Reinforcement Learning and Collusion

Clemens Possnig
University of Waterloo
August 23, 2023
Link to current version

Abstract. This paper presents an analytical characterization of the long run policies learned by algorithms that interact repeatedly. These algorithms update policies which are maps from observed states to actions. I show that the long run policies correspond to equilibria that are stable points of a tractable differential equation. As a running example, I consider a repeated Cournot game of quantity competition, for which learning the stage game Nash equilibrium serves as non-collusive benchmark. I give necessary and sufficient conditions for this Nash equilibrium not to be learned. These are requirements on the state variables algorithms use to determine their actions, and on the stage game. When algorithms determine actions based only on the past period's price, the Nash equilibrium can be learned. However, agents may condition their actions on richer types of information beyond the past period's price. In that case, I give sufficient conditions such that the policies converge with positive probability to a collusive equilibrium, while never converging to the Nash equilibrium.

JEL classification. 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 and patience.

1. Introduction

More and more companies are using algorithms to try to optimize sales and increase profits. Such algorithms take market data to determine current price or quantity levels, updating in real-time. Software firms such as solutions.ai and Quicklizard state that a well-performing algorithm needs to "*deliver real-time insights based on market signals, competitive intelligence and changes in customer preferences (...)*" and "*trigger repricing of items based on a criteria such as (...) competitor price changes (...)* ". What outcomes can we expect when algorithms compete against each other?

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. This is a concern for consumer welfare. Moreover, legal systems are not currently adapted to deal with the kind of tacit collusion that might result from algorithmic competition. An antitrust regulator would need to understand how competing algorithms might lead to inefficient outcomes, depending on the market conditions, the nature of the competitive environment, and the details of the algorithms themselves. This paper seeks to better understand these issues.

I first introduce a model of reinforcement learning algorithms playing a market game such as Cournot quantity competition repeatedly. These 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 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 payoffs satisfy the previous requirement.

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 characterised in Abreu, Pearce, and Stacchetti (1986). I provide sufficient conditions such that this scheme will be learned with positive probability.

Characterisation of long-run behavior. The focus of this paper is actor-critic reinforcement learning. These algorithms keep track of an estimated performance criterion (the "critic", 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. As a result, the class of algorithms studied here have the ability to learn repeated game strategies (e.g., take the observable to be a summary of the history of the interaction), in contrast to the stage-game (myopic) strategies more commonly studied in the literature on learning in games.

When policies are updated with a decreasing stepsize[3] over time, and the performance criterion estimator is well-behaved, I show that the resulting policy iteration can be analyzed as a noisy discretization of a tractable differential equation. Under a well-studied special case in my class known as actor-critic Q-learning[4] (ACQ), this differential equation is a repeated game version of best-response dynamics. I obtain these results following the method of stochastic approximation (V. S. Borkar (2009)), which I extend to characterize the limiting behavior of competing algorithms.

Suppose that after a large enough number of iterations, the performance criterion estimator differs from the true criterion at most by a bounded, smooth bias term[5]. I then show that attractors of the underlying differential equation will be learned with positive probability, while unstable points will not be learned[6]. In the remainder of the paper, I use ACQ as the running example. In that case, if the long run behavior of algorithms converges to a point, that point must be a Markov-perfect equilibrium (MPE) of the repeated stage

[3]Stepsizes signify the impact innovations have on policies in the algorithmic updating rule. They must satisfy the Robbins-Monro condition commonly invoked in the computer science literature. See V. S. Borkar (2009), Chapter 2.

[4]See Dutta and Upreti (2022), Grondman et al. (2012) for relevant surveys.

[5]The bias term is a modelling decision inspired by the fact that for many real-world performance criterion estimators, bias is unavoidable due to function approximation (Fujimoto, Hoof, and Meger (2018)). Also, often convergence proofs are lacking, in which case the fitness of an algorithm is shown by it doing better at benchmark tasks than previous algorithms. C.f. the discussion in Chapter 9 of François-Lavet et al. (2018). My results imply that the analysis of long run behavior is robust to well-behaved, non-vanishing bias.

[6]An equilibrium of a differential equation is attracting ("stable") if trajectories of the differential equation that reach a neighborhood of that equilibrium converge to it. In contrast, an equilibrium is repelling ("unstable") if trajectories close to that equilibrium can be repelled from it and will not converge to it.

game. The implication of this characterisation is that it becomes necessary to understand what it means for a given MPE to be stable (i.e. attracting).

Learning Nash. Intuitions about stability for more general MPEs can be gained from the study of the repeated static Nash equilibrium. I show that in order to learn the stage game Nash equilibrium, details of the state variables algorithms keep track of are essential. I provide the first step at a novel categorization of the "coordinative ability" granted to learning algorithms by their state variables. This is done by studying what kinds of state variables allow for algorithms to learn the stage game Nash equilibrium. To the best of my knowledge, this paper is the first to uncover this aspect of cooperative ability.

A definition of state variables comes with the space of their realizations and a transition function that pins down how the states evolve over time as a function of actions taken by the algorithms. Call a policy space the space of decision rules mapping from a given state variable to actions. Different policy spaces may support different MPEs, but stage game equilibria are MPEs under any policy space, making comparative statics exercises viable.

Stability of an MPE is the characterising factor that enables us to tell whether a given MPE will be observed. Sufficient conditions for an MPE to be learned can therefore be constructed based on eigenvalue properties of the MPE. It turns out that the conditions for stability come down to a comparison between the slope of the (static) stage-game best response and a growth rate of transition probabilities with respect to action deviations. When transition probabilities are more sensitive than a cutoff defined through the static best response slope, the stage game equilibrium will be rendered unstable, and therefore not learned.

Using the example of repeated Cournot competition, I compare two binary state variables that in a sense have opposing transition features. Suppose that there is a finite set of possible price realizations, and conditional on the aggregate quantity produced, a price is drawn randomly and independently every period. I first consider simple state variables that feature *state-independent transitions*. This holds when past period's price observations are taken as state variable, since prices are drawn independently every period conditional on aggregate quantities. Given this state variable, I prove that a condition on market fundamentals is necessary and sufficient for a given stage game Nash equilibrium to be learned. In other words, sensitivity of transition probability does not matter in this case, but only the

slope of the stage game best response.[7] If the above bound on myopic best responses is satisfied for a given stage game Nash equilibrium, I call it *statically stable*. If for a given state variable the Nash equilibrium is stable under the resulting state-dependent best-response dynamics (and therefore learned with positive probability), I call it *dynamically stable*.

When state transitions are state-independent, information about the state evolution is irrelevant when determining whether a stage game equilibrium will be learned or not. Thus, under such state variables, static and dynamic stability is always the same. In contrast, I then characterise a state variable I call *direction-switching* so that when policies condition on those, the stage game Nash equilibrium will not be learned. A Nash equilibrium can thus be both statically stable and dynamically unstable.

Learning to collude. Next, I apply my characterisation to study collusion in the workhorse model of Cournot-competition. I provide conditions on payoffs and observables under which collusive equilibria exist that are attracting, and therefore will be learned with positive probability. Through an approximation exercise I show that these collusive equilibria are closely related to optimal imperfect monitoring equilibria of the bang-bang kind, as characterised in Abreu, Pearce, and Stacchetti (1986).

Exemplifying this approach, I show for a class of payoff functions that there exists a simple binary state variable (falling into the class of direction switching states) for which there exists a collusive equilibrium that is attracting. This collusive equilibrium has the carrot-and-stick property, where in one state, low quantities are played, which are supported by high punishment quantities in the other state. In addition, for this class of payoff functions it is true that the unique stage game Nash equilibrium is statically stable and at the same time dynamically unstable. The result then ties in with my second contribution discussed above: when Nash is not learned, then what is learned can likely be a collusive outcome.

Finally, I provide a numerical example featuring the above properties: there is an attracting, collusive equilibrium as well as a unique stage game Nash equilibrium that is unstable. I verify in a simulation that ACQ learners initialized in a neighborhood of the collusive equilibrium will indeed converge to it. I also simulate these algorithms initialized

[7]This condition also determines the stability of that equilibrium under myopic best-response dynamics. This refers to classical best-response dynamics that consider the learning of stage game strategies. In common textbook-versions of the Cournot game there is a unique, interior, and symmetric stage game Nash equilibrium that satisfies this condition (e.g. linear demand and convex cost, under some boundary conditions preventing the monopoly equilibrium to exist). A long history of research has established that many learning dynamics converge uniquely to this equilibrium, when learning to play myopic, stage game strategies. This is also true for fictitious play and myopic best-response dynamics (c.f. Milgrom and Roberts (1990)).

close to the stage game Nash equilibrium, and show how they not only do not converge to that equilibrium, but instead move to a neighborhood of the collusive equilibrium.

The results presented here potentially allow empirical researchers and industry regulators to understand what conditions of the market and features of algorithms lead to a greater likelihood of collusive behavior when competing firms use learning algorithms. My characterisation implies that once one knows the state variables of the algorithms, and the payoff environment of the game, one can determine whether any MPE in the policy space will be learned with positive probability by checking its stability property, which comes down to an eigenvalue-condition[8]. Using data on the firm's payoff structure, market demand, and the states kept track of by the algorithms, one can device a test to see whether a given equilibrium can be learned with positive probability.

Related Literature

Broadly speaking, this project speaks to results in the fast growing literature on algorithmic collusion, the theory of learning in games, as well as the study of asymptotic behavior of algorithms in the computer science literature.

Firstly, the literature on algorithmic collusion has received increasing attention in recent years. Assad et al. (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. Important recent work in the area of algorithmic collusion includes Lamba and Zhuk (2022), Z. Y. 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. Their relation to this work is discussed more thoroughly in Section 6.

Secondly, this paper connects to a long history of 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

[8]Specifically, one needs to linearize the state-dependent best response dynamics at the equilibrium. If all eigenvalues' real part is strictly smaller than 0, the equilibrium is attracting, if some are strictly larger than 0, it is repelling. If some are equal 0, the equilibrium is called a *center* in the literature and more analysis has to be done to determine whether it may be learned or not, but this is a non-generic knife-edge situation.

(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 RL I study has the ability to learn *repeated game strategies*, i.e. strategies that condition on summaries of the history of the game. 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, commonly called "model-free". The running example of ACQ learning considered in the main body of this paper is part of this special case. Such algorithms do not carry a model of opponent behavior and incentives, and also no model of their environment and own payoffs. Thus, this class can be seen as models of players following adaptive uncoupled learning rules as defined in Hart and Mas-Colell (2003). Further foundational papers in this literature include Milgrom and Roberts (1990) , Fudenberg and 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 for example V. S. 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).

Finally, this paper can be interpreted as casting RL competition as an equilibrium selection mechanism. The classical literature was developed as a model to understand how rational players may learn to play Nash equilibria, whereas here I consider real economic agents that happen to be algorithmic and show that their behavior can be understood through the theory of learning in games. Interestingly, among the repeated game equilibrium selection criteria known to me there exists none that exclude the stage game Nash equilibrium even when it is unique, which suggests that the selection ability of competing RL delivers new insights. I refer to Fudenberg and Levine (2009) for a thorough review of issues regarding the theory of learning in games, including algorithmic learning and applications of stochastic approximation.

This paper is structured as follows: In section 2 I give a brief introduction to RL, via the classical example of single-agent Markov-decision problems (MDPs). In section 3 I define the general economic environment our algorithms will play on, as well as ACQ learning, an important element of my RL class that will serve as the running example of the paper. I provide general limiting results in section 4. In section 5, I apply the results of the previous

section to a repeated Cournot game, and give a numerical examples with simulations in the end of the section. Section 6 concludes and discusses related literature more thoroughly.

This papers' structure is designed to maintain a logical progression across its sections. Sections 3-4 give general long-run characterisations for a class of algorithms. Building upon these results, section 5 employs this framework to examine its application in a Cournot game.

For readers more interested in the economic issue of collusion among algorithms, section 5 can be read independently of the previous sections. Sections 3-4 cater more towards readers seeking comprehensive statistical treatments of algorithmic updating processes.

Since this paper relies on technical methods that would overwhelm the main body, some sections are moved to the appendix. Appendix A characterises the full algorithm class that can be considered by this paper. Appendix B gives technical results regarding the determination of asymptotic stability of equilibria under ACQ learning, while Appendix **??** gives most of the proofs of the results stated in the paper.

2. Reinforcement Learning

This section gives a short introduction to reinforcement learning (RL) by ways of the example of an agent solving a multi-armed bandit problem. For a thorough introduction, consider Sutton and Barto (2018).

Consider an agent choosing actions $a \in X$ to repeatedly. There is a state variable $s \in S$ so that in every possible s, the agent may find it best to choose different a. Given s, the agent's expected payoff from choosing a is denoted $u(a, s)$. The agent discounts the future with $\delta \in (0, 1)$, and aims to find a policy $\rho : S \mapsto X$ that maximizes future expected discounted payoffs

$$W(s_0) = \mathbb{E} \sum_t \delta^t u_t,$$

where u_t is the payoff realization in period t. When the distribution over states and other randomness affecting the payoffs is known, the agent can solve the problem of maximizing W by computing the value function

$$V(s) = \max_{a \in X} \Big\{ u(a, s) + \delta \mathbb{E}\big[V(s') \big| \, a, s\big] \Big\}.$$

In practice, information about u and transition probabilities may be hard to come by. This is where RL methods can be useful.

RL algorithms are updating rules meant for the learning of optimal policies or value functions for a given problem. Such algorithms are commonly used to solve Markov decision

problems (MDPs). In general, RL updating rules move policies towards actions that have performed well in the past (i.e., such actions are *reinforced*), and away from actions that perform poorly, based on some performance criterion, e.g. W.

A well-known algorithm the agent could use in this context is Q-learning as introduced by C. J. C. H. Watkins (1989). The algorithm estimates a function $Q : S \times A \mapsto \mathbb{R}$, which is supposed to find the target implicitly defined as

$$Q^*(s,a) = u(a,s) + \delta \mathbb{E}\big[\max_{a' \in X} Q^*(s',a') \big| \, a, s\big]. \tag{1}$$

This Q- function is related to the value function by $V(s) = \max_{a \in X} Q^*(s,a)$. The Q-function can thus be seen as a function evaluating the expected payoff from selling a in current state s and playing optimally afterwards. One may therefore use this function to evaluate one-shot-deviations. Accordingly, Q^* is a helpful tool for decision makers, since it allows to read-off the optimal policy ρ^* simply by maximizing Q^* in every state.

C. J. C. H. Watkins (1989) then proposed a RL algorithm that estimates Q^*. In the language introduced earlier, the algorithm takes estimates of Q^* as the relevant performance criterion. This algorithm is celebrated due to its simplicity as well as minimal information requirement. One can use the algorithm without any knowledge of a payoff function and transition function, thus falling into the class of 'model-free' algorithms.

For simplicity let S, A be finite, so Q^* is a matrix. In the end of each period, the payoff realization u_t, current state s_t, current action taken a_t, and the next state s_{t+1} are observed. The algorithm takes some initial value Q_0, and then updates the following way:

$$Q_{t+1}(s,a) = \begin{cases} Q_t(s,a) + \beta_t \Big[u_t + \delta \max_{a' \in X} Q_t(s_{t+1}, a') - Q_t(s,a) \Big] & \text{if } s_t = s,\ a_t = a \\ Q_t(s,a) & \text{otherwise} \end{cases}, \tag{2}$$

where $\beta_t \geq 0$ is a (possibly stochastic) sequence of numbers converging to zero. Importantly, notice that Q-learning does not specify a policy, just a performance criterion. Convergence results on Q_t give requirements on how often actions are selected over time, but generally the updating rule is agnostic about how actions a_t are sampled in every period. As an agent who cares about behaving optimally, a clear exploration-exploitation tradeoff arises in this problem: should one follow the currently-believed optimal action, or try to find actions that may perform better? A common, basic sampling method is known as ε-greedy:

Fix a small $\varepsilon \in (0,1)$. In every period, the decision maker takes the currently believed optimal action $\arg\max_{a'} Q_t(s_t, a')$ with probability $1 - \varepsilon$. With probability ε, she samples uniformly from A.

For a suitable sequence β_t, one can show that Q_t converges in probability to Q^* if states form a Markov chain controlled by a_t and actions are sampled ε-greedily (c.f. C. J. Watkins and Dayan (1992))[9]. Stationarity of the state-transitions conditional on a fixed policy ρ is an important ingredient of the standard convergence proof for Q- learning. If stationarity fails, one can imagine that learning of the correct Q^* may fail also.

3. The Multi-Agent Setting

Now imagine the player described above in fact faces multiple competitors in a market, which transforms our MDP into a game. Without any knowledge about their payoff function, state transitions, and opponents, the player may resort to Q-learning again. What if all players in the game apply this method to learn their optimal policies?

The purpose of this section is to characterize a class of algorithms for which it is possible to analytically describe the limiting policies resulting from agents in that class competing against each other. As will be seen in the following sections, this requirement will be an important part of the characterisation of the class of RL considered here. The class does not contain the above described simple Q-learning rule, but a common evolution of it, that explicitly adapts a policy function ρ_t at the same time as estimating the Q-function, making it an actor-critic Q-learning (ACQ) rule as described in the Introduction. In this section and the main body of the paperter, I focus on ACQ for clarity. The general class of algorithms is broader and fully defined in Appendix A.

In general I allow the algorithms to be model-free. This translates to a restriction on how performance criterion estimates are generated. Model-free algorithms maintain estimates of their performance criterion without explicitly modeling their own or their opponent behavior and payoffs. This serves as a minimal-information benchmark, which will be shown to be sufficient to lead to the emergence of collusion.[10] There are multiple reasons why this is a difficult situation for such agents when it comes to learning a good policy, as discussed in the introduction, and also in Hernandez-Leal et al. (2017). I will be abstract about the estimation of the performance criterion, and introduce a class of algorithms that perform reasonably well in the function approximation step, up to a well-behaved asymptotic bias term. I believe that allowing for an asymptotic bias significantly increases the number of learning algorithms that fall into our class of RL agents, due to the inherent problems these agents face while learning.

[9]This convergence result for single-agent problems has been studied extensively, and it holds generally as long as all actions and states are visited sufficiently often.

[10]Furthermore, model-free algorithms can be thought of as a tool for a firm that recently entered a new, dynamic market. Information on payoffs and market conditions may be hard to come by, so such a firm may resort to an algorithm that has minimal information requirements.

The results in this section are concerned with algorithms that learn to play continuous action policies[11]. In that case unbiasedly estimating a value function becomes a difficult task even when stationarity of the environment is satisfied. Commonly in such situations algorithms use some form of parametric function approximation to generate an estimate, which can introduce bias. Often this involves deep neural networks due to their flexibility and scaleability. I refer to François-Lavet et al. (2018) for a thorough introduction to state of the art RL techniques and a deeper dive into issues of biased estimation of value functions and their gradients. I will show that the bias I allow in this class does not affect the main results developed in the next section.

There are n algorithms indexed by i, each having as action space a compact interval X_i, with profile space $X = \times_i X_i$. A finite state space S with $|S| = L$[12] comes with a transition probability function $T : S^2 \times X \mapsto (0, 1)$ where I will maintain throughout the paper that each state space considered is irreducible, as specified below. 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 $u^i : X \times S \mapsto \mathbb{R}$, $\mathcal{C}^2$[13] in X, and common discount factor $\delta \in (0, 1)$.

Algorithms update a policy function $\rho_t^i : S \mapsto X_i$, the long-run behavior of which is the object of our interest. Since states are discrete, policy profile $\rho_t \in \bar{X} = X^{nL}$ can be represented as a vector in $\mathbb{R}^{nL}$.

Assumption 1. *For all $\rho \in \bar{X}$, the Markov chain induced by $T_{ss'}[\rho(s)]$ is irreducible and aperiodic.*[14]

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

[11] This is not as restrictive as might seem. When playing discrete action games, RL algorithms commonly play on the mixed policy space, for example learning to play 'softmax' strategies of the form $\mathbb{P}[q|\, s] = \frac{exp(Q(s,q))}{\sum_{q'} exp(Q(s,q'))}$. This again falls into our continuous control scenario.

[12] While it may be possible to carry out an analogous characterisation of long-run policies under compact interval domains, the interpretability of the results would likely suffer. RL algorithms commonly used in the case of interval-state spaces take the policy to be a parametric function of the state, and optimize the parameters rather than the policy itself (c.f. Sutton and Barto (2018), Chapter 13), which introduces an issue of interpretability. At the same time, since this section is not concerned with speed of convergence or computational constraints, one can always take a fine enough discretization of an interval domain and the analysis in this section applies.

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

[14] For Definitions see e.g. Appendix A in Puterman (2014)

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

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \tag{3}$$

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

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

$$B_S^i(\rho^{-i}) = \arg\max_{\rho \in \bar{X}_i} W^i(\rho, \rho^{-i}, s_0), \tag{4}$$

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 dynamic policies since given a Markov stationary opponent profile ρ^{-i} there must be a Markov stationary best response.

Definition 1. *Define*

(i) $E_S \subset \bar{X}$ *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.* $\rho^{*i} \in B_S^i(\rho^{*-i})$ *for all* i*.*

(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.*

By definition, if $\rho^* \in E_S$ is a differential Nash equilibrium then 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 on the underlying payoff environment, the following assumption is introduced:

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

(i) Given state space S*, stationary equilibrium profiles* $\rho^* \in \bar{X}$ *exist. Call the set of such equilibria* E_S*.*

(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 an interior static Nash equilibrium given $u(a, s)$ for all $s \in S$. 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.

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

state space domain. When an algorithm is defined in our class, it comes with a finite state variable S as a primitive. I will take S as an exogenous object chosen by whoever initialized the algorithm. Importantly, I will assume throughout that the state variable and current state s is a common observable to all algorithms competing. The state variable can be interpreted as a model of what kind of information the RL is allowed to condition their policy on.

Now to state the running example of RL studied here. Assume that each algorithm uses the following adaptive rule to update their policy, which is known as actor-critic Q learning (ACQ):[15][16]

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

$$\rho_{t+1}^i(s) \in \rho_t^i(s) + \alpha_t \Big[\arg\max_{a' \in X} Q_t^i(s, a') - \rho_t^i(s) + M_{t+1}^i \Big], \tag{5}$$

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'[17][18]*.*

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

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

the correct 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

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

[15] I focus on the algorithm in Definition 2 because it forms the basis of many well-behaved real world algorithms, see for example Fujimoto, Hoof, and Meger (2018) who introduce an algorithm based on ACQ used in real-world applications. Other algorithms of interest that can be accommodated include gradient-type algorithms. A full exposition can be found in Appendix A.

[16] Notice that Definition 2 does not exclude the case in which the function to be approximated is fully known, or there is no bias term. 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.

[17] Since our main interest is in algorithms used under incomplete knowledge of the environment, the non-vanishing variance of M_{t+1} can be motivated constructively by a need to explore the policy space due to estimation requirements on the one hand, and residual randomness due to the fact that performance criterion Q^* is being estimated. 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).

[18] Notice that I use '$\in$' instead of '$=$' above, since I allow for the possibility of the argmax having multiple values. If that is indeed the case, I allow the algorithm to pick arbitrarily, which will not affect the limiting characterisation in ways that matter, as will be seen in section 4.

Q_t^i is motivated from stationary MDPs as introduced in subsection 2. It is important to note the use of this estimator in the Multi-agent case faced here imposes an implicit behavioral assumption on each algorithm. Suppose that $Q_t^i(s,a) = Q^{i*}(s,a,\rho_t^{-i})$, i.e. the estimator is perfectly correct. Then what the agent computes in their updating step (5) is a best response in stationary strategies *supposing that the opponents hold their current profile* ρ_t^{-i} *fixed forever into the future*. Having read that every algorithm uses (5) to update their policies, this supposition is clearly incorrect. However, firstly as stepsizes for ρ_t^i decrease, computing Q^* can be seen as an approximation to the true future expected value that would take into account evolving ρ_t^{-i}. Secondly, as stated before, this section is not concerned with a normative theory of how an optimal algorithm should behave. Rather, the interest is in developing a model that is realistic enough while staying analysable, and forms an informational lower benchmark on algorithms in the sense that they can be allowed to be model-free. As will be shown in section 5, the assumptions made here will be sufficient to allow for collusive and other interesting behavior to emerge.

The following assumption ensures that Q_t^i tracks the correct function Q^{i*} well when t is large enough. The classical Q- estimator (2) defined to motivate Q-learning will not be enough for this to be true, as it requires discretization of the continuous action space and may run into issues due to the underlying non-stationarity of the problem. However, more involved estimation schemes exist for which Q_t^i can be shown to track Q^{i*}, as shown e.g. in Possnig (2022).

Assumption 3. *For each i there exists a bounded function $g^i(s,a,\rho^{-i})$, $\mathcal{C}^2$ in a,ρ^{-i} such that we can define*

$$\chi_t^i \equiv \sup_{(s,a)\in S\times X,} \left\| Q_t^i(s,a) - Q^{i*}(s,a,\rho_t^{-i}) - g^i(s,a,\rho_t^{-i}) \right\|.$$

Assume for each i:

(i)

$$\chi_t^i = O_P(t^{-\frac{1}{2}}),$$

(ii)

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

In words, estimators Q_t^i converge uniformly to a biased version of Q^{i*}. Point (i) bounds the convergence speed by $t^{-\frac{1}{2}}$, and point (ii) ensures that large errors have negligible mass,

which is important in the approximation results established in the next section. For appropriate function approximation schemes, the convergence speed can also be shown to satisfy point (i).

Assumption 3 puts discipline on the limiting difference between Q_t and Q^*. Importantly, it is assumed that there exists a well-behaved function g, independent of t, that represents this limiting difference in expectation in the long run. Furthermore, this limiting difference, or asymptotic bias, is a function of t only through its dependence on current period's profile ρ_t. Thus, Assumption 3 allows for an asymptotic bias term in the Q_t estimation. Results in the long run characterisations of subsection 4 impose this bias to be small enough so that limits of ρ_t can be inferred via a model in which $g = 0$. This robustification means it is sufficient for researchers to verify smoothness and bound a possible asymptotic bias, without needing to know the specific functional form of g. As stated in the beginning of this section, g is introduced for the sake of realism and in order to significantly increase the number of RL algorithms that can be analysed in this paper.

Assumption 3 is not trivial, since it sweeps away the issue of non-stationarity when it comes to estimating Q_t discussed before. However, firms expecting to compete in a non-stationary environment can be readily assumed to prefer algorithms that can satisfy this Assumption over basic Q_t algorithms as in (2). There exist however more involved algorithms that can adapt to both of these issues, as shown in Possnig (2022).

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 represenation (Robbins and Monro (1951)). The assumption constrains the speed of convergence of α_t, needing to balance the averaging out of errors (i.e. be 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 Michel Benaïm and Faure (2012). For a more general discussion of sufficient conditions for bounded iterates, see V. S. 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. V. S. Borkar (2009)) to connect the long-run behavior of ρ_t to limiting sets of solutions to an underlying differential equation. Given Assumption 3, one can convince oneself that this differential equation will have to do with the computation of a best response. This is indeed the case, as will become clear shortly.

4. Long Run Behavior

Throughout, maintain Assumptions 2 - 5.

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

$$L_{S,g} = \bigcap_{t \geq 0} \overline{\{\rho_s \,|s \geq t\}},$$

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

I write S, g as subscript to underline the dependence of the limiting set on the state space S and bias function g, both of which are implied by the specification of the algorithms in use. As the characterisations 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* $\lambda \in \Lambda$.

As mentioned in the discussion following Assumption 3, I will now show that as long as the asymptotic bias term g is sufficiently bounded, one can analyse limits $L_{S,g}$ equivalently by analysing the special case $L_{S,0}$, i.e. a situation where $g = 0$ everywhere. For $\gamma > 0$, let $\mathcal{B}^k_\gamma$ be the set of C^k functions with bounded derivatives :

$$\mathcal{B}^k_\gamma = \Big\{ g : \bar{X} \mapsto \mathbb{R}^{nL} \mid \sup_{x \in \bar{X}} \|g(x)\| + \sum_{j=1}^{k} \sup_{x \in \bar{X}} \|D^j g(x)\| \leq \gamma \Big\}, \tag{6}$$

where $D^j g$ represents the j'th derivative.

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

$$F^S_B(\rho) = \bar{B}_S(\rho) - \rho, \tag{7}$$

as the state dependent best response dynamics, where I take $\bar{B}_S(\rho)$ to be the stacked version of $B_S^i(\rho^{-i})$ over i.

Proposition 1. *Let $\rho^* \in \mathcal{U}_S$ be asymptotically stable for F_B^S. Then for all γ small enough and all $g \in \mathcal{B}_\gamma^1$ there is a profile ρ^g such that*

(1) $\sup_{g \in \mathcal{B}_\gamma^1} |\rho^g - \rho^*| \to 0$ *as* $\gamma \to 0$.

(2) $\mathbb{P}[L_{S,g} = \{\rho^g\}] > 0$.

Proof Sketch of Proposition 1

The full proof for this and the following Propositions can be found in Appendix **??**.

Firstly, I make a general connection between the recursion in (5) and the differential inclusion F_B^S. This follows from celebrated results in stochastic approximation theory. One can relate a time-interpolated version of the recursion ρ_t to solutions of the differential inclusion

$$\dot{\rho} \in F_g(\rho(t)) \equiv conv[F_B^S(\rho(t))] + g(\rho(t)),$$

where for any set B, $conv[B]$ represents the convex closure.

Since the best-response may be multi-valued, solutions to this inclusion are not guaranteed. However, assumptions on the regularity of F_B^S (which comes down to a linear growth condition) 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 (5) looks similar to a discrete time approximation to a time-derivative. The idea then is to show that the time-interpolated version of ρ_t indeed must stay close, with probability one, to solutions of an underlying differential inclusion. The limiting behavior of ρ_t can then be deduced from a subset of the limiting behaviors of the differential inclusion above.

The proof of the Proposition then establishes a firm connection between ρ^* and ρ^g. I use a more general version of the inverse function theorem to show that since $g(\rho)$ is a well behaved, differentiable bias term, for every ρ^* there is a unique rest point ρ^g. Further, stability of ρ^* must carry over to stability of ρ^g. Once it is established that ρ_t tracks solutions to the above inclusion over time, it makes sense that attracting points of the differential system will also attract ρ_t over time.

Since I allow for the estimation of performance criterion Q^* used by each algorithm to be biased, when considering the long run of ρ_t one may only see ε-equilibria of the game, defined below:

Definition 5. *A profile ρ is an ε-equilibrium if for all players i all individual profiles $\rho' \in \bar{X}$ and states $s \in S$*

$$W^i(\rho, s) \geq W^i(\rho', \rho^{-i}, s) - \varepsilon.$$

Corollary 1. *Let $\rho^* \in E$ be asymptotically stable for F_B^S. Then for all γ small enough and all $g \in \mathcal{B}_\gamma^1$ there is a $\bar{\varepsilon} > 0$ and a profile ρ^g such that*

(1) ρ^g is an ε-equilibrium for all $\varepsilon \geq \bar{\varepsilon}$.

(2) $\sup_{g \in \mathcal{B}_\gamma^1} |\rho^g - \rho^*| \to 0$ as $\gamma \to 0$.

(3) $\mathbb{P}[L_{S,g} = \{\rho^g\}] > 0$.

Notice that the stability property of an equilibrium depends on the performance criterion F_B^S used by the underlying algorithm, and is not affected by the bias term g as long as it is well-behaved, i.e. $g \in \mathcal{B}_\gamma^1$. The stability of ρ^* itself depends further on the state space observed by the algorithms. I therefore emphasize this dependence by writing F_B^S as the best response dynamics defined on state space S.

Proposition 2. *Let $\rho^* \in \mathcal{U}_S$ be linearly unstable for F_B^S. Then for all γ small enough and all $g \in \mathcal{B}_\gamma^1$ there is an open neighborhood U_γ with $\rho^* \in U_\gamma$ such that*

$$\mathbb{P}[L_{S,g} \in U_\gamma] = 0.$$

Proof Sketch of Proposition 2

Firstly, as in the proof of Proposition 1, I establish a one to one relationship between the stability properties of ρ^* and the rest points ρ^g. ρ^g being unstable hyperbolic implies that there exists an unstable manifold that ρ^g lies on, which acts as a repeller to the differential inclusion F_g. I go on to show that due to the instability of ρ^g and nonvanishing variance of M_{t+1}, no matter how close the algorithm updates come to ρ^g, and no matter how large t is, there is always a high probability that ρ_t lands on the unstable manifold and therefore must move away from ρ^g. Finally I show the existence of a neighborhood U_γ. I show that due to the hyperbolicity of ρ^*, ρ^g, there is a neighborhood U around ρ^g with $\rho^* \in U$ such that ρ^g is the only internally chain transitive set within U. Recall that ρ^* is not internally chain transitive for the perturbed system F_g, and the result follows.

Corollary 1 and Proposition 2 show the full potential of our characterisation. 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_B^S, their policy will only stay close to ρ^* if the dynamics of F_B^S are somehow robust to deviations. This robustness is implied by asymptotic stability, and broken by unstable equilibria.

There is a caveat here however: Corollary 1 does not state that all limiting points in $L_{S,g}$ will be equilibria of the game. Depending on details of the environment, 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 Proposition 1 to the case of attracting cycles as in Faure and Roth (2010), and there exist results considering linearly unstable cycles (Michel Benaïm and Faure (2012)) that suggest one may extend Proposition 2 to such linearly unstable cycles also.[19]

5. Learning to Collude

In this section, I study a repeated Cournot game played by RL algorithms falling into the family of ACQ learners, as introduced in the previous section. The characterization provided in section 4 allows to analyze the long run behavior of ACQ learners playing this game. As a non-collusive benchmark, and to introduce intuitions, I ask when the static Nash equilibirum (Cournot equilibrium) can be learned. 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 only, 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 payoffs satisfy the previous requirement.

Recall that in section 4, I established a link between state dependent best-response dynamics F_B^S as defined in (7), and long-run behavior of ACQ learners. In section 4 (Corollary 1 and Proposition 2), I show the following:

"A given markov perfect equilibrium profile will be learned by ACQ learners with positive probability if it is attracting under F_B^S, and with zero probability if it is unstable under F_B^S."

Therefore, this section will consider the existence and stability of equilibria under F_B^S in a repeated Cournot game more closely. This game can be shown to satisfy Assumptions 1 and 2 as a consequence of the model setup and Lemma 1. It follows that whenever an ACQ algorithm satisfies Assumptions 3 and 4, the long run characterisations of section 4 apply.

The game is set up as follows:

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

- 2 agents $i \in \{1, 2\}$.
- Stochastic binary price outcome $Y \in \mathbf{Y} = \{P_L, P_H\}$.
- Quantity choice $q \in X = [0, M]$ for some large $M > 0$, with aggregate quantity Q.
- Probability of P_L given Q:

$$\mathbb{P}[Y = P_L | \, Q] = h(Q),$$

  with $h'(Q) \geq 0$.
- Expected price conditional on Q:

$$Y(Q) = P_L h(Q) + P_H(1 - h(Q)).$$

- Twice differentiable cost function $c(q)$.
- Stage game payoff for $i \in \{1, 2\}$

$$u^i(q_1, q_2) = Y(Q)q_i - c(q_i),$$

  with $Q = q_1 + q_2$.

Throughout, let $S_0 = \{1\}$ be the trivial state space. $F_B^{S_0}$ then simplifies to the classical stage game strategy based best response dynamics, which I sometimes refer to as 'myopic' best response dynamics, given the repeated nature of the interaction at hand. Under $F_B^{S_0}$, it is well known that under general conditions on $Y(Q)$, there is a unique Nash equilibrium that is globally attracting (Milgrom and Roberts (1990)).

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

$$P_{sB}(q_1, q_2) = \mathbb{P}\big[s' = B | \, s; q_1, q_2\big],$$

the transition probability to move to state B given current state s and quantity choices q_i in state s. Throughout, I maintain Assumption 1 as done in previous section. Also assume that

$$P_{sB}(q_1, q_2) = \mathbb{P}\big[s' = B | \, s; q_1 + q_2\big],$$

for all s, q_i, i.e. transition probabilities only depend on aggregate quantities. I will therefore sometimes write $P_{ss'}(q_1, q_2) = P_{ss'}(Q)$ with $Q = q_1 + q_2$.

Throughout, let $\rho^i : S \mapsto X$ be each player's policy, and recalling the definition of W^i in (3), note that in the binary case one can derive

$$W^i(\rho, A) = \omega^{-1}\Big[(1-\delta P_{BB}(\rho))u^i\big(\rho^i(A), \rho^{-i}(A)\big) + \delta P_{AB}(\rho)u^i\big(\rho^i(B), \rho^{-i}(B)\big)\Big],$$
$$W^i(\rho, B) = \omega^{-1}\Big[\delta(1-P_{BB}(\rho))u^i\big(\rho^i(A), \rho^{-i}(A)\big) + (1-\delta(1-P_{AB}))u^i\big(\rho^i(B), \rho^{-i}(B)\big)\Big], \tag{8}$$

where

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

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.[20]

In what follows I will eventually focus on symmetric equilibria, and therefore drop the i- superscript for all objects, fixing our attention on player 1's payoffs. Suppose ρ^{*1} is an interior best response to ρ^2 (i.e. $\rho^{*1} \in BR^1_S(\rho^2)$, as defined in (4)) for which local optimality conditions hold with a negative definite Hessian. One can then use the implicit function theorem to find the derivative of 1's best response with respect to ρ^2, which will be an essential building block in finding stability conditions of an equilibrium. Since policies are vectors in X^2, from now on I will use the conventions $\rho^i(s) = \rho^i_s \in X$ for all i, s.

$$J(\rho^{*1}, \rho^2) = \begin{bmatrix} \frac{\partial \rho^{*1}_A}{\partial \rho^2_A} & \frac{\partial \rho^{*1}_A}{\partial \rho^2_B} \\ \\ \frac{\partial \rho^{*1}_B}{\partial \rho^2_A} & \frac{\partial \rho^{*1}_B}{\partial \rho^2_B} \end{bmatrix}. \tag{9}$$

In the following, to further ease notation I will adopt the following conventions:

- $u^s = u(\rho_s, \rho_s)$, for $s \in S$.
- $u^s_i = \frac{\partial u^s}{\partial q_i}$ and $u^s_{ij} = \frac{\partial u^s_i}{\partial q_j}$, for $i, j = 1, 2$, $s \in S$.
- $P'_{sB} = \frac{\partial P_{sB}}{\partial q_1} = \frac{\partial P_{sB}}{\partial q_2}$ for all s and analogously for P''_{sB} where the equality comes from the fact that P_{sB} only depends on aggregate quantities.

[20] Uniqueness is implied by our irreducibility Assumption 1.

More explicitly, one can then write

$$\begin{aligned}
\frac{\partial \rho_A^{*1}}{\partial \rho_A^2} &= -1 + \frac{\omega^{-1}\delta P'_{AB}(u_2^A - u_1^A) + u_{11}^A - u_{12}^A}{\omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A}, \\
\frac{\partial \rho_A^{*1}}{\partial \rho_B^2} &= \frac{\omega^{-1}\delta P'_{AB}(u_1^B - u_2^B)}{\omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A}, \\
\frac{\partial \rho_B^{*1}}{\partial \rho_A^2} &= \frac{\omega^{-1}\delta P'_{BB}(u_2^A - u_1^A)}{\omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B}, \\
\frac{\partial \rho_B^{*1}}{\partial \rho_B^2} &= -1 + \frac{\omega^{-1}\delta P'_{BB}(u_1^B - u_2^B) + u_{11}^B - u_{12}^B}{\omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B}.
\end{aligned} \tag{10}$$

Now I introduce more structure on the Cournot payoff function that is natural and will be maintained throughout the section.

Definition 6. *Say that the payoff function $u(q_1, q_2)$ is regular if*

(i) $u_1(0,0) > 0$.
(ii) $c(0) = 0$, $c'(0) > 0$, $c''(q) \geq 0$ *for all* $q \in X$.
(iii) $Y'(2q) < 0$ *for all* $q < M$.
(iv) There exists $K \in (0, M)$ *with* $u_1(0, 2K) < 0$ *and such that*

$$\max_{q \leq 2K, q' \leq 2K - q} Y'(q + q') + qY''(q + q') \leq 0.$$

Definition 6 is slightly weaker than standard assumptions made for the Cournot game. Points (i-iii) are standard assumptions to be expected from a Cournot game. Point (i) makes the problem interesting, point (ii) is a natural assumption on the cost function, point (iii) a natural assumption on the inverse demand. Point (iv) represents a small deviation from the norm only in that I allow for the quantity representing the second derivative of marginal revenue to be positive for large quantities in X, which will give us flexibility to later on support a simple, symmetric binary-state collusive equilibrium. At the same time it is enough to have u be quasi-concave as will be shown below. The assumption is weaker than the commonly made assumption "$Y'(Q) + qY''(Q) \leq 0$" for all q, Q (e.g. Hahn (1962)).

Lemma 1. *Suppose u is regular. Then under a boundary restriction there exists a unique Nash equilibrium q_N, which is symmetric and statically stable.*

As stated before, when u is regular, the unique Nash equilibrium is globally attracting under myopic best-response dynamics, and therefore if RL played on the trivial state space S_0, they would converge to q_N with probability 1. I show next that even though that is true, binary state variables exist so that when RL condition their policies on them, they will not learn the statically stable Nash equilibrium:

First, consider the matrix of best-response derivatives under a binary state variable when the static Nash equilibrium is played: $\rho_s^i = q_N$ for both i, s. I call this policy ρ_N.

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

where ω_N signifies evaluation of ω at ρ_N and I use that $BR'_N = -\frac{u_{12}^N}{u_{11}^N}$. Thus, one can read off the tradeoffs faced by an agent who, starting at the static Nash equilibrium ρ_N, adjusts best responses to a deviation by the opponent. The agent has tradeoff between following incentives about payoffs *today* (static incentives), represented by BR'_N, and dynamic incentives considering effects on continuation payoffs, represented by$J_N - BR'_N I_2$. The dynamic incentives allow for additional interpretation:

Note that

$$\frac{\delta P'_{AB}(\rho_N)}{\omega_N} = \frac{\partial \ln\left(1-\delta+\delta(P_{AB}(\rho_N)+P_{BA}(\rho_N))\right)}{\partial \rho_A};$$
$$\frac{\delta P'_{BA}(\rho_N)}{\omega_N} = \frac{\partial \ln\left(1-\delta+\delta(P_{AB}(\rho_N)+P_{BA}(\rho_N))\right)}{\partial \rho_B},$$

thus the factor multiplying $\frac{u_2^N}{u_{11}^N}$ 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. In what follows, I write these dynamic incentives more compactly as

$$D_A = D_A^P D_A^u = \frac{\delta P'_{AB}(\rho_N)}{\omega_N}\frac{u_2^N}{u_{11}^N}; \quad D_B = D_B^P D_B^u = \frac{\delta P'_{BA}(\rho_N)}{\omega_N}\frac{u_2^N}{u_{11}^N},$$

so that the dynamic incentives can be split into a factor originating from the above described sensitivity of transition probabilities D_s^P, multiplying the weighted effect of a perturbation to future payoffs $D_A^u = D_B^u = D^u$.

As noted before, as $\delta \to 1$, $\omega_N^{-1}(1-\delta P_{BB}) \to \mu_A(\rho_N) = \frac{P_{BA}(\rho_N)}{P_{AB}(\rho_N)+P_{BA}(\rho_N)}$, the stationary probability of visiting $A \in S$ when ρ_N is played forever. On the other side, $D_A^P = D_B^P = 0$ when $\delta = 0$. Stability of the static Nash equilibrium is impacted by this quantity in a straightforward manner:

Proposition 3. *Let u be regular and consider arbitrary transition probabilities $P_{ss'}$ for a binary state variable. Then ρ_N is dynamically unstable (i.e. unstable w.r.t. F_B^S) if and only if*

$$\left|BR'_N + D_A + D_B\right| > 1.$$

Proof. As discussed in Appendix B, one needs to linearize best responses at ρ_N to determine the stability of that profile. We need to characterise the eigenvalues $\lambda_{1,2}$ of J_N. 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 $\lambda_2 = BR'_N + D_A + D_B$. Regularity gives that $|\lambda_1| < 1$, so that $|\lambda_2| > 1$ appears as the condition in the Proposition. □

Proposition 3 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. On the one side, market conditions matter through the size of the slope of the static best response BR'_N today and the weighted effect that an opponent's deviation has on stage game payoffs u in the future. On the other side, fixing the market conditions, state variables matter: $D^P_A + D^P_B$ is the total sensitivity of transition probabilities with respect to policy ρ. In words, this quantity represents the aggregate effect of a marginal change in policy ρ on transition probabilities, which in turn control the correlation structure over states. For algorithms to avoid the static Nash equilibrium, only the magnitude of this sensitivity matters: For any payoff function u of bounded derivatives, there is a threshold so that once $|D^P_A + D^P_B|$ 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, governed by $P_{ss'}$.

Corollary 2. *Let u be regular. There exists a pair $z^*_1 < 0 < z^*_2$ so that ρ_N is dynamically stable if and only if*

$$D^P_A + D^P_B \in (z^*_1, z^*_2),$$

where

$$z^*_1 = -\frac{1 + BR'_N}{D^u}; \quad z^*_2 = \frac{1 - BR'_N}{D^u},$$

and

$$D^P_A + D^P_B = \delta \frac{P'_{BA} + P'_{AB}}{1 - \delta + \delta\big(P_{AB}(\rho_N) + P_{BA}(\rho_N)\big)}.$$

Recall that regularity of u implies that $BR'_N \in (-1, 0)$. Therefore, the following forces are at work in stabilizing the equilibrium: Firstly, the more negative BR'_N, the smaller is $|z^*_1|$, and the easier is it for negative $D^P_A + D^P_B$ to be be below z^*_1 so that Nash is

unstable. This follows since BR'_N is negative from the start, and so if dynamic incentives work in tandem with the static incentive, perturbations can accumulate easily. This force must be seen as relative to D^u; even if transition probabilities are very sensitive ($|D^P_A + D^P_B|$ is large), they only matter for best responses if they lead to a sizeable payoff effect relative to static incentives. Analogously, z^*_2 is smaller the closer BR'_N is to zero, since positively sensitive transition probabilities must surpass the static incentive which works in the opposite direction.

Note that Proposition 3 and the above Corollary can be generalized to Nash equilibria of payoff functions u unrelated to the Cournot game studied here, given twice differentiability of u at the equilibrium.

5.1. Two Special State Variables

So far, state variable S has been allowed to exist without any connection to the underlying stage game u other than through its correlation to aggregate quantities Q. In what follows, I introduce two state variables that correlate to Q through the channel of price outcome Y, and lead to two extreme results. In the first case, static Nash will be dynamically stable for any state variable falling in that general family; in the second case, static Nash can be dynamically unstable, and the existence of collusive equilibria that are attracting is possible.

First, consider the state variable where $s_t = Y_{t-1}$, the past period's price realization.

Definition 7. *A public* 1R*-policy can be defined as policy* $\rho : \{P_L, P_H\} \mapsto X$*, so that states are price realizations representing last periods observed price. This can equivalently be defined as having a state realizations* $\mathbf{Y}$ *with transition function* $T(s, P) \in \mathbf{Y}$ *such that* $T(s, P) = P$ *for all* $s \in \mathbf{Y}$*, and all price observations* $P \in \mathbf{Y}$*.*

For this family of policies, one can show the following:

Corollary 3. *Let* ρ_N *be the* 1R*-policy that plays stage game Nash quantity* q_N *in every state. Then* ρ_N *is dynamically stable if and only if* q_N *is statically stable.*

This result follows from Corollary 2, since under 1R-policies at ρ_N, we must have that $D^P_A + D^P_B = 0$.[21] Note that under 1R-policies, $P_{AB} = 1 - P_{BA} = \mathbb{P}[P_H]$. Thus, $P'_{AB}(\rho_N) + P'_{BA}(\rho_N)$=0. In fact, for any policy that plays the same quantity in all states, when the state is past period's price it must be that $D^P_A + D^P_B = 0$. In general this comes from the fact that, for every given current state, conditional distributions over future states are the same. I call this quality of a state variable 'state-independent transitions' (SIT).

[21]The result holds more generally in the case of finitely many prices (more than 2), an analysis of which can be found in an online appendix available upon request.

Transitions of a state variable can readily be depicted in a transition diagram. Figure 5.1 depicts the underlying transition diagram when $s_t = Y_{t-1}$, with state A corresponding to $Y_{t-1} = P_L$, and state B corresponding to $Y_{t-1} = P_H$.

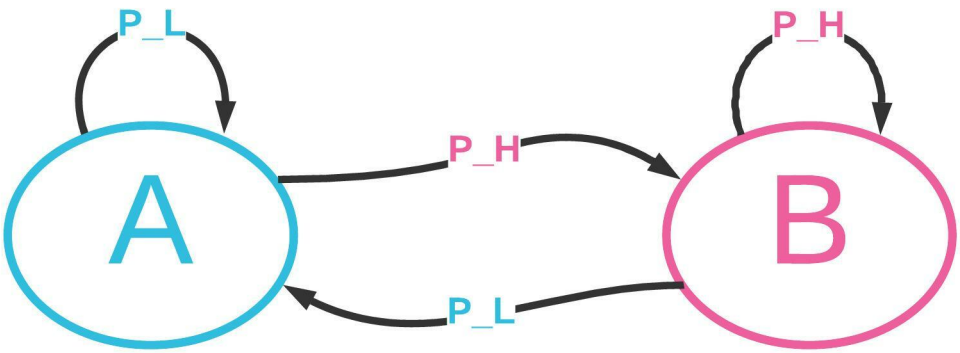

FIGURE 1. Transition Diagram: SIT

In contrast, the following is a state variable under a more involved transition profile, which I denote *direction-switching* (DS):

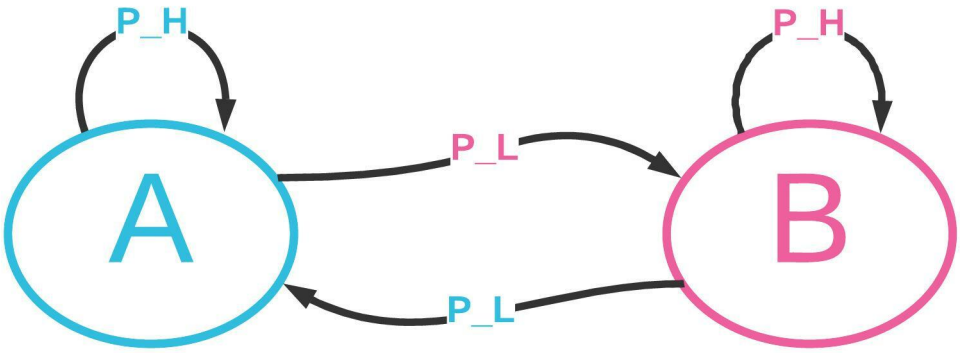

FIGURE 2. Transition Diagram: DS

A realized price P_L represents a *switch-signal*, while realizing P_H represents a *remain-signal*. Thus, states evolve from s to s' according to transition probability function

$$P_{AB}(Q) = \mathbb{P}[P_L \mid Q] = P_{BA}(Q).$$

More generally one can define policies following state variables with transition probabilities having the above property:

Definition 8. *Say a binary state policy is a* DS*-policy ('DS' for direction-switching) if the underlying state transitions are irreducible and* $P_{AB}(Q) = P_{BA}(Q)$ *holds for all* Q*. Denote the* DS*- state space as* S^**.*

In words, 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. Notice that this affects dynamic incentives in an interesting manner: for a given quantity Q, marginal deviations affect expected continuations in the opposite direction, depending on the current state. This fact introduces an essential difference in how states A, B are interpreted, even when $\rho_A = \rho_B$ is played. Notice that in this case, $D_A^P = D_B^P$, so that

$$D_A + D_B = 2\delta \frac{P'_{AB}(\rho_A)}{\omega_N} D^u,$$

with $P'_{AB}(\rho_A) = h'(2\rho_A)$. Let ζ_N be the DS-policy playing q_N in all states (call it ζ_N, to differentiate from ρ_N under 1R policies). It follows from Corollary 2 that there should exist conditional price distributions $h(Q)$ so that ζ_N will be dynamically unstable and therefore not learned by our RL. However, the issue is not immediate since $h(Q)$ plays a role both in construction of u and transition probabilities $P_{ss'}(Q)$. In the following, I show that there is a set of regular payoff functions such that ζ_N is indeed statically stable, but dynamically unstable. Moreover, this family of regular payoff functions will also allow for the existence of collusive equilibria.[22]

For the sake of analytical tractability, I introduce a global shape restriction on h so that the resulting payoffs are regular, but together with S^*, will allow for a collusive equilibrium to exist. To this end, impose for h (and therefore $Y(Q)$) to take an S-shaped form:

Definition 9. *Fix* $M > 0$*. Let* $\mathcal{G}^o$ *be the space of monotone increasing, twice differentiable functions* $h : [0, M] \mapsto [0, 1]$*, s.t.* $\underline{h}'' = \min_{Q\in[0,M]} h''(Q) \in (-\infty, 0)$*. Define a space of S-shaped functions with lower bounded second derivative as*

$$\mathcal{G} = \Big\{ h \in \mathcal{G}^o \text{ s.t. } \exists \tau \in (0, \frac{M}{2}) : h''(0) > -\underline{h}'', h''(2\tau) = 0$$
$$h''(Q) > 0 \,\forall\, Q \in [0, 2\tau),\ h''(Q) < 0 \,\forall\, Q \in (2\tau, M],$$
$$-\underline{h}'' < \frac{h'(2\tau)}{\tau} \Big\}.$$

Proposition 4. *There exists* $h \in \mathcal{G}$*,* $P_H > P_L \geq 0$ *and a convex* $c(q)$ *such that resulting* u *is regular,* ζ_N *is dynamically unstable and there exists a symmetric equilibrium* σ *with* $0 < \sigma_A < q_N < \sigma_B$*.*

[22] This result holds under more general state-and price spaces which are non-binary, as established in Appendix **??** .

The intuition for how S^* can support a simple collusive scheme as σ can be found by thinking through incentives to deviate at the equilibrium: in state A, quantities lower than q_N are to be played. Statically, by the strategic-substitutes nature of the Cournot game, one wants to deviate upwards. However, increasing quantities in state A increases the likelihood of realizing P_L, which in turn would lead the system to move to state B, which is undesirable. In state B, punishingly high quantities are to be played. Here, a player statically would want to deviate to lower quantities. However, that would increase the likelihood of realizing P_H, which would lead to a repetition of state B in the following period; again an undesirable outcome. Thus, collusion is reinforced.

Notice that by construction, S^* is symmetric in the sense that for both states A, B, observing P_H leads to staying in the given state, whereas observing P_L implies leaving the state. It is then no surprise that given that payoffs are also symmetric, Proposition 4 also implies that there exists a another symmetric collusive equilibrium σ', satisfying $\sigma'_A = \sigma_B, \sigma'_B = \sigma_A$.

We see from the above and the following subsection that SIT-state variables and DS-state variables lead to starkly different outcomes. 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:

Corollary 4. *Suppose $h \in \mathcal{G}$, $P_H > P_L \geq 0$ and $c(q)$ are such that the conditions of Proposition 4 hold. Then ρ_N is the unique symmetric equilibrium under* 1R*-policies.*

5.2. Stable Collusion

The stability of the collusive equilibrium is determined by local conditions at the equilibrium. The dynamic instability of the Nash equilibrium is sufficient for the existence of collusion, but neither necessary for the existence, nor is it sufficient for stability of the collusive equilibrium. However, 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 $\Pi(q_1, q_2) = Y(Q) - c'(q_1)$ be the marginal profit as computed by a price-taker. Notice that by construction of the Cournot-payoff function, $u_1(q_1, q_2) - u_2(q_1, q_2) = \Pi(q_1, q_2)$, which is true for all q and therefore also true for $\Pi'(q_1, q_2) \equiv \frac{\partial \Pi(q_1, q_2)}{\partial q_1}$. This definition allows one to write, for any interior equilibrium profile σ as constructed in Proposition 4,

$$
\begin{aligned}
\frac{\partial \rho_A^{*1}}{\partial \rho_A^2} &= -1 + \phi_A^{-1}\Big[\Pi'_A - \omega^{-1}\delta P'_{AB}\Pi_A\Big],\\
\frac{\partial \rho_A^{*1}}{\partial \rho_B^2} &= \phi_A^{-1}\omega^{-1}\delta P'_{AB}\Pi_B,\\
\frac{\partial \rho_B^{*1}}{\partial \rho_A^2} &= \phi_B^{-1}\omega^{-1}\delta P'_{BA}\Pi_A,\\
\frac{\partial \rho_B^{*1}}{\partial \rho_B^2} &= -1 + \phi_B^{-1}\Big[\Pi'_B - \omega^{-1}\delta P'_{BA}\Pi_B\Big],
\end{aligned}
\tag{11}
$$

where as before, s-subscripts denote evaluation at q_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 to re-write determinant and trace of $J(\sigma,\sigma)$ using (11) under the assumption that $\Pi'(q_1,q_2) < 0$ for all $q_i \in X$. First, define for $s, s' \in S$:

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

which can be interpreted as a ratio of elasticities of ω versus Π_s with respect to q_s. To save notation, write $J^* = J(\sigma,\sigma)$. Then:

$$
\begin{aligned}
tr\big(J^*\big) &= -2 + \frac{\Pi'_A}{\phi_A}\big[1 - R_A\big] + \frac{\Pi'_B}{\phi_B}\big[1 - R_B\big];\\
det\big(J^*\big) &= \left[1 - \frac{\Pi'_A}{\phi_A}\frac{\Pi'_B}{\phi_B}\right] - \frac{\Pi'_A}{\phi_A}\big[1 - R_A\big]\left[1 - \frac{\Pi'_B}{\phi_B}\right] - \frac{\Pi'_B}{\phi_B}\big[1 - R_B\big]\left[1 - \frac{\Pi'_A}{\phi_A}\right].
\end{aligned}
\tag{12}
$$

Notice that for the stage game as constructed in Proposition 4, $\phi_s < u^s_{11}$ 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 $\Pi(q_1,q_2)$. It is clear from (12) 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:

Lemma 2. *Consider an interior, symmetric equilibrium under a binary state variable, $\sigma = (q_A, q_B)$ with $q_A < q_B$ as constructed in Proposition 4. Suppose $\frac{\Pi'_s}{\phi_s} \in (0,1)$ for both s. Then if*

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

σ is asymptotically stable.

Proof. Firstly, as shown in Appendix B, stability of σ is equivalent to

$$\left|tr\left(J^*\right)\right| - det\left(J^*\right) < 1. \tag{13}$$

Then, note from (12) 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 (13). The last equation then determines stability through the term $\left[1 - R_A - R_B\right]$. □

Relationship to the Best Equilibrium

One might wonder about the relationship between a binary-state collusive equilibrium as constructed in Proposition 4 and the best possible payoff a player can achieve in a repeated game of imperfect public monitoring. Using the insights of Abreu, Pearce, and Stacchetti (1990) (henceforth APS), a useful link can be made between best equilibria and binary-state policies as defined in the previous discussions.

First, let $\Gamma = \langle u^1, u^2 \rangle$ 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 $b_t^i = \{q_i^s\}_{0<s<t}$ be the private memory of a player's own actions, and define $B_t^i = X^t$ as the set of those at period t. Now, a strategy of player i at period t can be written as map $\sigma_t^i : B_t \times B_t^i \mapsto X$. 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 $b_t^1 = b_t^2$, 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 SE can be supported by a bang-bang solution, under their setting. Their setting differs from the one of this section on two counts: APS requires uncountable signals under an absolutely continuous distribution, and finite (but arbitrarily many) actions.

By allowing for public randomization in the definition of the strategies in Σ, the uncountable signal requirement can be satisfied. An approximation argument can then 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 $X_K = \{x_1, ..., x_K\} \subset X$ such that $\max_{0<k,k'\leq K}\{|x_k - x_{k'}|\} \leq \frac{1}{K}$, and such that $q_N \in X_K$ for all $K > 0$. Let the restricted game Γ_K^∞ be the repeated game where players are constrained to choose actions from X_k. Let E, E_K be the sets of SSEs of $\Gamma^\infty, \Gamma_K^\infty$ respectively. The restriction that $q_N \in X_K$ ensures that E_K is nonempty for all $K > 0$.

Define $V = \{U(\sigma) : \sigma \in E\}$. Let $V_K = \{U(\sigma_K) : \sigma_K \in E_K\}$. For Γ_K^∞, allowing for public randomization means that APS' results can be applied, which specifically gives us that V_K is compact and can be implemented by a bang-bang strategy that only ever plays two actions in X_K.

Let $\bar{\sigma}_K \in E_K$ be such that $U(\bar{\sigma}_K) = \overline{V}_K = \max V_K$. By APS, $\bar{\sigma}_K$ can always be chosen to be a bang-bang profile. Note that any such $\bar{\sigma}_K \in E_K$ is a ε-equilibrium of Γ^∞:

Proposition 5. *For any $\varepsilon > 0$ there exists $K_\varepsilon > 0$ such that for all $K \geq K_\varepsilon$,*

(i) $\bar{\sigma}_K \in E_K$ is an ε-equilibrium of Γ^∞,

(ii) $\bar{\sigma}_K \in E_K$ satisfies $U(\bar{\sigma}_K) = \overline{V}_K$,

(iii) There exists a binary state Variable $S_K = \{A, B\}$ with transition probability functions $P_{ss'}^K(q_1, q_2)$ so that policies $\rho_K^i : S_K \mapsto X_K$ constitute an ε-equilibrium of $W^i(\rho, s)$ as defined in (8).

Proposition 5 tells us that at the very least there exist binary state variables such that, when algorithms condition their policies on those (as set up in the beginning of this section), and are constraint to choose quantities in finite set X_K, the best equilibrium of the repeated game is a candidate for possible long run behaviors of the algorithms. Whether it can be learned with positive probability or not would then depend on its stability properties. Of course, the caveat here is that a continuum of actions X is necessary for the stability approach applied in this paper to go through (differentiability in actions being the necessary requirement). It remains an open question whether the best SSE payoff $\sup V$ of Γ^∞ can be approximated by a sequence of $\overline{V}_K$, which would go beyond the scope of this paper. If such an approximation result were to hold, one can argue that indeed also for the unrestricted choice case (actions in X), there exist binary state variables that support the best payoff and therefore may be learned by the RL agents defined here.

To provide a visualization of the results discussed so far, the following section shows simulation results which can be interpreted using the the theory developed in this and the previous section.

5.3. Numerical Example and Simulations

I construct a piecewise-linear version of $h(Q)$, prices P_L, P_H and a convex cost function $c(q)$ for which there exists a unique stage game Nash equilibrium q_N that is statically stable, but dynamically unstable, and there exists a stable symmetric collusive equilibrium. The following is an overview of a numerical example for which these properties are satisfied.

Fix a discount factor $\delta = 0.98$. All numbers given in the example are rounded to two decimal points. Given domain $X = [0, M]$, derivative parameters $0 < h'_A, h'_B < h'_N$, and cutoffs $x = [x_1, x_2, x_3, x_4] > 0$, I define the set of piecewise linear functions $\hat{\mathcal{G}}$ so that $h \in \hat{\mathcal{G}}$ if and only if one can write

$$h(Q) = \begin{cases} \underline{h}(Q) & Q \in [0, x_1) \\ \underline{h}(x_1) + h'_A(Q - x1) & Q \in [x_1, x_2) \\ h(x_2) + h'_N(Q - x_2) & Q \in [x_2, x_3) \\ h(x_3) + h'_B(Q - x_3) & Q \in [x_3, x_4) \\ \overline{h}(Q) & Q \in [x_4, M] \end{cases},$$

where $\underline{h} \in [0, 1)$ is strictly increasing, with $\underline{h}'(0) = 0$, $\underline{h}'(x_1) = h'_A$, $\underline{h}'(x_4) = h'_B$ and $\underline{h}''(Q) > 0$ for all $Q \in [0, x_1]$, and $\overline{h} \in [0, 1)$ is strictly increasing, with $\overline{h}'(M) = 0$, $\overline{h}''(x_4) = 0$ and $\overline{h}''(Q) < 0$ for all $Q \in (x_4, M)$. Elements of $\hat{\mathcal{G}}$ are therefore piecewise-linear versions of elements of $\mathcal{G}$. The idea is that this construction facilitates a numerical example, while still allowing for existence of a collusive equilibrium using similar intuition as in Proposition 4.

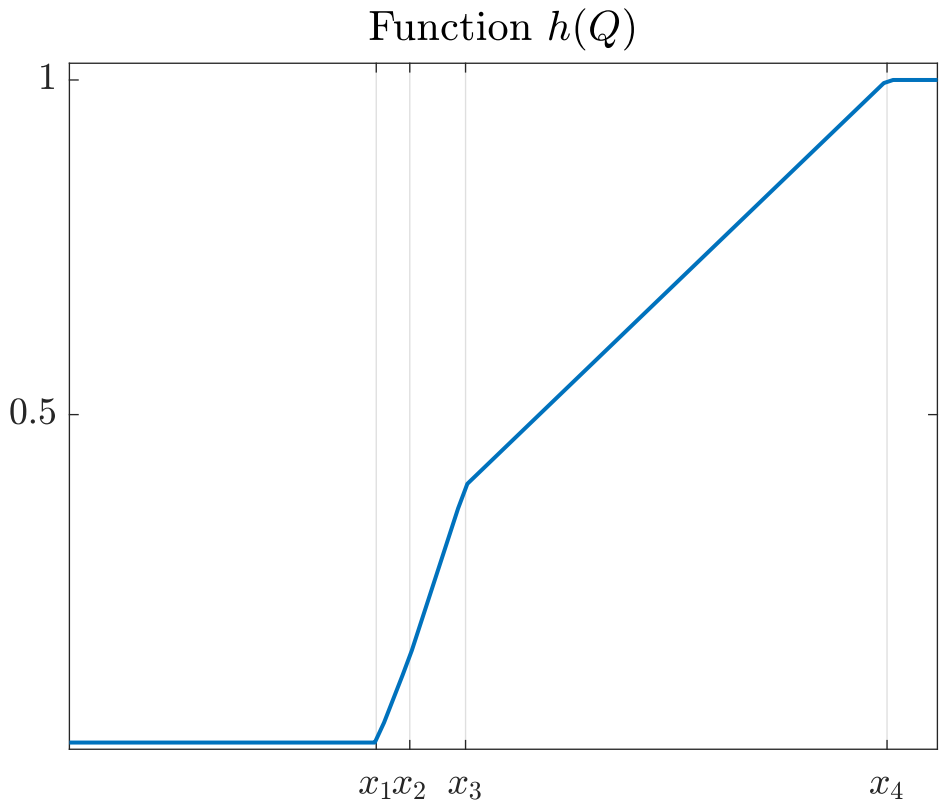

FIGURE 3. Location of cutoffs x_i for piecewise linear $h(Q)$.

Given $h(Q)$, prices and a cost function for which the properties stated in the beginning of this subsection are satisfied, one can plot the stage game best response and its inverse to verify the uniqueness of the static Nash equilibrium:

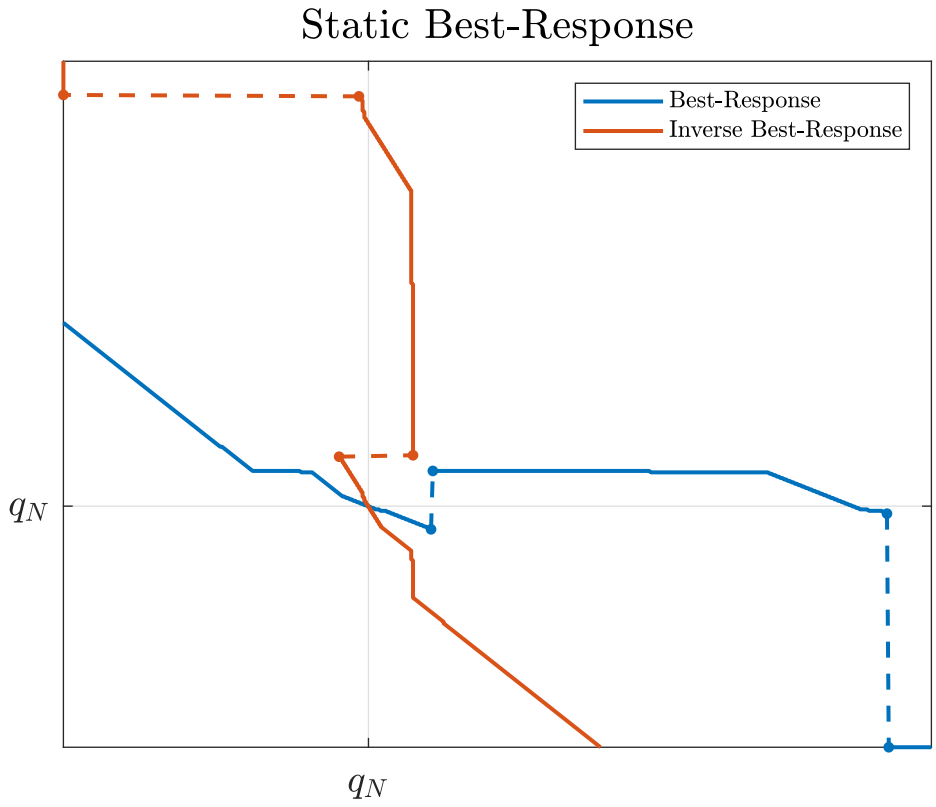

FIGURE 4. Best Response Intersection

This plot shows the unique intersection of static best-response functions in the numerical example, where $q_N = 1.58$. Best responses jump upwards at a value larger than q_N. This is due to the non-concavity introduced by the requirement $h'_B < h'_N$, which enforces the S-shape of the piecewise-linear $h(Q)$ function. Best responses jump to zero at a large quantity, where the interior local maximum has a negative value.

One can verify numerically that under this example,

$$-\frac{u_{12}^N}{u_{11}^N} = -0.5; \quad D_A + D_B = 2.28,$$

which implies from Proposition 3 that the Nash equilibirum is statically stable but indeed dynamically unstable. At the same time, this numerical example supports a pair of symmetric, collusive equilibria σ, σ' with $\sigma_A = 1.52 < \sigma_B = 1.94$ (rounded to two decimal points) and the quantities flipped for σ'. (see the discussion after Proposition 4. As laid out in detail in Appendix B, the stability of this equilibrium is verified by checking the

eigenvalues of the linearized system $F_B^{S^*}$ at the equilibrium. In this case, the largest eigenvalue is -0.5, implying that all eigenvalues are strictly negative, implying the stability of the collusive equilibrium.

I finish by providing a simulation study to visualize these results. This simulation study should be seen as a device to get intuitions about the system dynamics after many iterations of the algorithm have passed. The characterisation of long-run behavior given in subsection 4 is used here: instead of simulating the estimation part of Q_t of the algorithm given in Definition 2, I take Assumption 3 seriously, set bias term $g^i = 0$ for all i, and simulate iteration (5) in the following way:

For $i \in \{1, 2\}$ and all s,

$$\rho^i_{t+1}(s) \in \rho^i_t(s) + \alpha_t \Big[\arg\max_{q' \in X} Q^{i*}(s, q', \rho^{-i}_t) - \rho^i_t(s) + M^i_{t+1} \Big], \tag{14}$$

where $\alpha_t = t^{-0.6}$ satisfies the Robbins-Monro Assumption 4, and $M^i_{t+1} \sim N(0, .25)$ is an i.i.d mean-zero Normal noise variable with variance 0.25. Notice that (14) replaces Q_t given in (5) by its estimation target Q^*. Thus, this iteration represents a noisy discretization of $F_B^{S^*}$ rather than a simulation of a feasible model-free algorithm. As the results in subsection 4 tell us, for algorithms in the class studied in this section this simulation will give us an equivalent representation of long-run trajectories of ρ_t to a full simulation of (5) when t is large. Running a more in-depth simulation experiment including the estimation part of Q_t will be an insightful object of further investigation.

Note that the long-run characterisations given in subsection 4 are local in nature: if the iteration ρ_t at some point t enters a basin of attraction for a given stable equilibrium ρ^*, then the iteration will converge to that equilibrium with large probability.[23] One can not hope here to compute the exact basins of attraction for each equilibrium as such an exercise would go beyond the scope of this paper. However, any basin of attraction for a stable equilibrium must at the very least contain a small neighborhood of that equilibrium. I will use such a small neighborhood to initialize our simulations in this experiment.

In each simulation exercise, I run 96 separate simulations, and each for $25,000$ periods. Thus, simulations are potentially stopped before they've noticeably converged to a point. The idea is to take the following figures as relative to each other: each exercise was done for the same number of iterations, but the results differ starkly across exercises. As will be seen, depending on the state variables of the algorithms involved, iterations move closer to the

[23] In fact, approximations of this probability can be made given the neighborhood of ρ^* the iteration finds itself in, see for example Thoppe and V. Borkar (2019). This leads to an interesting avenue of future research, which will allow a study of the distribution of outcomes possible given a set of competing algorithms in the class described in this paper. Once a distribution over outcomes can be characterised, modeling a strategic interaction involving choosing algorithms will become feasible.

equilibrium the neighborhood of which they started at, or move away from it, confirming the theory developed in this paper.

First, I consider the result given in Corollary 3. Since in this example, the Nash equilibrium is statically stable, its repetition under 1R-policies ρ_N is also stable. Thus, one would expect that once algorithms using 1R-state variables come close to the Nash equilibrium, they should stay close to it forever, and in the long run converge to it. This is what is evidenced by Figure 5.3. Since the state space is binary, the two algorithms' policies can be represented as points in the X^2-plane. I now plot simulation outcomes in this plane, so that each simulation run is represented by two points in the plane spanned by $\rho(A), \rho(B)$.

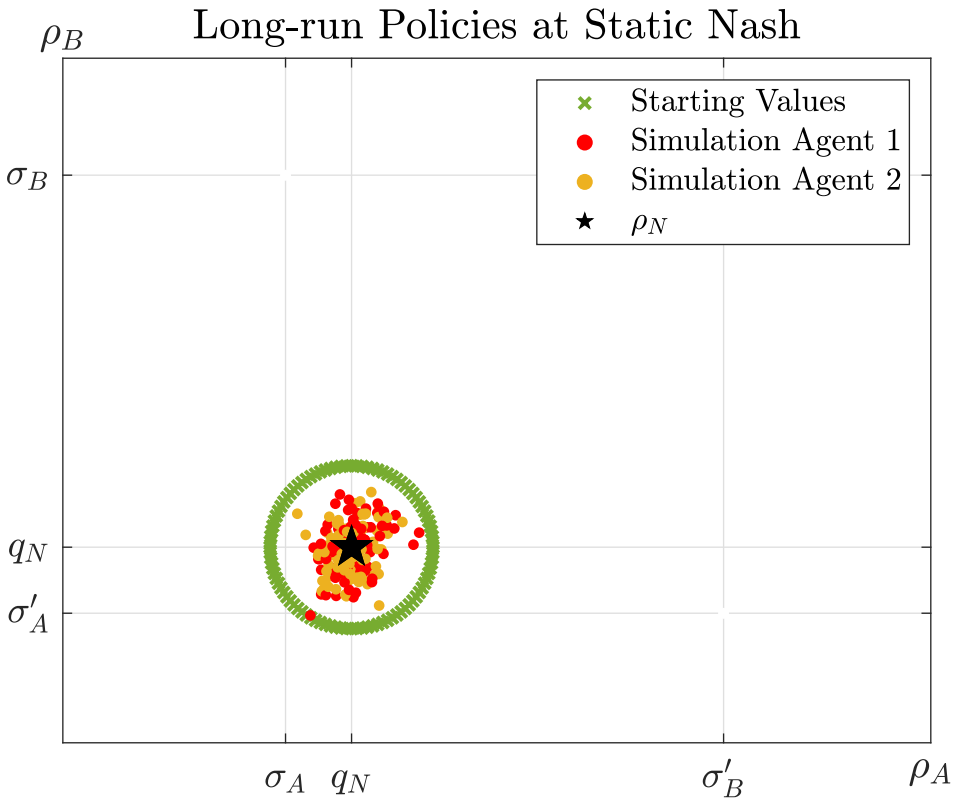

FIGURE 5. 1R-policies

The final policy profiles of 96 simulation runs of $25,000$ iterations each are shown in Figure 5.3. Simulations are started in a circular neighborhood of ρ_N, with a radius of $.025\|\rho_N\|$. One red and one orange dot represent a policy profile at the end of a simulation run. σ, σ' signify the collusive equilibria, which in this case have not been approached. All simulations remained in the neighborhood, as should be expected given the stability of ρ_N.

Now contrast this result with an analogous study given DS-policies for a state variable evolving as shown in Figure 5.3. Even though the neighborhood of starting values used in this scenario is the same as under 1R-policies, the picture is starkly different:

Recall that I denote the repetition of q_N under S^* as ζ_N. Since q_N is dynamically unstable under S^*-policies, no matter how close the starting values of the iteration are, the iteration must be pushed away from ζ_N as shown in the proof of Proposition 2. However, in the case of this example, it is not only true that the iteration is pushed away, but also that it is pulled towards the collusive equilibrium σ. This indicates that the basin of attraction for the collusive equilibrium in this example is not confined to a small neighborhood of the equilibrium but in fact quite large. This scenario also underlines the weight of consideration that should be given to the analysis of policy spaces given two competing algorithms. Even if one forced algorithms to initialize very close to a Cournot equilibrium, they can, given the right state variable, approach a collusive equilibrium instead. As this example supports a pair of collusive equilibria, the resulting figure shows how roughly half the simulation runs end up in the North-West of the plane, while the other half approached the South-East.

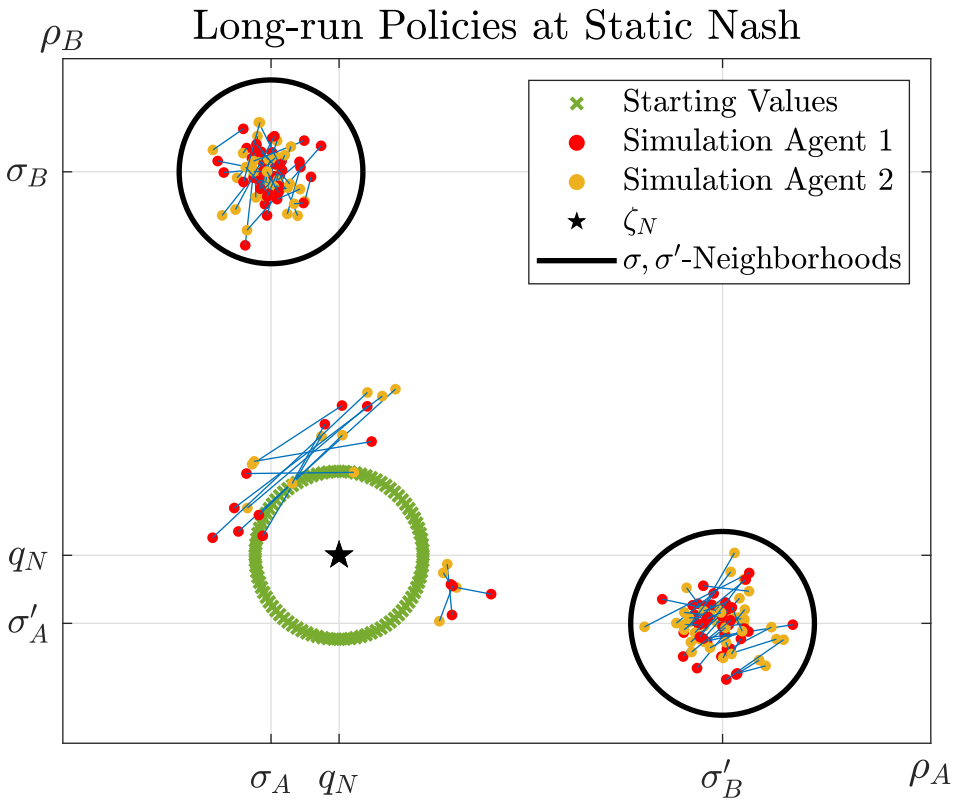

FIGURE 6. S^*-policies

The final policy profiles of 96 simulation runs of 25,000 iterations each are shown in Figure 5.3. Simulations are started in a circular neighborhood of ζ_N, with a radius of $.025\|\zeta_N\|$. One red and one orange dot represent a policy profile at the end of a simulation run. All simulations left the neighborhood of starting values, with most conglomerating at one of the two collusive equilibria σ, σ'. To see that outcomes indeed approached σ, σ', two dots connected by a line represent a single simulation outcome. The black circles represent neighborhoods of σ, σ' with radius $.25\|\sigma\|$.

With a similar exercise it can be seen that the collusive equilibria indeed attract the algorithm iterations if starting values are analogously defined as for the two above discussed simulations:

Figure 5.3 shows how after starting in a neighborhood of the collusive equilibrium σ, iterations stayed there for the course of the simulation. An analogous picture can be generated when initializing in a neighborhood of σ'.

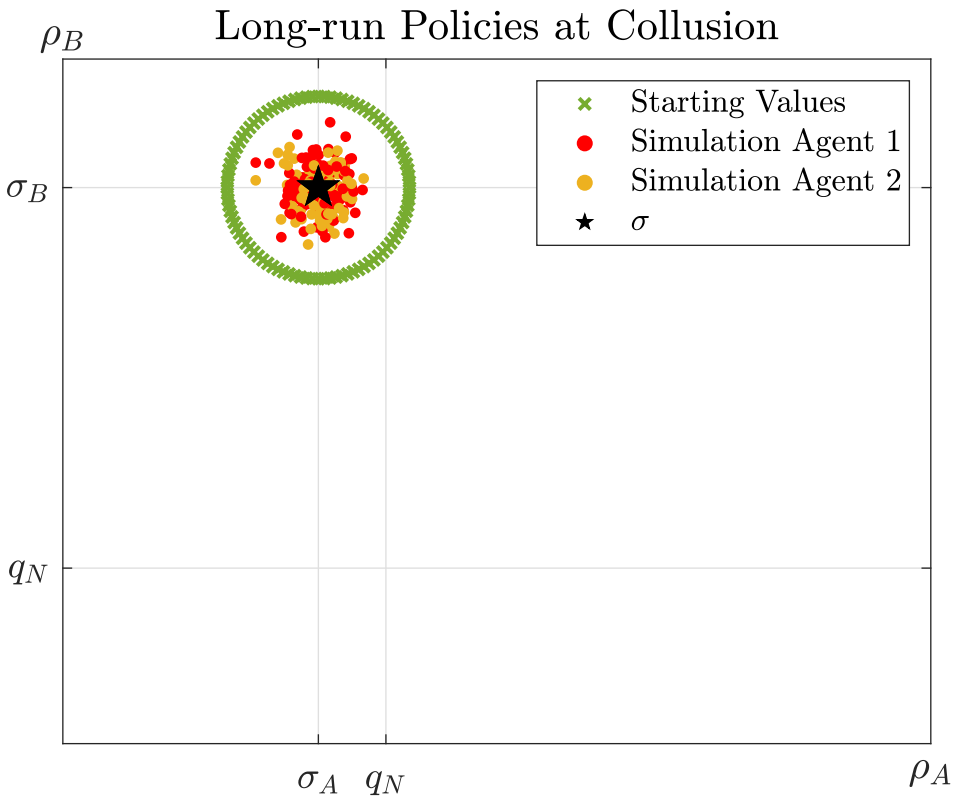

FIGURE 7. S^*-policies, initialized locally

The final policy profiles of 96 simulation runs of 25,000 iterations each are shown in Figure 5.3. One red and one orange dot represent a policy profile at the end of a simulation

run. Simulations are started in a circular neighborhood of σ, with a radius of $.025\|\sigma\|$. All simulations remained in the neighborhood, as should be expected given the stability of σ.

Finally, one might wonder about the global properties of the dynamical system generated from this numerical example. It happens to be the case that, when simulations are initialized at a radius of $.95q_N$, i.e. from values approaching the boundary of the action set X, the above results are unaffected: under 1R-policies, all simulation runs converge to the static Nash equilibrium ρ_N, while under S^*-policies, none of the simulation runs converge to ρ_N; all of them conglomerating at either σ or σ'.

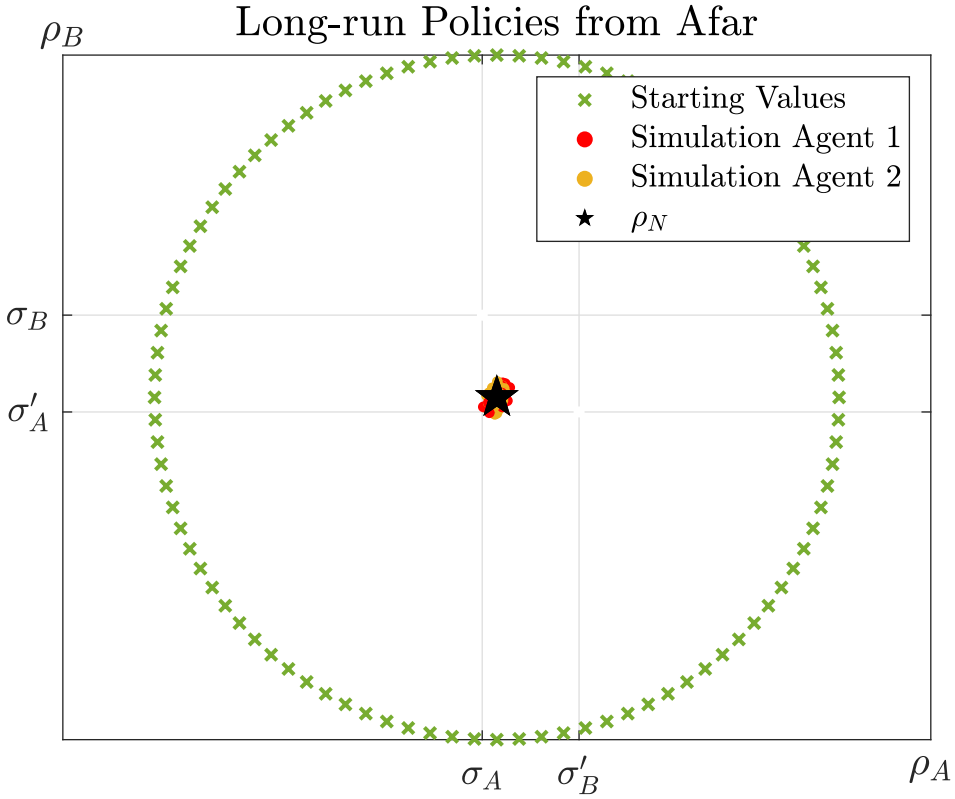

FIGURE 8. 1R-policies, initialized globally

The final policy profiles of 96 simulation runs of $25,000$ iterations each are shown in Figure 5.3. Simulations are started in a circular neighborhood of ρ_N, with a radius of $.95\|\rho_N\|$. One red and one orange dot represent a policy profile at the end of a simulation run. σ, σ' signify the collusive equilibria, which in this case have not been approached. All simulation runs conglomerated at ρ_N.

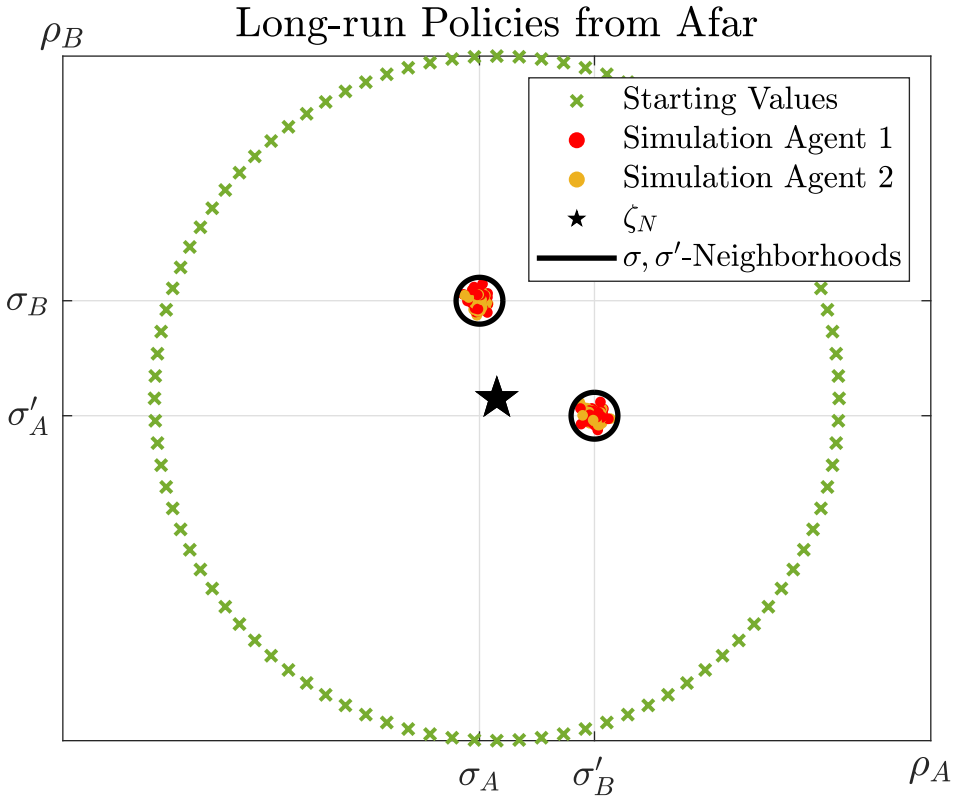

FIGURE 9. S^*-policies, initialized globally

The final policy profiles of 96 simulation runs of $25,000$ iterations each are shown in Figure 5.3. Simulations are started in a circular neighborhood of ζ_N, with a radius of $.95\|\zeta_N\|$. One red and one orange dot represent a policy profile at the end of a simulation run. All simulations conglomerated at one of the two collusive equilibria σ, σ'. The black circles represent neighborhoods of σ, σ' with radius $.25\|\sigma\|$.

6. Conclusion

This paper considers the long-run behavior of a class of RL algorithms and shows that one can interpret said long-run behavior by considering the stability of repeated game equilibria according to an underlying differential equation. By ways of the application of collusion in repeated games, I observe 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 characterisation 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 within the family allowed, and evaluate long-run outcomes in the chosen setting.

The characterisation allows to distinguish whether a given equilibrium will be learned with positive, or with zero probability. This is the current state of the art of the stochastic approximation approach applied here; in the future it will be interesting to look into an approach that allows to evaluate the relative likelihood of observing one equilibrium versus another. Such an improved characterisation will allow for the study of a meta-game. In such a meta-game, firms will choose algorithms (say, the state variable used by the algorithm, the stepsize sequence, or other details of the updating rule). Using this improved characterisation, firms can evaluate their expected profits from a given algorithm profile. That way, it will be possible to employ a Nash equilibrium analysis of a meta-game of choosing algorithms.

Since the algorithms in the class considered require to be given a fixed policy space on which to learn, an interesting comparative statics exercise that comes out of this project is a first step in categorizing dynamic policies by how amenable they are in allowing the learning of cooperative behavior. I introduce the categories 1-recall-policies and direction-switching (DS) policies as a first step in this endeavour. I also introduce the terms static and dynamic stability, in order to analyse the ability of RL to learn repeated stage game Nash equilibria in the context of the policy space that these RL learn on. When considering a statically stable Nash equilibrium, one can ask:

"How does the ability of the algorithms to learn a stage game equilibrium change when the state space their policies are allowed to condition on is changed?"

The resulting categorization of state-dependent policies is an important area of future research. For now, it gives us an idea of what restrictions an antitrust authority might want to impose on the information RL are allowed to condition their policies on.

Furthermore, my analysis generates testable conditions on the payoff functions RL face so that collusion or the stage game Nash will be learnable. Since the conditions only depend on market fundamentals, this can be affected by market interventions and therefore pose another viable channel for antitrust regulations.

A more precise understanding of the range of payoffs supportable in the long-run by competing RL is another area of interesting future research. Once a better understanding is achieved of the distribution over possible outcomes given a set of competing algorithms, one can construct a hyper-game of choosing algorithms (or their parameters), which will go a long way in the study of algorithmic collusion.

6.1. Discussion of related Literature

Firstly, Banchio and Mantegazza (2022) also consider a characterisation of competing RL algorithms and apply it to games of economic interest. The class of algorithms they study intersects with the class studied in this paper, but there are important differences. It is unclear that their approach can accommodate actor-critic approaches that are featured here, as such approaches require a separate estimation technique that can introduce dependence of policy parameters on histories of past observations. This is important, since the actor-critic feature allows us to consider closely the learning of repeated game strategies, which is not featured in the focus of Banchio and Mantegazza (2022).

There is a recent theoretical literature on stylized models of algorithmic competition. Lamba and Zhuk (2022) study how algorithms may learn to collude. They look at a stylized model of algorithmic competition, in which an algorithm is represented by a policy mapping from opponent actions to actions, which can be revised less frequently than actions are taken. They show that no equilibrium of that game is fully competitive. Salcedo (2015) goes along a similar direction, with an algorithm being an automaton strategy that can only be revised less frequently than actions can be taken.

Another paper of stylized algorithmic competition is Z. Y. Brown and MacKay (2021). They focus on the frequency with which algorithms can update prices, and let algorithms of different adjustment speeds compete against each other. When frequency abilities are asymmetric among algorithms, equilibrium outcomes can be collusive. Interestingly, when firms can choose algorithms (i.e. their adjustment frequency), the equilibrium features asymmetric frequencies.

The works mentioned above focus on different aspects of frequency of adjustment as a stylized feature of algorithmic updates. This paper shows a channel that has not been explored much in this literature: the role of state variables in the ability of algorithms to learn collusion. This could be an interesting new starting point for a study of stylized algorithms. Moreover, the works above abstract away from issues of learning and estimation, which is in contrast to this paper. An interesting aspect of learning present here is the importance of stability of equilibria in determining what can be learned. Stability of equilibria is tightly connected to dynamic reactions to imprecisions and mistakes (perturbations), which are present when learning and estimation are part of algorithmic updates.

Johnson, Rhodes, and Wildenbeest (2020) look into platform design under algorithmic sellers. They investigate differing policies implemented by a platform designer wishing to promote competition or raise own profits. They include a simulation study of Q-learning algorithms under different policy designs; clearly, results in this paper can be applied to

study related RL algorithms under any given platform policy. Once a more tight characterisation of the distribution over outcomes supported by a profile of algorithms is in place, one can go a step further and attempt to find the optimal platform policy for any given algorithm profile in my class.

In connection to the theory of learning in games, this paper speaks to the learning of repeated game strategies by making an important connection: I shed light on the ability of behavioral players to learn to play equilibria of a repeated game other than the static equilibrium of the underlying stage game. Since the players I consider learn policies on a fixed policy space, one may recast their payoffs as expected discounted payoffs based on stationary policy profiles that have to live in that policy space. See the definition of best response under a given state variable (4). Taking that view, one can say that algorithms in my class learn to play Nash equilibria of a repeated stage game with multi-dimensional continuous actions (which are precisely the policies in the policy space). In this sense my analysis ties neatly into classical analysis of the theory of learning in games with minimal information requirements.

There is now a growing area of research lying on the intersection of the theory of learning in games from the economics point of view, and the asymptotic theory of algorithmic learning from the computer science side. Leslie, Perkins, and Xu (2020)'s paper is an example of a paper intended more for economists, while applying language also common to the computer science literature. They consider zero-sum Markov games and construct an updating scheme related to best response dynamics that converges to equilibria of the game. As they also keep track of separate policy and value function updates, their scheme falls into the class of actor-critic learning rules generally, while not falling into the class considered in this paper due to important assumptions on the updating speed differential between policy and performance criterion used there.

Leslie and Collins (2006) introduce what they call "generalized weakened fictitious play" (GWFP), an adaptive learning process the limits of which can be related to classical continuous time fictitious play (G. W. Brown (1951)), or stochastic ficitious play (c.f. Hofbauer and Sandholm (2002)), depending on details of the process. Their framework allows to conclude asymptotic behavior of learning processes once one has shown that the process is a GWFP process. They show that GWFP converges in games that have the fictitious play property. Notably that class includes zero-sum games, submodular games, and potential games.

One can interpret results in this paper as showing that a subclass (ACQ) of the RL I consider can be seen as a GWFP process. Therefore, one can apply Leslie and Collins (2006) to conclude the limiting behavior of that process in games with the fictitious play property.

However, there are many repeated games of interest that do not have this property; notably standard repeated oligopoly (Cournot) games where agents learn repeated game strategies. I analyse the learnability of collusion in oligopoly games more seriously, and therefore give a more detailed analysis of limiting behavior in a class of games not known to have the fictitious play property. I do this by taking seriously the fact that GWFP can in general be defined to learn repeated game strategies, which to the best of my knowledge has so far only been considered under the restriction of Markov strategies for stochastic games.

This paper connects also to a growing strand of the computer science literature establishing convergence proofs in multi-agent algorithmic environments. The paper in that area closest to this one is Mazumdar, Ratliff, and Sastry (2020). They establish a connection between gradient-based learning algorithms for continuous action games and asymptotic stability of equilibria of the underlying game. While nested in our RL class, the updating rules that Mazumdar, Ratliff, and Sastry (2020) consider implicitly assume that algorithms observe each other's per period policies, or at least observe an unbiased estimator of their per-period value function gradient. I argue that this assumption is difficult to satisfy, especially in the case of continuous action games. In a companion paper (Possnig (2022)), I give low-level sufficient conditions on independent algorithms so that a weakened version of this assumption goes through. My results suggest that Mazumdar, Ratliff, and Sastry (2020)'s results are robust to the type of bias in the gradient estimation that my RL class allows. Furthermore, this paper focuses on the possibility of RL to learn history-dependent repeated game strategies, which is not the explicit goal of Mazumdar, Ratliff, and Sastry (2020).

Other papers related to asymptotic analysis of multi-agent systems commonly focus on developing a specific algorithm that behaves well in some metric, allow communication across algorithms, require information on the primitives of the game, or do not ask about the nature of the limiting points. Notably, Ramaswamy and Hullermeier (2021) give a thorough analysis of deep learning techniques for Q-functions using gradient updates, without considering stability properties of rest points. Others focus on specific classes of games, for example zero sum games (Sayin et al. (2021)) and show convergence of multi-agent learning there.

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.

Abreu, Dilip, David Pearce, and Ennio Stacchetti (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, M (1999). "Dynamics of Stochastic Approximation, Le Seminaire de Probabilite', Springer Lecture Notes in Mathematics". In:

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

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, George W (1951). "Iterative solution of games by fictitious play". In: *Act. Anal. Prod Allocation* 13.1, p. 374.

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.

Chicone, Carmen (2006). *Ordinary differential equations with applications*. Vol. 34. Springer Science & Business Media.

Dutta, Debaprasad and Simant R Upreti (2022). "A survey and comparative evaluation of actor-critic methods in process control". In: *The Canadian Journal of Chemical Engineering*.

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.

François-Lavet, Vincent et al. (2018). "An introduction to deep reinforcement learning". In: *arXiv preprint arXiv:1811.12560*.

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.

Fujimoto, Scott, Herke van Hoof, and David Meger (July 2018). "Addressing Function Approximation Error in Actor-Critic Methods". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1587–1596. URL: https://proceedings.mlr.press/v80/fujimoto18a.html.

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

Grondman, Ivo et al. (2012). "A survey of actor-critic reinforcement learning: Standard and natural policy gradients". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6, pp. 1291–1307.

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.

Hernandez-Leal, Pablo et al. (2017). "A survey of learning in multiagent environments: Dealing with non-stationarity". In: *arXiv preprint arXiv:1707.09183*.

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 and Edmund J Collins (2006). "Generalised weakened fictitious play". In: *Games and Economic Behavior* 56.2, pp. 285–298.

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.

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.

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.

Ramaswamy, Arunselvan and Eyke Hullermeier (2021). "Deep Q-Learning: Theoretical Insights from an Asymptotic Analysis". In: *IEEE Transactions on Artificial Intelligence*.

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*.

Sayin, Muhammed et al. (2021). "Decentralized Q-learning in zero-sum Markov games". In: *Advances in Neural Information Processing Systems* 34.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Thoppe, Gugan and Vivek Borkar (2019). "A concentration bound for stochastic approximation via Alekseev's formula". In: *Stochastic Systems* 9.1, pp. 1–26.

Watkins, Christopher JCH and Peter Dayan (1992). "Q-learning". In: *Machine learning* 8.3, pp. 279–292.

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*.

Appendix A. The Algorithm Class

In this section I provide the general reinforcement learning family the analysis of sections 3-4 applies to. Assume there are N algorithmic agents. Agents observe states on some fixed, finite state space S with $|S| = L$, and make per period choices (actions) in compact interval X_i. Let $\bar{X}_i = X_i^L$, with policy profile space $\bar{X} = \times_{i \in I} \bar{X}_i$. Agents then follow a fixed rule (algorithm) to update their strategy profiles over time.

Definition 10. *Each agent updates their policy according to the following adaptive procedure:*

$$\rho_{t+1}^i \in \rho_t^i + \alpha_t \left[F^i(\rho_t) + B_t^i \right],$$

where $\alpha_t > 0$ is a decreasing stepsize sequence, $F(\rho_t)$ is a (possibly multi-valued) mapping, and B_t^i represents a (possibly multi-valued) error term.

I stack the above iteration over i to get to the representation of study:

$$\rho_{t+1} \in \rho_t + \alpha_t \left[F(\rho_t) + B_t \right]. \tag{15}$$

I write '$\in$' instead of '$=$' above to allow for multi-valued mappings as can be the case when F^i represents an argmax, which corresponds to Q-iterations as in definition 2. The class of RL algorithms studied here is determined by restrictions on $F(\rho)$ and B_t^i. Whenever there is multi-valuedness, I allow the algorithm to pick arbitrarily. In our limiting characterisation this will show up as possibility of multiple solutions, which will not affect the limiting statements as shown in section 4. Throughout, impose Assumptions 4 and 5. The following are two important examples of what behavior B_t can be allowed to take:

(1) $B_t = 0$ and $F(\rho)$ is a Lipschitz-continuous function, we are in the familiar territory of Robbins-Monro algorithms for which the asymptotic behavior is well known (see chapter 2 in V. S. Borkar (2009)).

(2) B_t is a martingale-difference noise with respect to some filtration $\mathcal{F}_t$, with bounded second moment. This error term could be the result from an estimation method to estimating $F(\rho)$ consistently. This scenario can again be readily analysed using the methods developed in V. S. Borkar (2009), chapter 2.

Considering the iteration (15), we can see that $F(\rho_t)$ features importantly as a mapping that provides the reinforcement of the iteration profile ρ_t. In many scenarios, $F(\rho)$ represents a performance criterion based on market and opponent conditions that are not known to the algorithm designer and must be estimated. $F(\rho)$ thus becomes an estimation target, and B_t can then be seen as the resulting error term. If $F(\rho)$ were fully known, as is allowed by definition, one can set $B_t = 0$ for all t.

First, I introduce the class of performance criteria $F(\rho)$, and what kinds of approximation methods can be considered here:

Definition 11 (Candidate performance criteria)**.** *Define the set $\mathcal{M}^1$ of (possibly multivalued) maps G with domain $X \subseteq \mathbb{R}^k$ and range $\mathcal{P}[R]$ for $R \subseteq \mathbb{R}^k$ s.t.*

- *$G(x) \subset R$ is convex, compact valued.*
- *There exists $c > 0$ such that $\sup\{\|y\| : y \in G(x)\} \leq c(1 + \|x\|)$ for all $x \in X$, i.e. linear growth.*
- *There is a union of connected sets $C_k \subseteq X$ of positive measure, $\mathcal{U}_S = \bigcup_k C_k$, such that $G(x)$ is single-valued and $\mathcal{C}^1$ for $x \in \mathcal{U}_S$.*

Remark 1. *I allow for multi-valuedness to be able to handle to common learning scheme of actor-critic Q-learning, which maintains estimates of the argmax of a value function as introduced in section 3. Note however that, with some abuse of notation, $\mathcal{C}^1 \subset \mathcal{M}^1$.*

Now, define the distance between points x and sets A as

$$d(x, A) = \inf_{x' \in A} \|x - x'\|.$$

Now we are ready to consider the definition of function approximators to which this analysis applies.

Definition 12 ($\mathcal{C}^1$ Approximation)**.**
Let Y be some space of observations (datasets) D_t to be used to approximate a mapping. Given $\gamma > 0$, say that a function approximation operator $\mathcal{A}_g : \mathcal{M}^1 \times Y \mapsto \mathcal{M}^1$ is a $\mathcal{C}^1$ Approximation of a performance criterion $F \in \mathcal{M}^1$ if there is a bias function $g \in \mathcal{B}^1_\gamma$ and an integer $N > 0$ such that one can write for all $t \geq N$:

(i) For all $\rho \in X$,

$$\mathcal{A}_g[F, D_t](\rho) = F_g(\rho) + \delta_t,$$

where $F_g(\rho) \in \mathcal{M}^1$ such that

$$\sup_{z \in F_g(\rho)} d\big(z, F(\rho)\big) < \gamma,$$

(ii) For all $\rho \in \mathcal{U}_S$,

$$\mathcal{A}_g[F, D_t](\rho) = F(\rho) + g(\rho) + \delta_t,$$

with $g \in \mathcal{B}^1_\gamma$,

(iii) There is an increasing sequence of σ-algebras $\mathcal{F}_t$ such that

$$\delta_t - \mathbb{E}[\delta_t]$$

is a martingale difference sequence given $\mathcal{F}_t$,

(iv)

$$\sup_t \mathbb{E}\left[\|\delta_t\|^2\right] < \infty.$$

(v) There exists a sequence $\zeta_t \geq 0$ satisfying Robbins-Monro's condition (Assumption 4) such that

$$\lim_{t\to\infty} \left\| \sum_{k=t}^{\infty} \zeta_k \mathbb{E}[\delta_k] \right\| = 0.$$

One can interpret $g(\rho)$ as representing the bias part of the function approximation, and δ_t as a random variable such that $\mathbb{E}[\|\delta_t\|^2 \,|\, \mathcal{F}_t]$ represents the variance part given some increasing sequence of σ-fields $\mathcal{F}_t$ generated by datasets $D_t \in Y$ and histories of ρ_t. Points (iv) and (v) bound the variance and speed of convergence of the error term δ_t to ensure that our characterisation technique goes through.

In the case of classical model-free Q learning as in subsection 2, D_t only needs to consist of $(s_k, a_k, r_k, s_{k+1})_{k=1}^t$, i.e. past observations of states, actions, payoffs, state transitions, and the initial Q_0.

Generally one can think of $\mathcal{A}_g[F, D_t](\cdot)$ as a parametric or non-parametric function approximation to the performance criterion of interest F, with bounded errors that can be approximated by a small $\mathcal{C}^1$ function after enough data (large n) has been accumulated. Fix small $\gamma > 0$ and observation spaces Y^i. We can now state the following assumption that, together with definitions 11 and 12 characterizes the algorithm class that can be studied here.

Assumption 6.

(i) Let the bias functions $g^i \in \mathcal{B}^1_\gamma$.

(ii) Let $D^i_t \in Y^i$ be a sequence of datasets.

(iii)

$$B^i_t = \mathcal{A}^i_{g^i}[F^i, D^i_t](\rho_t) - F^i(\rho_t) + M^i_{t+1},$$

where $\mathcal{A}_g[F, D_t]$ is a $\mathcal{C}^1$ Approximation of performance criterion $F(\rho) \in \mathcal{M}^1$. Notice that accordingly, B^i_t can be a point or a compact convex set.

(iv) Stacked version of B^i_t:

$$B_t = \mathcal{A}_g[F, D_t](\rho_t) - F(\rho_t) + M_{t+1}.$$

(v) $\mathcal{F}_t$ is the σ-field generated by $\{\rho_t, D_t, M_t, \rho_{t-1}, D_{t-1}, M_{t-1}..., \rho_0, D_0, M_0\}$, i.e. all the information available to the updating rule at a given period t.

(vi) M_{t+1} *is a Martingale-difference noise. There is* $0 < \bar{M} < \infty, q \geq 2$ *such that for all* t

$$\mathbb{E}[M_{t+1} \,|\mathcal{F}_t] = 0; \quad \mathbb{E}[\|M_{t+1}\|^q \,|\mathcal{F}_t] < \bar{M} \; a.s.$$

(vii) Whenever $\rho_t \in \mathcal{U}_S$,

$$\Omega_t \equiv \mathbb{E}[M_{t+1} M_{t+1}' \,|\mathcal{F}_t],$$

where Ω_t *is symmetric positive definite for all* t.

(viii) Write $\delta_t = \mathcal{A}_g[F, D_t](\rho_t) - F(\rho_t)$. *Then* M_{t+1}, δ_t *are conditionally uncorrelated and* $M_{t+1} + \delta_t - \mathbb{E}[\delta_t]$ *is a martingale difference conditional on* $\mathcal{F}_t$.

Gradient-type Algorithms

Here I give a brief overview of the kind of gradient-type algorithms that are included in our class. First, few definitions are in order to properly understand this class:

For any $i \in I$, let $\bar{X}_{-i} = \times_{j \neq i} \bar{X}_j$. Recall that expected future discounted payoffs $W^i(\rho^i, \rho^{-i}, s_0)$ given stationary strategy profiles $[\rho^i, \rho^{-i}] \in \bar{X}$ are defined as:

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \tag{16}$$

where the expectation is made over the state transitions.

Then define

$$\bigtriangledown W^i(\rho^i, \rho^{-i}, s_0) \in \mathbb{R}^k,$$

as the gradient with regards to policies of agent i's long term payoff evaluated at $[\rho^i, \rho^{-i}]$. By abuse of notation, write $\bigtriangledown W(\rho)$ as the stacked gradients of all agents, where without much loss one can suppress the dependence on initial states due to our assumption on irreducibility 1. It is without much loss since stability properties of any differential Nash equilibrium will be independent of the initial state under irreducibility, and those properties are the focus of the rest of the paper.

Now define for $\rho \in \bar{X}$

$$F_D^S(\rho) = \bigtriangledown W(\rho), \tag{17}$$

as the state dependent gradient dynamics. Take an iteration ρ_t and its respective function estimation target F as denoted in (15). If $F = F_D^S$, we will call the RL iteration 'Gradient Equivalent'.

For Gradient Equivalent iterations, if there is no asymptotic bias in the estimation of the gradient ($g = 0$), our results match to the results in Mazumdar, Ratliff, and Sastry (2020), but note that we study the possibility of repeated game strategies, which is not

explicitly done there. Further, as noted in the introduction, our results extend Mazumdar, Ratliff, and Sastry (2020) to the more commonly observed situation of non-vanishing biased function estimators.

Remark 2 (A note on the case $B_t^i = 0$)**.**
As stated in the discussion below Definition 10, we do allow for deterministic updates representing the case where everything is known to the algorithm, i.e. $B_t^i = 0$*. By definition this implies that the nonvanishing variance condition of* M_{t+1} *mentioned above cannot hold. However, it is then possible to proof a similar result as Proposition 2 by measuring the set of initial values* ρ_0^i *that would allow for the process* ρ_t *to converge to an unstable restpoint* ρ^**. By definition of instability, the paths that could attract* ρ_t *to* ρ^**, if they exist, must be a lower-dimensional subspace of profile space* $\bar{X}$*. The statement one can then make is that the probability that a randomly chosen starting profile* ρ_0 *will converge to unstable* ρ^* *must equal* 0*, since any lower dimensional subspace of* $\bar{X}$ *has measure zero.*

Appendix B. Best Response Dynamics: Stability

I give here detailed results that allow the stability analysis for binary state profiles given two players, as outlined in section 5. I assume notation and nomenclature developed in that section.

Let $S = \{A, B\}$ be any binary state space. For a given policy-profile $\alpha, \beta \in X^2$, write best replies as $b_1(\beta) = (b_1^A, b_1^B)^\top \in BR_S^1(\beta)$, and $b_2(\alpha) = (b_2^A, b_2^B)^\top \in BR_S^2(\alpha)$. I consider the stability of rest points for the state-dependent best response dynamics under S, F_B^S (see (7)), given the stacked policies $\sigma \in X^4$:

$$\dot{\sigma}_t = F_B^S(\sigma_t) = \begin{bmatrix} b_1(\sigma_{t,3}, \sigma_{t,4}) \\ b_2(\sigma_{t,1}, \sigma_{t,2}) \end{bmatrix} - \sigma_t. \tag{18}$$

Suppose σ^* is an interior rest point $\in \mathcal{U}_S$. Then asymptotic stability of σ^* can be determined by linearizing the system and showing that all its eigenvalues have negative real parts. Let $X(\sigma^*)$ be the linearized system:

$$X(\sigma^*) = \begin{bmatrix} -I_{2x2} & J_1(\sigma^*) \\ J_2(\sigma^*) & -I_{2x2}, \end{bmatrix} \tag{19}$$

where I_2 is the 2-dimensional identity matrix and

$$J_i(\sigma^*) = \begin{bmatrix} \frac{\partial b_i^X}{\partial \beta_A} & \frac{\partial b_i^X}{\partial \beta_B} \\ \\ \frac{\partial b_i^B}{\partial \beta_A} & \frac{\partial b_i^B}{\partial \beta_B} \end{bmatrix},$$

for $i \in \{1, 2\}$.
This linearization has a special structure one canexploit:

Remark 3. *Suppose A, B, C, D are square matrices of same dimension, s.t. $CD = DC$. Let*

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

Then one can show

$$det(T) = det(AD - BC).$$

We can use this the following way: consider the characteristic equation of $X(\sigma^*)$:

$$ch(\lambda) = det(X(\sigma^*) - \lambda I_{4x4}).$$

Then all eigenvalues are characterised as the zeros of $ch(\lambda)$. Remark 3 tells us that

$$ch(\lambda) = det(J_1 J_2 - (1 + \lambda)^2 I_{2x2}).$$

That is, if μ is an eigenvalue of $J_1 J_2$, then $\pm\sqrt{\mu} - 1$ is an eigenvalue of $X(\sigma^*)$.
Note: $J_1 J_2$ is the matrix of derivatives one gets when considering the derivatives of an iterated application of best responses:

$$b_1(b_2(\sigma_1, \sigma_2)) - (\sigma_1, \sigma_2)^\top \tag{20}$$

with respect to σ. We can then interpret stability graphically as a scenario in which (20) doesn't grow above the 45-degree line. This can be translated to eigenvalues being less than 1, which from the above is equivalent to considering asymptotic stability of $X(\sigma^*)$.

When considering symmetric equilibria, one cango even further:

Remark 4. *Suppose A, B are square matrices of the same dimension. Let*

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

Then one can show

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

Now, in a symmetric equilibrium σ^*, we have $b_1(\sigma^*) = b_2(\sigma^*)$. Further, since we have symmetric payoff functions, we have $J_1 = J_2 = J$ as the matrix of derivatives of the best reply function. We can then apply Remark 4 to our system and arrive at the conclusion.

Firstly, given square matrix A, define Λ as the set of eigenvalues of the A. Then define

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

as the spectral radius of A.

Lemma 3. *Suppose $\alpha^* = \beta^* = \sigma^*$ is an interior, symmetric equilibrium. Let $\bar{\kappa}$ be the real part of the spectral radius of M. Then σ^* is asymptotically stable if $\bar{\kappa} < 1$, and unstable if $\bar{\kappa} > 1$.*

Proof. Using Remark 4, we get that

$$ch(\lambda) = det(M - (1+\lambda)I_2)det(M + (1+\lambda)I_2).$$

Thus, if μ is an eigenvalue of M, 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. □

Appendix C. Proofs

Proof of Proposition 1

The proofs are given in terms of the general algorithm class treated in this paper, introduced in Appendix A. Throughout we pick a sequence α_t s.t. Assumption 4 holds and $\alpha_t \leq \zeta_t$ holds for all $t \geq 0$. Under Assumption 3, it is quick to check that the ACQ iteration in Definition 2 falls into our class. First we prove the following result that employs known techniques from stochastic approximation theory.

First, a few definitions are in order. Take a correspondence $G(x) \in \mathcal{M}^1$, where we recall the domain being $X \subseteq \mathbb{R}^k$ for some $k \geq 1$. The following Definition can be found in Michel Benaïm, Hofbauer, and Sorin (2005):

Definition 13.

(1) Given a set $A \in X$ 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 $\dot{x} \in G(x)$[24], 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 $A \in X$ is said to be internally chain transitive (ICT) if A is compact and $x \hookrightarrow_A y$ holds for all $x, y \in A$.

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

Importantly, these 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 6. *With probabilitiy one, $L_{S,g}$ is an ICT set of the differential inclusion*

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

where $F_g(\rho(t)) \in \mathcal{M}^1$ s.t. $\sup_{z \in F_g(\rho)} d(z, F(\rho)) < \gamma$ holds for all ρ, and particularly

$$F_g(\rho(t)) = F(\rho(t)) + g(\rho(t))$$

whenever $\rho(t) \in \mathcal{U}_S$.

Proof. The algorithm (15) can be written as

$$\rho_{n+1} = \rho_n + \alpha_n \left[F_g(\rho_n) + \delta_n + M_{n+1} \right], \tag{21}$$

where $\delta_n = \mathcal{A}_g[F, D_n](\rho_n) - F_g(\rho_n)$.

We can now show first that iteration 21 is a perturbed solution to $\dot{\rho} \in F_g(\rho(t))$ as defined in Definition II in Michel Benaïm, Hofbauer, and Sorin (2005). Following the proof of their proposition 1.3, we only need to take care of the additional term δ_n present in iteration 21.

Following the notation in Hofbauer and Sandholm (2002), introduce:

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

Define for $k > n$:

$$\Psi_n^k = \left\| \sum_{i=n}^{k-1} \alpha_{i+1} \delta_{i+1} \right\|.$$

Then for proposition 1.3 in Hofbauer and Sandholm (2002) to hold, it suffices to show that, for all $T > 0$

$$\sup_{n \leq k \leq m(\tau_n + T) - 1} \Psi_n^k \to_P 0, \tag{22}$$

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

$$\Psi_n^k \leq \sum_{i=n}^{k-1} \alpha_{i+1} \left\| \delta_{i+1} \right\|,$$

and the right hand side is increasing in k, so that

$$\sup_{n\leq k\leq m(\tau_n+T)-1}\Psi_n^k\leq\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\big\|\delta_{i+1}\big\|.$$

Now we get that, for any $\varepsilon>0$,

$$\mathbf{P}\left(\sup_{n\leq k\leq m(\tau_n+T)-1}\Psi_n^k>\varepsilon\right)\leq\mathbf{P}\left(\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\big\|\delta_{i+1}\big\|>\varepsilon\right).$$

Next, recall from definition 12 that $\|\delta_n\|=O_P(n^{-\frac{1}{2}})$. Fix $\varepsilon'>0$. By definition there exists $M_{\varepsilon'}>0$ and $N>0$ such that for all $n>N$:

$$\mathbf{P}\Big(\|\delta_n\|>M_{\varepsilon'}n^{-\frac{1}{2}}\Big)<\varepsilon'.$$

Define

$$\mathbf{1}_n=\mathbf{1}\Big\{\|\delta_n\|>M_{\varepsilon'}n^{-\frac{1}{2}}\Big\}.$$

Thus,

$$\|\delta_n\|\leq M_{\varepsilon'}n^{-\frac{1}{2}}+\mathbf{1}_n\|\delta_n\|.$$

So we can write:

$$\mathbf{P}\left(\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\big\|\delta_{i+1}\big\|>\varepsilon\right)\leq\mathbf{P}\left(\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\mathbf{1}_{i+1}\big\|\delta_{i+1}\big\|+J_n>\varepsilon\right),$$

where

$$J_n=M_{\varepsilon'}\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}(i+1)^{-\frac{1}{2}}.$$

Notice that $J_n\to0$ as $n\to\infty$, as long as $\frac{\alpha_n}{n^{-\frac{1}{2}}}\to0$, which is in line with assumption 4. We can thus take n large enough to have $\varepsilon>J_n$, and then write

$$\begin{aligned}\mathbf{P}\left(\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\mathbf{1}_{i+1}\big\|\delta_{i+1}\big\|+J_n>\varepsilon\right)&\leq\frac{1}{\varepsilon-J_n}\mathbf{E}\left[\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\mathbf{1}_{i+1}\big\|\delta_{i+1}\big\|\right]\\&\leq\frac{\varepsilon'}{\varepsilon-J_n}\sum_{i=n}^{m(\tau_n+T)-1}\alpha_{i+1}\mathbf{E}\Big[\big\|\delta_{i+1}\big\|^2\Big]^{\frac{1}{2}}\\&\leq\frac{\varepsilon'}{\varepsilon-J_n}C_\delta T,\end{aligned}$$

where the first inequality follows from Markov's inequality, and the second one from Cauchy-Schwarz. Since δ_i are square-integrable by definition 12, there exists $0 < C_\delta < \infty$ so that the last inequality follows from the definition of τ_n. Since ε' is arbitrary, it can be made arbitrarily small and we have that indeed (22) holds.

Thus, ρ_n is a perturbed solution to $\dot{\rho} \in F_g(\rho(t))$ as defined in Definition II in Hofbauer and Sandholm (2002). The result then follows from theorem 3.6 in Hofbauer and Sandholm (2002).

□

Since payoffs are differentiable around ρ^*, point 1 follows as long as ρ^g and ρ^* are close. For point 2, we will prove something more general: as long as ρ^* is hyperbolic (c.f. Definition 4), point 2 holds.

This follows because when ρ^* is hyperbolic, there is a neighborhood U around 0 such that F has a differentiable inverse on U. Next, note that ρ^g solves

$$F(\rho^g) + g(\rho^g) = 0.$$

Since $\|g\|_1 \leq \gamma$, for γ small enough, $F(\rho^g) \in U$ must hold. Then there is some $L_{F^{-1}} > 0$ such that

$$\begin{aligned} \|\rho^g - \rho^*\| &= \|F^{-1}(F(\rho^g)) - F^{-1}(0)\| \\ &\leq L_{F^{-1}} \|F(\rho^g)\| \leq L_{F^{-1}} \gamma, \end{aligned}$$

where the first inequality follows because F^{-1} is differentiable and $F(\rho^*) = 0$, and the second by the definition of $F(\rho^g)$. Since the right hand side is independent of g, the bound is uniform.

For point 3, we first need to verify that all ρ^g close enough to ρ^* must also be asymptotically stable. The next Lemma gives a more general result:

Lemma 4. *Suppose ρ^* is hyperbolic. Then the eigenvalues of $DF_g(\rho^g)$ converge to the eigenvalues of $DF(\rho^*)$ uniformly over $g \in \mathcal{B}^1_\gamma$ as $\gamma \to 0$. Thus, for small enough γ, ρ^g has the same stability properties as ρ^*.*

Proof. I will show that eigenvalues of a hyperbolic matrix $DF(\rho^*)$ vary continuously in $\mathcal{C}^1$ perturbations g to F.

Proposition 2.18 in Palis Jr, Melo, et al. (1982) shows that eigenvalues vary continuously for any matrix A. Thus, if $\|DF(\rho^*) - DF_g(\rho^g)\|$ is small enough, the eigenvalues of the two

matrices must be close to each other. Now write

$$\begin{aligned}\|DF(\rho^*) - DF_g(\rho^g)\| &= \|DF(\rho^*) - DF(\rho^g)\| + \|Dg(\rho^g)\| \\ &\leq \|DF(\rho^*) - DF(\rho^g)\| + \gamma,\end{aligned}$$

where the equality follows from the definition of F_g. Since DF is continuous, and $\rho^g \to \rho^*$ uniformly for $g \in \mathcal{B}^1_\gamma$ as $\gamma \to 0$ (see above proof of point 2), we get that

$$\sup_{g \in \mathcal{B}^1_\gamma} \|DF(\rho^*) - DF_g(\rho^g)\| \to 0$$

as $\gamma \to 0$. Then applying Proposition 2.18 in Palis Jr, Melo, et al. (1982) finishes the result. □

Since we know that all ρ^g must be asymptotically stable for γ small enough, one can apply Faure and Roth (2010) (Thm 2.8). It suffices to verify that our process satisfies their attainability condition:

Definition 14. *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 (15). Then we need that the basin of attraction of an attractor has nonempty intersection with $Att(X)$. This can be verified:

Lemma 5. *Let B be a basin of attraction of an attractor A for F_g. Suppose $\rho_t \in \bar{X} \setminus B$. Then there exists $s > n$ such that $\rho_s \in B$ with positive probability.*

Proof. Since t 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:

$$M_z \in \frac{z - \rho_t}{\alpha_t} - F_g(\rho_t),$$

since F_g might be multi-valued.

Since $z \in int(E)$ by definition, M_z is in the support of M_{t+1} for every t. 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 t, since it is in the support of M_{t+1}. (if we allow $s > n + 1$, we may be able to increase the measure but we only need it to be positive.) □

All other conditions that are sufficient for the algorithm 15 to converge to the attractor hold by Assumption 6.

■

Proof of Proposition 2

Notice first that the following analysis is local to the rest points in E_S, which by assumption on $\mathcal{U}_S$ is also where F, F_g are single valued. Solution curves are unique whenever they intersect $\mathcal{U}_S$.

The proof will use the Hartman-Grobman Theorem (c.f.Chicone (2006), Theorem 4.8), which connects the flow of a nonlinear ODE in the neighborhood of a hyperbolic rest point to the flow of a linearized ODE. Since it works fully locally, our analysis only requires that $F(\rho)$ be single valued and $\mathcal{C}^1$ in U_{ρ^*}, and we can allow $F(\rho)$ to be multivalued otherwise.

First, define invariant sets for given differential equations:

Definition 15. *Let $z(t, z_0)$ be the solution so some given differential equation $\dot{z} = f(z)$ with initial value z_0. Then a set S*

- *is invariant for f, if $z(t, z_0) \in S$ holds for all $t \in \mathbb{R}$ and all $z_0 \in S$.*
- *isolated invariant for f if there is an open set N such that $S \subset N$ and*

$$S = \{z' : z(t, z') \in N \, \forall t \in \mathbb{R}\}.$$

Given a $g \in \mathcal{B}^1_\gamma$, we know from Proposition 6 that only ICT sets (recall Definition 13) subset of a neighborhood of ρ^g are candidates to being limiting points of the algorithm (15). The singleton $\{\rho^g\}$ is an ICT set, and we show first that this cannot be a limiting set of the algorithm. Then we go on to show that for small engough γ, no other ICT sets can exists in a neighborhood around ρ^*, which finishes the proof.

1) $\{\rho^g\}$ cannot be a limiting set.

Note that by Lemma 4, there are $\gamma > 0$ small enough such that all ρ^g are linearly unstable just as ρ^*. We can thus apply Michel Benaïm and Faure (2012), Theorem 3.12 to prove $\mathbb{P}[L_{S,g} = \rho^g] = 0$ first:

We can show that the sufficient conditions for this hold by definition of our algorithm under Assumption 6. First we need that Hypothesis 2.2 and 3.6 of Michel Benaïm and Faure (2012) hold. It is quick to check that these hypotheses would hold true if it were the case that $\mu_n \equiv \mathbb{E}[\delta_n] = 0$ for all n.[25] In our case, point (v) of Definition 12 is required additionally. To see this, following the arguments in Michel Benaïm and Faure (2012) one needs to bound

$$\mathbb{P}\Big(\sup_{h \in [0,T]} \|X(\tau_n + h) - \phi_h(X(\tau_n))\| \geq \varepsilon \mid \mathcal{F}_t \Big) \tag{23}$$

[25] For example by applying Proposition 2.16 of Faure and Roth (2010)

where $X(\tau_n)$ is the linear interpolation of ρ_n and $\phi_h(x)$ is the flow of $\dot{\rho} \in F_g(\rho)$ starting at x, carried forward by h periods. τ_n is as defined in the proof of Proposition 6. By following an argument analogous to the proof of Proposition 4.1 in Benaïm (1999), bounding the random variable inside the above measure leads to all terms present there, with one additional term owing to the presence of μ_n. This additional term is bounded by

$$K_n = \sup \left\{ \left\| \sum_{i=n}^{k-1} \alpha_{i+1}\mu_{i+1} \right\| : k = n+1, ..., m(\tau_n + T) \right\},$$

with $m(\cdot)$ as defined in the proof of Proposition 6. However, this term is deterministic - and by point (v) of Definition 12 must vanish with large n. Thus, one can write

$$\mathbb{P}\Big(\sup_{h\in[0,T]} \|X(\tau_n + h) - \phi_h(X(\tau_n)) \geq \varepsilon \mid \mathcal{F}_t \Big)$$
$$\leq \mathbb{P}\Big(\sup_{h\in[0,T]} \|X(\tau_n + h) - \phi_h(X(\tau_n)) \geq \varepsilon - K_n \mid \mathcal{F}_t \Big),$$

where K_n vanishes, so the bounding function as required in Hypothesis 2.2 and 3.6 of Michel Benaïm and Faure (2012) can be found e.g. by applying Proposition 2.16 of Faure and Roth (2010). Finally, note that the conditions and analysis sufficient for the proof of Michel Benaïm and Faure (2012)'s Theorem 3.12 are local with respect to ρ^g. Thus, the fact that F_g is globally potentially multivalued is of no importance, since in a small enough neighborhood around ρ^g it must be single-valued and $\mathcal{C}^1$.

2) No other ICT sets exist in a neighborhood of ρ^* and ρ^g.

We will prove that there are no other invariant sets in such a neighborhood. Since ICT sets are subsets of invariant sets, this will complete the proof.

We can use Hartman-Grobman to show that there are open neighborhoods N_g, N_0 with $\rho^* \in N_0, \rho^g \in N_g$ such that ρ^*, ρ^g are isolated invariant sets in their respective neighborhoods. These neighborhoods are nontrivial for all γ small enough, which follows from both ρ^*, ρ^g being hyperbolic:

By Hartman-Grobman and hyperbolicity there exists a homeomorphism H on a neighborhood $N \subseteq U_{\rho^*}$ of ρ^* with $H(\rho^*) = \rho^*$ such that

$$H(\phi(t, \rho)) = \psi(t, H(\rho)),$$

where $\phi(t, \cdot)$ is a solution (flow) to the differential inclusion $\dot{\rho} \in conv[F(\rho)]$, and $\psi(t, \cdot)$ is the solution to the ODE $\dot{y} = DF(\rho^*)(y - \rho^*)$. Given a neighborhood $U \subseteq N$ of ρ^*, define

$$inv(U) = \{\rho \in U : \phi(t, \rho) \in U \, \forall t \in \mathbb{R}\}.$$

We will show that $\rho^* = inv(U)$, and therefore it is isolated invariant.
Notice that $inv(U)$ can be rewritten as

$$inv(U) = \{y \in H(U) : H^{-1}(\psi(t,y)) \in U \, \forall t \in \mathbb{R}\} = \{y \in H(U) : \psi(t,y) \in H(U) \, \forall t \in \mathbb{R}\},$$

since H is bijective. We know that ρ^* is an isolated invariant set for the linear ODE solution $\psi(t,y) = Ce^{tDF(\rho^*)}y + \rho^*$. Thus, we must also have that

$$inv(U) = \rho^*,$$

and ρ^* is isolated invariant set for $\phi(t,\rho)$.

Since ρ^g are hyperbolic for γ small enough, an analogous argument gives us that ρ^g are isolated invariant also. Let N_g be the neighborhood on which the homeomorphism is defined that connects flows of F_g to flows of the linearized system $DF_g(\rho^g)$. By definition, $\rho^g \in N_g$, and we know that ρ^g is isolated invariant in N_g. We are left to show that for γ small enough, for all $g \in \mathcal{B}^1_\gamma$, $\rho^* \in N_g$:

To prove this, we will argue that each N_g contains a ball $B^g_z(\rho^g)$, for which the radius $z > 0$ can be lower bounded by a number that depends only on the eigenvalues of $DF(\rho^*)$ and γ. First we need an auxiliary Lemma to show how eigenvalues of $DF_g(\rho^g)$ vary continuously in γ. First some more notation:

For small enough γ, all ρ^g are hyperbolic when $g \in \mathcal{B}^1_\gamma$. Fix such a g. Define $\rho_l > 0$ to be the smallest positive eigenvalue of $DF_g(\rho^g)$, and $\rho_u < 0$ be the largest negative eigenvalue of $DF_g(\rho^g)$. Now let $a_g \in (0,1)$ be any number such that

$$\max\{e^{\rho_u}, e^{-\rho_l}\} < a_g < 1.$$

For the original system $DF(\rho^*)$, let $a_0 \in (0,1)$ be any such number.

Lemma 6. *For any $\delta > 0$ with $a_0 < 1 - \delta$ there exists $\overline{\gamma} > 0$ such that for all $\gamma \in (0, \overline{\gamma}]$, there is a set of $\{a_g\}_{g \in \mathcal{B}^1_\gamma}$ as defined above with*

$$\sup_{g \in \mathcal{B}^1_\gamma} |a_g - a_0| < \delta.$$

Proof. Apply Lemma 4. Since there is a one-to-one mapping between eigenvalues and $\{e^{\rho_u}, e^{-\rho_l}\}$, one can find numbers a_g. The result follows. □

Given this continuity in eigenvalues, we can prove the following Lemma to finish our result:

Lemma 7. *Suppose ρ^* is hyperbolic for F. Fix a small $\underline{z} > 0$. Then there is $\bar{\gamma}$ such that for all $\gamma \leq \bar{\gamma}$, and all $g \in \mathcal{B}^1_\gamma$, there is $B^g_z(\rho^g) \subseteq N_g$ with $z \geq \underline{z}$.*

Proof. For small enough γ, all ρ^g are hyperbolic when $g \in \mathcal{B}^1_\gamma$. Fix such a g. Given some $\varepsilon > 0$, let r_ε be defined as

$$\sup\{r > 0 : \|\rho - \rho^g\| < r;\ \|DF_g(\rho) - DF_g(\rho^g)\| < \varepsilon\}.$$

Since DF_g is continuous, $r_\varepsilon > 0$ must hold. Pick $a_g \in (0,1)$ as defined previously. Then define

$$\overline{\varepsilon}_g = \frac{1 - a_g}{a_g} > 0.$$

By Lemmas 4.3 and 4.4 of Palis Jr, Melo, et al. (1982), $B_{r_\varepsilon}(\rho^g) \subseteq N_g$, if $\varepsilon < \overline{\varepsilon}_g$.

We are left to show that r_ε can be made to depend only on the eigenvalues of $DF(\rho^*)$ and γ. Notice that small enough $\underline{z} > 0$ pins down the $\delta > 0$ referred to in Lemma 6: Let

$$\hat{z}(\overline{\gamma}) = \inf_{\gamma \in (0,\overline{\gamma}]} \inf_{g \in \mathcal{B}^1_\gamma} \overline{\varepsilon}_g.$$

For $\delta > 0$ small enough, choose $\overline{\gamma} > 0$ such that Lemma 6 holds. It follows from the Lemma that $\hat{z}(\overline{\gamma}) > 0$. Then any $\underline{z} < \hat{z}(\overline{\gamma})$ satisfies our conditions and the conclusion follows. □

Now recall that by the proof of Proposition 1 point 2, $\rho^g \to \rho^*$ uniformly over $g \in \mathcal{B}^1_\gamma$ as $\gamma \to 0$. Thus there is γ small enough for which $\sup_{g \in \mathcal{B}^1_\gamma} |\rho^g - \rho^*| < \underline{z}$ and therefore $\rho^* \in N_g$ for all $g \in \mathcal{B}^1_\gamma$. Let $U_\gamma = \cap_{g \in \mathcal{B}^1_\gamma} N_g$. Since ρ^g for $g \in \mathcal{B}^1_\gamma$ are isolated invariant in U_γ by construction, the result follows.

■

Proof of Lemma 1

First, let $Q = q_1 + q_2$ and write

$$\begin{aligned}
u_1(q, q_2) &= Y(Q) + Y'(Q)q - c'(q);\\
u_{12}(q, q_2) &= Y'(Q) + Y''(Q)q;\\
u_{11}(q, q_2) &= 2Y'(Q) + Y''(Q)q - c''(q) = u_{12}(q, q_2) + Y'(Q) - c''(q).
\end{aligned}$$

From the above we see that since $P' < 0, c'' > 0$, we always have $u_{11}(q, q_2) < u_{12}(q, q_2)$.

Definition 6 implies that $u_1(0, q_2)$ strictly decreases over $q_2 \in X$ so that there exists unique $M^* < 2K$ with $u_1(0, M^*) = 0$. For $q_2 \leq 2K$, (iv) implies that $u_{12}(q, q_2) < 0$ for all $q \in [0, 2K - q_2]$. It follows that $u_1(0, q_2) > 0$ for all $q_2 < M^*$ and for all such q_2 there must be a unique $q^*(q_2) < 2K - q_2$ s.t. $u_1(q^*(q_2), q_2) = 0$. In addition, $u_1(q, q_2) > 0$ for $q \in [0, q^*(q_2))$ and $u_1(q, q_2) < 0$ for $q \in (q^*(q_2), 2K - q_2]$. Also note that $u_1(q, q_2) < 0$ for all $q \in [\max\{0, 2K - q_2\}, M]$ since $u_1(0, 2K) < 0$.

We can conclude from this that for all $q_2 \in X$ there is a unique best response $q^*(q_2)$ that is pinned down by first order conditions whenever $q_2 \leq M^*$, and equals zero otherwise.

Whenever $q_2 \leq M^*$, it must be that $u_{11}(q^*(q_2), q_2) < 0$ since $q^*(q_2) + q_2 < 2K$ and by convexity of c. It follows that we can find the derivative of the best response in q_2 for all $q_2 < M^*$ by the implicit function theorem, which allows us to show

$$\frac{\partial q^*(q_2)}{\partial q_2} = -\frac{u_{12}(q^*(q_2), q_2)}{u_{11}(q^*(q_2), q_2)} \in (-1, 0),$$

since $0 > u_{12}(q^*(q_2), q_2) > u_{11}(q^*(q_2), q_2)$ must hold. Finally, for there to be a unique interior Nash equilibrium we only need the following boundary condition to be satisfied: $q^*(0) < M^*$. In that case, 0 is not a best response to the monopoly quantity $q^*(0)$. This together with the fact that $\frac{\partial q^*(q_2)}{\partial q_2} \in (0, -1)$ for all $q_2 < M^*$ implies that there must be a unique interior Nash equilibrium (one can see this by imagining the best response and inverse best response plotted in 2D). It must be symmetric since the payoff functions (and therefore best response functions) are symmetric.

Static stability of this equilibrium (i.e. w.r.t. $F_B^{S_0}$) follows from the eigenvalues of the linearization of $F_B^{S_0}(q_N)$. A detailed exposition can be found in Appendix B. The relevant condition for stability comes down to $\frac{\partial q^*(q_2)}{\partial q_2} \in (-1, 0)$, which we have shown above. The intuition in the static case is can be exemplified in the following way: suppose 2's strategy is perturbed from q_N by a small amount, and players apply best responses to adjust thereafter. Then the best-response derivative tells us that 1 would react by moving in the opposite direction, but always by an amount *smaller* than the initial perturbation of 2. Continuing, 2 must react to 1's reaction again by moving in the opposite direction, and by a smaller amount than 1. The result is the well known cobweb-like path back to the Nash equilibrium. ■

Proof of Proposition 4

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

Lemma 8. *Suppose $h \in \mathcal{G}$. Then there exist parameters $P_H > P_L \geq 0$ and a convex cost function $c(q)$ such that the resulting stage game payoffs $u(q_1, q_2)$ are regular.*

Proof. By definition of $\mathcal{G}$, $\exists! D \in (\tau, \frac{M}{2})$ such that $-\underline{h}'' = \frac{h'(2D)}{D}$. Since h is strictly increasing, there exist $P_H > 0 > \big(P_L - P_H\big)$ such that

$$h(0) < \frac{P_H}{-\big(P_L - P_H\big)} < h(2D). \tag{24}$$

Recall that, for any cost function $c(q)$,

$$u_1(q, q) = P_H + \big(P_L - P_H\big)h(2q) + \big(P_L - P_H\big)h'(2q)q - c'(q),$$

and therefore the above implies that there exists a $c(q)$ such that $u_1(0,D) < 0 < u_1(0,0)$ (pinned down only by $c'(0)$). Then since $u_1(0,\hat{q})$ strictly decreases in $\hat{q} \in [0,M]$, there exists a unique $M^* \in (0,2D)$ such that $u_1(0,M^*) = 0$. Finally to check whether Definition 6 (iv) holds, note that

$$P'(q+\hat{q}) + qP''(q+\hat{q}) = \big(P_L - P_H\big)\big(h'(q+\hat{q}) + h''(q+\hat{q})q\big),$$

$P'(q+\hat{q}) + qP''(q+\hat{q}) \leq 0$ holds for all $q+\hat{q} \leq 2\tau$, since $h''(q+\hat{q}) \geq 0$ then. By definition of $\mathcal{G}$, we get

$$0 = \big(h'(2D) + \underline{h}''D\big) \leq \big(h'(Q) + h''(Q)q\big),$$

holds for all $q \in [\tau, D]$ and $Q \in [2\tau, 2D]$, since $h'(Q)$ is decreasing on that interval. The result follows. □

Now we need the following observations based on the definition of W in (3):

$$\begin{aligned}
W_1 &= \omega^{-1}(1-\delta P_{BB})\Big[\omega^{-1}\delta P'_{AB}(u^B - u^A) + u_1^A\Big],\\
W_2 &= \omega^{-1}(\delta P_{AB})\Big[\omega^{-1}\delta P'_{BB}(u^B - u^A) + u_1^B\Big],\\
W_{11} &= -2\omega^{-1}\delta P'_{AB}W_1 + \omega^{-1}(1-\delta P_{BB})\Big[\omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A\Big],\\
W_{22} &= 2\omega^{-1}\delta P'_{BB}W_2 + \omega^{-1}(\delta P_{AB})\Big[\omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B\Big],\\
W_{12} &= \omega^{-1}\delta\Big[P'_{AB}\frac{1-\delta P_{BB}}{\delta P_{AB}}W_2 - P'_{BB}\frac{\delta P_{AB}}{1-\delta P_{BB}}W_1\Big],\\
W_{13} &= W_{11} + \omega^{-1}(1-\delta P_{BB})\Big[\omega^{-1}\delta P'_{AB}(u_1^A - u_2^A) + u_{12}^A - u_{11}^A\Big],\\
W_{24} &= W_{22} + \omega^{-1}(\delta P_{AB})\Big[\omega^{-1}\delta P'_{BB}(u_2^B - u_1^B) + u_{12}^B - u_{11}^B\Big],\\
W_{14} &= -\omega^{-1}\delta P'_{BB}\frac{\delta P_{AB}}{1-\delta P_{BB}}W_1 + \omega^{-1}(1-\delta P_{BB})\omega^{-1}\delta P'_{AB}\Big[\omega^{-1}\delta P'_{BB}(u^B - u^A) + u_2^B\Big]\\
&= W_{12} + \omega^{-1}(1-\delta P_{BB})\omega^{-1}\delta P'_{AB}(u_2^B - u_1^B),\\
W_{23} &= \omega^{-1}\delta P'_{AB}\frac{1-\delta P_{BB}}{\delta P_{AB}}W_2 - \omega^{-1}(\delta P_{AB})\omega^{-1}\delta P'_{BB}\Big[\omega^{-1}\delta P'_{AB}(u^B - u^A) + u_2^A\Big]\\
&= W_{12} + \omega^{-1}(\delta P_{AB})\omega^{-1}\delta P'_{BB}(u_1^A - u_2^A).
\end{aligned} \tag{25}$$

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. If a policy is optimal, it must be optimal given any starting state s, and therefore one can characterize it through FOCs equivalently for any starting s.

We now prove some helpful Lemmas. First to save notation, let $\Delta = P_L - P_H$.

Lemma 9. *Suppose $h \in \mathcal{G}$ and*

$$h(2\tau)+h'(2\tau)\tau < h(2D).$$

Then for all neighborhoods $\mathcal{N}$ of τ there exist $P_H > 0 > \Delta$ and a convex $c(q)$ such that $q_N \in \mathcal{N}$. In particular, there exist $P_H > 0 > \Delta$ and a convex $c(q)$ such that $q_N = \tau$.

Proof. As argued in Lemma 8, there exist $P_H > 0 > \Delta$ and a convex $c(q)$ such that

$$h(0) < h(2\tau)+h'(2\tau)\tau < \frac{P_H - c'(0)}{-\Delta} < h(2D).$$

Thus the same arguments as in Lemma 8 can be applied to see that there is still a unique q_N Nash equilibrium. We haven't made any assumptions on $c(q)$ except for convexity and the possible range of $c'(0)$. If we pick $c'(\tau) > c'(0)$ such that

$$h(2\tau)+h'(2\tau)\tau = \frac{P_H - c'(\tau)}{-\Delta},$$

it follows that $u_1(\tau,\tau)=0$, i.e. $q_N=\tau$. The result follows. □

We can use Lemma 9 to make our analysis cleaner. As long as h satisfies the condition of the Lemma, q_N can be treated as a primitive of the model, replacing τ.

Lemma 10. *Suppose $h \in \mathcal{G}$ and $P_H > 0 > \Delta$ and a convex cost function $c(q)$ such that Lemma 8 holds. Then for all $\hat{q} \in [0, M^*]$ there exists a unique $q^*(\hat{q}) \in [0, M^*]$ such that*

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

If in addition $h'(0) = h'(M) = 0$ and

$$h(2\tau)+h'(2\tau)\tau < h(2D),$$

and for all $q \in [D, M]$,

$$3\Delta h'(2q) + 2\Delta h''(2q)q \leq c''(q),$$

then there exist $P_H > 0 > \Delta$ *and a convex* $c(q)$ *that satisfy Lemma 8 such that*

- *For all* $q \in (0, q_N]$ *there exists a unique* $\hat{q} \in [q_N, M)$ *such that*

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

- *For all* $q, \hat{q} \in (0, M)$

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

has a unique zero at $q = \hat{q}$.

Proof. For the first claim, recall from the proof of Lemma 8 that $\hat{q} \leq M^* < 2D$ implies that $u_{11}(q, \hat{q}) < 0$ for all $q \in [0, 2D - \hat{q}]$ by definition of $\mathcal{G}$. Then by construction of M^*, $u_1(0, \hat{q}) > 0$ holds for all $\hat{q} \in [0, M^*)$. Additionally, $u_1(q, \hat{q}) < 0$ holds for all $q \in [2D-\hat{q}, M]$, and there must be a unique zero $q^*(\hat{q}) \in (0, D - \hat{q})$.

For the second claim, recall from Lemma 8 that $u_1(q, q)$ is strictly decreasing for all $q \in [0, D]$. From Lemma 9 we get that we can take q_N arbitrarily close to τ. First let us fix $P_H, \Delta, c'(\tau)$ so that $q_N = \tau$. We also have that $h'(2q)$ is strictly increasing for $q \in [0, q_N)$, and strictly decreasing for $q \in (q_N, M]$. Finally, we have that $u_1(q_N, q_N) = 0$, so that the fraction must be strictly decreasing for $q \in [0, D]$. Note that we have so far only imposed two point conditions on convex $c(q)$, for $c'(0), c'(\tau)$. By assumption of the Lemma, for all $q \in [D, M]$,

$$3\Delta h'(2q) + 2\Delta h''(2q)q \leq c''(q)$$

then we have that $u_{11}(q, q) + u_{12}(q, q) \leq 0$ for all $q \in [0, M]$ and the fraction $\frac{u_1(q,q)}{h'(2q)}$ is monotone in q. Then, recall $h'(0) = h'(M) = 0$. So for any $q > 0$, no matter how close to 0, we can find $\hat{q} \in (q_N, M)$ so that the claim holds: increasing $\hat{q}$ to M would send the fraction to $-\infty$ after all.

For the third claim, consider three cases:

Case 1: $\hat{q} \leq q_N$.

Notice that $\hat{q} < q_N$ implies $u_1(\hat{q}, \hat{q}) > 0$, and as shown for the first claim, $u_1(q, \hat{q})$ is monotone decreasing on the candidate solutions $q \in [0, q^*(\hat{q}))$. But by construction,

$$\begin{aligned} u_1(2q_N - \hat{q}, \hat{q}) &= P_H + \Delta h(2q_N) + \Delta h'(2q_N)(2q_N - \hat{q}) - c'(2q_N - \hat{q}) \\ &= \Delta h'(2q_N)(q_N - \hat{q}) + c'(q_N) - c'(2q_N - \hat{q}) < 0, \end{aligned}$$

and thus it must be that $q^*(\hat{q})+\hat{q} < 2q_N$. Thus, $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is strictly decreasing on $q \in (0, q^*(\hat{q}))$. By monotonicity there can only be one solution, $q = \hat{q}$.

Case 2: $\hat{q} \in (q_N, D]$.

Firstly, consider $q \in [q^*(\hat{q}), \hat{q}]$. No smaller q is a candidate, since u_1 would change sign. Firstly in case $\hat{q} \leq Q_N$, $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is strictly decreasing on $q \in (q^*(\hat{q}), Q_N - \hat{q}]$. That is because for all such $q, \hat{q}$, $h''(q+\hat{q}) \geq 0$ holds. Then take $q \in (\max\{0, Q_N - \hat{q}\}, D]$. By definition of D, $u_{11}(q,\hat{q}) < 0$ for all such q and we get that $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is further strictly decreasing on $q \in (\max\{0, Q_N - \hat{q}\}, D]$.

Now suppose that $M^* > D$, and consider $q \in (D, M^* - \hat{q}]$. Then as shown in the proof of Lemma 8, $u_{11}(q,\hat{q}) < 0$ for all such $q, \hat{q}$ and the fraction is monotone still.

Finally, notice that we can write

$$\frac{u_1(q,\hat{q})}{h'(q+\hat{q})} = \frac{P_H + \Delta h(q+\hat{q}) - c'(q)}{h'(q+\hat{q})} + \Delta q. \tag{26}$$

Recall by definition of $M^* < 2D$, we have $P_H + \Delta h(q+\hat{q}) - c'(q) < 0$ for all $q \in [0, M]$ if $q + \hat{q} \geq M^*$. So we can write

$$\frac{\partial \frac{u_1(q,\hat{q})}{h'(q+\hat{q})}}{\partial q} = \frac{\big[\Delta h'(q+\hat{q}) - c''(q)\big]h'(q+\hat{q}) - h''(q+\hat{q})\big(P_H + \Delta h(q+\hat{q}) - c'(q)\big)}{h'(q+\hat{q})^2} + \Delta,$$

which is negative whenever $q + \hat{q} \geq M^*$. So both in the case where $D > M^*$, and when considering $q \in (M^* - \hat{q}, M]$, we still have that $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is strictly decreasing. All together it follows that, for any $\hat{q} \in (q_N, D]$, $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is strictly decreasing for $q \in [q^*(\hat{q}), M]$, as required.

Case 3: $\hat{q} \in (M^*, M]$.

Taking equation (26) and the above argument together implies that again, $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is strictly decreasing, for any $q \in [0, M]$, since $M^* > Q_N$ holds (i.e. $h''(q+\hat{q}) \leq 0$). The claim follows.

Note that by assumption, u_{ij}, h'' are continuous for all $i, j, \in \{1, 2\}$ in all their arguments. We can therefore find a neighborhood N' of τ such that for all $P_H > 0 > \Delta$ and a convex $c(q)$ that give $q_N \in N'$, the Lemma still goes through. □

Now for the proof of the Proposition:

Assume, as in Lemma 10, that $h'(0) = h'(M) = 0$ and

$$h(2\tau) + h'(2\tau)\tau < h(2D).$$

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

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

From now on, to save notation, I write h_s to denote evaluation of $h()$ at $q_s, s \in \{A, B, N\}$. By Lemma 10 we have that for any $q_A \in (0, q_N]$ there exists a unique $q_B \in [q_N, M)$ such that

$$\frac{u_1^A}{h'(Q_A)} + \frac{u_1^B}{h'(Q_B)} = 0.$$

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

$$z'(q_A) = -\frac{h'_B}{h'_A} \frac{u_{11}^A + u_{12}^A - 2h''_A \frac{u_1^A}{h'_A}}{u_{11}^B + u_{12}^B + 2h''_B \frac{u_1^B}{h'_B}}. \tag{27}$$

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

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

Firstly, recall that as in Lemma 9 we have that for $q > 0$ small enough, $u_1(q, q) > 0$ must hold. Now consider $\Psi(q_A)$:

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

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

Next, and analogously, take $q_A \in (q_N, M)$ to be large. In that case we let $y(q_A) = z^{-1}(q_A) < q_N$ be the inverse solution that equalizes first order conditions. Then if $q_A < M$ 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, note that

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

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

$$\begin{aligned}
&1 + \frac{u_{12}^N - 2\omega^{-1}\delta h'_N u_2^N}{u_{11}^N} < 0 \\
\Leftrightarrow \quad & h'_N \Delta h'_N q_N - h_N \Delta h'_N < h_N\big[2\Delta h'_N - c''_N\big] + R_1 + R_2 \\
\Leftrightarrow \quad & \Delta h'_N\big[h'_N q_N - 3h_N\big] < -h_N c''_N + R_1 + R_2,
\end{aligned}$$

where $R_1 = 2\Delta h_N q_N h''_N$ vanishes for q_N close enough to τ, and $R_2 = \frac{1-\delta}{2\delta}(u_{12}^N + u_{11}^N)$ vanishes as $\delta \to 1$. Then for $\delta < 1$ close enough to 1 and q_N close enough to τ, the condition stated in the Proposition is sufficient for $\Psi'(q_N) > 0$.
This together with $\Psi(q) > 0$ for q small, $\Psi(q) < 0$ for q large, allows us to use the intermediate value theorem. It gives us that there exists $q_A < q_N < q_B$ such that $W_1(\sigma) = W_2(\sigma) = 0$ for $\sigma = (q_A, q_B, q_A, q_B)$.

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 (25) that $W_{12} = 0$, so the Hessian must be diagonal at σ. A sufficient condition for negative definitess then is $h''_A > 0 > h''_B$ and $u^A > u^B$. The first one follows since $q_A < q_N < q_B$, the second one follows from the first order conditions:

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

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

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

holds for all $q \in S_A$. Equivalently, we can show that $q = q_A$ is the unique solution to the first order condition of this problem with respect to q, and that boundary conditions are satisfied so that the maximizer can only be interior. Taking derivatives, we get

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

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

Since the Hessian is negative definite at q_A, q_A, $H_1^A(q_A, q_A) = \frac{\partial H^A(q,q_A)}{\partial q}\Big|_{q=q_A} < 0$. Recall that in the proof of Lemma 10 we we showed that q_A is the only solution to $H^A(q, q_A) = 0$, but also that $\frac{u_1(q,q_A)}{h'(q+q_A)}$ is strictly decreasing over $q \in [0, q^*(q_A)]$. Thus, $H^A(0, q_A) > 0$ and $H^A(M/2, q_A) < 0$ must hold and q_A is globally optimal.

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

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

Where again by the negative definite Hessian, we have $H_1^B(q, q_B) < 0$. Then in the proof of Lemma 10 we show that $\frac{u_1(q,q_A)}{h'(q+q_A)}$ is strictly decreasing over $q \in [q^*(q_B), M/2]$. The result follows as above: q_B is globally optimal.

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

■

Proof of Corollary 4

First, since we are restricting to symmetric equilibria, it is sufficient to consider two cases: $u^A \lessgtr u^B$.

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

Recall that state A corresponds to P_L. As laid out in the proof of Proposition 4, we can write an agent's FOC for the problem of best responding in the following way:

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

where we plug in the fact that $P_{AB}(Q) = 1 - h(Q)$. For both equations, the leading term is strictly positive since $h'(Q) > 0$ for all interior Q. It follows that $u_1^s < 0$ must hold for both s.

In the proof of Lemma 10 I show that for $q_N \sim \tau$, (which we assume throughout, as in Proposition 4), we have that $\frac{u_1(q,q)}{h'(2q)}$ is strictly decreasing for all $q \in [0, M]$. At the same time, $u(q, q)$ is strictly decreasing when $q \geq q_N$, which is necessary for $u_1(q, q) < 0$. Thus, for case (i) it must be that $q_A > q_B$, but since $\frac{u_1(q,q)}{h'(2q)}$ is strictly decreasing, there exists no such pair q_A, q_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 Proposition 5

For point (i):

Using Lipschitz properties implied by the differentiability of price distribution and stage-game payoffs, I will show that profile σ_K can at most violate incentive constraints of Γ^∞ by an amount bounded by $\frac{1}{K}$.

Recall that in this analysis, players are allowed to use a public randomization device so as to make APS result's applicable. Specifically, the public signal ζ used by players must satisfy (A2 in APS): $\zeta \in \Omega \subseteq \mathbb{R}^a$, for some $a \geq 1$, being absolutely continuously distributed with p.d.f $g(\cdot, q)$, for all $q \in X^2$.

I will construct a public randomization device (PRD) that allows for players to both condition actions on the realization of the price signal $Y \in \mathbf{P} = \{P_L, P_H\}$, and the realization of the public randomization.

Let $\Omega = [0,1]$. Call the PRD ζ, so that ζ is unconditionally uniform on $[0,1]$, but conditional on price realization, ζ will realize above or below a cutoff:

$$Y = P_L \Rightarrow g(z,Q) = \frac{1}{Pr[P_L \mid Q]}\mathbf{1}\{z \in [0, Pr[P_L \mid Q]]\}; \tag{28}$$

$$Y = P_H \Rightarrow g(z,Q) = \frac{1}{Pr[P_H \mid Q]}\mathbf{1}\{z \in [Pr[P_L \mid Q], 1]\}. \tag{29}$$

Thus, given Q, a realization of ζ below or above $Pr[P_L \mid Q]$ allows players to condition continuation values on price realization, as well as conditional public randomizations. ζ satisfies APS's assumptions A2, A3 by construction.

Since $\bar{\sigma}_K$ is a bang-bang profile, it is pinned down by two quantities $\overline{q}_K, \underline{q}_K$, and two binary partitions of Ω. Pin those partitions down by specifying $\overline{\Omega} \subset \Omega, \underline{\Omega} \subset \Omega$ in the following way:

Whenever σ_K specifies to play $\overline{q}_K$ in period t, realizing $\zeta \in \overline{\Omega}$ implies $\underline{q}_K$ will be played in $t+1$, and whenever σ_K specifies to play $\underline{q}_K$ in period t, realizing $\zeta \in \underline{\Omega}$ implies $\underline{q}_K$ will be played in $t+1$. In keeping with APS' construction and notation, let $\overline{V}_K, \underline{V}_K$ maximal an minimal values of SSEs in E_K. Then we can write them as

$$\begin{aligned}
\overline{V}_K &= (1-\delta)u(\overline{q}_K, \overline{q}_K) + \delta\left[\overline{V}_K + \int_{\overline{\Omega}} \left(\underline{V}_K - \overline{V}_K\right)dz\right], \\
\underline{V}_K &= (1-\delta)u(\underline{q}_K, \underline{q}_K) + \delta\left[\overline{V}_K + \int_{\underline{\Omega}} \left(\underline{V}_K - \overline{V}_K\right)dz\right].
\end{aligned} \tag{30}$$

We can use this to define incentive constraints through one-shot deviations, using (28):

$$\begin{aligned}\overline{V}_K \geq\ & (1-\delta)u(q',\overline{q}_K) \\ & + \delta\big(\underline{V}_K - \overline{V}_K\big)\left[h(q'+\overline{q}_K)\int_{\overline{\Omega}\cap[0,h(q'+\overline{q}_K)]}\frac{1}{h(q'+\overline{q}_K)}dz\right] \\ & + \delta\big(\underline{V}_K - \overline{V}_K\big)\left[\big(1-h(q'+\overline{q}_K)\big)\int_{\overline{\Omega}\cap[h(q'+\overline{q}_K),1]}\frac{1}{1-h(q'+\overline{q}_K)}dz\right]\end{aligned}$$

holds for all $q' \in X_K$ since $\sigma_K \in E_K$. To see that σ_K is an ε-equilibrium of Γ^∞, it suffices to show that the right hand side above cannot grow too much by allowing for deviation in all of X. One can bound the following:

$$\begin{aligned}\max_{q'\in X}\max_{q'_K\in X_K}\Bigg|&(1-\delta)\big(u\big(q',\overline{q}_K\big)-u\big(q'_K,\overline{q}_K\big)\big) \\ &+\delta\big(\underline{V}_K-\overline{V}_K\big)\left[\int_{\overline{\Omega}\cap[0,h(q'+\overline{q}_K)]}dz-\int_{\overline{\Omega}\cap[0,h(q'_K+\overline{q}_K)]}dz\right] \\ &+\delta\big(\underline{V}_K-\overline{V}_K\big)\left[\int_{\overline{\Omega}\cap[h(q'+\overline{q}_K),1]}dz-\int_{\overline{\Omega}\cap[h(q'_K+\overline{q}_K),1]}dz\right]\Bigg|\end{aligned}$$

which is bounded above by

$$\begin{aligned}&\max_{q'\in X}\max_{q'_K\in X_K}\left\{(1-\delta)L_u|q'-q'_K|+2\delta\big(\overline{V}_K-\underline{V}_K\big)\int_{[m_1,m_2]}dz\right\} \\ &\qquad\leq(1-\delta)L_u\frac{1}{K}+2\delta\big(\overline{V}_K-\underline{V}_K\big)L_h\frac{1}{K},\end{aligned}$$

where L_u, L_h are maximal lipschitz constants for u, h, and with $m_1 = \min\{h(q'_K+\overline{q}_K), h(q'+\overline{q}_K)\}$ and $m_2 = \max\{h(q'_K+\overline{q}_K), h(q'+\overline{q}_K)\}$. The result follows: for any $\varepsilon > 0$, choose K large enough so that the worst incentive violation can be bounded by ε. An analogous argument holds for $\underline{V}_K, \underline{q}_K$.

Point (ii) follows by construction of $\bar{\sigma}_K$.

As for point (iii), using the PRD defined with (28), one can construct a binary state variable that supports $\bar{\sigma}_K$. Define a state variable $S_K \in \{A, B\}$ such that

$$\begin{aligned}P_{AB}(Q) &= Pr[\zeta \in \overline{\Omega} \mid Q], \\ P_{BA}(Q) &= Pr[\zeta \in \big\{\Omega \setminus \underline{\Omega}\big\} \mid Q].\end{aligned} \tag{31}$$

Then, using this state variable as the binary state underlying a binary policy ρ as in section 5, it is quick to check that indeed $W^i(\rho, s)$ as defined in (8) represents long term

expected discounted payoffs when players follow a bang-bang strategy with $\overline{q} = \rho(A)$, $\underline{q} = \rho(B)$ and partitions for a PRD support $\overline{\Omega}, \underline{\Omega}$. Then, when $\rho(A), \rho(B)$ are chosen to equal $\overline{q}_K, \underline{q}_K$, the resulting profile will be a ε equilibrium of the game with long-run payoffs defined through $W^i(\rho, s)$, by an argument analogous to the one above.

■