty = \max_i \sum_j |\boldsymbol{A}_{ij}|$. Then,

$$\begin{aligned}|\lambda| &\leq \left\|\boldsymbol{A} - \boldsymbol{B}\left(\boldsymbol{D} - \lambda \boldsymbol{I}_k\right)^{-1} \boldsymbol{C}\right\|_\infty \\ &\leq \|\boldsymbol{A}\|_\infty + \|\boldsymbol{B}\|_\infty \|\boldsymbol{C}\|_\infty \left\|\left(\boldsymbol{D} - \lambda \boldsymbol{I}_k\right)^{-1}\right\|_\infty .\end{aligned}$$

Similarly, we have, as long as $\lambda \notin eig\left(\boldsymbol{A}\right)$,

$$\begin{aligned}|\lambda| &\leq \left\|\boldsymbol{D} - \boldsymbol{B}\left(\boldsymbol{A} - \lambda \boldsymbol{I}_m\right)^{-1} \boldsymbol{C}\right\|_\infty \\ &\leq \|\boldsymbol{D}\|_\infty + \|\boldsymbol{B}\|_\infty \|\boldsymbol{C}\|_\infty \left\|\left(\boldsymbol{A} - \lambda \boldsymbol{I}_m\right)^{-1}\right\|_\infty .\end{aligned}$$

We need to accurately bound the inverse terms. As we seek an approach that relates small $|\lambda|$ with small enough $|\partial_{p_1} g(a; p)|$, write $\boldsymbol{D} = h\boldsymbol{D}$, for some $h \neq 0$. $\lambda \neq 0$ by genericity. Write $\lambda\left(\boldsymbol{I}_k - \frac{h}{\lambda}\boldsymbol{D}\right) = (-\lambda)\boldsymbol{I}_k + \boldsymbol{D}$. For $|h| < |\lambda|$ small enough, we will have $\|\frac{h}{\lambda}\boldsymbol{D}\|_\infty < 1$. Then, an application of the von Neumann series approximation gives

$$\frac{1}{|\lambda|}\left\|\left(\boldsymbol{I}_k + \frac{h}{\lambda}\boldsymbol{D}\right)^{-1}\right\|_\infty \leq \sum_{k=0}^{\infty}\left\|\frac{h}{\lambda}\boldsymbol{D}\right\|_\infty^k = \frac{1}{|\lambda|}\frac{1}{1 - \frac{h\|\boldsymbol{D}\|_\infty}{|\lambda|}} = \frac{1}{|\lambda| - h\|\boldsymbol{D}\|_\infty}.$$

An analogous bound can be constructed for $\boldsymbol{A} - (\lambda)I_m$. By genericity, $eig(\boldsymbol{A}) \cap eig(\boldsymbol{D}) = \emptyset$. Then, for all $\lambda \in eig(\boldsymbol{M})$, when $|h| > 0$ small enough,

$$|\lambda| \leq \max\left(\|\boldsymbol{A}\|_\infty, \|\boldsymbol{D}\|_\infty\right) + \|\boldsymbol{B}\|_\infty \|\boldsymbol{C}\|_\infty \frac{1}{|\lambda|}.$$

Writing $\eta_1 = \max\left(\|\boldsymbol{A}\|_\infty, \|\boldsymbol{D}\|_\infty\right)$ and $\eta_2 = \|\boldsymbol{B}\|_\infty \|\boldsymbol{C}\|_\infty$, we can solve the above as an equality. This leads to a quadratic equation in $|\lambda|$, to larger root of which equals

$$\frac{\eta_1}{2} + \sqrt{\left(\frac{\eta_1}{2}\right)^2 + \eta_2}.$$

The result follows. $\square$

Now consider the general situation in which $O_S = \{s_1, ..., s_K\}$. Fixing the profile $(\boldsymbol{\rho}, \boldsymbol{\gamma}) \in \overline{Y}$, it is useful to consider the vector formulations:

$$\tilde{\boldsymbol{W}} = [W(\boldsymbol{\rho}, \boldsymbol{\gamma}, s_1), ..., W(\boldsymbol{\rho}, \boldsymbol{\gamma}, s_K)]^\top ,$$
$$\boldsymbol{U} = [\pi(\rho(s_1), \gamma(s_1)), ..., \pi(\rho(s_K), \gamma(s_K))]^\top ,$$

to write

$$\tilde{\boldsymbol{W}} = (1-\delta)\boldsymbol{U} + \delta \boldsymbol{T}\tilde{\boldsymbol{W}} \Longleftrightarrow \tilde{\boldsymbol{W}} = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}(1-\delta)\boldsymbol{U},$$

where $\boldsymbol{T} = (T_{kk'})_{k,k' \in \{1,...,K\}}$ is the Markov transition matrix given the fixed profile $(\boldsymbol{\rho}, \boldsymbol{\gamma})$. Note that for all $\delta < 1$, $\boldsymbol{I}_K - \delta\boldsymbol{T}$ is an M-matrix. The inverse of $\boldsymbol{I}_K - \delta\boldsymbol{T}$ exists and has all elements non-negative. As a result, we then have that all rows of $[\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}$ sum to $\frac{1}{1-\delta}$. In the remainder of this section, to save notation write $\boldsymbol{\rho}_k = \rho(s_k)$, and similarly for $\boldsymbol{\gamma}$. Counting arguments of W as the first K arguments referring to own strategy $\boldsymbol{\rho}$, the next K arguments referring to $\boldsymbol{\gamma}$, indicate derivatives and cross-derivatives of W with respect to a specific argument $1 \le j \le 2K$ using a subscript j. Then we have:

Corollary 1. *The derivatives of vector $\tilde{\boldsymbol{W}}$ can be written as, for $i \le K < j$:*

$$\tilde{\boldsymbol{W}}_i = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}\tilde{\boldsymbol{W}} + [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}(1-\delta)\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\rho}_i}$$
$$\tilde{\boldsymbol{W}}_j = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\gamma}_{j-K}}\tilde{\boldsymbol{W}} + [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}(1-\delta)\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\gamma}_{j-K}}$$
$$\tilde{\boldsymbol{W}}_{ii} = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial^2 \boldsymbol{T}}{(\partial \boldsymbol{\rho}_i)^2}\tilde{\boldsymbol{W}} + [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}(1-\delta)\frac{\partial^2 \boldsymbol{U}}{(\partial \boldsymbol{\rho}_i)^2} - 2[\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}\tilde{\boldsymbol{W}}_i$$
$$\tilde{\boldsymbol{W}}_{ij} = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial^2 \boldsymbol{T}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\gamma}_{j-K}}\tilde{\boldsymbol{W}} + [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}(1-\delta)\frac{\partial^2 \boldsymbol{U}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\gamma}_{j-K}}$$
$$+ [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\gamma}_{j-K}}\tilde{\boldsymbol{W}}_i + [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}\tilde{\boldsymbol{W}}_j.$$

Proof. This follows from some matrix algebra, importantly using the following fact:
For a matrix function $\boldsymbol{X}$ of variable y, let $\partial\boldsymbol{X}$ be the partial derivative of $\boldsymbol{X}$ with respect to y. Then $\partial(\boldsymbol{X}^{-1}) = -(\boldsymbol{X}^{-1})(\partial\boldsymbol{X})(\boldsymbol{X}^{-1})$. □

If $\tilde{\boldsymbol{W}}_i = 0_K$, we can further simplify these:

$$\begin{aligned}
\tilde{\boldsymbol{W}}_j &= [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} (1-\delta) \left[\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\gamma}_{j-K}} - \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\rho}_{j-K}} \right] \\
\tilde{\boldsymbol{W}}_{ii} &= [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} \delta \frac{\partial^2 \boldsymbol{T}}{(\partial \boldsymbol{\rho}_i)^2} \tilde{\boldsymbol{W}} + [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} (1-\delta) \frac{\partial^2 \boldsymbol{U}}{(\partial \boldsymbol{\rho}_i)^2} \\
\tilde{\boldsymbol{W}}_{ij} &= [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} \delta \frac{\partial^2 \boldsymbol{T}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\gamma}_{j-K}} \tilde{\boldsymbol{W}} + [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} (1-\delta) \frac{\partial^2 \boldsymbol{U}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\gamma}_{j-K}} \\
&\quad + [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} \delta \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i} \tilde{\boldsymbol{W}}_j.
\end{aligned} \tag{6}$$

B.B. **Proof of Theorem 3**

To determine the stability of ρ_N, we need to compute the eigenvalues of the linearized best-response dynamics at ρ_N. Since by assumption, states are irreducible , we can fix an arbitrary initial state s_1 when computing first order conditions to pin down best responses. The implicit function theorem and (6) can then be used to find the gradient of the best response.

First, write $W_{ij} = \tilde{\boldsymbol{W}}_{ij}(s_1)$ for all $i \leq K$, $j \leq 2K$. Then we can write the Hessian as $\boldsymbol{H} = diag\,(W_{ii})$, the diagonal matrix having the second derivatives W_{ii} for $i = 1, ..., K$ on the diagonal. $\boldsymbol{H}$ is diagonal by the derivation of W_i: write

$$W_i = [\boldsymbol{I}_K - \delta \boldsymbol{T}]_1^{-1} \left[\delta \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i} \tilde{\boldsymbol{W}} + (1-\delta) \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\rho}_i} \right],$$

where for any matrix $\boldsymbol{A}$ we write $\boldsymbol{A}_i$ as the i'th row of $\boldsymbol{A}$. Then taking another derivative with respect to a variable $j \leq K$:

$$W_{ij} = \left(\partial\, [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1}\right) [\boldsymbol{I}_K - \delta \boldsymbol{T}]\, W_i + [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1} \left[\delta \frac{\partial^2 \boldsymbol{T}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\rho}_j} \tilde{\boldsymbol{W}} + (1-\delta) \frac{\partial^2 \boldsymbol{U}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\rho}_j} \right]. \tag{7}$$

Notice that $\frac{\partial^2 \boldsymbol{T}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\rho}_j}$, $\frac{\partial^2 \boldsymbol{U}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\rho}_j}$ are matrices of all zeros if $i \neq j$ and $i, j \leq K$. Then plugging in that $W_i = 0$, we get that indeed $W_{ij} = 0$ whenever $i \neq j$ and $i, j \leq K$. So H must be diagonal.

Now define $\boldsymbol{R} = [W_{ij}]_{i \leq K < j}$ as the matrix of cross derivatives between an agent's own strategy $\rho(s_i)$ and an opponent's strategy $\gamma(s_{j-K})$. Then we can define, using the implicit function theorem, the best response derivative matrix as

$$\boldsymbol{M} = -\boldsymbol{H}^{-1}\boldsymbol{R}.$$

Since we evaluate this at ρ_N, we can make multiple observations that will greatly simplify the structure of $\boldsymbol{M}$:

Firstly, long term payoffs $\tilde{\boldsymbol{W}} = \boldsymbol{\pi}^N$, a $K-$vector equal to π_N, the static Nash payoff in all elements, since ρ_N prescribes the same action in each state.

Secondly, by the nature of p_N, $\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\rho}_i} = 0$ for all $i \leq K$.

Now note that since by definition each row of $\boldsymbol{T}$ sums to one, and therefore each row of $\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}$ and $\frac{\partial^2 \boldsymbol{T}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\rho}_j}$ must sum to zero. Therefore, at ρ_N, we can simplify the elements of $\boldsymbol{H}, \boldsymbol{R}$ to

$$W_{ii} = \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]_1^{-1} (1-\delta) \frac{\partial^2 \boldsymbol{U}}{(\partial \boldsymbol{\rho}_i)^2},$$
$$W_{ij} = \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]_1^{-1} (1-\delta) \left[\frac{\partial^2 \boldsymbol{U}}{\partial \boldsymbol{\rho}_i \partial \boldsymbol{\gamma}_{j-K}} + \delta \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i} \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]^{-1} \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\gamma}_{j-K}}\right],$$

where for any matrix A we write A_i as the i'th row of A. Let $\boldsymbol{e}_i$ be the K-vector that is one in entry i, and zero in all others. Using the fact that ρ_N is constant for all states, we can write this down in the more simple form

$$W_{ii} = \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]_1^{-1} (1-\delta) \pi_{11}^N \boldsymbol{e}_i,$$
$$W_{ij} = \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]_1^{-1} (1-\delta) \left[\pi_{12}^N \boldsymbol{e}_i + \delta \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i} \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]^{-1} \pi_2^N \boldsymbol{e}_i\right],$$

if $i = j - K$, and

$$W_{ij} = \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]_1^{-1} (1-\delta) \left[\delta \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i} \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]^{-1} \pi_2^N e_{j-K}\right],$$

otherwise, following the notation used in section III. To save notation, write $Z = \left[\boldsymbol{I}_K - \delta \boldsymbol{T}\right]^{-1}$. Then,

$$\frac{W_{ij}}{W_{ii}} = \begin{cases} \frac{1}{\pi_{11}^N} \left[\pi_{12}^N + \pi_2^N \delta \sum_{k=1}^K T'_{s_i s_k} Z_{k,i}\right] & \text{if } i = j - K \\ \frac{1}{\pi_{11}^N} \left[\pi_2^N \delta \sum_{k=1}^K T'_{s_i s_k} Z_{k,j-K}\right] & \text{o.w.} \end{cases}$$

Here T' is the transition-derivative matrix, where each row i corresponds to the derivative of row i of T with respect to ρ_i, all evaluated at ρ_N.

For the proof of point (1), we will upper bound eigenvalues of this system. Note from the above that we can write

$$\boldsymbol{M} = -\frac{\pi_{12}^N}{\pi_{11}^N}\boldsymbol{I}_K + \frac{\pi_2^N}{\pi_{11}^N}\boldsymbol{B},$$

where terms in $\boldsymbol{B}$ depend on $\boldsymbol{T}, \boldsymbol{Z}$. If $\pi_2^N = 0$, the bound is trivial. Assume $\pi_2^N > 0$ (positivity follows from Assumption 3). $\boldsymbol{M}$'s simple form implies that $\boldsymbol{v} \in \mathbb{R}^K$ is an eigenvector of $\boldsymbol{M}$ if and only if it is an eigenvector of $\boldsymbol{B}$. It follows that eigenvalues $\lambda \in eig(\boldsymbol{M})$ and $\mu \in eig(\boldsymbol{B})$ are related through the equation

$$\lambda = -\frac{\pi_{12}^N}{\pi_{11}^N} + \frac{\pi_2^N}{\pi_{11}^N}\mu.$$

As $-\frac{\pi_{12}^N}{\pi_{11}^N} \in (0, 1)$ by Assumption 3, it is sufficient to bound $\frac{\pi_2^N}{\pi_{11}^N}\mu$.

From this, we can derive that $|\lambda| < 1$ is equivalent to

$$\mu \in \left(\frac{\pi_{12}^N}{\pi_2^N} + \frac{\pi_{11}^N}{\pi_2^N}, \frac{\pi_{12}^N}{\pi_2^N} - \frac{\pi_{11}^N}{\pi_2^N}\right), \tag{8}$$

an interval that contains 0. Note that $\boldsymbol{B} = \delta\boldsymbol{T}'\boldsymbol{Z}$. Increasing $|\boldsymbol{T}'|$ in absolute value by a scalar c will increase $|\mu|$ by that scalar c. Generically in the space of transition probability matrices, there exists at least one eigenvalue $\mu \in eig(B)$ s.t. $\mu \neq 0$. Without loss, let $\mu > 0$. Then for any $c > 0$, $c\mu \in eig(c\boldsymbol{B})$. For any $\boldsymbol{T}'$ finite, $c\boldsymbol{T}'$ corresponds to another transition probability matrix, as summing to zero over rows is still satisfied, and the perturbation only needs to be carried out at the point $\boldsymbol{\rho}_N$. One can directly construct a transition probability matrix $\boldsymbol{D}$ such that $\boldsymbol{D} = \boldsymbol{T}$ at $\boldsymbol{\rho}_N$, and $D' = cT'$, by writing for $\boldsymbol{\rho}$ close enough to $\boldsymbol{\rho}_N$: $\boldsymbol{D}(\boldsymbol{\rho}) = \boldsymbol{T} + (\boldsymbol{\rho} - \boldsymbol{\rho}_N)c\boldsymbol{T}'$. Hence, one can find a transition probability matrix with $|\boldsymbol{T}'|$ small enough such that all eigenvalues μ will ensure $|\lambda| < 1$. Converseley, taking c large enough will lead to some eigenvalue of $\boldsymbol{B}$ large enough s.t. $|\lambda| > 1$ for some associated eigenvalue λ.

For point (1), interpret c as scaling down $\sup_{a\in\Omega}\left|\frac{\partial}{\partial P}g(a;P)\big|_{P=P_N}\right|$, which in turn scales down $|\boldsymbol{T}'|$. For point (2), let c scale up $\inf_{a\in\Omega}\left|\frac{\partial}{\partial P}g(a;P)\big|_{P=P_N}\right|$. Finally, recall that we have

$\pi_{12} \geq 0 > \pi_{11}$ under Assumption 3. Thus, $|\pi_{12}^N + \pi_{11}^N| < \pi_{12}^N - \pi_{11}^N$, so that the negative lower bound of the interval in (8) is closer to zero than the positive upper bound. This allows to conclude that for points (1) and (2), the cutoff C stated in the Theorem will be the same: for (1), all eigenvalues must satisfy $|\lambda| < 1$ - having all $|\mu|$ small enough to lie above the absolute value of the lower bound is sufficient. On the other hand, for instability, at least one λ must satisfy $|\lambda| > 1$. Finding one negative μ small enough is then enough, which allows to check the same lower bound. Note that the sign of μ can always be imposed by a perturbation using scalar c as indicated above, since $\boldsymbol{\rho}_N$ remains an equilibrium no matter the sign of transition probabilities.

For point (3), note first that we can write $\pi_1 = X(\boldsymbol{p})\left(1 + \xi_o(\boldsymbol{p})\frac{p_1 - c}{p_1}\right)$, so that at p_N, we have $-\xi_o(p_N) = \frac{1}{L_N}$, where $L_N = \frac{p_N - c}{p_N}$ is the Lerner index at p_N. Define

$$\xi_{o,1}(\boldsymbol{p}) = \frac{\partial}{\partial p_1}\xi_o(\boldsymbol{p}), \qquad \xi_{o,2}(\boldsymbol{p}) = \frac{\partial}{\partial p_2}\xi_o(\boldsymbol{p}).$$

Then,

$$\begin{aligned}
\pi_2 &= \xi_c(\boldsymbol{p})X(\boldsymbol{p})\frac{p_1 - c}{p_2} \\
\pi_{11} &= \xi_o(\boldsymbol{p})\frac{X(\boldsymbol{p})}{p_1}\pi_1 + \xi_{o,1}\frac{p_1 - c}{p_1}X(\boldsymbol{p}) + \xi_o(\boldsymbol{p})\frac{c}{p_1^2}X(\boldsymbol{p}) \\
\pi_{12} &= \xi_c(p)\frac{X(\boldsymbol{p})}{p_2}\pi_1 + \xi_{o,2}(p)\frac{p_1 - c}{p_1}X(\boldsymbol{p}).
\end{aligned}$$

At p_N, these simplify to

$$\begin{aligned}
\pi_2 &= \xi_c(p_N)X(p_N)L_N \\
\pi_{11} &= X(p_N)\left(\xi_{o,1}(p_N)L_N + \xi_o(p_N)\frac{\partial}{\partial p_1}L_N\right) \\
\pi_{12} &= X(p_N)\xi_{o,2}(p_N)L_N.
\end{aligned}$$

Now consider the lower bound for (8) above. It is clear that the constant C identified for points (1),(2) is proportional to this lower bound. Using the above, and as $\pi_{12} + \pi_{11} \leq 0$ by Assumption 3, we can write the magnitude of the lower bound. Define $G_N = \frac{\partial}{\partial p_1}L_N/L_N$ as growth rate of the Lerner index at p_N. Then,

$$\begin{aligned}\frac{|\pi_{11}^N + \pi_{12}^N|}{\pi_2^N} &= \frac{-1}{\xi_c(p_N)}\left(\xi_o(p_N)G_N + \xi_{o,1}(p_N) + \xi_{o,2}(p_N)\right) \\ &= \frac{1}{\xi_c(p_N)}\left(|\xi_o(p_N)|G_N - d\xi_o(p_N)\right).\end{aligned}$$

The conclusion follows: the bound grows as ξ_c falls. G_N grows as c grows, $p_N - c$ falls, p_N falls. The bound grows as $|\xi_o(p_N)|$ grows. In a balanced market, $d\xi_o \leq 0$. Here the bound grows as $|d\xi_o|$ grows. In unbalanced markets, the bound grows as $|d\xi_o|$ falls. In general, as the market becomes 'more balanced', the bound grows. ■

B.C. **Proof of Theorem 4**

For point (1), recall that $\boldsymbol{J}$ refers to the $K \times K$ linearization of $F_S(\rho)$ at $\bar{\boldsymbol{\rho}}$. Let $O_{S_A} \subset O_S$ refer to those states s for which $\bar{\boldsymbol{\rho}}(s) = p_A$, and let $O_{S_B} = O_S \setminus O_{S_A}$. Note that as $\bar{\boldsymbol{\rho}}$ supports $\boldsymbol{\rho}^*$, we may write $\boldsymbol{J}$ just as matrix $\boldsymbol{M}$ in the statement of Lemma 3, where k corresponds to $k = |O_{S_A}|$, and $m = |O_{S_B}|$. The eigenvalue bound of Lemma 3 then applies to $\boldsymbol{J}$.

We can then follow a similar argument as in the proof of Theorem 3. As derived in (6), entries of $\boldsymbol{J}$ grow in magnitude firstly with transition probability terms, which are pinned down by the density derivatives $\left|\frac{\partial}{\partial P} g(a;P)\big|_{P=P_s}\right|$, $\left|\frac{\partial^2}{(\partial P)^2} g(a;P)\big|_{P=P_s}\right|$, $s \in \{A, B\}$. Secondly, terms of $\boldsymbol{J}$ grow with payoff terms $|\pi_i^s|, |\pi_{ij}^s|$ for $1 \leq i,j \leq 2$, $s \in \{A, B\}$. For any such payoff terms, we can push down density derivatives to a small enough value so that stability is accomplished. This follows, as in the extreme case where $g(a;p)$ does not vary with p, the best SSE satisfies $\boldsymbol{\rho}^*(s) = p_N$ for all s trivially. Then, the bound collapses to the bound of Theorem 3.

As for point (2), recall that for any matrix, the sum of eigenvalues equals the trace of the matrix. In this case, using (6) we can write

$$tr(\boldsymbol{J}) = -K + \sum_{k=1}^{K} \frac{\pi_{11}^k - \pi_{12}^k + [\boldsymbol{I}_K - \delta \boldsymbol{T}]_1^{-1} \delta \frac{\partial \boldsymbol{T}}{\partial \rho_k} e_k \left(\pi_1^k - \pi_2^k\right)}{\delta \frac{\partial^2 \boldsymbol{T}}{(\partial \rho_k)^2} e_k W + \pi_{11}^k},$$

where π_i^k, π_{ij}^k refer to the derivatives of π evaluated at $\bar{\boldsymbol{\rho}}_k$, and e_k is the K-dimensional column vector equal to 1 at at element k, and zeros everywhere else. From this it is clear that $|tr(\boldsymbol{J})|$ can be made to grow by growing $|\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_k}|$. The result follows. ■

B.D. **Proof of Lemma 1**

Given Assumption 3, the Nash equilibrium must be interior. (3) ensures that best responses are unique, wile (4) and (5) imply that the slope of the best response is less than one for all $p_{\text{-}i} \geq 0$. This implies that the Nash equilibrium p_N is unique, and symmetry of the payoffs implies that this equilibrium is symmetric. For static best response dynamics, it is well known that these conditions imply global attraction to the Nash equilibrium (Milgrom and Roberts 1990). As for gradient dynamics, it is straightforward to show that p_N must be asymptotically stable: let $\boldsymbol{M}_G, \boldsymbol{M}_B$ be the linearization of gradient dynamics and best response dynamics, respectively at p_N. Then by symmetry of payoffs, we can write $\boldsymbol{M}_G = -\pi_{11}^N \boldsymbol{M}_B$. Any eigenvalue $\lambda \in eig(\boldsymbol{M}_G)$ is such that $-(\lambda/\pi_{11}^N) \in eig(\boldsymbol{M}_B)$. As $-\pi_{11}^N > 0$, all eigenvalues of $\boldsymbol{M}_G$ are negative if and only if all of $\boldsymbol{M}_B$ are, so that stability carries over. ■

Proof of Proposition 1

For any interior $\boldsymbol{\sigma} \in E^*(T_{ss'})$, let $\boldsymbol{J}(\boldsymbol{\sigma})$ be the 2×2 matrix of best response derivatives, i.e. the Jacobian of the best response function at the equilibrium $\boldsymbol{\sigma}$, $B_S^1(\boldsymbol{\sigma}_2)$. We require the following additional assumption:

Assumption 6. *For all $T_{ss'} \in \mathcal{T}$, all $\boldsymbol{\sigma} \in E^*(T_{ss'})$ are interior, with negative definite Hessian, and all eigenvalues of $\boldsymbol{J}(\boldsymbol{\sigma})$ are different from* 1.

For any fixed $T_{ss'}$, the above assumption is a generic property over the space of regular payoff functions. The assumption has additional strength as it imposes that given the regular stage game payoff function π, there exists no $T_{ss'} \in \mathcal{T}$ that could lead to a singular Hessian at some equilibrium, or a $\boldsymbol{J}(\boldsymbol{\sigma})$ with eigenvalue equal to 1. For any discretization Y_L, define $W^L(\boldsymbol{\rho}, T) : Y^2 \times \mathcal{T} \to \mathbb{R}$ as restriction of the payoff function to Y_L, given some T.

$$W^L(\boldsymbol{\rho}, T) = W(f^L(\boldsymbol{\rho}), T),$$

where

$$f^L(\boldsymbol{\rho}) = \arg \min_{\boldsymbol{\rho}' \in Y_L^2} \left\| \boldsymbol{\rho} - \boldsymbol{\rho}' \right\|,$$

for any norm on Y^2, the projection of ρ onto discrete space Y_L. For every sequence Y_L there is an associated sequence $\alpha_L(T)$ with

$$\alpha_L(T) = \max_{\boldsymbol{\rho} \in Y^2} \left\| W^L(\boldsymbol{\rho}, T) - W(\boldsymbol{\rho}, T) \right\|.$$

Continuity of W and the Lipschitz property of density $g(a; p)$ implies that $\alpha_L(T) \to 0$ for all $T \in \mathcal{T}$. Write $\alpha_L(Y_L, T)$ for a sequence of α_L given a fixed sequence of discretizations and transition function T. Say that a discretization sequence Y_L is *covering* if $\alpha_L(Y_L, T) \to 0$ (and $p_N \in Y_L$). From now on fix a covering sequence of discretizations Y_L and transition probability T.

Notice that $E_L(T)$ is closed-valued, trivially by finiteness of Y_L. Furthermore, $E^*(T)$ is closed-valued: W is continuous, Y compact, and thus Berge gives us that the best-response correspondence is closed and compact-valued. Then, applying the closed-graph theorem gives us that the equilibrium set $E^*(Y)$, as a set of fixed points of a closed and compact correspondence, must be closed. To get to claim (1), I will show that any converging sequence $\boldsymbol{\rho}_L \in E_L(T)$ has its limit in $E^*(T)$, and any $\boldsymbol{\rho} \in E^*(T)$ has a converging sequence in $E_L(T)$ approaching it. In other words, upper and lower hemicontinuity properties hold for the equilibrium correspondence along sequences of covering discretizations.

Lemma 4. *For all covering sequences Y_L,*

$$\lim_{K \to \infty} H\left(E_L(T), E^*(T)\right) = 0,$$

where $H(\cdot, \cdot)$ is the Hausdorff-distance.

Proof. We first show upper hemicontinuity in K. Suppose u.h.c. is not satisfied. Then there exists a subsequence $\boldsymbol{\rho}_{L_t} \in E_{L_t}(T)$ with $\boldsymbol{\rho}_{L_t} \to_t \bar{\boldsymbol{\rho}} \notin E^*(T)$. The converging subsequence exists since Y^2 is compact. To ease notation, re-define $L = L_t$ for the rest of the proof. Not being an equilibrium, we have that there exists $\boldsymbol{\rho}_T \neq \bar{\boldsymbol{\rho}}$ that maximizes the deviation payoff

$$\Delta_T = W(\boldsymbol{\rho}_{T,i}, \bar{\boldsymbol{\rho}}_{\text{-}i}, T) - W(\bar{\boldsymbol{\rho}}, T) > 0.$$

Pick $\varepsilon \in (0, \Delta_T)$. By convergence of $\boldsymbol{\rho}_L$, and by continuity of W, we have that there exists $L_{1,T}$ such that for all $L \geq L_{1,T}$,

$$\left| W(\boldsymbol{\rho}_{T,i}, \boldsymbol{\rho}_{L,\text{-}i}, T) - W(\boldsymbol{\rho}_{T,i}, \bar{\boldsymbol{\rho}}_{\text{-}i}, T) \right| \leq \frac{\varepsilon}{3}. \tag{9}$$

By the same argument, there is a $L_{2,T}$ s.t. for all $L \geq L_{2,T}$,

$$\left| W(\boldsymbol{\rho}_L, T) - W(\bar{\boldsymbol{\rho}}, T) \right| \leq \frac{\varepsilon}{3}. \tag{10}$$

Furthermore, we can always choose $\bar{L}_T \geq \max\{L_{1,T}, L_{2,T}\}$ large enough so that $\alpha_L \leq \frac{\varepsilon}{3}$, implying

$$\left| W(f^L(\boldsymbol{\rho}_{T,i}), \boldsymbol{\rho}_{L,\text{-}i}, T) - W(\boldsymbol{\rho}_{T,i}, \boldsymbol{\rho}_{L,\text{-}i}, T) \right| \leq \frac{\varepsilon}{3}. \tag{11}$$

Take $L \geq \bar{L}_T$. Define the best deviation under the discrete game as

$$\hat{\boldsymbol{\rho}}_L = \arg \max_{\boldsymbol{\rho} \in Y_L^2 \setminus \boldsymbol{\rho}_L} W(\boldsymbol{\rho}, \boldsymbol{\rho}_{L,\text{-}i}, T).$$

Now we have

$$\begin{aligned} W(\hat{\boldsymbol{\rho}}_L, \boldsymbol{\rho}_{L,\text{-}i}, T) - W(\boldsymbol{\rho}_L, T) &\geq W(f^L(\boldsymbol{\rho}_{T,i}), \boldsymbol{\rho}_{L,\text{-}i}, T) - W(\boldsymbol{\rho}_L, T) \\ &= W(\boldsymbol{\rho}_{T,i}, \boldsymbol{\rho}_{L,\text{-}i}, T) - W(\boldsymbol{\rho}_L, T) + \beta_{1,L} \\ &= W(\boldsymbol{\rho}_{T,i}, \bar{\boldsymbol{\rho}_{\text{-}i}}, T) - W(\bar{\boldsymbol{\rho}}, T) + \beta_{1,L} + \beta_{2,L} + \beta_{3,L} \\ &\geq \Delta_T + \beta_{1,L} + \beta_{2,L} + \beta_{3,L}, \end{aligned}$$

where $\beta_{1,L}$ corresponds to the projection error (11), and $\beta_{2,L}, \beta_{3,L}$ correspond to (9),(10) respectively. We have that $|\beta_{i,L}| \leq \frac{\varepsilon}{3}$, and thus

$$W(\hat{\boldsymbol{\rho}}_{L,i}, \boldsymbol{\rho}_{L,\text{-}i}, T) - W(\boldsymbol{\rho}_L, T) \geq \Delta_T - \varepsilon > 0,$$

implying that $\boldsymbol{\rho}_L \notin E_L(T)$, a contradiction.

For lower hemicontinuity, Assumption 6 imposes that for all equilibria in $E^*(T)$ for all players, Hessians at the equilibrium are negative definite. Thus, small deviations must lead to strict payoff loss. Fix T, then the proof is via contradiction: there exists some equilibrium $\boldsymbol{\rho}^* \in E^*(T)$ that is not approximated by any sequence of equilibria in $E_L(T)$. The proof can be done analogously to the one above; defining $\Delta_T > 0$ as the best deviation payoff:

$$\Delta_T = \sup_{\boldsymbol{\rho} \in Y^2 \setminus \boldsymbol{\rho}} W(\boldsymbol{\rho}, \boldsymbol{\rho}^*_{\text{-}i}, T) - W(\boldsymbol{\rho}^*, T) > 0.$$

Since $\Delta_T > 0$, we can find a fine enough discretization s.t. $\boldsymbol{\rho}^*$ can be approximated arbitrarily closely, in which case incentives must also align, by continuity of W. The contradiction follows as before. $\square$

Continuity of the equilibrium correspondence gives us that for all $\varepsilon > 0$ there is $K > 0$ large enough so that

$$\|V_L(T) - V^*(T)\| < \varepsilon,$$

with $V_L(T), V^*(T)$ being the maximal payoff over the equilibrium sets $E_L(T), E^*(T)$.

To make judgements about $\sup_{T \in \mathcal{T}} V_L(T)$, a uniform continuity property of $V_L(T)$ will be useful. By Assumption 6, all equilibria in E_L and E^* are hyperbolic, for K large enough. Hyperbolicity implies that an implicit function theorem holds: For any $\boldsymbol{\rho} \in E_L$, there exists neighborhoods $\mathcal{N}_{\boldsymbol{\rho}}, \mathcal{N}_T$ of ρ, T and a continuous map $h : \mathcal{N}_T \to \mathcal{N}_{\boldsymbol{\rho}}$ such that $h(T) \in E_L(T)$ for all $T \in \mathcal{N}_T$. Thus, the equilibrium correspondences $E_L(T), E^*(T)$ are continuous in T for all K large enough.

As $W(\boldsymbol{\rho}, T)$ is continuous both in $\boldsymbol{\rho}$ and T[17], and equilibrium correspondences are continuous in T, Berge's Maximum Theorem applies. We have that $V_L(T)$ is continuous in T. Moreover, as the payoff functions $W(\boldsymbol{\rho}, T)$ are bounded, twice differentiable in $\boldsymbol{\rho}$ and transition probabilities T (evaluated at $\boldsymbol{\rho}$), these payoff functions are Lipschitz both in $\boldsymbol{\rho}$ and T. Then, $V_L(T)$ are bounded, Lipschitz as well for K large enough. This follows first from continuity of $E^*(T)$, and then by the Lipschitz property of $W(\boldsymbol{\rho}, T)$. Finally, boundedness and the Lipschitz property imply that $V_L(T)$ are equicontinuous in T.

By the Arzelá-Ascoli Theorem, boundedness and equicontinuity of $V_L(T)$ implies the existence of a subsequence K_m so that V_{K_m} converges uniformly to some V. Pointwise convergence of $V_L(T)$ implies that this limit satsifies $V = V^*$. A simple contradiction argument with another application of Arzelá-Ascoli shows that indeed $V_L(T)$ converges uniformly to V^*.

[17] Regarding functions T, consider the sup-norm as metric on $\mathcal{T}$.

Twice differentiability of T and boundedness implies, by the Arzelá-Ascoli Theorem, the relative compactness of $\mathcal{T}$. Hence, we have that $V_L(T)$ converges uniformly to V^* over a relatively compact set $\mathcal{T}$. It follows that

$$\lim_{K\to\infty}\sup_{T\in\mathcal{T}} V_L(T) = \sup_{T\in\mathcal{T}} V^*(T).$$

■

B.E. **Proof of Proposition 2**

Point (i) follows directly from Theorem 3. Point (ii) is an application of the analysis in Kandori 1992.

B.F. **Proof of Proposition 3**

We are proving the following statement, restated in more technical terms:

Proposition 4. *Take any state variable S, and any symmetric $\boldsymbol{\rho}^* \in E_S$. Consider best response dynamics $F_{B,S}$. When all eigenvalues of $J(\boldsymbol{\rho}^*)$ are real,*

(1) $\boldsymbol{\rho}^*$ is asymptotially stable given $\bar{\alpha} \in (0,1)$ if and only if it is asymptotically stable given $\bar{\alpha} = 1$.

(2) When $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$, $\boldsymbol{\rho}^*$ is asymptotially stable given $\bar{\alpha} \in (0,1)$ if and only if it is asymptotically stable given $\bar{\alpha} = 1$ also under gradient dynamics $F_{G,S}$.

(3) When $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$, $\boldsymbol{\rho}^*$ is asymptotially stable when one player uses best response, and the other uses gradient dynamics if and only if $\boldsymbol{\rho}^*$ it is asymptotically stable given $\bar{\alpha} = 1$ under $F_{B,S}$.

When there are $K \geq 1$ states, for any $\alpha \in (0,1]$ we can write the linearization $\boldsymbol{M}_B(\alpha)$ of F_S at symmetric $\boldsymbol{\rho}^* \in E_S$ as follows

$$\boldsymbol{M}_B(\alpha) = \begin{bmatrix} -\alpha \boldsymbol{I}_K & \alpha \boldsymbol{J} \\ \boldsymbol{J} & -\boldsymbol{I}_K \end{bmatrix},$$

where $\boldsymbol{J}$ is the best-response derivative matrix at $\boldsymbol{\rho}^*$ of players 1 and 2, by symmetry. As $-\boldsymbol{I}_K$ and $\boldsymbol{J}$ commute, the characteristic equation of $\boldsymbol{M}_B(\alpha)$ can be written as

$$\begin{aligned} char(\lambda) &= \det\left(\alpha\boldsymbol{J}\boldsymbol{J} - (\alpha+\lambda)(1+\lambda)\boldsymbol{I}_K\right) \\ &= \det\left(\alpha^{\frac{1}{2}}\boldsymbol{J} - ((\alpha+\lambda)(1+\lambda))^{\frac{1}{2}}\,\boldsymbol{I}_K\right)\det\left(\alpha^{\frac{1}{2}}\boldsymbol{J} + ((\alpha+\lambda)(1+\lambda))^{\frac{1}{2}}\,\boldsymbol{I}_K\right). \end{aligned}$$

Thus, for any μ such that $|\mu| \in eig(\boldsymbol{J})$, $\lambda_{1,2} \in eig\left(\boldsymbol{M}_B(\alpha)\right)$ where $\lambda_{1,2}$ are the solutions to

$$\lambda^2 + (1+\alpha)\lambda + \alpha(1-\mu^2) = 0,$$

i.e.

$$\lambda_{1,2} = -\frac{1+\alpha}{2} \pm \sqrt{\left(\frac{1+\alpha}{2}\right)^2 - \alpha(1-\mu^2)}.$$

$\boldsymbol{M}_B(1)$ has all eigenvalues negative if and only if $|\mu| < 1$. $\lambda < 0$ if and only if

$$\alpha(1-\mu^2) > 0,$$

which is equivalent to $|\mu| < 1$. Thus, stability under $\alpha = 1$ carries over to all $\alpha \in (0,1)$.

As for point (2), note that the Hessian of $W(\boldsymbol{\rho})$ at $\boldsymbol{\rho}_N$ equals $\pi_{11}^N \boldsymbol{I}_K$. An analogous argument to the proof of Lemma 1 can be used to show that $\boldsymbol{\rho}_N$ is asymptotically stable under gradient learning if and only if it is under F_S. Thus, the above conclusion remains under gradient learning for $\boldsymbol{\rho}_N$. Similarly, note that for point (3), at $\boldsymbol{\rho}_N$ we can consider the difference between $F_{S,B}$ and $F_{S,G}$ as being down to scaling every value in $F_{S,B}$ by the constant $-\pi_{11}^N \boldsymbol{I}_K$, which one can interpret as some $\alpha > 0$, and reach the required conclusion.

Regarding other symmetric equilibria $\boldsymbol{\rho}^* \in E_S$, the connection between gradient learning and best response dynamics is more tenuous. For any symmetric $\boldsymbol{\rho}^* \neq \boldsymbol{\rho}_N$, we may not write

$$\boldsymbol{M}_G = \nu \boldsymbol{M}_B,$$

for some scalar $\nu \neq 0$. In general, we'd have

$$\boldsymbol{M}_G = \boldsymbol{D}\boldsymbol{M}_B,$$

where $\boldsymbol{D} \geq 0$ is a diagonal matrix of positive, but varying, entries. In this case, it is easy to construct examples where $\boldsymbol{M}_B$ has negative eigenvalues only, but $\boldsymbol{M}_G$ does not. ■

B.G. **Proof of Theorem 5**

We start by a derivation analogous to the proof of Theorem 3, using (6). First, as transitions of $O_{S_{PR}}$ and O_{S_D} are independent, it is useful to look into properties of the joint transition matrix $\boldsymbol{M} \in [0,1]^{KL}$, which now represents an irreducible Markov chain over states $r \in R = O_{S_D} \times O_{S_{PR}}$. Note that $\boldsymbol{M}$ is a function of policies ρ through transition function T over S_{PR}. With some abuse of notation, let $W(\boldsymbol{\rho}, r)$ be the expected discounted payoffs of player 1 in this setting, given initial state $r \in R$. Given this notation of states, we denote F_R as the best-response dynamic solutions of which ACQ learners would converge to in this setting. As in the proof of Theorem 3, we can derive, at an equilibrium $\boldsymbol{\rho}^*$,

$$W_{ij} = [\boldsymbol{I}_K - \delta \boldsymbol{M}]^{-1} \left[\delta \frac{\partial^2 \boldsymbol{M}}{\partial \rho_i \partial \rho_j} \tilde{\boldsymbol{W}} + (1-\delta) \frac{\partial^2 \boldsymbol{U}}{\partial \rho_i \partial \rho_j} \right],$$

for $1 \leq i, j \leq KL$, using the notation as before, noting that now W, U are KL-dimensional vectors. Order states r so that $r_1, \ldots, r_K = (s_D^1, s_1), \ldots, (s_D^1, s_K)$, i.e. s_D is kept fixed while $s_k \in O_{S_{PR}}$ advances. Then, at $\boldsymbol{\rho}_N$ it follows that $U_N = U(\boldsymbol{\rho}_N)$ is such that the first K elements equal $\pi_N^1 = \pi(p_N^1, s_D^1)$, etc, until the last K elements equal $\pi_N^L = \pi(p_N^L, s_D^L)$.

To simplify these derivatives, the following will be helpful: Letting $\boldsymbol{T}_D, \boldsymbol{T}$ be the transition matrices of $O_{S_D}, O_{S_{PR}}$ respectively, due to the ordering over states we can write $\boldsymbol{M} = \boldsymbol{T}_D \otimes \boldsymbol{T}$, where $\otimes$ is the Kronecker product. As the spectral radius of $\delta \boldsymbol{M}$ is $\delta < 1$ ($\boldsymbol{M}$ is a row-stochastic matrix), we can apply the geometric series expansion and properties of the Kronecker product to get

$$[\boldsymbol{I} - \delta \boldsymbol{M}]^{-1} = \sum_{q=1}^{\infty} \delta^q \boldsymbol{M}^q = \sum_{q=1}^{\infty} \delta^q \left(\boldsymbol{T}_D^q \otimes \boldsymbol{T}^q \right).$$

Let $1 \leq \alpha, \beta \leq L$ be indices for $K \times K$ blocks of $[I - \delta \boldsymbol{M}]^{-1}$, which we denote $[I - \delta \boldsymbol{M}]^{-1}_{\alpha,\beta}$. Note that for each such block, rows sum to the same constant. Let $\mathbf{1}_K$ be a K dimensional column of ones. Row sums can be written as

$$[\boldsymbol{I}-\delta\boldsymbol{M}]^{-1}_{\alpha,\beta}\,\mathbf{1}_K=\sum_{q=1}^{\infty}\delta^q\left(\boldsymbol{T}^q_{D,\alpha,\beta}\otimes\boldsymbol{T}^q\right)\mathbf{1}_K=\sum_{q=1}^{\infty}\delta^q\boldsymbol{T}^q_{D,\alpha,\beta}\mathbf{1}_K=C_{\alpha,\beta}\mathbf{1}_K,$$

as rows of $\boldsymbol{T}^q$ sum to one for each $0\leq q$, and where $C_{\alpha,\beta}=\frac{1}{1-\delta\boldsymbol{T}_{D,\alpha,\beta}}$. As also $\boldsymbol{U}_N$ is constant across each K-element block, it follows that $\boldsymbol{W}=[\boldsymbol{I}-\delta\boldsymbol{M}]^{-1}\boldsymbol{U}_N$ is constant across each K-element block. This significantly simplifies the cross-derivatives required for the Jacobian of F_R to be analysed here. Derivatives of $\boldsymbol{M}$ such as $\frac{\partial^2\boldsymbol{M}}{\partial\rho_i\partial\rho_j}$ sum to zero across each K-element block, which implies that $\frac{\partial^2\boldsymbol{M}}{\partial\rho_i\partial\rho_j}W=0_{KL}$. We can thus write the derivatives of W evaluated at $\boldsymbol{\rho}_N$ in manner analogous to Theorem 3, taking r_1 as initial state without loss:

$$\begin{aligned}
W_{ii}&=[\boldsymbol{I}_K-\delta\boldsymbol{M}]^{-1}_1(1-\delta)\frac{\partial^2\boldsymbol{U}_N}{(\partial\rho_i)^2},\\
W_{ij}&=[\boldsymbol{I}_K-\delta\boldsymbol{M}]^{-1}_1(1-\delta)\left[\frac{\partial^2\boldsymbol{U}_N}{\partial\rho_i\partial\boldsymbol{\gamma}_j}+\delta\frac{\partial\boldsymbol{M}}{\partial\rho_i}[\boldsymbol{I}_K-\delta\boldsymbol{M}]^{-1}\frac{\partial\boldsymbol{U}_N}{\partial\boldsymbol{\gamma}_j}\right],
\end{aligned}$$

where γ_j is some opponent strategy with $1\leq i,j\leq KL$. For each such i,j, let $s_D(i),s_D(j)$ be the associated state $s_D\in O_{S_D}$. With some abuse of notation, let $\pi^N_{11}(i)=\frac{\partial^2\boldsymbol{U}_N}{(\partial\rho_i)^2}$, $\pi^N_{12}(i)=\frac{\partial^2\boldsymbol{U}_N}{\partial\rho_i\partial\boldsymbol{\gamma}_j}$, and $\pi^N_2(j)=\frac{\partial\boldsymbol{U}_N}{\partial\boldsymbol{\gamma}_j}$ similarly to Theorem 3. Then, letting $\boldsymbol{e}_i$ be KL-dimensional columns which are zero everywhere but in element i,

$$\begin{aligned}
W_{ii}&=[I-\delta\boldsymbol{M}]^{-1}_1(1-\delta)\pi^N_{11}(i)\boldsymbol{e}_i,\\
W_{ij}&=[I-\delta\boldsymbol{M}]^{-1}_1(1-\delta)\left[\pi^N_{12}(i)\boldsymbol{e}_i+\delta\frac{\partial\boldsymbol{M}}{\partial\rho_i}[I-\delta\boldsymbol{M}]^{-1}\pi^N_2(i)\boldsymbol{e}_i\right],
\end{aligned}$$

if $i=j$, and

$$W_{ij}=[I-\delta\boldsymbol{M}]^{-1}_1(1-\delta)\left[\delta\frac{\partial\boldsymbol{M}}{\partial\rho_i}[I-\delta\boldsymbol{M}]^{-1}\pi^N_2(j)\boldsymbol{e}_j\right],$$

otherwise, following the notation used in section III. To save notation, write $\boldsymbol{Z}=[\boldsymbol{I}-\delta\boldsymbol{M}]^{-1}$. Then,

$$\frac{W_{ij}}{W_{ii}}=\begin{cases}\frac{1}{\pi^N_{11}(i)}\left[\pi^N_{12}(i)+\pi^N_2(i)\delta\sum_{\ell=1}^{KL}\boldsymbol{\zeta}_{r_ir_\ell}\boldsymbol{Z}_{\ell,i}\right] & \text{if } i=j\\ \frac{1}{\pi^N_{11}(i)}\left[\pi^N_2(j)\delta\sum_{\ell=1}^{KL}\boldsymbol{\zeta}_{r_ir_\ell}\boldsymbol{Z}_{\ell,j}\right] & \text{o.w.}\end{cases}$$

where $\boldsymbol{\zeta}_{rr'} = \frac{\partial}{\partial \rho(r)} \boldsymbol{M}_{rr'}$. For the proof of point (1), we will upper bound eigenvalues of this system. Note from the above that we can write $\boldsymbol{J}$, the Jacobian of F_R, as

$$\boldsymbol{J} = (\boldsymbol{E}_1 \otimes \boldsymbol{I}_K) + (\boldsymbol{E}_2 \otimes \boldsymbol{I}_K)\, \delta \boldsymbol{\zeta} \boldsymbol{Z}\, (\boldsymbol{E}_3 \otimes \boldsymbol{I}_K)\,,$$

where $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3$ are L dimensional diagonal matrices with values equal to $-\frac{\pi_{12}^N(s_D^m)}{\pi_{11}^N(s_D^m)}$, $\frac{-1}{\pi_{11}^N(s_D^m)}$, $\pi_2^N(s_D^m)$, respectively over $1 \leq m \leq L$.

Now to the proof of point (1). As in the proof of Theorem 4, recall that the spectral radius κ of $\boldsymbol{J}$ can be upper bounded by any matrix norm of $\boldsymbol{J}$. Hence,

$$\begin{aligned}
\kappa &\leq \|\boldsymbol{E}_1\|\|\boldsymbol{I}_K\| + \delta\|\boldsymbol{E}_2 \otimes \boldsymbol{I}_K\|\|\boldsymbol{\zeta}\|\|\boldsymbol{Z}\|\|\boldsymbol{E}_3 \otimes \boldsymbol{I}_K\| = \|\boldsymbol{E}_1\| + \frac{\delta}{1-\delta}\|\boldsymbol{E}_2\|\|\boldsymbol{E}_3\|\|\boldsymbol{\zeta}\| \\
&= \lambda_1 + \lambda_2\lambda_3 \frac{\delta}{1-\delta}\|\zeta\|,
\end{aligned}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the largest absolute entries of the diagonal matrices η_1, η_2, E_3. Recall that we define, analogously to Theorem 3,

$$\begin{aligned}
\pi_2^N(s_D^m) &= X^m \xi_c^m L_N^m \\
\pi_{11}^N(s_D^m) &= X^m \left(\xi_{o,1}^m L_N^m + \xi_o^m \frac{\partial}{\partial p_1} L_N^m \right) \\
\pi_{12}^N(s_D^m) &= X^m \xi_{o,2}^m L_N^m,
\end{aligned}$$

where superscript m denotes payoff terms given state s_D^m, all evaluated at $p_N(s_D^m)$, which is dropped for ease of notation. Then,

$$\begin{aligned}
\lambda_1 &= \max_m \left| \frac{\xi_{o,2}^m}{\xi_{o,1}^m + \xi_o^m G_N^m} \right| \\
\lambda_2 &= \max_m \left(X^m \left| \xi_{o,1}^m L_N^m + \xi_o^m \frac{\partial}{\partial p_1} L_N^m \right| \right)^{-1} \\
\lambda_3 &= \max_m X^m \left| \xi_c^m L_N^m \right|.
\end{aligned}$$

For the first insight, given the upper bound on κ above, note that $\|\boldsymbol{\zeta}\|$ shrinks with $\sup_{a\in\Omega} \left| \frac{\partial}{\partial P} g(a;P)\big|_{P=P_N} \right|$, and the result follows. Point (1) then follows directly also from this upper bound.

For point (2), recall that for any matrix, the trace equals the sum of its eigenvalues. In this case,

$$tr(\boldsymbol{J}) = Ktr(\boldsymbol{E}_1) + \delta \sum_{m=1}^{L} \boldsymbol{E}_{2,m} \boldsymbol{E}_{3,m} (\boldsymbol{\zeta Z})_m,$$

where $\boldsymbol{E}_{2,m}, \boldsymbol{E}_{3,m}$ are m-th elements of $\boldsymbol{E}_2, \boldsymbol{E}_3$, and $(\boldsymbol{\zeta Z})_m$ is the m-th diagonal $K \times K$ block of $\boldsymbol{\zeta Z}$. Thus, for all $\boldsymbol{\zeta Z}$ there are elements of $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3$ large enough in absolute value so that $|tr(\boldsymbol{J})|$ is large, which can only be true if at least one eigenvalue of $\boldsymbol{J}$ is large. ■

References

Abreu, Dilip, David Pearce, and Ennio Stacchetti (1986). "Optimal cartel equilibria with imperfect monitoring". In: *Journal of Economic Theory* 39.1, pp. 251–269.

– (1990). "Toward a theory of discounted repeated games with imperfect monitoring". In: *Econometrica: Journal of the Econometric Society*, pp. 1041–1063.

Assad, Stephanie et al. (2024). "Algorithmic pricing and competition: empirical evidence from the German retail gasoline market". In: *Journal of Political Economy* 132.3, pp. 723–771.

Banchio, Martino and Giacomo Mantegazza (2022). "Games of Artificial Intelligence: A Continuous-Time Approach". In: *arXiv preprint arXiv:2202.05946*.

Benaïm, M (1999). "Dynamics of Stochastic Approximation, Le Seminaire de Probabilite', Springer Lecture Notes in Mathematics". In.

Benaïm, Michel and Mathieu Faure (2012). "Stochastic approximation, cooperative dynamics and supermodular games". In: *The Annals of Applied Probability* 22.5, pp. 2133–2164.

Borkar, Vivek S (2009). *Stochastic approximation: a dynamical systems viewpoint*. Vol. 48. Springer.

Brown, Zach Y and Alexander MacKay (2021). *Competition in pricing algorithms*. Tech. rep. National Bureau of Economic Research.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, et al. (2020). "Artificial intelligence, algorithmic pricing, and collusion". In: *American Economic Review* 110.10, pp. 3267–97.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicoló, et al. (2021). "Algorithmic collusion with imperfect monitoring". In: *International journal of industrial organization* 79, p. 102712.

Cartea, Álvaro et al. (2022). "Algorithms can Learn to Collude: A Folk Theorem from Learning with Bounded Rationality". In: *Available at SSRN 4293831*.

Faure, Mathieu and Gregory Roth (2010). "Stochastic approximations of set-valued dynamical systems: Convergence with positive probability to an attractor". In: *Mathematics of Operations Research* 35.3, pp. 624–640.

Fudenberg, Drew and David M Kreps (1993). "Learning mixed equilibria". In: *Games and economic behavior* 5.3, pp. 320–367.

Fudenberg, Drew, David Levine, and Eric Maskin (1994). "The Folk Theorem with Imperfect Public Information". In: *Econometrica* 62.5, pp. 997–1039.

Fudenberg, Drew and David K Levine (2009). "Learning and equilibrium". In: *Annu. Rev. Econ.* 1.1, pp. 385–420.

Gaunersdorfer, Andrea and Josef Hofbauer (1995). "Fictitious play, Shapley polygons, and the replicator equation". In: *Games and Economic Behavior* 11.2, pp. 279–303.

Gilboa, Itzhak and Akihiko Matsui (1991). "Social stability and equilibrium". In: *Econometrica: Journal of the Econometric Society*, pp. 859–867.

Hart, Sergiu and Andreu Mas-Colell (2003). "Uncoupled dynamics do not lead to Nash equilibrium". In: *American Economic Review* 93.5, pp. 1830–1836.

Johnson, Justin, Andrew Rhodes, and Matthijs R Wildenbeest (2020). "Platform design when sellers use pricing algorithms". In: *Available at SSRN 3753903*.

Kandori, Michihiro (1992). "The use of information in repeated games with imperfect monitoring". In: *The Review of Economic Studies* 59.3, pp. 581–593.

Klein, Timo (2021). "Autonomous algorithmic collusion: Q-learning under sequential pricing". In: *The RAND Journal of Economics* 52.3, pp. 538–558.

Lamba, Rohit and Sergey Zhuk (2022). "Pricing with algorithms". In: *arXiv preprint arXiv:2205.04661*.

Leslie, David S and Edmund J Collins (2003). "Convergent multiple-timescales reinforcement learning algorithms in normal form games". In: *The Annals of Applied Probability* 13.4, pp. 1231–1251.

Leslie, David S, Steven Perkins, and Zibo Xu (2020). "Best-response dynamics in zero-sum stochastic games". In: *Journal of Economic Theory* 189, p. 105095.

Loots, Thomas and Arnoud V. den Boer (2023). "Data-driven collusion and competition in a pricing duopoly with multinomial logit demand". In: *Production and Operations Management* 32.4, pp. 1169–1186.

Mazumdar, Eric, Lillian J Ratliff, and S Shankar Sastry (2020). "On gradient-based learning in continuous games". In: *SIAM Journal on Mathematics of Data Science* 2.1, pp. 103–131.

Meylahn, Janusz M and Arnoud V. den Boer (2022). "Learning to collude in a pricing duopoly". In: *Manufacturing & Service Operations Management* 24.5, pp. 2577–2594.

Milgrom, Paul and John Roberts (1990). "Rationalizability, learning, and equilibrium in games with strategic complementarities". In: *Econometrica: Journal of the Econometric Society*, pp. 1255–1277.

– (1991). "Adaptive and sophisticated learning in normal form games". In: *Games and economic Behavior* 3.1, pp. 82–100.

Palis Jr, J, W de Melo, et al. (1982). "Geometric Theory of Dynamical Systems". In.

Papadimitriou, Christos and Georgios Piliouras (2018). "From nash equilibria to chain recurrent sets: An algorithmic solution concept for game theory". In: *Entropy* 20.10, p. 782.

Puterman, Martin L (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *The annals of mathematical statistics*, pp. 400–407.

Salcedo, Bruno (2015). "Pricing algorithms and tacit collusion". In: *Manuscript, Pennsylvania State University*.

Schulman, John et al. (2017). "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347*.

Watkins, Christopher John Cornish Hellaby (1989). "Learning from delayed rewards". In.