Reinforcement Learning and Collusion

Clemens Possnig

University of Waterloo

September 18, 2024

Abstract. This paper presents an analytical characterization of the long run policies learned by algorithms that interact repeatedly. The algorithms observe a state variable and update policies to maximize long term discounted payoffs. I show that their long run policies correspond to equilibria that are stable points of a tractable differential equation. As an example, I consider a repeated Cournot game, for which learning the stage game Nash equilibrium serves as non-collusive benchmark. I give necessary and sufficient conditions for this Nash equilibrium to be learned. State variables play an important role: With the previous period's price as a state variable, the Nash equilibrium can be learned. On the other hand, I present richer types of state variable, under which the Nash equilibrium will never be learned, while collusive equilibria may be learned. State variables exist that enable the learning of the best strongly symmetric equilibrium of nearby discretized repeated games.

JEL classification. C62, C73, D43, D83.

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

I thank my committee members Li Hao, Vitor Farinha Luz, and Michael Peters for years of guidance and conversations. I am grateful to 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 EC 22, GTA22, 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 an University of Waterloo SSHRC Institutional Grant (SIG). .

1. Introduction

More and more companies are using artificial intelligence (AI)-based tools to try to optimize sales and increase profits. Such algorithms take market data to determine current price or quantity levels, updating in real-time. Algorithms can help firms adapt to rapidly changing market environments, and potentially better serve their markets. However, recent empirical[1] and simulation-based[2] studies show that algorithms may learn to collude. Which algorithms, and which markets, are likely to support such outcomes?

While the folk theorem tells us the set of possible payoffs rational players may achieve in a repeated interaction, the outcomes to algorithmic learning may not even constitute a subset of these. This paper is concerned with the analytical properties of these outcomes for a common family of reinforcement learning (RL) algorithms.

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

Next, I use this characterization to study whether the algorithms can learn to repeat the static Nash equilibrium, which we can think of as the non-collusive benchmark. It turns out that the answer depends on what state variables these algorithms keep track of, and how these states evolve as a function of past prices and quantities. For instance, in the case where the state variable is the past period's price alone, learning the static Nash equilibrium comes down to a condition on the stage game payoff function alone. In contrast, I construct a richer state variable under which the static Nash equilibrium may not be learned, even if it had been learned under the past period's price state.

Finally, I study the channels through which the algorithms learn to collude. The rich state variable I constructed supports a symmetric binary-state equilibrium that in one state

[1]Studying the German gasoline retail market, Assad et al. (2020) observe that after a critical mass of firms deployed pricing algorithms, profit margins rose by 28%.

[2]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.

plays collusive, low quantities, and high punishment quantities in the other. Through an approximation exercise, I show that such collusive equilibria are closely related to optimal imperfect monitoring equilibria of the bang-bang kind, as characterized in Abreu, Pearce, and Stacchetti (1986). I provide sufficient conditions for this scheme to be learned with positive probability.

Related Literature

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

Firstly, the literature on algorithmic collusion has received increasing attention in recent years. Assad et al. (2020) 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 interesting phenomena regarding the learning of cooperation for a class of RL algorithms. Meylahn and V. den Boer (2022), Loots and denBoer (2023) use ODE methods related to the ones used in this paper to prove that specific algorithms can learn to collude in a pricing game. 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 a 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 finite history of the repeated interaction.

My class contains algorithms that impose little informational assumptions as a special case, known as "model free". Such algorithms do not carry a model of opponent behavior, and also no model of their environment and own payoffs. Thus, this class falls into the family of adaptive uncoupled learning rules as defined in Hart and Mas-Colell (2003). To the best of my knowledge, the study of uncoupled learning to collude in an oligopoly game based on the canonical game of Abreu, Pearce, and Stacchetti (1986) is new to this paper. Further foundational papers in this literature include Milgrom and Roberts (1990) , Fudenberg and Levine (2009), Gaunersdorfer and Hofbauer (1995), and many more.

Thirdly, this paper makes use of an extensive body of research related to stochastic approximation theory (see e.g. Borkar (2009)) and hyperbolic theory (Palis Jr, Melo, et al. (1982)). 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).

2. Multi-Agent Learning

This section introduces the updating rule (algorithm) and main assumptions used as running example in this paper. The algorithm is known as actor-critic Q-learning (ACQ). These algorithms keep track of an estimated performance criterion (the "critic", or Q-function, essentially a value function) and a policy function (the "actor") that is updated towards the maximizer of the performance criterion. The policy is a mapping from observables (states), such as past prices or other market data, to actions, e.g. prices or quantities. A main advantage of ACQ over the simpler and more commonly known Q-learning (Watkins (1989)) is that it directly applies to continuous-action problems, which are the focus of this paper. The results presented in this paper are not unique to the case of a Q-function used as the critic. A broader characterisation of algorithms for which the results stated here hold, is

relegated to the online appendix. Actor-critic reinforcement learning, which is a superset of the class studied here, has become popular in the reinforcement learning commmunity, due to better performance stemming from variance reduction and higher flexibility than pure critic- or actor-based methods (Q-learning being a critic-based method). See e.g. the substantial popularity of PPO (Schulman et al. (2017)).

There are N algorithms indexed by i, each having as action space a compact interval $\mathbf{X}_i$, with profile space $\mathbf{X} = \times_i \mathbf{X}_i$. A state variable S taking values in space $\mathbf{S}$ with $|\mathbf{S}| = L$ comes with a transition probability function, twice differentiable in $\mathbf{X}$, $P : \mathbf{S}^2 \times \mathbf{X} \mapsto [0, 1]$. Furthermore, after defining its transition probability function, I will refer to a state space S keeping implicitly in mind that it comes with its own transition probability. Each algorithm has a payoff function $u^i : \mathbf{X} \times \mathbf{S} \mapsto \mathbb{R}$, $\mathcal{C}^2$[3] in $\mathbf{X}$, and common discount factor $\delta \in (0, 1)$.

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

Algorithms update a policy function $\rho_t^i : S \mapsto \mathbf{X}_i$. Since states are finite, policy profile $\rho_t \in \overline{\mathbf{X}} = \mathbf{X}^{NL}$ can be represented as a vector in $\mathbb{R}^{NL}$.

Assumption 1. *For all $\rho \in \overline{\mathbf{X}}$, the Markov chain induced by $P_{ss'}[\rho(s)]$ is irreducible and aperiodic.*[4]

In fact, one can view such a policy as a stationary Markov strategy given state space S. Further, define $\overline{\mathbf{X}}_i = \mathbf{X}_i^L$, and $\overline{\mathbf{X}}_{-i} = \times_{j \neq i} \overline{\mathbf{X}}_j$.

Expected future discounted payoffs $W^i(\rho^i, \rho^{-i}, s_0)$ can be defined given stationary policy profiles $[\rho^i, \rho^{-i}] \in \overline{\mathbf{X}}$:

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \tag{1}$$

where the expectation is taken over the randomness in the stage game payoffs and state transitions.

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

[4]For definitions, see e.g. Appendix A in Puterman (2014)

Then define $B_S^i(\rho^{-i})$ as the optimal policy for i given a profile $\rho^{-i} \in \overline{\mathbf{X}}_{-i}$, chosen from the constraint set of stationary, S-state policies:

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

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-dependend policies since given a Markov stationary opponent profile ρ^{-i} there must be a Markov stationary best response. In what follows, write $\bar{B}_S(\rho)$ as the stacked best response correspondence over i.

Definition 1. *Define*

(i) $E_S \subset \overline{\mathbf{X}}$ *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* ρ^* *s.t.* $\rho^* \in \bar{B}_S(\rho^*)$*.*

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

If $\rho^* \in E_S$ is a differential Nash equilibrium, there is an open neighborhood U_{ρ^*} of ρ^* such that best responses must be single valued for all $\rho \in U_{\rho^*}$. Let $\mathcal{U}_S = \bigcup_{E_S} U_{\rho^*}$. Given these definitions of the underlying payoff environment, the following assumption is introduced:

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

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

(ii) There exist $\rho^* \in E_S$ *that are differential Nash equilibria.*

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

Assume that each algorithm uses ACQ to update their policy:

Definition 2. *Each algorithm i updates policies ρ_t^i according to*

$$\rho_{t+1}^i(s) \in \rho_t^i(s) + \alpha_t \Big[\arg\max_{a' \in \mathbf{X}} Q_{t+1}^i(s, a') - \rho_t^i(s) + M_{t+1}^i \Big], \tag{3}$$

where $\alpha_t > 0$ is a sequence of stepsizes converging to zero and M_{t+1}^i is an i.i.d, zero-mean, bounded variance noise generated as a means of exploring the policy space, commonly referred to as "parameter noise exploration".[5]

$Q_t^i(s, a)$ is an estimator[6] of

$$Q^{i*}(s, a, \rho_t^{-i}) = u(a, s) + \delta \mathbb{E}\big[\max_{a' \in \mathbf{X}} Q^{i*}(s', a', \rho_t^{-i}) \big| \, a, s \big],$$

the action-value Q^*-function conditional on i's opponents playing profile ρ_t^{-i} forever into the future. This Q^* is related to W through the equation

$$\max_{a' \in \mathbf{X}_i} Q^{i*}(s, a', \rho^{-i}) = \max_{\rho \in \overline{\mathbf{X}}_i} W^i(\rho, \rho^{-i}, s).$$

In what follows, whenever it is clear from context, write $Q_t^{i*} = Q^{i*}(s, a, \rho_t^{-i})$. Note that Q_t^{i*} is of interest in the reinforcement learning community as it pins down an optimal policy in stationary environments.

This paper remains agnostic about the specificities of the value function estimation part of the algorithms. The goal is to gain insights about what can be learned as long as the function approximation step is reasonably well behaved, a property to be defined below. The following assumption ensures that maximizers of Q_t^i track the maximizers of the correct function Q_t^{i*} well when t is large enough. The classical Q- estimator will not be enough for this to be true, as it requires discretization of the continuous action space. However, more involved estimation schemes exist for which Q_t^i can be shown to track Q_t^{i*}, as shown e.g. in Possnig (2022).

[5]For continuous action problems, various methods of exploration have been suggested, the version of parameter noise introduced here being one that is adopted frequently in the literature and allows for especially clean analytical results (see Plappert et al. (2017), and Yang et al. (2021) for a comprehensive survey).

[6]Notice that Definition 2 does not exclude the case in which the function to be approximated is fully known. The results thus include the case where agents know their value functions and follow a simple heuristic in updating their payoffs, taking as an input the current strategies of their opponent.

For a concrete example, consider, for some compact $\Theta \subset \mathbb{R}^\ell$, $1 \leq \ell < \infty$:

$$\mathcal{Q} \subseteq \left\{ Q : \mathbf{S} \times \mathbf{X} \times \Theta \mapsto \mathbb{R} \right\}$$

as the parametrized space of functions used to estimate Q^{i*}. Thus, $\theta_t \in \Theta$ becomes the parameter to estimate in order to find Q_t^{i*}, and we write $Q_t(s,a) = Q(s,a,\theta_t)$. While the more general case is treated in the online appendix, here we assume $Q^{i*}(\cdot,\cdot,\rho_t^{-i}) \in \mathcal{Q}$ for all ρ_t^{-i}. Define the supremum distance between estimator and target as

$$\chi_t^i \equiv \sup_{(s,a)\in\mathbf{S}\times\mathbf{X},} \left\| Q_{t+1}^i(s,a) - Q^{i*}(s,a,\rho_{t+1}^{-i}) \right\|.$$

As ultimately we are interested in the consistency properties of maximizers of Q_t, introduce the following notation: define for any $Q \in \mathcal{Q}$, $A_s(Q) = \arg\max_{a\in\mathbf{X}} Q(s,a)$. Let $A(Q) = [A_s(Q)]_{s=1}^L$. Now, let

$$d(A(Q_t), A(Q_t')) = \sup_{(s,b)\in\mathbf{S}\times A_s(Q_t)} \inf_{b'\in A_s(Q_t')} \|b - b'\|,$$

be the worst-case distance between maximizers of $Q_t, Q_t' \in \mathcal{Q}$. Note that $d(A(Q_t), A(Q_t')) = 0$ whenever $A(Q_t) \subseteq A(Q_t'))$. By definition, we have $A(Q_t^{i*}) = B_S^i(\rho_t^{-i})$. The algorithm's goal is to find a best response, so it is not necessary to find the full set of equally valuable maximizers at each step.

Assumption 3. *Assume for each i:*

(i) There exists $\beta > 1, C > 0, T > 0$ such that for all $t \geq T$,

$$d(A(Q_t^i), A(Q_t^{i*})) \leq C\left(\chi_t^i\right)^{\frac{1}{\beta}},$$

(ii)

$$E\left[\chi_t^i\right] = o(b_t^\beta),$$

where $b_t \to 0$ satisfies $\lim_{t\to\infty} \frac{\alpha_t}{b_t} = 0$, α_t being the stepsize in Definition 2.

(iii)

$$\sup_t \mathbb{E}\left[(\chi_t^i)^{2\beta}\right] < \infty,$$

(iv) Define $\mathcal{F}_t$ as the σ-algebra generated from $\{\rho_t, Q_t, M_t, \rho_{t-1}, Q_{t-1}, M_{t-1}..., \rho_0, Q_0, M_0\}$. For all $t < t'$, $\left(\chi_t^i\right)^{\frac{1}{\beta}}, \left(\chi_{t'}^i\right)^{\frac{1}{\beta}}$ are uncorrelated given $\mathcal{F}_t$.

In words, estimators Q_t^i converge uniformly in mean to Q_t^{i*}, with maximizers of Q_t^i converging at a related rate. Point (i) imposes a direct relationship between the uniform convergence of Q_t, and its maximizers. In the case of parametric function spaces such as $\mathcal{Q}$, this relationship commonly holds. Here in that case, χ_t^i can be written in terms of distance between estimated and "true" parameter, and upper hemicontinuity of $A(Q)$ implies that the upper bound in (i) is of the same order of magnitude as χ_t^i.[7] Point (ii) bounds the convergence speed in mean, and point (iii) ensures that large errors have negligible mass, which is important in the approximation results established in the next section. Point (iv) ensures that one can bound the variance of tail-sums of χ_t^i, which one can think of as the accumulated estimation error. $\mathcal{F}_t$ can be thought of as the information available to the algorithmic updating rule at a given period t. Such increasing sequences of σ-algebras are commonly employed in the analysis of stochastic difference equations such as (3).

For the stepsizes α_t I maintain the following:

Assumption 4. *Robbins-Monro Condition on stepsizes: $\alpha_t \to 0$ with*

$$\sum_{t=0}^{\infty} \alpha_t = \infty; \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty.$$

This assumption takes its name from the celebrated Robbins-Monro algorithm representation (Robbins and Monro (1951)). The assumption constrains the speed of convergence of α_t, needing to balance the averaging out of errors (i.e. converge fast enough), versus moving slowly enough to ensure sufficient exploration of the policy space.

Throughout the rest of the paper, I impose the following assumption on the iteration ρ_t:

Assumption 5. *Iterates stay bounded almost surely:*

$$\sup_t \|\rho_t\| < \infty, \ a.s..$$

Even though commonly made, Assumption 5 is often difficult to verify. It is common for authors to give all their results conditioning on the event that 5 holds, see for example

[7] More broadly, (i) is inspired by set-valued estimation results, e.g. conditions C.1 and C.2 in Chernozhukov, Hong, and Tamer (2007), adapted to this setting of maximization under time-dependent target.

Benaïm and Faure (2012). For a more general discussion of sufficient conditions for bounded iterates, see Borkar (2009), Chapter 2.

With Assumptions 3 and 4 in place, I will show that one can apply results from stochastic approximation theory (see e.g. Borkar (2009)) to connect the long-run behavior of ρ_t to limiting sets of solutions to an underlying differential equation.

3. Long Run Behavior: Main Results

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

Definition 3. *Take the algorithm from Definition 2. The limit set is defined as*

$$L_S = \bigcap_{t\geq 0} cl\Big(\{\rho_\ell \,|\ell \geq t\}\Big),$$

the set of limits of convergent subsequences ρ_{t_k}.

I write S as subscript to underline the dependence of the limiting set on the state space S. As the characterizations introduced here will require properties of a differential equation, I present next some useful definitions:

Definition 4. *Given some ODE* $\dot{\rho} = f(\rho)$, *let* ρ^* *be a rest point of* $f(\rho)$. *Let* $\Lambda = eigv[Df(\rho^*)]$ *the set of eigenvalues of the linearization of* f *at* ρ^*. *For a complex number* z, *let* $\mathbf{Re}[z] \in \mathbb{R}$ *be the real part.* ρ^* *is*

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

To save notation, define for $\rho \in \overline{\mathbf{X}}$

$$F_S(\rho) = co\Big(\bar{B}_S(\rho) - \rho\Big), \tag{4}$$

as the state dependent best response dynamics, where I take $\bar{B}_S(\rho)$ to be the stacked version of $B^i_S(\rho^{-i})$ over i, and $co(\cdot)$ represents the convex hull.

Theorem 1. *Let $\rho^* \in \mathcal{U}_S$ be asymptotically stable for F_S. Then*

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

Proof Sketch of Theorem 1

The full proof for this and the following Theorems can be found in Appendix A.

Firstly, I make a connection between the recursion in (3) and the differential inclusion F_S. One can relate a time-interpolated version of the recursion ρ_t to solutions of the differential inclusion

$$\dot{\rho} \in F_S(\rho(t)),$$

Due to the nature of stepsizes α_t and the presence of mistakes, the time-interpolation of recursion (3) will approximate a convex hull of the best response. Since the best-response may be multivalued, solutions to this inclusion are not guaranteed. However, assumptions on the regularity of F_S (which comes down to a linear growth condition, see Assumption 6 (i) in the Appendix) allow us to show that there is a global solution in the sense of Filipov (1988). When considering that the updating rate α_t converges to zero, one may convince oneself that the recursion in (3) looks similar to a discrete approximation to a time-derivative. The idea is to show that the time-interpolated version of ρ_t must stay close, almost surely, to solutions of $F_S(\rho)$. Attracting points of the differential inclusion are then natural candidates to also attract ρ_t.

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

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

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

Proof Sketch of Theorem 2

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

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

There is a caveat here, however: Theorem 1 does not state that all limiting points in $L_{S,g}$ 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.[8] Notice that this observation implies that the Folk theorem is neither necessary nor sufficient in describing the possible payoffs achieveable by learning algorithms.

4. Learning to Collude

In this section, I exemplify the potential of my characterisation via a repeated Cournot game played by RL algorithms falling into the family of ACQ learners. This game can be shown to satisfy Assumptions 1 and 2. It follows that whenever an ACQ algorithm satisfies Assumptions 3 and 4, the long run characterizations of section 3 apply.

The game is set up in line with Abreu, Pearce, and Stacchetti (1986)'s oligopoly game: There are two agents, $i \in \{1, 2\}$. Actions are chosen as quantities $x \in \mathbf{X} = [0, M]$ for

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

some large $M > 0$, with aggregate quantity X. I will sometimes write $X \in \mathbf{X}$, in the understanding that the actual space of aggregate quantites is $[0, 2M]$. The price outcome is stochastic, $y \in \mathbf{Y} = [0, \overline{Y}]$, continuously distributed conditional on X. The conditional price density is denoted $g(y; X)$ with full support on $\mathbf{Y}$, $\mathcal{C}^2$ in X for almost all X. Let the expected price conditional on X be

$$Y(X) = \int_{\mathbf{Y}} yg(y; X)dy.$$

Stage game payoffs are symmetric[9] for $i \in \{1, 2\}$:

$$u^i(x_i, x_{-i}) = Y(X)x_i - c(x_i),$$

with $c(x)$ a convex, twice differentiable cost function.

Due to symmetry, write $u = u^i$ whenever it is clear from context.

Definition 5. *Say that the payoff function $u(x_1, x_2)$ is regular if*

(i) $\frac{\partial}{\partial x_1} u^1(0, 0) > 0$.

(ii) $c(0) = 0$, $c'(0) > 0$, $c''(x) \geq 0$ *for all* $x \in \mathbf{X}$.

(iii) $Y'(2x) < 0$ *for all* $x < M$.

(iv) For all $x, x' \in \mathbf{X}$

$$Y'(x + x') + xY''(x + x') \leq 0.$$

(v) $\arg\max_{x \in \mathbf{X}} u(x, x_M) > 0$, *where* $x_M = \arg\max_{x \in \mathbf{X}} u(x, 0)$.

Definition 5 follows standard assumptions made in the Cournot game (e.g. Hahn (1962)). For point (v) note that it rules out the boundary equilibrium, the unique Nash equilibrium (x_N, x_N) being interior.

[9]Symmetry is not necessary for the results, but saves on notation.

4.1. Binary State Variables

To start, I will derive the objects relevant for stability analysis given a general commonly observed binary state variable S with $\mathbf{S} = \{A, B\}$. Define for any $s \in \mathbf{S}$, and $x_i \in \mathbf{X}$:

$$P_{sB}(x_1, x_2) = \mathbb{P}\big[s' = B \big| \, s; x_1, x_2\big],$$

the transition probability to move to state B given current state s and quantity choices x_i in state s. Also assume that

$$P_{sB}(x_1, x_2) = \mathbb{P}\big[s' = B \big| \, s; x_1 + x_2\big],$$

for all s, x_i, i.e. transition probabilities only depend on aggregate quantities. I will therefore commonly write $P_{ss'}(x_1, x_2) = P_{ss'}(X)$ with $X = x_1 + x_2$. Let $\rho^i : \mathbf{S} \mapsto \mathbf{X}$ be each player's policy, and recalling the definition of W^i in (1), note that in the binary case one can derive

$$\begin{aligned} W^i(\rho, A) &= \omega^{-1}(\rho)\Big[(1 - \delta P_{BB}(\rho))u^i\big(\rho^i(A), \rho^{-i}(A)\big) + \delta P_{AB}(\rho)u^i\big(\rho^i(B), \rho^{-i}(B)\big)\Big], \\ W^i(\rho, B) &= \omega^{-1}(\rho)\Big[\delta(1 - P_{BB}(\rho))u^i\big(\rho^i(A), \rho^{-i}(A)\big) + (1 - \delta(1 - P_{AB}))u^i\big(\rho^i(B), \rho^{-i}(B)\big)\Big], \end{aligned} \tag{5}$$

where

$$\omega(\rho) = \Big[1 + \delta(P_{AB}(\rho) - P_{BB}(\rho))\Big].$$

Thus, W^i is a convex combination of stage game payoffs u^i over the two states, with weights being a function of transition probabilities. Notably, as $\delta \to 1$, these weights will converge to the unique stationary distribution over states given the policy profile ρ.[10]

Let S_0 with $|\mathbf{S}_0| = 1$ be the trivial state variable. F_{S_0} then simplifies to the classical stage game strategy based best response dynamics. Under F_{S_0}, it is well known that for regular u, the unique Nash equilibrium x_N is globally attracting (Milgrom and Roberts (1990)). If ACQ-learners (and many other agents) played on the trivial state space S_0, they would converge to x_N with probability 1. I show next that even though that is true, a larger

[10] Uniqueness is implied by irreducibility (Assumption 1).

family of binary state variables exist so that when they are used, ACQ learners will not learn this Nash equilibrium. This inspires the following definition:

Definition 6. *Say a given equilibrium x^* of u is **statically stable** if it is stable under F_{S_0}. For a given state variable S with $|\mathbf{S}| \geq 2$, say that ρ^* with $\rho^*(s) = x^* \; \forall \; s$ is **dynamically stable** (under S) if it is stable under F_S.*

To see how a statically stable equilibrium may not be dynamically stable, focus on a binary state variable. First, consider the incentives faced by a rational agent playing a binary Markov strategy at the stage game Nash equilibrium, $\rho^i(s) = x_N$ for both i, s. I call this policy ρ_N. Given regular u, note that for any twice differentiable interior transition probabilities, the Hessian of each player's optimization problem of maximizing (5) at ρ_N is negative definite. Thus, we can derive a player's best-response derivative at ρ_N.

Since the focus is on symmetric equilibria, I will drop the i- superscript for all objects, fixing our attention on player 1's payoffs. To further ease notation, I will adopt the following conventions:

- $u^s = u(\rho_s, \rho_s)$, for $s \in \mathbf{S}$.
- $u^s_k = \frac{\partial u^s}{\partial x_k}$ and $u^s_{kk'} = \frac{\partial u^s_k}{\partial x_{k'}}$, for $k, k' = 1, 2$, $s \in \mathbf{S}$.
- $P'_{sB} = \frac{\partial P_{sB}}{\partial x_1} = \frac{\partial P_{sB}}{\partial x_2}$ for all s and analogously for P''_{sB} where the equality comes from the fact that P_{sB} only depends on aggregate quantities.

Letting $b(\rho)$ be the best response to ρ, consider player 1's best-response derivative in state A to an incremental change in opponent's policy in state A:

$$\frac{\partial b(\rho_N)(A)}{\partial \rho_N(A)} = BR'_N + \frac{\delta P'_{AB}(\rho_N)}{\omega_N} \frac{u^N_2}{u^N_{11}},$$

where $b(\rho_N) = \rho_N$ is the best response according to long-run payoffs as derived in (5), $\omega_N = \omega(\rho_N)$ signifies evaluation of $\omega(\rho)$ at ρ_N, and I use that $BR'_N = -\frac{u^N_{12}}{u^N_{11}}$. The terminology BR'_N is used since indeed, $-\frac{u^N_{12}}{u^N_{11}}$ is the slope of the stage-game best response function evaluated at x_N. The agent has a tradeoff between following incentives about payoffs today (static incentives), represented by BR'_N, and dynamic incentives considering effects on continuation payoffs, represented by the second term. The factor multiplying $\frac{u^N_2}{u^N_{11}}$ can

be interpreted as the sensitivity of the sum of transition probabilities $P_{AB}(\rho_N) + P_{BA}(\rho_N)$ with respect to policy ρ, which I now denote as ζ_N.

Dynamic stability of the static Nash equilibrium is impacted by dynamic incentives in a straightforward manner:

Proposition 1. *Let u be regular and consider arbitrary transition probabilities $P_{ss'}$ for a binary state variable. Then ρ_N is dynamically unstable if and only if*

$$\left| BR'_N + \delta \frac{P'_{AB}(\rho_N) + P'_{BA}(\rho_N)}{\omega_N} \frac{u_2^N}{u_{11}^N} \right| > 1.$$

This leads to the following corollary:

Corollary 1. *Let u be regular. ρ_N is dynamically stable if and only if*

$$\zeta_N \in (z_1^*, z_2^*),$$

where

$$z_1^* = -\big(1 + BR'_N\big) \frac{u_{11}^N}{u_2^N}; \quad z_2^* = \big(1 - BR'_N\big) \frac{u_{11}^N}{u_2^N}.$$

Note that regularity of u implies $z_1^* < 0 < z_2^*$, and $BR'_N \in (-1, 0)$. This result1[11] uncovers the channels through which the static Nash equilibrium can be destabilized, and eventually through which algorithms in my class will learn to avoid this Nash equilibrium. Fixing the market conditions, state variables come into play through ζ_N. For any payoff function u of bounded derivatives, there is a threshold so that once ζ_N surpasses that threshold, static Nash will not be learned. The set of state variables that can render a static Nash equilibrium unstable is therefore quite large. This intuition then allows to separate two factors that determine whether the RL will learn to play static Nash: properties of stage game payoffs u, and properties of the state variable's distribution.

[11]Note that this result holds more generally under (not necessarily unique) Nash equilibria of payoff functions u unrelated to the Cournot game studied here, given twice differentiability of u at the equilibrium.

4.2. Two Fundamental Families of Binary State Variables

Here I introduce two families of binary state variables that correlate to X through the channel of price outcome Y. In the first case, static Nash will be dynamically stable for an important subset of that family; in the second case, static Nash can be dynamically unstable, and the existence of collusive equilibria that are attracting is possible, which are best strongly symmetric equilibria of the game restricted to any discrete subset of the action space $\mathbf{X}$.

First, consider the following state variable: fix price cutoffs $y_A, y_B \in \mathbf{Y}$. Then the state variable is defined via the transition function $f_{1R} : \{A, B\} \times \mathbf{Y} \times \mathbf{Y}^2 \mapsto \{A, B\}$:

$$f_{1R}(s_{t-1}, Y_{t-1}, (y_A, y_B)) = \begin{cases} A & \text{if } s_{t-1} = A \text{ and } Y_{t-1} \leq y_A \\ A & \text{if } s_{t-1} = B \text{ and } Y_{t-1} \leq y_B \\ B & \text{if } s_{t-1} = B \text{ and } Y_{t-1} > y_B \\ B & \text{if } s_{t-1} = A \text{ and } Y_{t-1} > y_A. \end{cases} \tag{6}$$

In other words, this state recalls whether the last period's price was low (state A) or high (state B), where the definition of low and high can depend on the present state through cutoffs y_A, y_B.

Definition 7. *A public binary* 1R*-policy (one-recall) is defined as policy $\rho : \{A, B\} \mapsto \mathbf{X}$, so that states evolve according to f_{1R}, given some cutoffs (y_A, y_B). Refer to state variables evolving according to f_{1R} as* 1R*-state variables. Call the policy "consistent" if $y_A = y_B$.*

Corollary 2. *Let ρ_N^{1R} be a consistent* 1R*-policy that plays stage game Nash quantity x_N in every state. Then, ρ_N^{1R} is dynamically stable if and only if x_N is statically stable.*[12]

This result follows from Corollary 1, since under consistent 1R-policies, we must have that $\zeta_N = 0$ whenever $y_A = y_B$. Note that under consistent 1R-policies, $P_{AB} = 1 - P_{BA} =$

[12]This result extends to the case of consistent 1R-policies of finitely many (> 2) price cutoffs. This follows from an iterated application of the Sherman-Morrision formula together with the matrix determinant lemma.

$\mathbb{P}[Y \leq y_A]$. Thus, $P'_{AB}(\rho_N) + P'_{BA}(\rho_N)$=0. In general, this comes from the fact that, for every given current state, conditional distributions over future states are the same.

In contrast, the following is a state variable resulting from the switch of two inequalities in the state dynamics, which I denote *direction-switching* (DS), under transition function f_{DS}:

$$f_{DS}(s_{t-1}, Y_{t-1}, (y_A, y_B)) = \begin{cases} A & \text{if } s_{t-1} = A \text{ and } Y_{t-1} > y_A \\ A & \text{if } s_{t-1} = B \text{ and } Y_{t-1} \leq y_B \\ B & \text{if } s_{t-1} = B \text{ and } Y_{t-1} > y_B \\ B & \text{if } s_{t-1} = A \text{ and } Y_{t-1} \leq y_A. \end{cases} \tag{7}$$

A realized price lower than the current cutoff y_s represents a *switch-signal*, while realizing a high price leads to no change in the state. More generally, one can define policies following state variables with transition probabilities having the above property:

Definition 8. *Say a public binary* DS*-policy is defined as a binary policy under a state variable following transition function f_{DS} for some cutoffs (y_A, y_B). Refer to state variables evolving according to f_{DS} as* DS*-state variables. Call it a consistent* DS*-policy if $y_A = y_B$.*

Note that under consistent DS-policies, the probability of reaching any state s conditional on being in A is complementary to the probability of reaching s conditional on being in B. This fact introduces an essential difference in how states A, B are interpreted, even when $\rho(A) = \rho(B)$ is played.

There is a set of regular payoff functions such that ρ_N^{DS} is statically stable, but dynamically unstable. Moreover, this family of regular payoff functions will also allow for the existence of collusive equilibria.

Definition 9. *Define the set of densities $\mathcal{G}$ so that all $g(y; X) \in \mathcal{G}$ satisfy:*

*(i) A monotone likelihood ratio property (*MLRP*):*

$$\eta(y, X) \equiv \frac{\partial \log\big(g(y; X)\big)}{\partial X} \tag{8}$$

is decreasing in y for all $X \in \mathbf{X}$, and for all $X \in int(\mathbf{X})$, $\eta(0, X) > 0 \geq \eta(\overline{Y}, X)$. Define $\overline{y}(X)$ such that $\eta(\overline{y}(X), X) = 0$.

(ii) *Let* $g(y;X) = \int_0^p g(y;X)dy$ *be the c.d.f. based on* $g(y;X)$. *For every* $X' \in int(\mathbf{X})$,

$$\frac{\partial}{\partial X} G(y;X)|_{y=\overline{y}(X')}$$

is quasi-concave in $X \in \mathbf{X}$, *with peak at* X'.

(iii) $\lim_{X\to 0} G_2(y;X) = 0 = \lim_{X\to M} G_2(y;X) = 0$ *for all* $y \in \mathbf{Y}$.

(iv)

$$\frac{\partial \log\big(-Y'(x+x')\big)}{\partial x} < \frac{1}{x}.$$

MLRP ensures that lower prices are more sensitive to changes in quantities than higher prices. Points (ii) and (iii) will allow to construct collusive equilibria from first order conditions. Point (iv) ensures that a strict version of Definition 5 (5) can be satisfied given this set of density functions.[13]. By carefully constructing $g(y;X)$ so as to have it loose sensitivity under large x, this can be made to hold. The numerical example in the online appendix verifies this.

Proposition 2. *For all* $g \in \mathcal{G}$ *there exist convex* $c(x)$ *such that the resulting* u *is regular. For a generic subset of* $\mathcal{G}$[14], *there exists a state variable in* S^* *such that for all* $\delta \in (0,1)$ *large enough,* ρ_N^{DS} *is dynamically unstable and there exists a symmetric equilibrium* σ *with* $0 < \sigma_A < x_N < \sigma_B$.

We see from the above and the following subsection that 1R-state variables and DS-state variables lead to starkly different outcomes under reinforcement learning. In the former, static Nash will always be learned with positive probability while in the latter, static Nash may never be learned, while collusion can be learned with positive probability. Furthermore, payoff functions resulting in collusive equilibria under DS-state can be such that the static Nash equilibrium is the unique symmetric equilibrium under 1R-state variables:

Lemma 1. *Suppose* $g \in \mathcal{G}$ *with regular* u *such that the conditions of Proposition 2 hold. Then* ρ_N *is the unique symmetric equilibrium under consistent* 1R*-policies.*

[13]Note that (i) implies that $Y'(x+x') < 0$ for all $x, x' \in \mathbf{X}$

[14]A growth rate condition local to X_N.

Thus, in an economy that supports collusion under a DS-state, the choice of state variable by firms that employ RL is all the more impactful.

Whether collusion will be learned, if it is a Nash equilibrium, still comes down to stability. Stability depends on quantities related to growth rates of transition probabilities and the stage game in manners analogous to the stability analysis of the static Nash equilibrium.

To see this, let $\Pi(x_1, x_2) = Y(X) - c'(x_1)$ be the marginal profit as computed by a price-taker. Notice that by construction of the Cournot-payoff function, $u_1(x_1, x_2) - u_2(x_1, x_2) = \Pi(x_1, x_2)$, which is true for all x and therefore also true for $\Pi'(x_1, x_2) \equiv \frac{\partial \Pi(x_1, x_2)}{\partial x_1}$. Note that $\Pi'(x_1, x_2) < 0$ for all $x_i \in \mathbf{X}$. First, define for $s, s' \in \mathbf{S}$:

$$R_s = \frac{\delta P'_{ss'}}{\omega} \frac{\Pi_s}{\Pi'_s},$$

where $\Pi_s = \Pi(\rho_1(s), \rho_2(s))$. This quantity can be interpreted as a ratio of elasticities of ω versus Π_s with respect to symmetric $\rho(s)$.

Lemma 2. *Consider an interior, symmetric equilibrium under a binary state variable, $\sigma = (x_A, x_B)$ with $x_A < x_B$ as constructed in Proposition 2. Then σ is asymptotically stable if*

$$0 \leq \min\{R_A, R_B\}, \text{ and } R_A + R_B \leq 1.$$

4.3. Relationship to the Best Equilibrium

While the characterisation in Theorem 1 holds for any finite state variable, more tractable results were achieved above under the restriction to binary state variables. At this point it becomes interesting to ask about the breadth of a theory under such a restriction. It turns out that we can connect known results on the best possible payoff a rational player can achieve in a repeated game of imperfect public monitoring (Abreu, Pearce, and Stacchetti (1990), henceforth APS), and binary-state collusive equilibria as constructed in Proposition 2.

First, let $\Gamma = \left(u^1, u^2\right)$ be the stage game as defined in the beginning of the section. Then one can define $\Gamma^\infty(\delta)$ as the infinite repetition of Γ where players discount expected long term payoffs by $\delta \in (0, 1)$. For any $0 < t$, define $b_t = \{Y_s\}_{0<s<t}$ to be a public history of the

game, with $B_t = \mathbf{Y}^t$ the set of possible public histories up to time t. Then let $b_t^i = \{x_i^s\}_{0<s<t}$ be the private memory of a player's own actions, and define $B_t^i = \mathbf{X}^t$ as the set of those at period t. Now, a strategy of player i at period t can be written as map $\sigma_t^i : B_t \times B_t^i \mapsto \mathbf{X}$. A strategy is then a sequence $\sigma^i = \{\sigma_t^i\}_{t>0}$, with the set of such sequences denoted Σ. In keeping with APS, we can define strongly symmetric sequential equilibria (SSE) of Γ^∞ as profiles $\sigma = \{\sigma_t^1, \sigma_t^2)\}_{t>0}$ with $\sigma_t^1(b_t, b_t^1) = \sigma_t^2(b_t, b_t^2)$ whenever $b_t^1 = b_t^2$ (i.e. strategies that are "public"), that are individually unimprovable for each player, with respect to their expected future discounted payoffs:

$$U^i(\sigma) = (1-\delta)\mathbb{E}\sum_{t>0}\delta^t u^i(\sigma_t) \geq U^i(\sigma', \sigma^{-i}) = (1-\delta)\mathbb{E}\sum_{t>0}\delta^t u^i(\sigma_t', \sigma_t^{-i}),$$

for any $\sigma' \in \Sigma$. APS provide a result stating that the best SSE can be supported by a bang-bang solution, under their setting. Their setting differs from the one of this section in that APS require finite (but arbitrarily many) actions, instead of a continuum as considered here.

An approximation argument can be made to approximate the best SSE of Γ^∞ by a sequence of best SSEs of repeated games with a finite, increasing number of actions.

Define the restricted action set $\mathbf{X}_K = \{x_1, ..., x_K\} \subset \mathbf{X}$ such that $\max_{0<k,k'\leq K}\{|x_k - x_{k'}|\} \leq \frac{1}{K}$, and such that $x_N \in \mathbf{X}_K$ for all $K > 0$. Let the restricted game Γ_K^∞ be the repeated game where players are constrained to choose actions from $\mathbf{X}_k$.

Under Γ_K^∞, APS' result applies: the payoff-maximizing SSE of Γ_K^∞ can be achieved by a symmetric bang-bang profile $\sigma_K \in \mathbf{X}_K^4$, with V_K defined as its value. Notice further that under $g \in \mathcal{G}$, MLRP gives us that σ_K can be implemented as a DS-policy for some thresholds $(y_A, y_B) \in \mathbf{Y}^2$. This follows from the fact that optimal punishment regions of the price space, as characterised in APS, are monotone (binary) partitions under the MLRP. Payoffs under the "good" state can be increased when the probability of punishment state is decreased; this can be done as long as incentives are preserved. Thus, starting from a price threshold $x \sim 0$, one can find the "punishment set" of prices in the good state by considering all prices below x. By the MLRP, sensitivity of the conditional p.d.f. is

maximal for the lowest prices, and decreases over prices, implying that this leads to the most efficient choice of punishment set in terms of probability of punishment.[15]

Define $W(\sigma, z)$ for symmetric profiles $\sigma \in \mathbf{X}^4$, thresholds $z \in \mathbf{Y}^2$ as long run binary state payoffs given those thresholds, abusing notation slightly from (5). Stretching notation slightly further, I will write $W(\sigma, \sigma', z)$ for profiles $(\sigma, \sigma') \in \mathbf{X}^4$ that are not necessarily symmetric. Define

$$E_K(z) = \left\{\sigma \in \mathbf{X}_K^2 \mid W(\sigma, z) \geq W(\sigma', \sigma, z) \; \forall \; \sigma' \in \mathbf{X}_K^2\right\},$$

the set of symmetric equilibria under any threshold $z \in \mathbf{Y}^2$. Let $E^*(z)$ be the corresponding symmetric equilibrium set when the full continuous $\mathbf{X}$ is available. $E_K(z), E^*(z)$ are nonempty due to the inclusion of x_N.

Thus, σ_k, V_K can be alternatively characterised as solution to the problem

$$V_K = \max_{\substack{\sigma \in E_K(z) \\ z \in \mathbf{Y}^2}} W(\sigma, z).$$

Analogously, define

$$V = \sup_{\substack{\sigma \in E^*(z) \\ z \in \mathbf{Y}^2}} W(\sigma, z). \tag{9}$$

Proposition 3.

(1) $V \geq \limsup_{K \to \infty} V_K$.

(2) If all $\sigma \in E^*$ *are strict,* $V = \lim_{K \to \infty} V_K$.

Define V^* to be the best SSE payoff among all SSE of Γ^∞. We have now shown the following:

Corollary 3. *There exists an SSE* σ *of* Γ^∞ *supported by a binary* DS*-policy under thresholds* z^**, such that* $V = W(\sigma, z^*)$*. It holds that*

(1) $V \leq V^*$.

(2) For any ε *there exists* $\bar{K}$ *such that for all* $K \geq \bar{K}$, $|V - V_K| < \varepsilon$.

[15]Note that MLRP is sufficient for the punishment regions to be pinned down by at most two thresholds. The result in this section readly extend to the case of finitely many thresholds required to pin down punishment and reward regions.

Corollary 3 tells us that there exist binary state variables such that if used by algorithms, they may learn to achieve the best DS-policy equilibrium of the continuous action game. If Lemma 2 holds for σ, then with positive probability, the algorithms' long run payoffs will be arbitrarily close to the best SSE payoffs overall, of any arbitrarily fine discretization of the action space. Hence, while it is not known whether algorithms will learn the best symmetric public monitoring equilibrium of the underlying continuous-action game, they may achieve payoffs arbitrarily close to the best public monitoring payoffs of arbitrarily close-by discrete games.

5. Conclusion

This paper considers the long-run behavior of a class of RL algorithms and shows how it can intepreted via the stability of repeated game equilibria according to an underlying differential equation. The application of collusion in repeated games is employed to show the usefulness of this framework: it allows one to consider comparative statics exercises on the long-run learning behavior of RL with respect to details of the game and algorithms.

The characterization of long-run behaviors serves as a methodology that can allow researches to pick a given interaction of interest, e.g. an auction, a stock market, or multilateral platform, then pick a class of algorithms, and evaluate long-run outcomes in the chosen setting.

An important insight from my analysis is the dependence of the attractability of a given equilibrium of the repeated game, on state variables observed by algorithms. This insight can serve as a starting point in efforts to curb algorithmic collusion.

Finally, I show that the best symmetric binary equilibrium learnable by the algorithms considered here will achieve payoffs arbitrarily close to the best symmetric imperfect public monitoring equilibrium of any discretization of the action space. While this insight doesn't answer whether there may exist better imperfect public monitoring equilibria of the continuous-action game, it does give some reassurance in terms of payoff-guarantees for the algorithms studied here.

Appendix A. Proofs

The following assumption summarizes assumptions made throughout the main text, for ease of readability of the proofs to follow. For notational ease, write $F(\rho) = F_S(\rho)$, as for all results to follow, state variables will be fixed.

The algorithm (3) can be written as

$$\rho_{n+1} \in \rho_n + \alpha_n \left[F(\rho_n) + \delta_n + M_{n+1} \right], \tag{10}$$

where $\delta_n^i = d(A(Q_t^i), A(Q_t^{i*}))$, with δ_n stacked over i. We switch to an identification of time periods by n in order to distinguish of the continuous timescale t used in the associated continuous time systems.

Assumption 6. *Let $\mathcal{F}_n$ be the σ-field generated by $\{\rho_n, Q_n, M_n, \rho_{n-1}, Q_{n-1}, M_{n-1} \ldots, \rho_0, Q_0, M_0\}$, i.e. all the information available to the updating rule at a given period n.*

(i) There exists $c > 0$ such that $\sup\{\|y\| : y \in F(\rho)\} \leq c(1 + \|\rho\|)$ for all $\rho \in \bar{\mathbf{X}}$.

(ii) M_{n+1} is a Martingale-difference noise. There is $0 < \bar{M} < \infty$ and $x > 2$ such that for all n

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

(iii) There exists a continuous function

$$\Omega : \mathcal{U} \mapsto O(\overline{\mathbf{X}}),$$

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

$$\mathbb{E}[M_{n+1} M_{n+1}' \,|\mathcal{F}_n] = \Omega(\rho_n),$$

whenever $\rho_n \in \mathcal{U}$.

(iv)

$$E\left[\|\delta_n\|\right] = o(b_n),$$

where $b_n \to 0$ satisfies $\lim_{n\to\infty} \frac{\alpha_n}{b_n} = 0$, α_n being the stepsize in Definition 2.

(v)

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

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

Point (i) is an assumption on the stage game and transition probabilities: best responses should not grow by undbounded amounts. This assumption is a sufficient condition for the existence of Filipov solutions (Filipov (1988) to the differential inclusion $\dot{\rho} \in F(\rho)$. Points $(ii), (iii)$ is are a weakening of the i.i.d. assumption made for simplicity in Definition 2.

Proof of Theorem 1

Proof. First, we prove the following result that employs known techniques from stochastic approximation theory.

The following Definition can be found in Benaïm, Hofbauer, and Sorin (2005, Section 3.3):

Definition 10.

(1) Given a set $A \in \mathbf{X}$ and $x, y \in A$, we write $x \hookrightarrow_A y$ if for every $\varepsilon > 0$ and $T > 0$, there exists an integer $n \in \mathbb{N}$, solutions $x_1, ..., x_n$ to $\dot{x} \in F(x)$[16], and real numbers $t_1, ..., t_n$ greater than T such that:

a) $x_i(s) \in A$ for all $0 \leq x \leq t_i$, and for all $i = 1, ..., n$,

b) $\|x_i(t_i) - x_{i+1}(0)\| \leq \varepsilon$ for all $i = 1, ..., n-1$,

c) $\|x_1(0) - x\| \leq \varepsilon$ and $\|x_n(t_n) - y\| \leq \varepsilon$.

(2) A set $A \in \mathbf{X}$ is said to be internally chain transitive (ICT) if A is compact and $x \hookrightarrow_A y$ holds for all $x, y \in A$.

One can think of chains as described in this definition as a generalization to periodic orbits of an ordinary differential equation (ODE), where solutions to the ODE are allowed to take on arbitrarily small jumps. This generalization turns out to be very useful in the description of long run behavior of discrete-time stochastic systems.

[16]Recall that $G(x)$ is an inclusion, so uniqueness of solutions cannot be guaranteed.

Importantly, ICT sets include rest points and limit cycles (if they exist). Consider Papadimitriou and Piliouras (2018) for an intuitive discussion. The following result shows why these sets are of importance in our analysis:

Proposition 4. *Almost surely, L_S is an ICT set of the differential inclusion*

$$\dot{\rho} \in F(\rho(t)).$$

Proof. We can now show first that iteration 10 is a perturbed solution to $\dot{\rho} \in F(\rho(t))$ as defined in Benaïm, Hofbauer, and Sorin (2005, Definition II). The approach is to construct a linear interpolation of (10), and show that this will shadow solutions to $\dot{\rho} \in F(\rho(t))$ asymptotically, for large enough n. Following the notation in Hofbauer and Sandholm (2002), introduce:

$$\tau_0 = 0; \quad \tau_n = \sum_{i=1}^{n} \alpha_i; \quad m(t) = \sup\{k \geq 0 : \tau_k \leq t\}.$$

Then, construct the interpolation as

$$X(\tau_n + s) = \rho_n + s\frac{\rho_{n+1} - \rho_n}{\alpha_{n+1}}, \quad s \in [0, \alpha_{n+1}]. \tag{11}$$

Following the proof of Hofbauer and Sandholm (2002, Proposition 1.3), we only need to take care of the additional term δ_n present in iteration 10.

We will consider the accumulated δ_n, M_{n+1} error terms. First, note that

$$\begin{aligned}
&\sup\left\{\left\|\sum_{i=n}^{k-1}\alpha_{i+1}\Big(\delta_{i+1} + M_{i+2}\Big)\right\| : \; k = n+1, \ldots, m(\tau_n + T)\right\} \\
&\leq \sup_{n \leq k \leq m(\tau_n+T)-1}\left\|\sum_{i=n}^{k}\alpha_{i+1}\Big(M_{i+2}\Big)\right\| + \sup_{n \leq k \leq m(\tau_n+T)-1}\left\|\sum_{i=n}^{k}\alpha_{i+1}\Big(\delta_{i+1}\Big)\right\| \\
&= R_n + \sup_{n \leq k \leq m(\tau_n+T)-1} \Psi_n^k.
\end{aligned}$$

By Assumption 6, R_n is a standard error term in stochastic approximation theory, satisfying the usual assumptions of Robbins-Monro algorithms with martingale difference noise. The sufficient conditions of Benaïm, Hofbauer, and Sorin (2005) are satisfied here, so it is

known that R_n converges almost surely to zero.[17] We need to take care of the additional term δ_n present in iteration 10. It suffices to show that, for all $T > 0$

$$\sup_{n \leq k \leq m(\tau_n + T) - 1} \Psi_n^k \to 0, \tag{12}$$

almost surely as $n \to \infty$. First, note that

$$\Psi_n^k \leq \sup_{n \leq k \leq m(\tau_n+T)-1} \left\| \sum_{i=n}^{k} \alpha_{i+1} \Big(\|\delta_{i+1}\| - \mathbb{E}\big[\|\delta_{i+1}\| \mid \mathcal{F}_{i+1} \big] \Big) \right\| + \sum_{i=n}^{m(\tau_n+T)-1} \alpha_{i+1} \mathbb{E}\|\delta_{i+1}\| \tag{13}$$

$$= R_{2,n} + K_n, \tag{14}$$

where $\mathcal{F}_i$ is the filtration defined in Assumption 6. Now, by Assumption 6 (vi) and square integrability of $\|\delta_n\|$, $R_{2,n}$ is the supremum on another martingale difference noise term with bounded variance, just as R_n. Thus, again for $R_{2,n}$ we have almost sure convergence to zero. As for K_n, recall from Assumption 6 (iv) that $\mathbb{E}\|\delta_n\| = o(b_n)$. Hence, there exists some $C_K > 0$ such that for all n large enough,

$$\sum_{i=n}^{m(\tau_n+T)-1} \alpha_{i+1} \mathbb{E}\|\delta_{i+1}\| \leq C_K \sum_{i=n}^{m(\tau_n+T)-1} \alpha_{i+1} b_{i+1} \leq \sum_{i=n}^{m(\tau_n+T)-1} \alpha_{i+1}^2,$$

by assumption that $\lim_{n\to\infty} \frac{\alpha_n}{b_n} = 0$. Thus, by square summability of α_i, the sum above must converge to zero in n, and therefore $K_n \to 0$ as well, and the result (12) follows.

Thus, ρ_n is almost surely a perturbed solution to $\dot{\rho} \in F_g(\rho(t))$. The result then follows from Hofbauer and Sandholm (2002, Theorem 3.6), which states that the set of convergent subsequences of any perturbed solution to F_g is an ICT set of F_g. □

Next, to prove convergence to an attractor $\{\rho^*\}$ with positive probability, a stronger result than Proposition 4 is first needed:

Assumption 7 (Condition 11, Faure and Roth (2010))**.** *There exists a map $\omega : \mathbb{R}_+^3 \mapsto \mathbb{R}_+$ such that*

(1) For any $\varepsilon > 0$, $T > 0$,

$$\mathbb{P}\left(\sup_{m' \geq n} \sup_{m' \leq k \leq m(\tau_{m'}+T)} \left\| \sum_{i=n}^{k-1} \alpha_{i+1} \Big(\delta_{i+1} + M_{i+2} \Big) \right\| > \varepsilon \,\middle|\, \mathcal{F}_n \right) \leq \omega(n, \varepsilon, T),$$

[17] See e.g. Faure and Roth (2010, Proposition 2.16).

almost surely in $\mathcal{F}_0$.

(2) $\lim_{n\to\infty} \omega(n, \varepsilon, T) = 0$.

Faure and Roth (2010, Proposition 2.16) states that Condition 11 above is satisfied for our M_{n+1} martingale difference sequence (i.e. if $\delta_n = 0$ for all n). I show next that this result extends to our case:

Lemma 3. *Condition* 11 *is satisfied under Assumption 6.*

Proof. Note first that

$$\begin{aligned}
&\left\| \sum_{i=n}^{k-1} \alpha_{i+1} \Big(\delta_{i+1} + M_{i+2} \Big) \right\| \\
&\leq \left\| \sum_{i=n}^{k-1} \alpha_{i+1} \Big(M_{i+2} \Big) \right\| + \left\| \sum_{i=n}^{k-1} \alpha_{i+1} \Big(\delta_{i+1} \Big) \right\| \\
&= R_n + \Psi_n^k,
\end{aligned}$$

similarly as stated in the proof above. For R_n, Faure and Roth (2010, Proposition 2.16) immediately applies, as it only requires the Robbins-Monro condition on α_n, and that Assumption 6 is satisfied for M_n. For Ψ_n^k, recall (13) from the previous proof. As noted there, $R_{2,n}$ is another bounded martingale-difference noise, so Faure and Roth (2010, Proposition 2.16) applies as well. Finally, K_n is a determinstic sequence converging to zero as shown in the previous proof, so that the probability of the term being larger than any fixed $\varepsilon > 0$ is always zero for large enough n. The result follows. □

Finally, Faure and Roth (2010, Theorem 2.15) states that if Assumption 7 is satisfied, $\mathbb{P}[L_S = \{\rho^*\}] > 0$ holds as long as $\{\rho^*\}$ is *attainable* by the process ρ_n:

Definition 11. *A point* p *is attainable if, for any* $n > 0$ *and any neighborhood* U *of* p

$$\mathbb{P}[\exists s \geq n : \rho_s \in U] > 0.$$

Let $Att(X)$ be the set of attainable points for algorithm (10). Then we need that the basin of attraction of an attractor has nonempty intersection with $Att(X)$. This can be verified:

Lemma 4. *Let B be a basin of attraction of an attractor A for $F(\rho)$. Suppose $\rho_n \in \overline{\mathbf{X}} \setminus B$. Then there exists $s > n$ such that $\rho_s \in B$ with positive probability.*

Proof. Since n is finite, to show existence we construct $s = n + 1$: For any $z \in B$, one can pin down the necessary shock M_z to reach it:

$$M_z \in \frac{z - \rho_n}{\alpha_n} - F(\rho_t),$$

since F might be multivalued.

By finiteness of M_z, M_z is in the support of M_{n+1} for every n. For any ball B_z around z, define

$$\mathbf{M}_z = \{M_{x'} : x' \in B_z\}.$$

$\mathbf{M}_z$ must have positive measure for all finite n, since it is in the support of M_{n+1}. (if we allow $s > n + 1$, we may be able to increase the measure, but we only need it to be positive.) □

Thus, Faure and Roth (2010, Theorem 2.15) applies, concluding this proof. □

Proof of Theorem 2]

Proof. We can apply Benaïm and Faure (2012, Theorem 3.12) to prove $\mathbb{P}[L_S = \{\rho^*\}] = 0$ in the following. The conditions and analysis sufficient for the proof of Benaïm and Faure (2012, Theorem 3.12) are local with respect to ρ^*. Thus, the fact that F is globally potentially multivalued is of no importance, since in a small enough neighborhood around ρ^* it must be single-valued and $\mathcal{C}^1$ (see Definition 1).

Benaïm and Faure's result is concerned with time-interpolations of stochastic differential inclusions $F(\rho)$ satisfying Assumption 6 (i), such as (11). Their Theorem 3.12 states, translated in terms of this paper, that under an Assumption the authors refer to as Hypothesis 2.2, and Assumption 6 $(ii), (iii)$, the result to be proved here holds true.

In fact, Benaïm and Faure (2012, Hypothesis 2.2) is equivalent[18] to Assumption 7, which was shown to hold for our algorithm in Lemma 3. Thus, the result applies, concluding the proof. □

[18]See Faure and Roth (2010, Remark 2.14)

Proof of Proposition 1

Proof. First, note the following fact about block symmetric matrices.

Remark 1. *Suppose A, B are square matrices of the same dimension. Let*

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

Then one can show

$$det(T) = det(A - B)det(A + B).$$

Given a square matrix A, define Λ as the set of eigenvalues of the A. Then define

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

as the spectral radius of A.

Lemma 5. *Suppose $\alpha^* = \beta^* = \sigma^*$ is an interior, symmetric equilibrium. Let $\bar{\kappa}$ be the real part of the spectral radius of $J(\sigma^*)$, the matrix of best response derivatives of a given player (by symmetry, the identity is irrelevant). Then σ^* is asymptotically stable if $\bar{\kappa} < 1$, and unstable if $\bar{\kappa} > 1$.*

Proof. Using Remark 1, we get that

$$ch(\lambda) = det(J(\sigma^*) - (1 + \lambda)I_2)det(J(\sigma^*) + (1 + \lambda)I_2).$$

Thus, if μ is an eigenvalue of $J(\sigma^*)$, then $\pm|\mu - 1|$ is an eigenvalue of $X(\sigma^*)$, and the conclusion follows, since asymptotic stability requires that all eigenvalues of $X(\sigma^*)$ have negative real parts. □

Hence, it is enough to characterize the eigenvalues $\lambda_{1,2}$ of the matrix of best-response derivatives of player 1 at symmetric Nash policies ρ_N:

$$J_N = \begin{bmatrix} BR'_N + \frac{\delta P'_{AB}(\rho_N)}{\omega_N}\frac{u_2^N}{u_{11}^N} & -\frac{\delta P'_{AB}(\rho_N)}{\omega_N}\frac{u_2^N}{u_{11}^N} \\ -\frac{\delta P'_{BA}(\rho_N)}{\omega_N}\frac{u_2^N}{u_{11}^N} & BR'_N + \frac{\delta P'_{BA}(\rho_N)}{\omega_N}\frac{u_2^N}{u_{11}^N} \end{bmatrix}.$$

We have that

$$\lambda_{1,2} = \frac{tr(J_N)}{2} \pm \sqrt{\frac{tr(J_N)^2}{4} - det(J_N)},$$

where $tr(\cdot), det(\cdot)$ represent trace and determinant. Thus, $\lambda_1 = BR'_N$, and $\lambda_2 = BR'_N + \delta \frac{P'_{AB}(\rho_N)+P'_{BA}(\rho_N)}{\omega_N} \frac{u_2^N}{u_{11}^N}$. Regularity gives that $|\lambda_1| < 1$, so that $|\lambda_2| > 1$ appears as the condition in the Proposition. □

Proof of Proposition 2]

Proof. First, we prove that given $\mathcal{G}$, u can be regular:

Lemma 6. *Suppose $g \in \mathcal{G}$. Then there exist a convex cost function $c(x)$ such that the resulting stage game payoffs $u(x_1, x_2)$ are regular.*

Proof. By definition of $\mathcal{G}$, we only need to construct the cost function $c(x)$ to satisfy Definition 5 points $(ii), (v)$. Fix some $g \in \mathcal{G}$. Now, pick a cost function satisfying $(i), (ii)$. Note that for x_M as defined in Definition 5 (v), it must be that $Y(x_M) > 0$. (See (i)). Thus, as long as $c'(0) < Y(x_M)$, we can guarantee that (v) holds. Finally, $\mathcal{G}$ satisfies Definition 5 $(iii), (iv)$ by Definition 9 (iv), so that there must be a unique interior Nash equilibrium (x_N, x_N), which is symmetric. □

Recall the following conventions:

- $u^s = u(\rho_s, \rho_s)$, for $s \in \mathbf{S}$.
- $u_k^s = \frac{\partial u^s}{\partial x_k}$ and $u_{kk'}^s = \frac{\partial u_k^s}{\partial x_{k'}}$, for $k, k' = 1, 2$, $s \in \mathbf{S}$.
- $P'_{sB} = \frac{\partial P_{sB}}{\partial x_1} = \frac{\partial P_{sB}}{\partial x_2}$ for all s and analogously for P''_{sB} where the equality comes from the fact that P_{sB} only depends on aggregate quantities.
- $G_2(y; X) \equiv \frac{\partial}{\partial X} g(y; X)$.

For every $X \in int(\mathbf{X})$, define $\overline{y}(X)$ such that $\eta(\overline{y}(X), X) = 0$, which exists by Definition 9 (i). Pick $y^* = \overline{y}(X_N)$ as a price cutoff, for $X_N = 2x_N$, given interior static Nash equilibrium x_N. Let (y^*, y^*) be the symmetric cutoff for a binary state variable following f_{DS} transitions. Thus, we construct a consistent DS-state variable with $P_{AB}(X) = P_{BA}(X) = Pr[p \leq y^* \mid X] = G(y^*; X)$, and fix this state variable throughout the remainder of the proof. Also, define $h(X) = P_{ss'}(X)$ for $s \neq s' \in \mathbf{S}$, to save notation. We now prove a helpful Lemma.

Lemma 7.

(1) There exists $M^* \leq M$ such that for all $\hat{x} \in [0, M^*]$ there exists a unique $x^*(\hat{x}) \in [0, M^*]$ such that

$$u_1(x^*(\hat{x}), \hat{x}) = 0.$$

(2) For all $x \in (0, x_N]$ there exists a unique $\hat{x} \in [x_N, M)$ such that

$$\frac{u_1(x,x)}{h'(2x)} + \frac{u_1(\hat{x},\hat{x})}{h'(2\hat{x})} = 0.$$

(3) For all $x, \hat{x} \in (0, M)$

$$\frac{u_1(q,\hat{x})}{h'(x+\hat{x})} - \frac{u_1(\hat{x},\hat{x})}{h'(2\hat{x})} = 0$$

has a unique solution at $x = \hat{x}$.

Proof. For the first claim, notice that Definition 9 (iv) implies that $u_{12}(x, x') < 0$ for all $x, x' \in \mathbf{X}$, and therefore best responses must be strictly decreasing whenever positive. Thus, there exists $M^* \leq M$ (choose M large enough) so that $u_1(0, M^*) = 0$, and $x = 0$ is the best response to $x' \geq M^*$.

For the second claim, note that convexity of c and $u_{12}(x, x') < 0$ implies that $u_1(x, x)$ is strictly decreasing for all $x \in \mathbf{X}$, crossing 0 at x_N. By construction of y^*, note that by Definition 9 (ii), $G_2(y^*; X) > 0$ for all $X \in int(\mathbf{X})$, with peak at X_N. Thus, the fraction $\frac{u_1(x,x)}{h'(2x)} \in (-\infty, \infty)$ is strictly decreasing over $x \in \mathbf{X}$, and the claim follows.

For the third claim, consider two cases:

Case 1: $\hat{x} \leq x_N$.

Notice that $\hat{x} < x_N$ implies $u_1(\hat{x}, \hat{x}) > 0$, and as shown for the first claim, $u_1(x, \hat{x})$ is monotone decreasing on the candidate solutions $x \in [0, x^*(\hat{x})]$. Larger x are not candidates, due to the sign change of u_1. In the following I will write $x^* = x^*(\hat{x})$ for brevity. Define $\overline{x} = x_N - \hat{x}$. Note that case 1 implies $x^* \geq x_N$, which in turn implies

$$Y(x^* + \hat{x}) + Y'(x^* + \hat{x})x^* \geq Y(X_N) + Y'(X_N)\overline{x},$$

by Definition 9 (iv), and since $\overline{x} \leq x_N$ in this case. Thus, $x^* \leq \overline{x}$, which implies that for all $x \in [0, x^*]$, $h'(x + \hat{x})$ is increasing. Thus, $\frac{u_1(q,\hat{x})}{h'(x+\hat{x})}$ is strictly decreasing on $x \in (0, x^*(\hat{x}))$. By monotonicity there can only be one solution, $x = \hat{x}$.

Case 2: $\hat{x} \in (x_N, M]$.

Here, note that $u_1(\hat{x}, \hat{x}) < 0$, and so all candidate solutions x must satisfy $x \in (x^*, M]$. For $\hat{x} \leq X_N$ we get analogously to above, that now $\overline{x} \leq x^*$. This implies $h'(x+\hat{x})$ is decreasing on the set of candidate solutions, and again we arrive at strict monotonicity of $\frac{u_1(q,\hat{x})}{h'(x+\hat{x})}$.

For $\hat{x} > X_N$, we have immediately that $h'(x + \hat{x})$ is decreasing for all $x \in \mathbf{X}$, and the result follows. □

Now we need the following observations based on the definition of W in (1):

$$
\begin{aligned}
W_1 &= \omega^{-1}(1 - \delta P_{BB})\Big[\omega^{-1}\delta P'_{AB}(u^B - u^A) + u_1^A\Big],\\
W_2 &= \omega^{-1}(\delta P_{AB})\Big[\omega^{-1}\delta P'_{BB}(u^B - u^A) + u_1^B\Big],\\
W_{11} &= -2\omega^{-1}\delta P'_{AB} W_1 + \omega^{-1}(1 - \delta P_{BB})\Big[\omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A\Big],\\
W_{22} &= 2\omega^{-1}\delta P'_{BB} W_2 + \omega^{-1}(\delta P_{AB})\Big[\omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B\Big],\\
W_{12} &= \omega^{-1}\delta\Big[P'_{AB}\frac{1 - \delta P_{BB}}{\delta P_{AB}} W_2 - P'_{BB}\frac{\delta P_{AB}}{1 - \delta P_{BB}} W_1\Big],
\end{aligned}
\tag{15}
$$

Then, an optimal, non-degenerate, interior strategy α^* must satisfy

$$
\begin{aligned}
W_1(\alpha^*, \beta) = 0 &\iff \omega^{-1}\delta P'_{AB}(u^B - u^A) + u_1^A &&= 0,\\
W_2(\alpha^*, \beta) = 0 &\iff \omega^{-1}\delta P'_{BB}(u^B - u^A) + u_1^B &&= 0,\\
W_{11}(\alpha^*, \beta) < 0 &\iff \omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A &&< 0,\\
W_{22}(\alpha^*, \beta) < 0 &\iff \omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B &&< 0.
\end{aligned}
$$

Notice that for all such α^*, we also have $W_{12}(\alpha^*, \beta) = 0$. This follows under irreducibility, since then initial states do not affect the optimal policy choice.

Now for the proof of the Proposition:

Firstly, note that finding interior σ such that $W_1(\sigma) = W_2(\sigma) = 0$ is equivalent to finding σ such that

$$W_1(\sigma) = 0; \quad \frac{u_1^A}{h'(X_A)} + \frac{u_1^B}{h'(X_B)} = 0.$$

From now on, to save notation, I write h_s to denote evaluation of $h()$ at $x_s, s \in \{A, B, N\}$. By Lemma 7 we have that for any $x_A \in (0, x_N]$ there exists a unique $x_B \in [x_N, M)$ such that

$$\frac{u_1^A}{h'_A} + \frac{u_1^B}{h'_B} = 0.$$

We will call such $x_B = z(x_A)$. By strict monotonicity we can apply the implicit function theorem to get

$$z'(x_A) = -\frac{h'_B}{h'_A} \frac{u_{11}^A + u_{12}^A - 2h''_A \frac{u_1^A}{h'_A}}{u_{11}^B + u_{12}^B + 2h''_B \frac{u_1^B}{h'_B}}. \tag{16}$$

It is then not surprising that at x_N, $z'(x_N) = -1$. Now define $\Psi(x_A) = W_1(x_A, z(x_A), x_A, z(x_A))$ as the first order condition of W with respect to x_A, substituting in $z(x_A)$ so that at every x_A,
$W_2(x_A, z(x_A), x_A, z(x_A)) = W_1(x_A, z(x_A), x_A, z(x_A))$ must hold. Thus, any zero of $\Psi(x_A)$ must set both first order conditions to zero.

Since ρ_N^{DS} is always a solution, we have that $\Psi(x_N) = 0$, i.e. one zero always exists. We will now show that for small x, $\Psi(x) > 0$ holds, while for large x, $\Psi(x) < 0$. The sufficient condition stated in this Proposition is then the condition ensuring $\Psi'(x_N) > 0$, which ensures that there must be another zero with $x_A < x_N$.

Firstly, recall that by regularity of u, for $x > 0$ small enough, $u_1(x, x) > 0$ must hold. Now consider $\Psi(x_A)$:

$$\Psi(x_A) > 0 \Leftrightarrow \omega^{-1}\delta h'(2x_A)(u^B - u^A) + u_1^A > 0.$$

Then since $h'(0) = 0$ we get that the first term must be dominated by the second term for $x_A > 0$ small enough, which is positive.

Next, and analogously, take $x_A \in (x_N, M)$ to be large. In that case, we let $y(x_A) = z^{-1}(x_A) < x_N$ be the inverse solution that equalizes first order conditions. Then if $x_A < M$ large enough, we get that the first term must be dominated by the second term since $h'(M) = 0$, and the second term is negative by definition of $D < M$. Finally to prove that $\Psi'(x_N) > 0$, note that

$$\begin{aligned}
\Psi'(x_N) &= W_{11}^N + W_{13}^N + W_{14}^N z'(x_N) = W_{11}^N + W_{13}^N - W_{14}^N \\
&= \omega^{-1}(1-\delta(1-h_N))\Big[u_{11}^N + u_{12}^N - \omega^{-1}\delta h'_N u_2^N + \omega^{-1}\delta h'_N u_2^N z'(x_N)\Big] \\
&= \omega^{-1}(1-\delta(1-h_N))\Big[u_{11}^N + u_{12}^N - 2\omega^{-1}\delta h'_N u_2^N\Big] \\
&= \omega^{-1}(1-\delta(1-h_N))u_{11}^N\Big[1 + \frac{u_{12}^N - 2\omega^{-1}\delta h'_N u_2^N}{u_{11}^N}\Big].
\end{aligned}$$

Since $u_{11}^N < 0$, we have $\Psi'(x_N) > 0$ if

$$\begin{aligned}
&1 + \frac{u_{12}^N - 2\omega^{-1}\delta h'_N u_2^N}{u_{11}^N} < 0 \\
\Leftrightarrow \quad & 2\omega^{-1}\delta h'_N u_2^N - u_{12}^N < u_{11}^N \\
\Leftrightarrow \quad & 2\delta h'_N u_2^N - \omega u_{12}^N < \omega u_{11}^N \\
\Leftrightarrow \quad & 2\delta h'_N u_2^N - 2\delta h_N u_{12}^N < 2\delta h_N u_{11}^N + (1-\delta)(u_{12}^N + u_{11}^N).
\end{aligned}$$

Thus we can write

$$\begin{aligned}
&1 + \frac{u_{12}^N - 2\omega^{-1}\delta h'_N u_2^N}{u_{11}^N} < 0 \\
\Leftrightarrow \quad & \frac{h'_N}{h_N} Y'_N x_N < 3Y'_N + 2Y''_N x_N - c''_N + R \\
\Leftrightarrow \quad & -\frac{h'_N}{h_N} < \frac{c''_N - 2Y''_N x_N}{Y'_N}\frac{1}{x_N} - \frac{3}{x_N} + R,
\end{aligned}$$

where for the last line, we used that $u_1^N = 0 \Rightarrow Y'_N x_N = c'_N - Y_N < 0$, and where $R = \frac{1-\delta}{2\delta}(u_{12}^N + u_{11}^N)$ vanishes as $\delta \to 1$. Now we need to show that this inequality can be satisfied for some $g \in \mathcal{G}$.

Lemma 8. *For any $g \in \mathcal{G}$ there exists $\tilde{g}$ differing from g only on a neighborhood of (y^*, X_N), so that under this $\tilde{g}$, $\Psi'(x_N) > 0$ holds.*

Proof. Assume we start from $g(y;X)$ such that the condition fails:

$$\frac{h'_N}{h_N} < -\frac{c''_N - 2Y''_N x_N}{Y'_N}\frac{1}{x_N} + \frac{3}{x_N}.$$

We will perturb $g(y;X)$ so as to flip the inequality in our benefit. To this end, we will use that the left hand side depends directly on the c.d.f. evaluated at a point y^*, while the right hand side depends only on an integral over all $p \in \mathbf{Y}$ of the c.d.f.. Note that

$$h'(X) = \int_0^{y^*} g_2(y;X)dp.$$

For a small neighborhood N_1 of y^*, let $\mu = \ell(N_1)$ be the Lebesgue-measure, and let μ_C be the Lebesgue-measure of $[\sup N_1, \overline{Y}]$. Define, for $\Delta > 0$,

$$\tilde{g}_2(y;X) = g_2(y;X) + \Delta \mathbf{1}\{p \in N_1,\ Q \in N_2\} - \Delta\frac{\mu}{\mu_C}\mathbf{1}\{p \geq \sup N_1,\ Q \in N_2\}, \tag{17}$$

where N_2 is a small neighborhood of X_N. Say that the perturbation is feasible if $\tilde{g}$ remains a density:

$$\tilde{g}(y;X) > 0 \,\forall\, p \in \mathbf{Y}; \quad \int_{\mathbf{Y}} \tilde{g}_2(y;X)dp = 0.$$

I will show that this perturbation is feasible for N_1 small enough relative to Δ, ensuring that $\tilde{g}(y;X) > 0$ remains true; the construction (17) ensures that $\tilde{g}_2$ integrates to zero. Define $[\underline{y}, \overline{y}] = N_1$, $[\underline{X}, \overline{X}] = N_2$. It follows that

$$\tilde{G}_2(y^*, X_N) = G_2(y^*, X_N) + \Delta(y^* - \underline{y}),$$

and

$$\tilde{G}(y^*, X_N) = G(y^*, X_N) + \Delta(y^* - \underline{y})(X_N - \underline{X}).$$

Let $\mu_1 = (y^* - \underline{y})$, $\mu_2 = (X_N - \underline{X})$, we can write

$$\frac{\tilde{h}'_N}{\tilde{h}_N} = \frac{h'_N + \Delta\mu_1}{h_N + \Delta\mu_1\mu_2}.$$

Now for the expected price:

$$\begin{aligned}\tilde{Y}_N &= \overline{Y} - \int_{\mathbf{Y}} \tilde{G}(p; X_N)dp \\ &= Y_N - \int_{N_1} \Delta(p - \underline{y})\mu_2 dp \\ = Y_N - \Delta\mu_2\big(\frac{1}{2}(\overline{y}^2 - \underline{y}^2) - \underline{y}(\overline{y} - \underline{y})\big) &= Y_N - \frac{1}{2}\Delta\mu_2\mu^2,\end{aligned}$$

where the first equality is due to integration by parts. Then,

$$\begin{aligned}\tilde{Y}'_N &= -\int_{\mathbf{Y}} \tilde{G}_2(p; X_N)dp \\ &= Y'_N - \frac{1}{2}\Delta\mu^2,\end{aligned}$$

and $\tilde{Y}''_N = Y''_N$ for all $Q \neq \underline{X}, \overline{X}$. We get that

$$\begin{aligned}&\frac{h'_N + \Delta\mu_1}{h_N + \Delta\mu_1\mu_2} < -\frac{c''_N - 2Y''_N x_N}{\tilde{Y}'_N}\frac{1}{x_N} + \frac{3}{x_N} \\ \Leftrightarrow &\frac{h'_N + \Delta\mu_1}{h_N + \Delta\mu_1\mu_2} < -\frac{c''_N - 2Y''_N x_N}{Y'_N - \frac{1}{2}\Delta\mu^2}\frac{1}{x_N} + \frac{3}{x_N}.\end{aligned}$$

If we choose Δ increasing, μ_1, μ_2 decreasing so that $\Delta\mu_1$ increases, while keeping $\Delta\mu^2$ and $\Delta\mu_1\mu_2$ constant, this inequality can be flipped. However, we need to ensure that in so doing, $\tilde{g}$ remains positive. Note

$$\tilde{g}(y; X) = g(y; X) + \Delta(X - \underline{X})\mathbf{1}\{p \in N_1, X \in N_2\} - \Delta\frac{\mu}{\mu_C}(X - \underline{X})\mathbf{1}\{p \geq \overline{y}, X \in N_2\},$$

and thus, for all $Q \in N_2$, the decrease in $\tilde{g}(y; X)$ can be controlled via $\Delta\mu_2$ and $\Delta\mu\mu_2$. We can always find three sequences $\Delta_j, \mu_{1,j}, \mu_{2,j} > 0$ for all j such that $\Delta_j\mu_{1,j}$ increases, $\Delta_j\mu_{1,j}^2$ decreases, $\Delta_j\mu_{2,j}$ is weakly increasing, and $\Delta_j\mu_{1,j}\mu_{2,j}$ decreases.[19]

By choosing these sequences as above, it follows that $\frac{\tilde{h}'_N}{\tilde{h}_N}$ increases, while keeping the right hand side above constant, and also keeping $\tilde{g}(y; X) > 0$ for all p, X. We have arrived at a $\tilde{g}(y; X) \in \mathcal{G}$ under which $\Psi'(x_N) > 0$. □

[19] E.g., for some $c_1, c_2, c_3 > 0$, let $\Delta_j = c_j j$, $\mu_{1,j} = c_2 j^{-b}$, $\mu_{2,j} = c_3 j^{-b}$, where $b \in (\frac{1}{2}, 1)$. Note also that N_1 can be chosen so that $\mu = \frac{1}{2}\mu_{1,j}$ for all j, preserving the same order of magnitude.

Now, $\Psi'(x_N) > 0$ together with $\Psi(x) > 0$ for x small, $\Psi(x) < 0$ for x large, allows us to use the intermediate value theorem. It follows that there exists $x_A < x_N < x_B$ such that $W_1(\sigma) = W_2(\sigma) = 0$ for $\sigma = (x_A, x_B, x_A, x_B)$.

We are left to show that this zero is a global maximizer. Firstly, we note that the Hessian at σ must be negative definite: we see from (15) that $W_{12} = 0$, so the Hessian must be diagonal at σ. A sufficient condition for negative definiteness then is $h''_A > 0 > h''_B$ and $u^A > u^B$. The first one follows given Definition 9 (ii) and since $x_A < x_N < x_B$, the second one follows from the first order conditions:

$$W_1 = 0 \;\Rightarrow\; u^A - u^B = \omega \frac{u_1^A}{\delta h'(X_A)} > 0.$$

Now we have that σ is a local max, and we can consider one-shot deviations to show that it is global. In state A, we need to show that

$$\begin{aligned}(1-\delta)u(x_A, x_A) + \delta\Big[W^A + h_A\big(W^B - W^A\big)\Big] \\ \geq (1-\delta)u(x, x_A) + \delta\Big[W^A + h(x + x_A)\big(W^B - W^A\big)\Big],\end{aligned}$$

holds for all $x \in \mathbf{X}$, where we take the shorthand W^S to indicate the value function at state s evaluated at the policy (x_A, x_B). Equivalently, we can show that $x = x_A$ is the unique solution to the first order condition of this problem with respect to x, and that boundary conditions are satisfied so that the maximizer can only be interior. Taking derivatives, we get

$$H^A(x, x_A) = (1-\delta)u_1(x, x_A) + \delta h'(x + x_A)\big(W^B - W^A\big).$$

By construction, $H^A(x_A, x_A) = 0$.

Since the Hessian is negative definite at x_A, x_A, $H_1^A(x_A, x_A) = \frac{\partial H^A(x, x_A)}{\partial x}\Big|_{q=x_A} < 0$. Recall that in the proof of Lemma 7 we showed that x_A is the only solution to $H^A(x, x_A) = 0$, but also that $\frac{u_1(x, x_A)}{h'(x+x_A)}$ is strictly decreasing over $x \in [0, x^*(x_A)]$. Thus, $H^A(0, x_A) > 0$ and $H^A(M/2, x_A) < 0$ must hold and x_A is globally optimal.

Now, in state B we do the analogous argument, take derivatives to get

$$H^B(x, x_B) = (1-\delta)u_1(x, x_B) - \delta h'(x + x_B)\big(W^B - W^A\big).$$

Where again by the negative definite Hessian, we have $H_1^B(x, x_B) < 0$. Then in the proof of Lemma 7 we show that $\frac{u_1(x,x_A)}{h'(x+x_A)}$ is strictly decreasing over $x \in [x^*(x_B), M/2]$. The result follows as above: x_B is globally optimal.

We have shown that playing $\sigma = (x_A, x_B)$ is the unique best reply to an opponent playing σ, and thus σ is a symmetric equilibrium as required. □

Proof of Lemma 1

Proof. First, since we are restricting to symmetric equilibria, it is sufficient to consider two cases: $u^A \lessgtr u^B$. Since we consider consistent 1R policies, let the unique threshold be x.

i) $u^A > u^B$.

Recall that state A corresponds to observing a price below x. As laid out in the proof of Proposition 2, we can write an agent's FOC for the problem of best responding in the following way:

$$W_1 = 0 \Leftrightarrow \frac{\delta h'(X_A)}{\omega}\big(u^A - u^B\big) + u_1^A = 0;$$
$$W_2 = 0 \Leftrightarrow \frac{\delta h'(X_B)}{\omega}\big(u^A - u^B\big) + u_1^B = 0,$$

where we plug in the fact that $P_{AB}(X) = 1 - h(X) = Pr[p > x]$. For both equations, the leading term is strictly positive, since $h'(X) > 0$ for all interior X (recall Definition 9 $(iii), (iv)$). It follows that $u_1^s < 0$ must hold for both s.

In the proof of Lemma 7 I show that we have that $\frac{u_1(x,x)}{h'(2x)}$ is strictly decreasing for all $x \in [0, M]$. At the same time, $u(x, x)$ is strictly decreasing for all x, which is necessary for $u_1(x, x) < 0$. Thus, for case (i) it must be that $x_A > x_B$, but since $\frac{u_1(x,x)}{h'(2x)}$ is strictly decreasing, there exists no such pair x_A, x_B to set $W_1 = W_2$. It follows that no such pair can be an equilibrium.

The case $u_A < u_B$ follows from an analogous argument. □

Proof of Lemma 2]

Proof. To save notation, write $J^* = J(\sigma, \sigma)$, where J is the matrix of best-response derivatives of player 1 at symmetric policies σ.

This definition allows one to write, for any interior equilibrium profile σ as constructed in Proposition 2,

$$\begin{aligned}
\frac{\partial \rho_A^{*1}}{\partial \rho_A^2} &= -1 + \phi_A^{-1}\Big[\Pi'_A - \omega^{-1}\delta P'_{AB}\Pi_A\Big],\\
\frac{\partial \rho_A^{*1}}{\partial \rho_B^2} &= \phi_A^{-1}\omega^{-1}\delta P'_{AB}\Pi_B,\\
\frac{\partial \rho_B^{*1}}{\partial \rho_A^2} &= \phi_B^{-1}\omega^{-1}\delta P'_{BA}\Pi_A,\\
\frac{\partial \rho_B^{*1}}{\partial \rho_B^2} &= -1 + \phi_B^{-1}\Big[\Pi'_B - \omega^{-1}\delta P'_{BA}\Pi_B\Big],
\end{aligned} \tag{18}$$

where ρ^{*1} indicates 1's best-response policy, s-subscripts denote evaluation at x_s, and

$$\begin{aligned}
\phi_A &= \omega^{-1}\delta P''_{AB}(u^B - u^A) + u^A_{11};\\
\phi_B &= \omega^{-1}\delta P''_{BB}(u^B - u^A) + u^B_{11},
\end{aligned}$$

Some tedious algebra then allows re-writing determinant and trace of $J(\sigma, \sigma)$ using (18), Then:

$$\begin{aligned}
tr\big(J^*\big) &= -2 + \frac{\Pi'_A}{\phi_A}\big[1 - R_A\big] + \frac{\Pi'_B}{\phi_B}\big[1 - R_B\big];\\
det\big(J^*\big) &= \left[1 - \frac{\Pi'_A}{\phi_A}\frac{\Pi'_B}{\phi_B}\right] - \frac{\Pi'_A}{\phi_A}\big[1 - R_A\big]\left[1 - \frac{\Pi'_B}{\phi_B}\right] - \frac{\Pi'_B}{\phi_B}\big[1 - R_B\big]\left[1 - \frac{\Pi'_A}{\phi_A}\right].
\end{aligned} \tag{19}$$

Notice that for the stage game as constructed in Proposition 2, $\phi_s < u^s_{11}$ holds, and therefore $\frac{\Pi'_s}{\phi_s} \in (0, 1)$ can be guaranteed as long as $u^s_{12} \leq 0$, since $\Pi'_s = u^s_{11} - u^s_{12}$. Sign and magnitude of R_s depend on local conditions of both transition probabilities and the stage game quantity $\Pi(x_1, x_2)$. It is clear from (19) that both trace and determinant depend crucially on the quantities R_s. Indeed, if R_A, R_B are not too negative, stability of σ follows:

Firstly, by Lemma 5, stability of σ is equivalent to

$$\left|tr\left(J^*\right)\right| - det\left(J^*\right) < 1. \tag{20}$$

Then, note from (19) that by the condition of the Proposition,

$$tr\left(J^*\right) < -R_A - R_B$$

and so for R_A, R_B not too negative, we must have that $tr\left(J^*\right) \leq 0$. Next note that we can write

$$det\left(J^*\right) = -tr\left(J^*\right) - 1 + \frac{\Pi'_A}{\phi_A}\frac{\Pi'_B}{\phi_B}\left[1 - R_A - R_B\right].$$

Thus, for R_A, R_B bigger than 0, the trace drops out in the condition in (20). The last equation then determines stability through the term $\left[1 - R_A - R_B\right]$. □

Proof of Proposition 3]

Proof. For any discretization $\mathbf{X}_K$, define $W^K(\sigma, z) : \mathbf{X}^2 \times \mathbf{Y}^2 \mapsto \mathbb{R}$ as restriction of the payoff function to $\mathbf{X}_K$:

$$W^K(\sigma, z) = W(f^K(\sigma), z),$$

where

$$f^K(\sigma) = \arg\min_{\sigma' \in \mathbf{X}_K^2} \left\|\sigma - \sigma'\right\|,$$

for any norm on $\mathbf{X}^2$, the projection of σ onto discrete space $\mathbf{X}_K$.

For every sequence $\mathbf{X}_K$ there is an associated sequence α_K with

$$\alpha_K = \max_{(\sigma,z)\in\mathbf{X}^2\times\mathbf{Y}^2} \left\|W^K(\sigma, z) - W(\sigma, z)\right\|.$$

Continuity of W implies that $\alpha_K \to 0$. Write $\alpha_K(\mathbf{X}_K)$ for a sequence given a fixed sequence of discretizations. Say that a discretization sequence $\mathbf{X}_K$ is *covering* if $\alpha_K(\mathbf{X}_K) \to 0$ (and $x_N \in \mathbf{X}_K$). From now on, fix some $z \in \mathbf{Y}^2$, and a covering sequence of discretizations $\mathbf{X}_K$.

Notice that $E_K(z)$ is closed-valued, trivially by finiteness of $\mathbf{X}_K$. Furthermore, $E^*(z)$ is closed-valued: W is continuous, $\mathbf{X}$ compact, and thus Berge gives us that the best-response correspondence is closed and compact-valued. Then, applying the closed-graph theorem gives us that the equilibrium set $E^*(z)$, as a set of fixed points of a closed and compact correspondence, must be closed. To get to claim (1), I will show that any converging sequence $\sigma_K \in E_K(z)$ has its limit in E^*. In other words, an upper hemicontinuity property holds for the equilibriunm correspondence along sequences of covering discretizations.

Lemma 9. *For all sequences* $\{\sigma_K\}$ *with* $\sigma_K \in E_K(z)$,

$$\alpha_K \to 0,\ \sigma_K \to \bar{\sigma} \ \Rightarrow\ \bar{\sigma} \in E^*(z).$$

Proof. Suppose not. Then there exists a subsequence $\sigma_{K_t} \in E_{K_t}(z)$ with $\sigma_{K_t} \to_t \bar{\sigma} \notin E^*(z)$. The converging subsequence exists since $\mathbf{X}^2$ is compact. To ease notation, re-define $k = k_t$ for the rest of the proof. Not being an equilibrium, we have that there exists $\sigma_z \neq \bar{\sigma}$ that maximizes the deviation payoff

$$\Delta_z = W(\sigma_z, \bar{\sigma}, z) - W(\bar{\sigma}, \bar{\sigma}, z) > 0.$$

Pick $\varepsilon \in (0, \Delta)$. By convergence of σ_K , and by continuity of W, we have that there exists $K_{1,z}$ such that for all $K \geq K_{1,z}$,

$$\left|W(\sigma_z, \sigma_K, z) - W(\sigma_z, \bar{\sigma}, z)\right| \leq \frac{\varepsilon}{3}. \tag{21}$$

By the same argument, there is a $K_{2,z}$ s.t. for all $K \geq K_{2,z}$,

$$\left|W(\sigma_K, \sigma_K, z) - W(\bar{\sigma}, \bar{\sigma}, z)\right| \leq \frac{\varepsilon}{3}. \tag{22}$$

Furthermore, we can always choose $\bar{K}_z \geq \max\{K_{1,z}, K_{2,z}\}$ large enough so that $\alpha_K \leq \frac{\varepsilon}{3}$, implying

$$\left|W(f^K(\sigma_z), \sigma_K, z) - W(\sigma_z, \sigma_K, z)\right| \leq \frac{\varepsilon}{3}. \tag{23}$$

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

$$\hat{\sigma}_K = \arg \max_{\sigma \in \mathbf{X}_K^2 \setminus \sigma_K} W(\sigma, \sigma_K, z).$$

Now we have

$$\begin{aligned} W(\hat{\sigma}_K, \sigma_K, z) - W(\sigma_K, \sigma_K, z) &\geq W(f^K(\sigma_z), \sigma_K, z) - W(\sigma_K, \sigma_K, z) \\ &= W(\sigma_z, \sigma_K, z) - W(\sigma_K, \sigma_K, z) + \beta_{1,K} \\ &= W(\sigma_z, \bar{\sigma}, z) - W(\bar{\sigma}, \bar{\sigma}, z) + \beta_{1,K} + \beta_{2,K} + \beta_{3,K} \\ &\geq \Delta + \beta_{1,K} + \beta_{2,K} + \beta_{3,K}, \end{aligned}$$

where $\beta_{1,K}$ corresponds to the projection error (23), and $\beta_{2,K}, \beta_{3,K}$ correspond to (21),(22) respectively. We have that $|\beta_{i,K}| \leq \frac{\varepsilon}{3}$, and thus

$$W(\hat{\sigma}_K, \sigma_K, z) - W(\sigma_K, \sigma_K, z) \geq \Delta - \varepsilon > 0,$$

implying that $\sigma_K \notin E_K$, a contradiction. □

To finish the proof, note that Lemma 9 implies that for any feasible α_K, $\lim_{K\to\infty} F_z(\alpha_K) \subseteq E^*(z)$, with $E^*(z)$ being the continuous-action version of the equilibrium set for fixed z. We get that

$$\limsup_{K\to\infty} V_K(z) \leq V_z,$$

with $V_K(z), V_z$ being the maximal payoff over the equilibrium sets $E_K(z), E^*(z)$. The inequality holds for every z, and therefore also holds when taking maximum over z on both sides, and claim 1 is proven.

For claim 2, the claim to prove is that when all equilibria in E^* are strict, lower hemicontinuity property holds for the sequence of equilibrium correspondences $E_K(z)$. Fix z, then the proof is via contradiction: there exists some strict equilibrium $\sigma \in E^*(z)$ that is not approximated by any sequence of equilibria in $E_K(z)$. The proof can be done analogously to the one above; defining $\Delta_z > 0$ as the best deviation payoff:

$$\Delta_z = W(\sigma, \sigma, z) - \max_{\mathbf{X}^2 \setminus \sigma} W(\sigma_z, \sigma, z) > 0.$$

Since $\Delta_z > 0$, we can find large enough discretizations s.t. σ can be approximated arbitrarily closely, in which case incentives must also align, by continuity of W. The

contradiction follows. Hence, together with Lemma 9, we get that

$$\lim_{K\to\infty} \sup_{\substack{\sigma_K \in F(\alpha_K) \\ z\in \mathbf{Y}^2}} W(\sigma_K, z) = \sup_{\substack{\sigma^* \in F(0) \\ z\in \mathbf{Y}^2}} W(\sigma^*, z) = \sup V.$$

□

References

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

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

Assad, Stephanie et al. (2020). "Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market". In.

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

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

Benaïm, Michel, Josef Hofbauer, and Sylvain Sorin (2005). "Stochastic approximations and differential inclusions". In: *SIAM Journal on Control and Optimization* 44.1, pp. 328–348.

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

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

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

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

Chernozhukov, Victor, Han Hong, and Elie Tamer (2007). "Estimation and confidence regions for parameter sets in econometric models 1". In: *Econometrica* 75.5, pp. 1243–1284.

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.

Filipov, Aleksei Fedorovich (1988). "Differential equations with discontinuous right-hand side". In: *Amer. Math. Soc*, pp. 191–231.

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

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

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

Hahn, Frank H (1962). "The stability of the Cournot oligopoly solution". In: *The Review of Economic Studies* 29.4, pp. 329–331.

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

Hofbauer, Josef and William H Sandholm (2002). "On the global convergence of stochastic fictitious play". In: *Econometrica* 70.6, pp. 2265–2294.

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

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

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

Leslie, David S, 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 denBoer (2023). "Data-driven collusion and competition in a pricing duopoly with multinomial logit demand". In: *Production and Operations Management* 32.4, pp. 1169–1186.

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

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

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

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

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

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

Plappert, Matthias et al. (2017). "Parameter space noise for exploration". In: *arXiv preprint arXiv:1706.01905*.

Possnig, Clemens (2022). "Learning to Best Reply: On the Consistency of Multi-Agent Batch Reinforcement Learning". URL: https://cjmpossnig.github.io/papers/marlbatchconv_CPossnig.pdf.

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

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

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

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

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

Yang, Tianpei et al. (2021). “Exploration in deep reinforcement learning: a comprehensive survey”. In: *arXiv preprint arXiv:2109.06668*.