MONITORING, MARKET PRIMITIVES, AND THE STABILITY OF ALGORITHMIC COLLUSION

CLEMENS POSSNIG

Abstract. This paper develops an analytical framework to study when sophisticated machine learning algorithms may learn to collude. 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. It is well known by now that these tools appear to have an inherent ability to collude.[1] However, we still lack a general theoretical understanding of the conditions under which these outcomes emerge. This paper addresses a fundamental question at the heart of the literature on algorithmic collusion:

Date: February 16, 2026.
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) .

[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. Assad et al. 2024 observe that after a critical mass of firms deployed pricing algorithms, profit margins rose by 28%, and price wars were implemented after some firms cut prices.

Which algorithmic architectures, and which market environments, are conducive to collusion?

I develop a unified framework to analyze the long-run behavior of a broad family of learning algorithms across diverse market structures. By characterizing the asymptotic properties of these systems tractably, I uncover a novel interplay between high-level algorithmic properties and market primitives. I demonstrate that the convergence to either a stage-game Nash equilibrium or a collusive outcome is governed by the interaction between the sensitivity of the algorithms' monitoring technology and the underlying market conditions, specifically price elasticities and markups.

The algorithms considered here fall within the foundational learning paradigm known as "reinforcement learning" (RL). RL algorithms update their policy, a mapping from some observed state variable to actions (prices), using observations of payoffs accrued over time. Often, such algorithms estimate a value function based on their current policy, and compute a best response, or gradient, to update their policy, and repeat.

The algorithms' state variable serves as a model of the data input that the RL's actions are conditioned on.[2] This data can act as a monitoring technology through the correlation between observations and price choices of competitors. Whether stage-game Nash or collusion can be learned by algorithms comes down to an interplay between the sensitivity of the monitoring technology, and market conditions derived from price elasticities and markups.

The intuition hinges on the stochastic nature of the algorithmic learning process. Any real-world application of algorithmic learning involves estimation, e.g. of expected payoffs, implying the possibility of making mistakes. Such mistakes induce perturbations in currently played policy profiles. Whether a given profile may be learned is due to its robustness to such mistakes.

I establish that for a perturbation to destabilize an equilibrium, three conditions must be met: (i) the shock must be detected through the monitoring technology, (ii) it must have a significant payoff impact, and (iii) the optimal algorithmic response must be sufficiently large to shift the system's trajectory. I show that when monitoring is highly sensitive, even minor perturbations can lead the system toward collusive outcomes, provided the market

[2]E.g., aggregate market conditions, time of day, sales volume, consumer data, etc.

conditions amplify the payoff relevance of these shocks. To the best of my knowledge, this is the first paper that uncovers the connection between monitoring technology of algorithms and market conditions as drivers for long-run outcomes of multi-agent learning settings.

I focus specifically on the "actor-critic" family of algorithms. Unlike "critic-only" methods such as Q-learning, actor-critic architectures decouple the performance evaluation (the critic) from the policy update (the actor). This separation provides a crucial variance-reduction property, allowing policy updates to remain robust to estimation errors in the value function. Actor-critic methods have become highly popular due to this desirable variance property, and are applied in large-scale, real-world tasks.[3]

The popularity of this algorithm family makes it particularly well-suited for evaluating regulatory interventions against algorithmic collusion. As a further contribution, I examine two distinct policy dimensions. First, leveraging the insight that monitoring facilitates collusion, I consider regulations that curb an algorithm's monitoring sensitivity. Here I show that such a policy will have a clear impact on improving the likelihood of stage-game Nash to be learned, while curbing the extent of possible collusive outcomes.

Second, I investigate the effects of increased competition in the algorithmic software market. While a monopolistic market may foster symmetry in algorithmic software produced—and thus coordination—competition may introduce heterogeneity. I analyze possible asymmetries along three dimensions: data inputs (state variables), learning rates, and critic profiles. In doing so, I offer the first steps at analysing the implications of asymmetry in the chosen state variables among competing algorithms, which is an important unresolved concern in the literature on algorithmic collusion.

I show that competition in the software market leads to overall ambiguous welfare effects. For instance, consider the transition from a symmetric to an asymmetric state regime (i.e. from a setting with a common state to one with asymmetric states), when there are two learning agents. Suppose the stage-game Nash equilibrium cannot be learned in the symmetric regime. Whether it will be learned depends on the information content of the new regime relative to the old. Specifically, if Nash cannot be learned under symmetry, it may be restored under asymmetry only if that regime sufficiently coarsens the information

[3]See for example Proximal Policy Optimization (PPO), Schulman et al. 2017.

available to one agent while leaving the other's information unchanged; otherwise, collusion is likely to persist. Secondly, I show that the learnability of symmetric equilibria is robust to asymmetry among learning rates, and restricting attention to stage game Nash, there is also robustness of my learnability to asymmetric critic terms.

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.

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 a 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 , it cannot accommodate Q-learning as a special case, nor is it a special case of Q-learning (ACQ, see discussion below). 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. While applying similar technical methods to establish convergence results, this paper takes the analysis further by asking what properties of the competitive environment of the algorithms make it so that some equilibria are attracting, and others are not.

Meylahn and V. den Boer 2022, Loots and V. den Boer 2023 use 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 determining 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 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.

Thirdly, this paper makes use of and generalizes an extensive body of research related to stochastic approximation theory (see e.g. Borkar 2009). The generalizations consider actor-critic learning under non-vanishing estimation bias, and 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. The Model

There are n agents 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, assumed to be twice differentiable in Y, $T : O_S^2 \times Y \to [0, 1]$. Each agent has a stage game payoff function $\pi^i : Y \times O_S \to \mathbb{R}$, $\mathcal{C}^2$ in Y, and common discount factor $\delta \in (0, 1)$.[4]

[4]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 $\mathcal{C}^i$.

The agents choose a policy function $\rho_i : O_S \to Y_i$. Since states are finite, policy profile $\boldsymbol{\rho} \in \overline{Y} = \times_i Y_i^K$ can be represented as a vector in $\mathbb{R}^{nK}$. Next, define $\overline{Y}_i = Y_i^K$, and $\overline{Y}_{-i} = \times_{j \neq i} \overline{Y}_j$. The following assumption on irreducibility is commonly made in RL applications and significantly simplifies notation.[5]

Assumption 1. *For all $\boldsymbol{\rho} \in \overline{Y}$, 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. The agents' objectives, expected discounted payoffs, are defined given stationary policy profiles $(\boldsymbol{\rho}_i, \boldsymbol{\rho}_{-i}) \in \overline{Y}$, and any initial state $s_0 \in O_S$:

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

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

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

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-dependent policies since given a Markov stationary opponent profile $\boldsymbol{\rho}_{-i}$ 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.

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. $\boldsymbol{\rho}^* \in \bar{B}_S(\boldsymbol{\rho}^*)$.

[5] When irreducibility fails, agents' play will lead to some absorbing subset of states, which depends on initial conditions.

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

(2) $\boldsymbol{\rho}^* \in E_S$ *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. Define the subset of such* $\boldsymbol{\rho}^*$ *as* $E_S^* \subseteq E_S$*.*

Given these definitions regarding the underlying payoff environment, assume:

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

(1) Given state variable S*,* E_S^* *is nonempty.*

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

II. Multi-Agent Learning

I study a family of RL algorithms that update policies $\boldsymbol{\rho}_t^i : O_S \to Y_i$ in discrete time. The algorithms keep track of an estimator for directions of improvement of W^i, denoted $\hat{\Psi}_t^i$. This setup allows considering a host of popular RL algorithms: For example, what is known as actor-critic Q (ACQ), for which $\hat{\Psi}_t^i = \hat{b}_t^i - \rho_t^i$ represents the difference between the current estimate $\hat{b}_t^i$ of best response $B_S^i(\boldsymbol{\rho}_{-i})$, and current policy $\boldsymbol{\rho}_t^i$. Secondly, actor-critic gradient (ACG) learning, for which $\hat{\Psi}_t^i$ is an estimate of $\nabla W^i(\boldsymbol{\rho}_t, s_0)$, the gradient of the value function with respect to policy $\boldsymbol{\rho}^i$. In general, I refer to the target of the critic estimator as a function $\Psi^i(\boldsymbol{\rho}) \in \mathbb{R}^K$, and $\Psi(\boldsymbol{\rho}) = (\Psi^1(\boldsymbol{\rho}), \dots \Psi^n(\boldsymbol{\rho})) \in \mathbb{R}^{nK}$ as a critic profile.

Assumption 3. *Say a critic profile* Ψ *is admissible if*

(1) $\Psi(\boldsymbol{\rho})$ *is almost everywhere continuously differentiable in the interior of* $\overline{Y}$*.*

(2) E_S *is a subset of the rest points of* $\Psi(\boldsymbol{\rho})$*.*

(3) Differential Nash equilibria are isolated rest points of $\Psi(\boldsymbol{\rho})$*, and* $\Psi(\boldsymbol{\rho}^*)$ *is continuously differentiable at* $\boldsymbol{\rho}^* \in E_S^*$*.*

Point (1) ensures the existence of unique solutions of an ordinary differential equation (ODE) based on the critic profile when initialized in the interior[7], which will be useful for

[7] As follows from the Picard-Lindelöf Theorem.

our characterisation of long-run behavior later on. Regarding the boundary of $\overline{Y}$, I will allow for the application of projection operators that ensure that $\boldsymbol{\rho}$ stays within $\overline{Y}$, as may be required under ACQ and ACG methods. Points (2), (3) ensure that Nash equilibria may be converged to by the algorithms, and that their stability properties can be analysed in a standard fashion. Note that this is an assumption on both the critic chosen for the algorithm, and the underlying game; indeed, this requires under ACQ-learning that best responses be almost everywhere differentiable. When the stage game is globally quasi-concave, and transition functions are twice differentiable as assumed here, standard results in optimization theory such as parametric transversality ensure almost everywhere differentiability.[8]

Each algorithm's policy updating rule comes with a stepsize sequence $\alpha_t^i \to 0$ with $t \to \infty$, which I write as follows given some initial $\boldsymbol{\rho}_0^i$:

$$\boldsymbol{\rho}_{t+1}^i = \boldsymbol{\rho}_t^i + \alpha_t^i \hat{\Psi}_t^i. \tag{3}$$

As discussed previously, the 'actor' $\boldsymbol{\rho}_t^i$ is updated using the 'critic' $\hat{\Psi}_t^i$, but updates are dampened with α_t^i. These stepsizes are assumed to satisfy the Robins-Monro conditions

$$\sum_{t\geq 0} \alpha_t^i = \infty; \qquad\qquad \sum_{t\geq 0} (\alpha_t^i)^2 < \infty,$$

which ensure that steps converge slowly enough so that policies may traverse the whole action space, while being fast enough so as to diminish any variance impact due to estimation errors of the critic $\hat{\Psi}_t^i$. Note that convergence of stepsizes to zero does not imply convergence of the policy updating scheme, as possible cyclical behavior is supportable when stepsize are not summable.[9]

The goal of this paper is to gain insights about what can be learned as long as the critic estimator is reasonably well-behaved, a property to be defined below. To this end, I follow a common approach in the literature of reinforcement learning in games, by imposing

[8] Assumption 3 (1) can be weakened to require differentiability only in a neighborhood of E_S^*, but would add distracting notation.

[9] In applications, stepsizes typically converge to a small, positive number. Some results shown later on qualitatively extend to this case, with appropriate adjustments. Theoretical results in the literature on reinforcement learning resort to assuming $\alpha_t^i \to 0$ as it allows for crisper predictions (c.f. Borkar 2009).

assumptions on the bias and variance components of $\hat{\Psi}^i$.[10] Note that well-behavedness in the critic estimation does not imply convergence of the algorithms to a Nash equilibrium. Write

$$\begin{aligned}\hat{\Psi}_t^i &= \Psi_t^i + \left(\hat{\Psi}_t^i - \Psi_t^i\right) \\ &= \Psi_t^i + \boldsymbol{d}_{t+1}^i + \boldsymbol{M}_{t+1}^i,\end{aligned}$$

where $\boldsymbol{d}_{t+1}^i = \mathbb{E}\left[\hat{\Psi}_t^i - \Psi_t^i \mid \mathcal{F}_t^i\right]$, represents a bias, and $\boldsymbol{M}_{t+1}^i = \left(\hat{\Psi}_t^i - \Psi_t^i\right) - \boldsymbol{d}_{t+1}^i$ a variance term of the estimation error given a natural filtration $\mathcal{F}_t^i$ that keeps track of payoff, state, and action histories of the algorithm.

Detailed sufficient conditions on stepsizes, bias, and error terms are relegated to Appendix A. Importantly, I will assume that the bias terms $\boldsymbol{d}_t^i$ converge to zero, in expectation, at a speed that is faster than α_t^i. Essentially, this says that errors in the critic estimation average out an order of magnitude faster than $\boldsymbol{\rho}_t^i$ moves; hence to the critic estimation, $\boldsymbol{\rho}_t$ appears essentially stationary, allowing for standard laws of large numbers to kick in.

The assumptions 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). Finally, I assume throughout that stepsizes of all agents lie within an order of magnitude of each other, which is further discussed later on.

Remark 1. *An important approach sufficient to satisfy these assumptions involves two-timescale learning methods, of which Chen et al. 2024 provide a recent study on stochastic zero-sum games. Here, critic terms are updated based on histories of payoff, state, and own-action observations (in case of a critic based on a value function, think of a Bellman iteration scheme) at a rate that is faster than the updating speed of the policy (controlled by α_t^i). When all agents update critic terms at sufficiently fast speeds relative to their policy updates, critic estimates will converge towards their targets while policies stay approximately stationary, ensuring the abovementioned bias terms $\boldsymbol{d}_t^i$ indeed converge faster than stepsizes.*

[10] See Mazumdar, Ratliff, and Sastry 2020 who make a stronger assumption, and Mertikopoulos, Hsieh, and Cevher 2024 for a recent work using this approach, giving examples satisfying their assumptions in 'critic-only' settings.

II.A. **Long Run Outcomes**

Now to state the characterisation of long-run learning outcomes for the algorithmic families discussed here. For a set A, let $cl\,(A)$ be its closure.

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

$$L_S(\boldsymbol{\rho}_0) = \bigcap_{t\geq 0} cl\left(\{\boldsymbol{\rho}_\ell \,|\ell \geq t\}\right),$$

the set of limits of convergent subsequences $\boldsymbol{\rho}_{t_k}$*, given some initial* $\boldsymbol{\rho}_0$*.*

The limit set represents 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.

Long run behavior of our algorithmic family is pinned down using the critic profile weighted by the limiting relative stepsize sequences of all algorithms; this is not surprising, as these relative stepsizes encode the relative updating speeds of the algorithms. The following two characterising results are straightforward applications of results in Borkar 2009 and Benaïm and Faure 2012 in our setting, and hence their proofs are moved to the online appendix. Their stochastic approximation theory allows us to connect long run behavior of the algorithms

to solutions of an ODE with right-hand side $\Psi_A(\boldsymbol{\rho}) = \Psi_A(\boldsymbol{\rho})$, where $\boldsymbol{A}$ is a diagonal matrix with entries equal to the abovementioned limiting relative stepsizes.

Theorem 1. *Let $\boldsymbol{\rho}^* \in E_S$ be asymptotically stable for the weighted admissible critic profile $\Psi_A(\boldsymbol{\rho})$. Then there is an open neighborhood $\mathcal{N}$ around $\boldsymbol{\rho}^*$ so that for all $\boldsymbol{\rho}_0 \in \mathcal{N}$,*

$$\mathbb{P}[L_S(\boldsymbol{\rho}_0) = \{\boldsymbol{\rho}^*\}] > 0.$$

For intuitions, stochastic approximation theory as developed in Borkar 2009 shows how one can connection the iteration of $\boldsymbol{\rho}_t$ to solutions of the ordinary differential equation

$$\dot{\boldsymbol{\rho}} = \Psi_A(\boldsymbol{\rho}).$$

The idea is to show that the time-interpolated version of $\boldsymbol{\rho}_t$ must stay close, with probability close to 1, to solutions of $\Psi_A(\boldsymbol{\rho})$. 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 $\Psi_A(\boldsymbol{\rho})$. Then there exists an open neighborhood $\mathcal{N}$ of $\boldsymbol{\rho}^*$ such that for all $\boldsymbol{\rho}_0 \in \bar{Y}$*

$$\mathbb{P}[L_S(\boldsymbol{\rho}_0) \subseteq \mathcal{N}] = 0.$$

Here, $\boldsymbol{\rho}^*$ being unstable implies that there are directions in the policy space around $\boldsymbol{\rho}^*$ which act as a repeller to the differential equation based on $\Psi_A(\boldsymbol{\rho})$. The noise generated by estimation errors of the algorithms ensures that, 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 $\boldsymbol{\rho}_t$ hits such directions and therefore must move away from $\boldsymbol{\rho}^*$.

Finally, note that Theorem 1 does not state that all elements of 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 (Benaïm and Faure 2012) that suggest one may extend Theorem 2 to such linearly unstable cycles also.[11] Notice that this observation implies that the Folk theorem is neither necessary nor sufficient in describing the possible payoffs achievable by learning algorithms.

Next, I will move to applying the results of this section in a foundational market competition game. I show how stability properties of repeated-game equilibria can be analysed via standard economic intuitions.

III. **Learning in the Bertrand Model**

I consider here as application a differentiated-goods Bertrand model (Bertrand model for short).

There are two firms, $i \in \{1,2\}$. Firms choose prices $p_i \in Y \subseteq \mathbb{R}_+$. Posted prices induce an aggregate demand shock observable by all. Let $\tilde{A} \in \Omega \subset \mathbb{R}_+$ be the random variable representing this shock, with Ω compact. Conditional on the price vector $\boldsymbol{p}$, $\tilde{A}$ has an absolutely continuous distribution with a density $g(a;P)$, where $P = p_1 + p_2$ is the aggregate price. Demand for each individual firm is then a random variable, denoted $\tilde{X}_i$, which depends on the other firm's actions only through $\tilde{A}$. The expectation of $\tilde{X}_i$ given $\boldsymbol{p}$ is written as $X_i(\boldsymbol{p}) = \mathbb{E}[\tilde{X}_i \mid \boldsymbol{p}]$.

Expected profits are 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, I 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 4 (Stage Game)**.** *For all* $\boldsymbol{p} \in \mathbb{R}^2_+$, $p_{-i} \geq 0$*:*

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

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

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

(4) $\left|\frac{\partial}{\partial p_i} X_i(\boldsymbol{p})\right| > \left|\frac{\partial}{\partial p_{-i}} X_i(\boldsymbol{p})\right|$

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

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

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

(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* $X_i(\boldsymbol{p}) = (D - bp_i + \gamma p_{-i})_+$, *where* $D > 0$, $b > \gamma > 0$. *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 accommodates our motivation above, and satisfies Assumption 4.*

Example 2. *A slight extension to the logit-demand model also accommodates 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. Let the aggregate shock* $\tilde{A} \in [0, Z]$ *for some* $Z > 0$ *represent the mass of active consumers. The distribution of* $\tilde{A}$ *depends on aggregate price* P *so that* $Z \exp(-\gamma P)$, *for some sensitivity parameter* $\gamma > 0$. *Conditional on participation, consumers choose between firms according to the standard multinomial logit model. Then* $X_i(\boldsymbol{p}) = \mathbb{E}[\tilde{A} \mid P] \exp(\mu_i - \beta_i p_i) / \left(\sum_{j=1}^{2} \exp(\mu_j - \beta_j p_j)\right)$, *where* $\mu_i, \beta_i > 0$.[12]

Furthermore, throughout this section I assume the following:

Assumption 5 (Symmetric Algorithms)**.** *Assume that*

(1) Both agents observe the same state variable S.

(2) $\frac{\alpha_t^1}{\alpha_t^2} \to 1$ *as* $t \to \infty$, *i.e. stepsizes of the two agents are asymptotically equal.*

This assumption allows narrowing the focus on sensitivities of the common state and market conditions in the results to follow. What may be learned when these assumptions fail is discussed in Section IV. Define the trivial state variable, which takes only one value (i.e. only learning of stage-game strategies is possible), as S_0.

[12] Assumption 4 needs to be weakened for this example. The remaining claims in this section carry over here as well for small enough γ, which ensures that for a large enough set of prices, the strategic substitutes property of the stage game is not lost.

Lemma 1. *Under Assumption 4, there is a unique interior Nash equilibrium $\boldsymbol{p}_N$, which is symmetric. Given state variable 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. Indeed, the stage game considered here under Assumption 4 is a classical example of robust convergence to static Nash in the literature on evolutionary learning. Once non-trivial states are possible, convergence to (the repetition of) static Nash becomes substantially harder to achieve.

III.A. **Static Nash**

Recall that I defined $T_{ss'}(p_1, p_2)$ as the twice continuously differentiable transition probability of moving to s' given current state s and choices p_i in state s. I further impose the following:

Assumption 6. *Assume that for all $s, s' \in O_S$, all $p_1, p_2 \in \mathbb{R}_+$, and all $i, j \in \{1, 2\}$:*

$$\frac{\partial}{\partial p_i} T_{ss'}(p_1, p_2) = \frac{\partial}{\partial p_j} T_{ss'}(p_1, p_2)$$
$$\frac{\partial^2}{(\partial p_i)^2} T_{ss'}(p_1, p_2) = \frac{\partial^2}{(\partial p_j)^2} T_{ss'}(p_1, p_2) = \frac{\partial^2}{\partial p_i \partial p_j} T_{ss'}(p_1, p_2).$$

I impose this assumption throughout the rest of the paper. A sufficient condition to satisfy this is to have transitions depend on the price vector only through an aggregate price index, i.e. as is true for aggregate shock $\tilde{A}$. I will therefore commonly write $T_{ss'}(p_1, p_2) = T_{ss'}(P)$ with $P = p_1 + p_2$. As an example, one may think of standard public monitoring schemes as considered in Abreu, Pearce, and Stacchetti 1986, where realizations of $\tilde{A}$ would induce state transitions. In such a setting, one could have a threshold rule for realizations of $\tilde{A}$ to induce state transitions. All results stated here do not require the additional assumption that the state be payoff relevant as in this example.

For the remainder of this section, if not further specified, results are stated regarding ACQ learners only. How insights extend to gradient learners will be discussed in Section IV.

To state the first main result of this section, I define a measure of monitoring sensitivity. This can be thought of as the sensitivity of transition probabilities in regard to deviations in price choices; if transitions are sensitive, agents are more likely to detect deviations.

Definition 4. *Given policy profile $\boldsymbol{\rho}$, define the monitoring sensitivity upper bound at $\boldsymbol{\rho}$ as*

$$\overline{R}(\boldsymbol{\rho}) = \max_{s,s' \in O_S} \left| T'_{ss'}(\boldsymbol{\rho}(s)) \right|,$$

where $T'_{ss'}(\boldsymbol{\rho}(s))$ refers to the derivative with respect to aggregate price. Similarly, the monitoring sensitivity lower bound at $\boldsymbol{\rho}$ is

$$\underline{R}(\boldsymbol{\rho}) = \min_{s,s' \in O_S} \left| T'_{ss'}(\boldsymbol{\rho}(s)) \right|.$$

Define the uniform versions of these as $\overline{R} = \sup_{\boldsymbol{\rho} \in \overline{Y}} \overline{R}(\boldsymbol{\rho})$, and $\underline{R} = \inf_{\boldsymbol{\rho} \in \overline{Y}} \underline{R}(\boldsymbol{\rho})$.

Clearly, the larger $\underline{R}$, the larger the effect a deviation has on future realizations of the state variable, and hence the higher the likelihood of being detected. Turning to market properties, define own-and-opponent's price elasticity 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 $L(p) = \frac{p-c}{p}$, 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

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

The more standard case $d\xi_o(\boldsymbol{p}) < 0$ 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 4.[13]

Theorem 3. *Consider a state variable S with $|O_S| = K > 1$ states, and let $\boldsymbol{\rho}_N : O_S \to \mathbb{R}_+$ be the policy that plays p_N in every state. Given π, there exists $0 < C_\pi < \infty$ such that*

(1) $\boldsymbol{\rho}_N$ will be learned with positive probability if $\overline{R} < C_\pi$.

(2) $\boldsymbol{\rho}_N$ will not be learned if $\underline{R} > C_\pi$.

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

C_π is proportional to

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

This finding has immediate policy relevance: Points $(1), (2)$ indicate that if regulators can limit the sensitivity of the data algorithms use, they might be able to promote more competitive outcomes (see Section IV).

The characterisation of C_π clarifies how market properties can lead to an increase in the set of transition probabilities that allow for the learning of $\boldsymbol{\rho}_N$. Overall, this point indicates an interesting dichotomy: on the one hand, weaker substitutability (low ξ_c) 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 result holds for *any* state variable in the context of this model.

In terms of intuitions, recall that robustness to perturbations of $\boldsymbol{\rho}_N$ with respect to $\Psi(\boldsymbol{\rho})$ determines whether this equilibrium may be learned by algorithms. Under ACQ learners, robustness to perturbations is analogous to robustness of the best response map to perturbations. Here, being able to detect likely deviations (i.e. having an effective monitoring technology) is necessary for perturbations to have any bite at all, as otherwise best responses would not react to a perturbation. Next, best responses depend on profits through the stage game. Hence, the properties C_π 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), while secondly own deviations must readily correct for perturbations, requiring a 'large enough' $|\xi_o|, |d\xi_o|$.

The conclusion is concerning, as achieving favorable market conditions for the learning of $\boldsymbol{\rho}_N$ may be difficult, weak subsitutes being commonly associated also with low own-price elasticity. An example of markets favorable to the learning of $\boldsymbol{\rho}_N$ according to C_π would be luxury goods markets with high brand recognition. Strong branding leads to weak substitutes, while luxury goods may be avoided by some consumer groups if prices are too high, which constitutes high ξ_o.

III.B. **Collusion**

Turning away from static Nash, 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.[14] The stability of such optimal collusive schemes then serves as a bound to the extent of possible collusion among algorithms.

However, the relationship between stability and monitoring sensitivity of collusive schemes is more complex than that of $\boldsymbol{\rho}_N$ derived previously; next, I show whether a collusive equilibrium may be learned depends crucially on the growth rate of $T'_{ss'}(\boldsymbol{\rho}(s))$. It is therefore necessary to introduce another sensitivity measure. Let $T''_{ss'}(\boldsymbol{\rho}(s))$ be the second derivative of transition probabilities with respect to aggregate price. Then define

$$\underline{R}^*(\boldsymbol{\rho}) = \min_{s,s' \in O_S} \frac{|T'_{ss'}(\boldsymbol{\rho}(s))|}{1 + |T''_{ss'}(\boldsymbol{\rho}(s))|}; \qquad \overline{R}^*(\boldsymbol{\rho}) = \max_{s,s' \in O_S} \frac{|T'_{ss'}(\boldsymbol{\rho}(s))|}{1 + |T''_{ss'}(\boldsymbol{\rho}(s))|},$$

and similar to before, let $\underline{R}^* = \inf_{\boldsymbol{\rho} \in \overline{Y}} \underline{R}^*(\boldsymbol{\rho})$, $\overline{R}^* = \sup_{\boldsymbol{\rho} \in \overline{Y}} \overline{R}^*(\boldsymbol{\rho})$.

Theorem 4. *For any binary state variable S and symmetric $\boldsymbol{\rho}^* \in E^*_S$, there exists $0 < C_{\pi,1}, C_{\pi,2} < \infty$ such that*

(1) $\boldsymbol{\rho}^*$ will be learned with positive probability if $\overline{R}^* < C_{\pi,1}$.

(2) $\boldsymbol{\rho}^*$ will not be learned if $\underline{R}^* > C_{\pi,2}$.

In contrast to $\boldsymbol{\rho}_N$, equilibria considered in Theorem 4 feature differing stage-game payoffs realized in different states. Upon detecting deviations, the learner will adjust their behavior; if the optimal adjustment is small, a perturbation may lead back to the equilibrium $\boldsymbol{\rho}^*$. When curvature is high, a small adjustment is enough to affect the distribution of future payoffs substantially; hence, the optimal adjustment is likely to be small.

This insight implies the possibility of a concerning scenario: When stage-game payoffs are equalized across states, as is true under $\boldsymbol{\rho}_N$, all terms featuring curvature of transition

[14] APS prove this given finite action sets. In the online appendix, I show how the result extends in an approximate sense to the continuous action case.

probabilities cancel (as changes in the distribution of payoffs don't matter when all payoffs are equalized). One may then have that $\underline{R}$ is large enough so that $\boldsymbol{\rho}_N$ may not be learned, while $\overline{R}^*$ is small enough so that highly collusive, binary-state equilibria are learned instead.

III.C. **Numerical Example**

I present here a brief numerical example satisfying the concerning scenario discussed above. Consider a stage game with linear demand as discussed in Example 1, where $D = 20, b = 1, \gamma = 1/2$, and $c = 2$. One can verify that in this model, $p_N = \frac{D-c}{2b-\gamma} = 12$. Suppose that agents discount time with $\delta = 0.9$. The agents are ACQ learners and commonly observe a binary state variable with $O_S = \{A, B\}$, where transition probability functions take a logit form. Specifically,

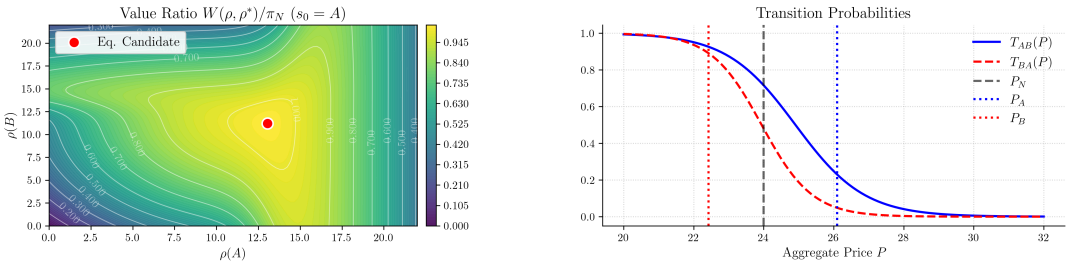
$$T_{AB}(P) = \left(1 + \exp\left(k_A\left(P - d_A\right)\right)\right)^{-1},$$
$$T_{BA}(P) = \left(1 + \exp\left(k_B\left(P - d_B\right)\right)\right)^{-1},$$

where $k_A = 1.02, d_A = 24.91, k_B = 1.38, d_B = 23.94$. I verify numerically (see figure 1a) that this set of transition functions supports two symmetric collusive equilibria $\boldsymbol{\rho}^{*(1)}, \boldsymbol{\rho}^{*(2)}$ where $\boldsymbol{\rho}^{*(1)}(A) \approx 13.05 > p_N > \boldsymbol{\rho}^{*(1)}(B) \approx 11.21$, and $\boldsymbol{\rho}^{*(2)}(A) \approx 11.54 < p_N < \boldsymbol{\rho}^{*(2)}(B) \approx 12.64$. Hence, one state realization supports a collusive state of high prices, and the other induces a punishment state needed to incentivize the collusive price.[15]

This setup confirms the intuitions discussed after Theorem 4: as shown in Figure 1b, transition functions are excessively steep around $\boldsymbol{\rho}_N$, and the static Nash equilibrium is therefore unstable under Ψ_B; therefore never learned by the ACQ learners. At the same time, curvature around $\boldsymbol{\rho}^{*(1)}, \boldsymbol{\rho}^{*(2)}$ is high enough for these transition functions so that these equilibria are attracting, therefore will be learned with strictly positive probability.[16] In

[15]Note that as both T_{AB}, T_{BA} are decreasing in P, this is intuitive: e.g. for $\boldsymbol{\rho}^{*(1)}$, at high prices in state A, optimal one-shot deviations ask for reduction in price, which leads to an increase in probability of the punishment state. On the other hand, in the punishment state B, optimal one-shot deviations require increases in price, which leads to an increase in the probability of remaining in B.

[16]Following Theorems 1, 2, I use the characterisation of Lemma 3 to compute the jacobian of $\Psi^1_B(\boldsymbol{\rho}_N)$ and $\Psi^1_B(\boldsymbol{\rho}^{*(j)})$ for $j \in \{1, 2\}$. The eigenvalues are $\Lambda_N \approx \{.25, 1.52\}$ for the former, and $\Lambda^{*(1)} \approx \{-.21, .41\}, \Lambda^{*(2)} \approx \{.05, .42\}$ for the latter, implying that $\boldsymbol{\rho}_N$ is repelling, while $\boldsymbol{\rho}^{*(1)}$ and $\boldsymbol{\rho}^{*(2)}$ attract (see Lemma 2).



(a) (b)

FIGURE 1. Figure (a) shows a contour plot in price space Y_1 of agent 1's value ratio $W(\boldsymbol{\rho}, \boldsymbol{\rho}^{*(1)}, s_0 = A)/\pi_N$ when 2 follows $\boldsymbol{\rho}^{*(1)}$. One can see that firstly, a global maximum is achieved at $\boldsymbol{\rho} = \boldsymbol{\rho}^{*(1)}$, and secondly that this maximum leads to a value $\approx 2.7\%$ above the static Nash value. The figure regarding $\boldsymbol{\rho}^{*(2)}$ looks qualitatively similar. For that equilibrium, $W(\boldsymbol{\rho}^{*(2)}, \boldsymbol{\rho}^{*(2)}, s_0 = B)/\pi_N \approx 1.02$. Figure (b) plots the transition probability functions over aggregate price, where one can verify the steepness around aggregate price P_N and higher curvature around P_A, P_B when $\boldsymbol{\rho}^{*(1)}$ is played.

the online appendix, I provide simulation results that verify that ACQ and ACG learners globally converge to $\boldsymbol{\rho}^{*(1)}$ or $\boldsymbol{\rho}^{*(2)}$, where either outcome is observed in $\approx 50\%$ of simulation runs. I further verify that this outcome continues to emerge in simulations under various cases of asymmetry discussed in the next section, including asymmetric critics and learning rates.[17]

IV. **Implications for Policy**

IV.A. **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 $\boldsymbol{\rho}_N$ 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 successful 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, desensitizing the density g of aggregate shock $\tilde{A}$ may also introduce the possibility of collusive outcomes being

[17]The example was generated using the python programming language and packages provided in Van Rossum, Drake, et al. 1995, Harris et al. 2020,Virtanen et al. 2020, and Hunter 2007.

learned. Nevertheless, less sensitive g can be interpreted as less accurate monitoring technology available to the agents; this in turn can only lead to (weakly) less concerning collusive possibilities, as shown below.

Formally, suppose initially that algorithms follow some state variable the transitions of which are pinned down through realizations of commonly observable $\tilde{A}$: there is a function $f : O_S \times \Omega \to O_S$ mapping current state and realization of $\tilde{A}$ to next state. I will refer to such state variables colloquially as 'state variable given $\tilde{A}$'. 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 $\tilde{C}(\beta) \sim (1 - \beta) \circ \tilde{A} + \beta \circ \tilde{U}$, i.e. $\tilde{C}$ is distributed as a convex combination of $\tilde{A}$ and $\tilde{U}$. The support of $\tilde{C}(\beta)$ is still compact and positive, just as for $\tilde{A}$, so that all previous results go through for state variables S given commonly observed $\tilde{C}(\beta)$. For $\beta = 0$, the unrestricted case is recovered, and $\beta = 1$ leads to $E_S = \{\boldsymbol{\rho}_N\}$ as the monitoring technology carries no information in such a case. For a state variable S given $\tilde{C}(\beta)$, and given profile $\boldsymbol{\rho}$, define $\Delta_p(\boldsymbol{\rho}) = \max_{s \in O_S} |p_N - (\rho^1(s) + \rho^2(s))/2|$ as the largest deviation of the mean price from the static Nash price.

Proposition 1. *Take any stage game and $\tilde{A}$ satisfying Assumption 4. For all state variables given $\tilde{C}(\beta)$*

(1) there exists $\bar{\beta}_1 \in (0, 1)$ s.t. $\boldsymbol{\rho}_N$ will be learned with positive probability whenever $\beta \geq \bar{\beta}_1$.

(2) For all $\eta > 0$, there exists $\bar{\beta}_2 \in (0, 1)$ s.t. $\Delta_p(\boldsymbol{\rho}^*) < \eta$ for all $\boldsymbol{\rho}^* \in E_S$ when $\beta \geq \bar{\beta}_2$.

Point (1) follows directly from the insights of Theorem 3. According to point (2), when transitions are insensitive enough, for all equilibria that may be learned, equilibrium prices must be close to the static Nash price. Hence, a garbling of the monitoring sensitivity has unambiguous welfare effects for consumers.

In a more general model where a state variable carries information not only about past choices of the agents but also about future market conditions, a restriction on the observability of $\tilde{A}$ also hampers the firms' ability to respond efficiently legitimate variations in market parameters. Forcing firms to set prices based on a noisier signal may introduce welfare loss

to producers. Such a more general model can be studied using this framework and the results qualitatively extend, as was shown in a previous version of this paper.

In the more general case, a regulator using this instrument 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 welfare analysis that balances these two competing effects, which is beyond the scope of this paper, but interesting to consider in future work.

IV.B. **Upstream Competition**

Here I consider an extension where the two agents may use algorithms that differ via three possible parameters: their observed state variable, their stepsizes α_t^i, or their critic function Ψ^i. This analysis can be interpreted as one of competition among algorithmic software providers selling learning algorithms to firms; in such a case one may more likely see asymmetry among algorithmic setups.

Asymmetric States. State variables can be thought of as summarizing the data that algorithms' policies can act upon. In a market competition, the data that one firm believes is important enough to use is likely highly correlated to the dataset other firms believe to be important. The case where firms observe a common state is the extreme limit of this scenario. Weakening this, one may believe that data is highly correlated, but differentiated only through *precision*. This is the setting considered here.

Suppose firm 1 purchased a coarser dataset than firm 2, and then implemented their algorithm. Hence, agent 2's policy has access to finer information than agent 1's policy. Label the state variables S_{coarse}, S_{fine} respectively.

Definition 5. *[Asymmetric States] Say that 2's state space is finer than 1's if $O_{S_{coarse}}$ is constructed as a coarsening of $O_{S_{fine}}$. In other words, $z_k \in O_{S_{coarse}}$ are constructed as partition elements $s(z_k) \subset O_{S_{fine}}$ that 2 can distinguish but 1 cannot, and transitions of S_{coarse} are constructed accordingly:*
Take a profile $\boldsymbol{\rho}$ in the asymmetric setting, where $\boldsymbol{\rho}^2$ is measurable with respect to states in $O_{S_{fine}}$, while $\boldsymbol{\rho}^1$ is only measurable with respect to $O_{S_{coarse}}$. Let $\mu(\boldsymbol{\rho})$ be the associated

stationary distribution over S_{fine}, and define $P_z = \sum_{s \in s(z)} \mu(s, \boldsymbol{\rho})$ as the stationary mass of states in $s(z)$. Let $T^{fine}_{ss'}(\boldsymbol{\rho})$ be the transition probability among states $s, s' \in O_{S_{fine}}$.

(1) For any $z, z' \in O_{S_{coarse}}$, define

$$T^{coarse}_{zz'}(\boldsymbol{\rho}) = \sum_{s \in s(z), s' \in s(z')} \frac{\mu(s, \boldsymbol{\rho})}{P_z} T^{fine}_{ss'}(\boldsymbol{\rho}),$$

the expected transition from $s(z)$ to $s(z')$ given the stationary distribution.

*(2) Let $v_k \in \{0,1\}^{K_2}$ be the vector indicating the set $s(z_k) \in O_{S_{fine}}$ given state $z_k \in O_{S_{coarse}}$. Define the matrix $\boldsymbol{Z} \in \{0,1\}^{K_2 \times K_1}$ consisting of vectors $\{v_k\}_{1 \leq k \leq K_1}$. For any $\eta \geq 0$, S_{fine} is η -**lumpable** with respect to S_{coarse} at $\boldsymbol{\rho}$ if*

$$\left\| \boldsymbol{T}^{fine}(\boldsymbol{\rho}) \boldsymbol{Z} - \boldsymbol{Z} \boldsymbol{T}^{coarse}(\boldsymbol{\rho}) \right\|_\infty < \eta.$$

When S_{fine} is 0-lumpable, this definition coincides with strongly lumpable transitions as coined in Kemeny, Snell, et al. 1969, who refer to finite Markov chains that can be 'lumped' into coarser partitions while retaining their Markov properties for any initial distribution. In essence, lumpability implies that within partition elements $s(z_k)$, states of S_{fine} are distributed conditionally uniformly, and hence carry no information beyond that which is contained in z_k. Next I show that this homogeneity restriction is sufficient to make claims about the learnability of $\boldsymbol{\rho}_N$ under various allocations of state variables.

In what follows I consider the three possible 'regimes', allocations of state variables, that Definition 5 generates: one where both agents have access to S_{coarse} only, one where 2 has access to S_{fine} but 1 doesn't, and on where both agents have access to S_{fine}. Note that, with some abuse of notation, the repetition of static Nash $\boldsymbol{\rho}_N$ is an equilibrium in all possible regimes.

Suppose a regulator where to try and achieve that $\boldsymbol{\rho}_N$ be learned by algorithms by breaking up a monopoly on the upstream software market. Suppose that currently, under the monopoly, $\boldsymbol{\rho}_N$ is not learned. The following proposition shows that (1) if the resulting competition among software providers leads to an asymmetric regime where one agent carries a more informative state variable than was true under the symmetric regime (and the other stays with the previous state), then $\boldsymbol{\rho}_N$ will also not be learned when upstream competition

is imposed. When the asymmetric regime is such that one agent's new state variable carries less information than before (2) (and the other stays with the previous state), all we can say is the opposite implication; when it was already possible to learn $\boldsymbol{\rho}_N$, it will remain possible to learn this also under upstream competition.

Proposition 2. *For any admissible critic profile Ψ, there is $\eta > 0$ small enough such that when S_{fine} is η-lumpable in a neighborhood of $\boldsymbol{\rho}_N$,*

(1) if $\boldsymbol{\rho}_N$ can be learned under asymmetric states, then it can be learned when both agents only have access to S_{coarse}.

(2) if $\boldsymbol{\rho}_N$ can be learned when both agents have access to S_{fine}, then it can be learned under asymmetric states.

Hence, conclusions on the landscape of possible equilibria learned change substantially depending on how coarse or fine state variables are that may be purchased after upstream competition has been imposed.

Other equilibria, with abuse of notation, may exist in multiple of these regimes.[18] Lumpability is not enough to make claims similar to the ones above in this case, where continuation values differ across states. In fact, as observed in the discussion of Theorem 4, stability properties of equilibria depend on the magnitude of optimal adjustments, which is determined by continuation values weighted by the curvature of transition probabilities. Even when states are lumpable, the distribution of this curvature across fine states with differing continuation values generally varies (even within a single partition element $s(z_k)$), which can break down the implications discovered in Proposition 2. The relationship of stability properties of equilibria different from $\boldsymbol{\rho}_N$ across regimes would be beyond the scope of this paper.

Asymmetric Stepsizes and Critics. Turning to the possibility of asymmetric stepsize schedules, recall that I assume throughout that stepsizes of all agents lie within an order of magnitude from each other (see Assumption 7 in the appendix). It turns out that, in the case of symmetric equilibria, all insights stated for equal stepsizes carry over to the more

[18]This has been studied in the literature on private monitoring, see e.g. Lehrer 1991, Kandori 1992,and Cherry and L. Smith 2010.

general case. Suppose that $\frac{\alpha_t^1}{\alpha_t^2} \to \bar{\alpha}$ as $t \to \infty$. Symmetry would be implied by $\bar{\alpha} = 1$. Finally, also define the system of asymmetric in critic functions

$$\dot{\boldsymbol{\rho}} = \Psi_{asym} \equiv \begin{bmatrix} B_S^1(\boldsymbol{\rho}_2) - \boldsymbol{\rho}_1 \\ \nabla W_S^2(\boldsymbol{\rho}_2, \boldsymbol{\rho}_1) \end{bmatrix},$$

where ∇W_S^2 refers to player 2's derivative vector of long run payoffs with respect to $\boldsymbol{\rho}_2$, given a symmetric state regime.

Proposition 3. *Take any state variable S, and any symmetric $\boldsymbol{\rho}^* \in E_S$.*

(1) When both agents are ACQ learners, $\boldsymbol{\rho}^*$ will be approached with positive probability given $\bar{\alpha} \in (0, 1)$ if this is true given $\bar{\alpha} = 1$.

(2) There is $0 < C_G < \infty$ such that when $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$ and $\bar{R} < C_G$, (1) holds also when both agents are ACG learners.

(3) There is $0 < C_A < \infty$ such that when $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$ and $\bar{R} < C_A$, $\boldsymbol{\rho}^*$ will be approached with positive probability given Ψ_{asym} if $\boldsymbol{\rho}^*$ will be approached with positive probability given $\bar{\alpha} = 1$ under ACQ learners.

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

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 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.

V. Conclusion

This paper shows how, for a large family of popular RL algorithms, the learnability of equilibria can be studied tractably. I uncover the dependence of the attractability of a given equilibrium of the repeated game on the interplay between state variables observed by algorithms, i.e. their implied monitoring technology, and market conditions. 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 and competition among algorithm designers, 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 applications. The method enables researchers 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 2006, Borkar 2009, Benaïm and Faure 2012. A thorough argument generalizing results further to the non-vanishing bias case can be found in the online appendix[19].

For all results to follow, state variables will be fixed. The stacking over i of the algorithm (3) can be written as

$$\boldsymbol{\rho}_{t+1} = \boldsymbol{\rho}_t + \alpha_t \boldsymbol{A} \left[\Psi(\boldsymbol{\rho}_t) + \boldsymbol{\delta}_t + \boldsymbol{M}_{t+1} \right], \tag{4}$$

where $\alpha_t = \alpha_t^1$ is agent 1's stepsize schedule, $\boldsymbol{A}$ is a diagonal matrix with i'th diagonal entry equal to $\lim_{t\to\infty} \frac{\alpha_t^i}{\alpha_t^1} \in (0, \infty)$. This then leads to our defined $\Psi_A = \Psi_A$. $\boldsymbol{\delta}_t$ is a vector of biases $\boldsymbol{\delta}_t^i$ stacked over i, and $\boldsymbol{M}_{t+1}$ a vector of errors stacked over as outlined in Section II.

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

Assumption 7. *Let $\mathcal{F}_t$ be the σ-field generated by $\{\boldsymbol{\rho}_t, \boldsymbol{\delta}_t, M_t, \boldsymbol{\rho}_{t-1}, \boldsymbol{\delta}_{t-1}, \boldsymbol{M}_{t-1} \ldots, \boldsymbol{\rho}_0, \boldsymbol{\delta}_0, \boldsymbol{M}_0\}$, i.e. all the information available to the updating rule at a given period t.*

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

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

(3) Ψ is admissible.

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

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

(5) There exists a continuous function

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

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

$$\mathbb{E}[\boldsymbol{M}_{t+1}\boldsymbol{M}'_{t+1} \,|\mathcal{F}_t] = \Omega(\boldsymbol{\rho}_t),$$

for all $\boldsymbol{\rho}_t \in \overline{Y}$.

(6)

$$\mathbb{E}\big[\|\boldsymbol{\delta}_t\| \mid \mathcal{F}_t\big] = o(b_t),$$

where $b_t \to 0$ satisfies $\max_i \lim_{t\to\infty} \frac{\alpha_t^i}{b_t} = 0$, α_t^i being i's stepsize.

(7)

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

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 $\dot{\boldsymbol{\rho}} = \Psi(\boldsymbol{\rho})$. 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 Ψ are well-behaved. It is a common assumption in stochastic approximation theory. Point (5) ensures that some variance in error terms remains for all t; this is satisfied e.g. if the estimation of Ψ 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. Point (7) is a further regularity condition on the bias term.

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

$$\dot{\boldsymbol{\rho}} = \Psi_A(\boldsymbol{\rho}(t)). \tag{5}$$

Then, to conclude convergence with positive probability to an attractor, one can apply Faure and Roth 2010, Theorem 2.15. For Theorem 2, see 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 2. *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

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

as the largest real part of eigenvalues of $\boldsymbol{A}$. For some policy profile $\boldsymbol{\rho}^* \in \overline{Y}$, Define $\boldsymbol{J}^i(\boldsymbol{\rho}^*)$ as the Jacobian of $B_S^i(\boldsymbol{\rho}^*)$, which is the matrix of best response derivatives of a given player. For symmetric $\boldsymbol{\rho}^*$, I drop the i superscript to save notation.

Lemma 2. *Suppose $\boldsymbol{\rho}^* \in E_S$ is a differential, symmetric Nash equilibrium. Then $\boldsymbol{\rho}^*$ is asymptotically stable when all agents are ACQ learners if $\kappa(\boldsymbol{J}(\boldsymbol{\rho}^*)) < 1$, and unstable if $\kappa(\boldsymbol{J}(\rho^*)) > 1$.*

Proof. Using Remark 2 leads to

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

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

Hence, it is enough to be concerned with the eigenvalues of individual jacobians of each player, when considering symmetric Nash equilibria and all agents use ACQ learners.

For the following arguments it will be useful to derive some notation concerning any state variable with $O_S = \{s_1, ..., s_K\}$. Fixing the profile $(\boldsymbol{\rho}, \boldsymbol{\gamma}) \in \overline{Y}$, it will be 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

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

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, 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 $1 \leq j \leq 2K$ using a subscript j. Then :

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

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

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

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

Now for a general result about the best response derivative matrix. For any matrix $\boldsymbol{A}$, write $\boldsymbol{A}_{i,:}$ as the i'th row, and $\boldsymbol{A}_{:,j}$ as the j'th column. Given profile $\boldsymbol{\rho}$, write $\pi^i = \pi(\rho(s_i))$ to save notation. Since by assumption, states are irreducible , fix an arbitrary initial state s_1 when computing first order conditions to pin down best responses. Write the Hessian as

$\boldsymbol{H} = diag\,(W_{ii})$. $\boldsymbol{H}$ is diagonal by the derivation of W_i: write

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

Then taking another derivative with respect to a variable $j \leq K$:

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

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 $\tilde{\boldsymbol{W}}_i = \tilde{\boldsymbol{W}}_j = 0$ for $i, j \leq K$, indeed $W_{ij} = 0$ whenever $i \neq j$ and $i, j \leq K$. So $\boldsymbol{H}$ 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 $\boldsymbol{\rho}(s_i)$ and an opponent's strategy $\gamma(s_{j-K})$. Then define, using the implicit function theorem, the best response derivative matrix as

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

Lemma 3. *At any interior, symmetric differential Nash equilibrium $\boldsymbol{\rho} \in E_S$, the best response derivative matrix $\boldsymbol{J}$ can be written as $\boldsymbol{J} = \boldsymbol{D} + \boldsymbol{B}$, where $\boldsymbol{D}$ is a diagonal matrix with entries*

$$D_{ii} = -1 + \frac{\pi^i_{11} - \pi^i_{12}}{\frac{\delta}{1-\delta} \frac{\partial^2 \boldsymbol{T}}{(\partial \boldsymbol{\rho}_i)^2}_{i,:} \tilde{\boldsymbol{W}} + \pi^i_{11}},$$

and $\boldsymbol{B}$ has entries

$$B_{ij} = \frac{\delta \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}_{i,:} [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1}_{:,j} \left(\pi^j_1 - \pi^j_2 \right)}{\frac{\delta}{1-\delta} \frac{\partial^2 \boldsymbol{T}}{(\partial \boldsymbol{\rho}_i)^2}_{i,:} \tilde{\boldsymbol{W}} + \pi^i_{11}}.$$

Proof. First, from (6) we have that

$$\begin{aligned} W_{ii} &= [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1}_{1,:} \left(\delta \frac{\partial^2 \boldsymbol{T}}{(\partial \boldsymbol{\rho}_i)^2} \tilde{\boldsymbol{W}} + (1-\delta) \pi^i_{11} e_i \right) \\ &= (1-\delta) [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1}_{1,:} \left(\frac{\delta}{1-\delta} \frac{\partial^2 \boldsymbol{T}}{(\partial \boldsymbol{\rho}_i)^2} \tilde{\boldsymbol{W}} + \pi^i_{11} e_i \right), \end{aligned}$$

Similarly, when $i = j - K$,

$$W_{ij} = W_{ii} + (1-\delta)\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,:}\left(\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1} e_j\left(\pi_2^j - \pi_1^j\right) + \left(\pi_{12}^j - \pi_{11}^j\right) e_j\right)$$

and otherwise

$$W_{ij} = (1-\delta)\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,:}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1} e_j\left(\pi_2^j - \pi_1^j\right).$$

As a result, write the best response derivative matrix $\boldsymbol{J}$ as $\boldsymbol{J} = \boldsymbol{D} + \boldsymbol{B}$, where $\boldsymbol{D}$ is a diagonal matrix with entries

$$D_{ii} = -1 + \frac{\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,:}\left(\left(\pi_{11}^i - \pi_{12}^i\right) e_i\right)}{\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,:}\left(\frac{\delta}{1-\delta}\frac{\partial^2 \boldsymbol{T}}{(\partial\boldsymbol{\rho}_i)^2}\tilde{\boldsymbol{W}} + \pi_{11}^i e_i\right)}.$$

To simplify further:

$$\begin{aligned} D_{ii} &= -1 + \frac{\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,i}\left(\left(\pi_{11}^i - \pi_{12}^i\right)\right)}{\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,i}\left(\frac{\delta}{1-\delta}\frac{\partial^2 \boldsymbol{T}}{(\partial\boldsymbol{\rho}_i)^2}_{i,:}\tilde{\boldsymbol{W}} + \pi_{11}^i\right)} \\ &= -1 + \frac{\left(\pi_{11}^i - \pi_{12}^i\right)}{\frac{\delta}{1-\delta}\frac{\partial^2 \boldsymbol{T}}{(\partial\boldsymbol{\rho}_i)^2}_{i,:}\tilde{\boldsymbol{W}} + \pi_{11}^i}, \end{aligned}$$

as $\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}$ is nonzero only in the i'th row, and similarly for the 2nd derivative. Analogously, $\boldsymbol{B}$ has entries

$$\begin{aligned} B_{ij} &= \frac{\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,:}\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1} e_j\left(\pi_1^j - \pi_2^j\right)}{\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{1,:}\left(\frac{\delta}{1-\delta}\frac{\partial^2 \boldsymbol{T}}{(\partial\boldsymbol{\rho}_i)^2}\tilde{\boldsymbol{W}} + \pi_{11}^i e_i\right)} \\ &= \frac{\delta\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\rho}_i}_{i,:}\left[\boldsymbol{I}_K - \delta\boldsymbol{T}\right]^{-1}_{:,j}\left(\pi_1^j - \pi_2^j\right)}{\frac{\delta}{1-\delta}\frac{\partial^2 \boldsymbol{T}}{(\partial\boldsymbol{\rho}_i)^2}_{i,:}\tilde{\boldsymbol{W}} + \pi_{11}^i}. \end{aligned}$$

□

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

Given Assumption 4, 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 $p_{-i} \geq 0$. This implies that the Nash equilibrium p_N is unique, and symmetry of the payoffs implies that this equilibrium is symmetric. Note that $\Psi(\boldsymbol{\rho})$ collapses to best response dynamics when ACQ learners are used. 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 (when ACG learners are used), it is straightforward to show that p_N must be asymptotically stable: let $\boldsymbol{M}_B, \boldsymbol{M}_G$ be the linearization of $\Psi_{S_0,B}$ and $\Psi_{S_0,G}$, respectively at p_N. Then by symmetry of payoffs, 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. ■

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

To determine the stability of $\boldsymbol{\rho}_N$, one needs to compute the eigenvalues of the linearized best response dynamics at $\boldsymbol{\rho}_N$. Define $\boldsymbol{M}(\boldsymbol{\rho})$ as the Jacobian of the system $\dot{\boldsymbol{\rho}} = \Psi(\boldsymbol{\rho})$ at $\boldsymbol{\rho}$, when ACQ learners are used. Since $\boldsymbol{M}(\boldsymbol{\rho})$ is evaluated at $\boldsymbol{\rho}_N$, multiple simplifications are possible, and when clear from context, I simplify notation to write $\boldsymbol{M} = \boldsymbol{M}(\boldsymbol{\rho}_N)$.

Firstly, long term payoffs $\tilde{\boldsymbol{W}} = \boldsymbol{\pi}^N$, a $K-$vector equal to π_N, the stage-game Nash payoff in all elements, since $\boldsymbol{\rho}_N$ 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 $\boldsymbol{\rho}_N$, I can simplify the elements of $\boldsymbol{H}, \boldsymbol{R}$ to

$$W_{ii} = [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1}_{1,:} (1-\delta) \frac{\partial^2 \boldsymbol{U}}{(\partial \boldsymbol{\rho}_i)^2},$$

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

To save notation, write $Z = [\boldsymbol{I}_K - \delta \boldsymbol{T}]^{-1}$. Then using Lemma 3, I get

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

Here T' is the transition-derivative matrix, where each row i corresponds to the derivative of row i of T with respect to $\boldsymbol{\rho}_i$, all evaluated at $\boldsymbol{\rho}_N$.

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

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

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 4). $\boldsymbol{M}$'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 4, it is sufficient to bound $\frac{\pi_2^N}{\pi_{11}^N}\mu$.

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

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

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. $\mathcal{R}e(\mu) \neq 0$. Without loss, let $\mathcal{R}e(\mu) > 0$. 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 $\boldsymbol{\rho}_N$: $\boldsymbol{D}_{ss'}(\boldsymbol{\rho}) = \boldsymbol{T}_{ss'} + (\boldsymbol{\rho}_s - p_N)c\boldsymbol{T}'_{ss'}$, for all $s, s' \in O_S$. Hence, one can find a transition probability matrix with $|\boldsymbol{T}'|$ small enough such that all eigenvalues μ will ensure

$|\lambda| < 1$. Conversely, 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 $\pi_{12} \geq 0 > \pi_{11}$ under Assumption 4. Thus, $|\pi_{12}^N + \pi_{11}^N| < \pi_{12}^N - \pi_{11}^N$, so that the positive upper bound of the interval in (8) is closer to zero than the negative lower 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 satisfy the the upper bound is sufficient. On the other hand, for instability, at least one λ must satisfy $|\lambda| > 1$. Finding one positive μ large enough is then enough, which allows checking 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 one 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, I have that $-\xi_o(p_N) = \frac{1}{L_N}$, where $L_N = \frac{p_N-c}{p_N}$ is the Lerner index at p_N. Define

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

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 4, I 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.D. **Proof of Theorem 4**

I fix the critic profile Ψ to be generated from ACQ learners. Consider the solution to the eigenvalue problem for binary $\boldsymbol{J}$:

$$\lambda_{1,2} = \frac{Tr(\boldsymbol{J}) \pm \sqrt{Tr(\boldsymbol{J})^2 - 4det(\boldsymbol{J})}}{2}.$$

Let $O_S = \{A, B\}$, and suppose first that $T''_{ss'}(\boldsymbol{\rho}(s)) \neq 0$. Define for $s, s' \in O_S$ with $s \neq s'$

$$\Phi_s = \frac{T'_{ss'}(\boldsymbol{\rho}(s))}{T''_{ss'}(\boldsymbol{\rho}(s))}.$$

Then

$$\begin{aligned}\boldsymbol{J} &= \begin{pmatrix} \Phi_A d_{A,1} + d_{A,2} & \Phi_A b_A \\ \Phi_B b_B & \Phi_A d_{B,1} + d_{B,2} \end{pmatrix} \\ &= \begin{pmatrix} \Phi_A & 0 \\ 0 & \Phi_B \end{pmatrix} \begin{pmatrix} d_{A,1} & b_A \\ b_B & d_{B,1} \end{pmatrix} + \begin{pmatrix} d_{A,2} & 0 \\ 0 & d_{B,2} \end{pmatrix},\end{aligned}$$

Next I ease notation to write $T'_{ss'}, T''_{ss'}$ as first and 2nd derivative of transition probabilities evaluated at the equilibrium policy profile, and similarly replace the notation from Lemma 3, π_1^i with π_1^s for the first derivative of stage game payoff evaluated at state s. Then, $\Phi_A = \frac{T'_{AB}}{T''_{AB}}$,

$\Phi_B = \frac{T'_{BA}}{T''_{BA}}$, and from Lemma 3 I can derive that

$$d_{A,1} = \frac{f\delta(\pi_2^A - \pi_1^A)}{f\delta(\pi^B - \pi^A) + \frac{\pi_{11}^A}{T''_{AB}}}; \qquad d_{A,2} = \frac{\pi_{11}^A - \pi_{12}^A}{f\delta T''_{AB}(\pi^B - \pi^A) + \pi_{11}^A} - 1;$$

$$d_{B,1} = \frac{f\delta(\pi_2^B - \pi_1^B)}{f\delta(\pi^A - \pi^B) + \frac{\pi_{11}^B}{T''_{BA}}}; \qquad d_{B,2} = \frac{\pi_{11}^B - \pi_{12}^B}{f\delta T''_{BA}(\pi^A - \pi^B) + \pi_{11}^B} - 1,$$

$$b_A = \frac{f\delta(\pi_1^B - \pi_2^B)}{f\delta(\pi^B - \pi^A) + \frac{\pi_{11}^A}{T''_{AB}}}; \qquad b_B = \frac{f\delta(\pi_1^A - \pi_2^A)}{f\delta(\pi^A - \pi^B) + \frac{\pi_{11}^B}{T''_{BA}}},$$

where $f = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}_{1,1} - [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}_{2,1} = [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}_{2,2} - [\boldsymbol{I}_K - \delta\boldsymbol{T}]^{-1}_{1,2} > 0$ is the difference in long-run discounted occupancy measure of state 1, i.e. A, when initialized at state 1 and state 2.

For point (1), I consider $|\Phi_A|, |\Phi_B|$ vanishing to zero. When these factors vanish, the eigenvalues of $\boldsymbol{J}$ converge to (the limit of) $\{d_{A,2}, d_{B,2}\}$.

As Φ_s vanishes, either $T'_{ss'}$ vanishes, $|T''_{ss'}|$ grows, or both. In the former case, as shown in the proof of Proposition 4, any equilibrium must approach $\boldsymbol{\rho}_N$, so that $\pi^B \approx \pi^A \approx \pi^N$, and $d_{s,2}$ approaches $-\frac{\pi_{12}^N}{\pi_{11}^N}$. $-\frac{\pi_{12}^N}{\pi_{11}^N} \in (-1, 1)$ by Assumption 4 (this is the derivative of the stage-game best response at static Nash). When $|T''_{ss'}|$ grows, $d_{s,2}$ shrinks from above to -1. By Assumption 4, I have that $\pi_{11}^s - \pi_{12}^s < 0$ for all $s \in O_S$. As $\boldsymbol{\rho}$ is a differential Nash equilibrium by assumption of the Theorem, it must be that the denominators of $d_{A,2}, d_{B,2}$ are strictly negative. Hence, $d_{s,2} > -1$ for $s \in O_S$, and again, $d_{s,2} \in (-1, 1)$ for $s \in O_S$ when $|T''_{ss'}|$ large enough.

As for point (2), note that $d_{A,1}d_{B,1} = b_A b_B$. Hence, the eigenvalue formula simplifies to

$$\begin{aligned}\lambda_{1,2} &= \frac{Tr(\boldsymbol{J}) \pm \sqrt{(\Phi_A d_{A,1} + d_{A,2} - \Phi_B d_{B,1} - d_{B,2})^2 + 4\Phi_A\Phi_B d_{A,1}d_{B,1}}}{2} \\ &= \frac{Tr(\boldsymbol{J})}{2} \pm \frac{|\Phi_A\Phi_B|}{2}\sqrt{\left(\frac{d_{A,1}}{\Phi_B} - \frac{d_{B,1}}{\Phi_A} + \frac{1}{\Phi_A\Phi_B}(d_{A,2} - d_{B,2})\right)^2 + 4\frac{d_{A,1}d_{B,1}}{\Phi_A\Phi_B}}\end{aligned}$$

The conclusion follows: as $|\Phi_A|, |\Phi_B|$ grow, the dominant term under the square root has a factor of $d_{A,1}d_{B,1} \neq 0$, by Assumption 4. For $|\Phi_A\Phi_B|$ large enough, at least one eigenvalue will be larger than 1 in absolute value, implying instability of $\boldsymbol{\rho}$.

If on the other hand $T''_{ss'}(\boldsymbol{\rho}(s)) = 0$ for some $s \in O_S$, replace Φ_s with $T'_{ss'}(\boldsymbol{\rho}(s))$ and adjust the definition of the terms in $\boldsymbol{J}$ accordingly. It is clear that the result still holds true in that case as well. ■

B.E. **Proof of Proposition 1**

Point (1) follows directly from Theorem 3. For point (2), I prove a more general result:

Proposition 4. *For all $K > 1$, state variables S with $|O_S| = K$, and $\eta > 0$, there exists $\varepsilon > 0$ s.t. $\Delta_p(\boldsymbol{\rho}^*) < \eta$ for all $\boldsymbol{\rho}^* \in E_S$ when $\overline{R} < \varepsilon$.*

Proof. Fix any state variable with state size $K > 1$. Consider first an interior equilibrium $\boldsymbol{\rho}$ different from $\boldsymbol{\rho}_N$, if it exists. The following is written as fixing agent 1, and dropping their identifier (the equilibrium may be asymmetric, but the derivation is analogous for the other agent). First order conditions must be satisfied, i.e. for every state s_i:

$$\frac{\delta}{1-\delta}\left(\frac{\partial}{\partial \rho_i}\boldsymbol{T}\right)_{i,:}\tilde{\boldsymbol{W}} + \pi_1^i = 0,$$

following the insights and notation developed in Lemma 3. Pushing $\overline{R}$ towards zero must move the first term of the summation towards zero, as all other terms are bounded. Hence for both agents, in all states i, π_1^i must get arbitrarily close to zero. By continuity of π_1, and uniqueness of p_N under Assumption 4, π_1^i close enough to zero, i.e. $\overline{R}$ small enough, this can only happen for $\boldsymbol{\rho}(s_i)$ close enough to p_N, concluding the claim.

Now consider an equilibrium, if it exists, where for some s_i, $\boldsymbol{\rho}(s_i) = 0$, so that

$$\frac{\delta}{1-\delta}\left(\frac{\partial}{\partial \rho_i}\boldsymbol{T}\right)_{i,:}\tilde{\boldsymbol{W}} + \pi_1^i \leq 0.$$

For $\overline{R}$ small enough, it must be true that $\pi_1^i \leq 0$ at that equilibrium, which violates Assumption 4 (2) - when $p_1 = 0$, 1 must find it strictly profitable to increase price no matter 2's choice. □

B.F. **Proof of Proposition 2**

Consider the Jacobian matrix of system (5), $\boldsymbol{M}_{coarse} \in \mathbb{R}^{2K_1 \times 2K_1}$ at $\boldsymbol{\rho}_N$ when both agents use S_{coarse}. Similarly let $\boldsymbol{M}_{asym} \in \mathbb{R}^{K_1+K_2 \times K_1+K_2}$ be the Jacobian in the asymmetric states system, when agents use the same profile of critics Ψ. I argue first that the eigenvalues $\Lambda(\boldsymbol{M}_{coarse}) \subseteq \Lambda(\boldsymbol{M}_{asym})$. Consider the upper right block matrix $\boldsymbol{J}^1_{asym} \in \mathbb{R}^{K_1 \times K_2}$ submatrix of $\boldsymbol{M}_{asym}$ corresponding to agent 1's derivatives with respect to agent 2's policy at $\boldsymbol{\rho}_N$, and similarly define $\boldsymbol{J}^2_{asym}$ as 2's derivatives with respect to agent 1's policy in the asymmetric regime. Then define $\boldsymbol{J}^1_{coarse} = \boldsymbol{J}^2_{coarse} \in \mathbb{R}^{K_1 \times K_1}$ for $\boldsymbol{M}_{coarse}$. Take any $z_k \in O_{S_{coarse}}$ and associated $s(z_k) = \{x_{k_1}, \dots, x_{k_m}\} \subset O_{S_{fine}}$. Then for any $1 \leq k' \leq K_1$ I have $\boldsymbol{J}^1_{coarse,k',k} = \sum_{i=1}^m \boldsymbol{J}^1_{asym,k',k_i}$, as the same action p_N is played in all states no matter if symmetric or asymmetric setting. The derivative represented by $\boldsymbol{J}^1_{coarse}$ in the symmetric setting can be represented as a sum of derivatives in $\boldsymbol{J}^1_{asym}$. Hence, there is a matrix $\boldsymbol{Z} \in \{0,1\}^{K_2 \times K_1}$ so that $\boldsymbol{J}^1_{coarse} = \boldsymbol{J}^1_{asym}\boldsymbol{Z}$.

To move to the relationship between $\boldsymbol{M}_{coarse}, \boldsymbol{M}_{asym}$, represent a direction of deviation by agent 2 in the symmetric system by vector $\boldsymbol{v}_2 \in \mathbb{R}^{K_1}$. This deviation can be imitated by agent 2 in the asymmetric system simply by fixing the direction across the finer states this agent has access to; this amounts to constructing the vector $\tilde{\boldsymbol{v}_2} \in \mathbb{R}^{K_2}$ that satisfies $\tilde{\boldsymbol{v}_2} = \boldsymbol{Z}\boldsymbol{v}_2$. Then, for any deviation $\boldsymbol{v}_1 \in \mathbb{R}^{K_1}$ of agent 1, I can relate overall deviations in the symmetric system $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ to deviations in the asymmetric system $(\boldsymbol{v}_1, \tilde{\boldsymbol{v}_2})$ via operator $\boldsymbol{G}$:

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{I}_{K_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z} \end{pmatrix}.$$

Then

$$\begin{pmatrix} \boldsymbol{v}_1 \\ \tilde{\boldsymbol{v}_2} \end{pmatrix} = \boldsymbol{G} \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{pmatrix},$$

and

$$\begin{pmatrix} \boldsymbol{v}_1 \\ \tilde{\boldsymbol{v}_2} \end{pmatrix} \boldsymbol{G} = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{pmatrix}.$$

Extending this logic to the Jacobians of the whole system, I have

$$\boldsymbol{M}_{asym}\boldsymbol{G} = \begin{pmatrix} -\boldsymbol{I}_{K_1} & \boldsymbol{J}^1_{coarse} \\ \boldsymbol{J}^2_{asym} & -\boldsymbol{Z} \end{pmatrix},$$

$$\boldsymbol{G}\boldsymbol{M}_{coarse} = \begin{pmatrix} -\boldsymbol{I}_{K_1} & \boldsymbol{J}^1_{coarse} \\ \boldsymbol{Z}\boldsymbol{J}^1_{coarse} & -\boldsymbol{Z} \end{pmatrix}.$$

In general, $\boldsymbol{J}^2_{asym} \neq \boldsymbol{Z}\boldsymbol{J}^1_{coarse}$. $\boldsymbol{Z}\boldsymbol{J}^1_{coarse}$ essentially 'lifts' best response derivatives of an agent with access to the coarse state to the best response derivatve of an agent with access to the fine state, assuming that such derivatives be *uniform* over the finer states this agent has access to (i.e. $s(z_k)$ for every z_k referring to a coarse state in $O_{S_{coarse}}$). This will introduce the sufficiency of strong lumpability discussed soon.

First, using the derivation of Lemma 3, and the simplifications implied by evaluating at $\boldsymbol{\rho}_N$, I get that in the notation of the Lemma, $\boldsymbol{J}^1_{coarse} = \boldsymbol{D}_{coarse} + \boldsymbol{B}_{coarse}$ where $\boldsymbol{D}_{coarse} = -\frac{\pi^N_{12}}{\pi^N_{11}}\boldsymbol{I}_{K_1}$, and

$$\boldsymbol{B}_{coarse} = -\delta\frac{\pi^N_2}{\pi^N_{11}}\nabla\boldsymbol{T}_{coarse}\left[\boldsymbol{I}_{K_1} - \delta\boldsymbol{T}_{coarse}\right]^{-1},$$

where $\nabla\boldsymbol{T}$ is the matrix where each row i is represented by the vector $\frac{\partial}{\partial\boldsymbol{\rho}(s_i)}\boldsymbol{T}_{i,:}$. For $\boldsymbol{J}^2_{asym} = \boldsymbol{D}_{asym} + \boldsymbol{B}_{asym}$, note first that a deviation of player one in state z_k to player two is a deviation in all states $s(z_k)$. Hence, $\boldsymbol{D}_{asym}\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{D}_{coarse}$, as the 'diagonal' term reflects best response derivatives with respect to opponent's actions in the matching state. For the terms in $\boldsymbol{B}_{asym}$, the proof of Lemma 3 reveals that best responses of player two will react to a deviation in $z_k \in O_{S_{coarse}}$, for state $s_i \in O_{S_{fine}}$ as if to a total deviation in all states $s(z_k)$. Hence I get

$$\boldsymbol{B}_{asym} = -\delta\frac{\pi^N_2}{\pi^N_{11}}\nabla\boldsymbol{T}_{fine}\left[\boldsymbol{I}_{K_2} - \delta\boldsymbol{T}_{fine}\right]^{-1}\boldsymbol{Z}.$$

I can then consider

$$\begin{aligned}\boldsymbol{J}^2_{asym} - \boldsymbol{Z}\boldsymbol{J}^1_{coarse} = &-\frac{\pi^N_{12}}{\pi^N_{11}}\left(\boldsymbol{Z} - \boldsymbol{Z}\right)\boldsymbol{D}_{coarse} \\ &- \delta\frac{\pi^N_2}{\pi^N_{11}}\left(\nabla\boldsymbol{T}_{fine}\left[\boldsymbol{I}_{K_2} - \delta\boldsymbol{T}_{fine}\right]^{-1}\boldsymbol{Z} - \boldsymbol{Z}\nabla\boldsymbol{T}_{coarse}\left[\boldsymbol{I}_{K_1} - \delta\boldsymbol{T}_{coarse}\right]^{-1}\right).\end{aligned}$$

Here, first suppose that strong lumpability holds between S_{fine} and S_{coarse}, i.e. I have that $\boldsymbol{T}_{fine}\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{T}_{coarse}$. Then $[\boldsymbol{I}_{K_2} - \delta\boldsymbol{T}_{fine}]\,\boldsymbol{Z} = \boldsymbol{Z}\,[\boldsymbol{I}_{K_1} - \delta\boldsymbol{T}_{coarse}]$. Pre- and - post multiplying both sides by inverses leads us to

$$[\boldsymbol{I}_{K_2} - \delta\boldsymbol{T}_{fine}]^{-1}\,\boldsymbol{Z} = \boldsymbol{Z}\,[\boldsymbol{I}_{K_1} - \delta\boldsymbol{T}_{coarse}]^{-1}.$$

Hence,

$$\begin{aligned}
&\nabla\boldsymbol{T}_{fine}\,[\boldsymbol{I}_{K_2} - \delta\boldsymbol{T}_{fine}]^{-1}\,\boldsymbol{Z} - \boldsymbol{Z}\nabla\boldsymbol{T}_{coarse}\,[\boldsymbol{I}_{K_1} - \delta\boldsymbol{T}_{coarse}]^{-1} \\
&= \nabla\boldsymbol{T}_{fine}\,[\boldsymbol{I}_{K_2} - \delta\boldsymbol{T}_{fine}]^{-1}\,\boldsymbol{Z} - \nabla\boldsymbol{T}_{fine}\boldsymbol{Z}\,[\boldsymbol{I}_{K_1} - \delta\boldsymbol{T}_{coarse}]^{-1} \\
&= 0.
\end{aligned}$$

Hence, $\boldsymbol{J}^2_{asym} = \boldsymbol{Z}\boldsymbol{J}^1_{coarse}$, and therefore $\boldsymbol{M}_{asym}\boldsymbol{G} = \boldsymbol{G}\boldsymbol{M}_{coarse}$. Now, if $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{M}_{coarse}$, I have for some scalar λ that $\boldsymbol{M}_{coarse}\boldsymbol{v} = \lambda\boldsymbol{v}$, and hence $\boldsymbol{G}\boldsymbol{M}_{coarse} = \boldsymbol{G}\lambda\boldsymbol{v}$, so that $\boldsymbol{M}_{asym}\boldsymbol{G}\boldsymbol{v} = \lambda\boldsymbol{G}\boldsymbol{v}$, i.e. $\boldsymbol{G}\boldsymbol{v}$ is an eigenvector of $\boldsymbol{M}_{asym}$ with the same eigenvalue. Hence, $\Lambda(\boldsymbol{M}_{coarse}) \subseteq \Lambda(\boldsymbol{M}_{asym})$.

Now, suppose that $\boldsymbol{T}_{fine}\boldsymbol{Z} \neq \boldsymbol{Z}\boldsymbol{T}_{coarse}$. Recall that eigenvalues are continuous in matrix entries. So, for all $\boldsymbol{M}_{coarse}$ where at least one eigenvalue is nonzero and all $\varepsilon > 0$ there is $\eta > 0$ such that when $\|\boldsymbol{T}_{fine}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{T}_{coarse}\|_\infty + \|\nabla\boldsymbol{T}_{fine}\boldsymbol{Z} - \boldsymbol{Z}\nabla\boldsymbol{T}_{coarse}\|_\infty < \eta$, I have

$$\max_{\lambda\in\Lambda(\boldsymbol{M}_{coarse})}\ \min_{\lambda'\in\Lambda(\boldsymbol{M}_{asym})} |\lambda - \lambda'| < \varepsilon,$$

and hence, when $\boldsymbol{M}_{coarse}$ is unstable, so must $\boldsymbol{M}_{asym}$ be.

When considering a regime where both agents have access to S_{fine} first, and comparing that to the asymmetric system, I can do an analogous derivation to conclude point (2): when $\boldsymbol{M}_{asym}$ is unstable, $\boldsymbol{M}_{fine}$ will be unstable also. ■

B.G. **Proof of Proposition 3**

I prove the following statement, restated in more technical terms. Refer to the critic profile approximated by ACQ learners as Ψ_B, and that approximated by ACG learners as Ψ_G.

Proposition 5. *Take any state variable S, and any symmetric $\boldsymbol{\rho}^* \in E_S$.*

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

(2) When $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$, $\boldsymbol{\rho}^*$ is asymptotically stable given $\bar{\alpha} \in (0,1)$ if it is asymptotically stable given $\bar{\alpha} = 1$ also under Ψ_G.

(3) When $\boldsymbol{\rho}^* = \boldsymbol{\rho}_N$, $\boldsymbol{\rho}^*$ is asymptotically stable under Ψ_{asym} if $\boldsymbol{\rho}^*$ it is asymptotically stable given $\bar{\alpha} = 1$ under Ψ_B.

When there are $K \geq 1$ states, for any $\alpha \in (0,1]$ I 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 $\kappa(\boldsymbol{M}_B(1)) < 1$ (recall the definition in Remark 2). Take $\mu = x + iy$. Plugging this into the above, one can derive that $\lambda_{1,2} < 0$ if

$$0 < (1+\alpha)^2(1 - x^2 + y^2) - 4\alpha x^2 y^2.$$

Recall that given $\kappa(M_B(1)) < 1$, I must have $|x| < 1$, and one can check that also the above inequality holds at $\alpha = 1$ and $\alpha = 0$ when $|x| < 1$. Finally, the second derivative in α of the right hand side is strictly positive. I get that the inequality can only flip for $\alpha \in (0,1)$ if the

right hand side takes a minimum within that interval. The extremum of this quadratic in α is

$$\alpha^* = \frac{x^2 y^2}{(1 - x^2 + y^2)} - 1.$$

$\alpha^* \in (0, 1)$ if and only if $x^2 y^2 > (1 - x^2 + y^2)$, which only holds when $|x| > 1$. Hence, the inequality holds for all $\alpha \in [0, 1]$ when $|x| < 1$.

As for point (2), note that according to Lemma 3, the diagonal of the Jacobian $\boldsymbol{J}_B$ of $\Psi_B^i(\boldsymbol{\rho}_N)$ equals $-\pi_{12}^N/\pi_{11}^N \in (-1, 0)$, the static best response slope of the stage game at the stage game Nash equilibrium. Furthermore, matrix $\boldsymbol{B}$ identified in Lemma 3 vanishes as $\bar{R}$ vanishes. Hence, for $\bar{R}$ small enough, $\boldsymbol{J}_B$ becomes a strictly diagonally dominant matrix with negative diagonal. Recall also that $\boldsymbol{D}\boldsymbol{J}_B = \boldsymbol{J}_G$, where $\boldsymbol{D}$ is a positive diagonal matrix (the negative of the Hessian of $W(\boldsymbol{\rho}_N)$). It is well known that strictly diagonally dominant matrices are 'D'-stable - i.e. signs of eigenvalues are preserved under premultiplication by a positive diagonal matrix. Thus, the above conclusion remains under gradient learning for $\boldsymbol{\rho}_N$ when $\bar{R}$ small enough.

Similarly, note that for point (3), at $\boldsymbol{\rho}_N$ one can consider the difference between Ψ_B and Ψ_{asym} as being down to scaling every value in $\Psi_{S,A}$ by the same matrix $\boldsymbol{D}$ as in the previous point, and reach the required conclusion again for small enough $\bar{R}$.

Regarding other symmetric equilibria $\boldsymbol{\rho}^* \in E_S$, the connection between Ψ_B and Ψ_G is more tenuous. As in that case one cannot guarantee the sign of the diagonal of $\boldsymbol{J}_B$ even when $\bar{R}$ is small, D-stability is not true, and examples can be easily generated where the stability conclusion between case (1) and (2), (3) fails. ■

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, Robert Clark, Daniel Ershov, and Lei Xu (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, Michel (2006). "Dynamics of stochastic approximation algorithms". In: *Seminaire de probabilites XXXIII*. Springer, pp. 1–68.

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, and Sergio Pastorello (2020). "Artificial intelligence, algorithmic pricing, and collusion". In: *American Economic Review* 110.10, pp. 3267–97.

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

Cartea, Álvaro, Patrick Chang, José Penalva, and Harrison Waldon (2022). "Algorithms can Learn to Collude: A Folk Theorem from Learning with Bounded Rationality". In: *Available at SSRN 4293831*.

Chen, Zaiwei, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman (2024). "Two-Timescale Q-Learning with Function Approximation in Zero-Sum Stochastic Games". In: *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 378–378.

Cherry, Josh and Lones Smith (2010). "Unattainable Payoffs for Repeated Games of Private Monitoring". In: *Available at SSRN 1427602*.

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.

Harris, Charles R, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. (2020). "Array programming with NumPy". In: *nature* 585.7825, pp. 357–362.

Hunter, John D (2007). "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.03, pp. 90–95.

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

Kandori, Michihiro (1992). "Social norms and community enforcement". In: *The Review of Economic Studies* 59.1, pp. 63–80.

Kemeny, John G, J Laurie Snell, et al. (1969). *Finite markov chains*. Vol. 26. van Nostrand Princeton, NJ.

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*.

Lehrer, Ehud (1991). "Internal correlation in repeated games". In: *International Journal of Game Theory* 19.4, pp. 431–456.

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.

Mertikopoulos, Panayotis, Ya-Ping Hsieh, and Volkan Cevher (2024). "A unified stochastic approximation framework for learning in games". In: *Mathematical Programming* 203.1, pp. 559–609.

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.

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, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov (2017). "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347*.

Van Rossum, Guido, Fred L Drake, et al. (1995). *Python reference manual*. Vol. 111. Centrum voor Wiskunde en Informatica Amsterdam.

Virtanen, Pauli, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature methods* 17.3, pp. 261–272.