

# REINFORCEMENT LEARNING AND COLLUSION

Clemens Possnig

*Vancouver School of Economics, University of British Columbia*

January 19, 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 conditions 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, conditioning actions on a richer type of information can preclude the Nash equilibrium from being reached. This type of information allows for the existence of a collusive equilibrium that will be learned with positive probability.

**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 an algorithm takes 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<sup>1</sup> and simulation-based<sup>2</sup> 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 this 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. For example, the rich state variable I constructed supports a symmetric binary-state equilibrium that in

---

<sup>1</sup>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%.

<sup>2</sup>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.

one state plays collusive, low quantities, and high punishment quantities in the other. This equilibrium 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 these algorithms update with a decreasing step size<sup>3</sup> 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<sup>4</sup> (ACQ), this differential equation is a state-dependent best-response dynamic. 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.

If after a long enough amount of play, the performance criterion estimator differs from the true criterion at most by a bounded, smooth bias term<sup>5</sup>, attractors of the differential equation will be learned with positive probability, while unstable points will not be learned<sup>6</sup>. 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 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.

---

<sup>3</sup>Stepsizes must satisfy the Robbins-Monro condition commonly invoked in the computer science literature. See V. S. Borkar (2009), Chapter 2.

<sup>4</sup>See Dutta and Upreti (2022), Grondman et al. (2012) for relevant surveys.

<sup>5</sup>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. This is important since bias terms are often hard to pin down, but it may be possible to verify whether the bias is well-behaved as required here.

<sup>6</sup>An equilibrium of a differential equation is attracting (“stable”) if once trajectories of the differential equation reach a neighborhood of that equilibrium, they converge to it. Conversely, an equilibrium is repelling (“unstable”) if no matter how close a trajectory is to that equilibrium, it can be repelled from it and will not converge to it.

attracting).

**Learning Nash.** 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 tight comparisons possible.

While my results are more general, they are best understood in the example of repeated Cournot competition. 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 these state variables, I prove that a condition on market fundamentals is necessary and sufficient for a given stage game Nash equilibrium to be learned. The condition is a bound on the slope of the myopic best response function evaluated at that stage game Nash equilibrium.<sup>7</sup> 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, as is true for “1-recall policies” using the past period’s price as above, 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

---

<sup>7</sup>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)).

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.

Direction-switching states are best explained in the example of a binary state space  $S = \{1, 2\}$ . Let the probability of transitioning from state to state be determined by the aggregate quantity  $Q$ . I say that this state space is direction-switching if the probability of transitioning from state 1 to state 1 increases in  $Q$ , while the probability of transitioning from state 2 to state 1 decreases in  $Q$ . Thus, the incentive direction to achieve a given state in the future “switches”. I show that under this setting, no matter if statically stable or not, there is a transition probability such that the stage game Nash equilibrium will be dynamically unstable and therefore not learned. These observations indicate an important difference between 1-recall and direction switching states in the coordinative ability granted to learning algorithms. The intuition comes down to how perturbations of policies close to the Nash equilibrium are translated dynamically when algorithms play. Under 1-recall policies, it turns out that such perturbations must collapse to trivial perturbations that mimic myopic behavior. Under direction-switching states, that is not so: Incentives based on expected future payoffs get reinforced, facilitating coordination. Importantly, direction-switching states support binary carrot-stick style collusive equilibria as discussed next.

**Learning to collude.** Finally, 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. My results 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<sup>8</sup>. Using data on the firm’s payoff structure, market demand, and the states kept track of by the algorithms, one can devise a test to see whether a given equilibrium can be learned with positive probability.

---

<sup>8</sup>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.

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. Next, 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 statically stable but dynamically 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 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.

## 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

(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. Since this paper relies on some quite 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 C 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 way 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  $q \in A$  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  $q$ . Given  $s$ , the agent's expected payoff from choosing  $q$  is denoted  $u(q, s)$ . The agent discounts the future with  $\delta \in (0, 1)$ , and aims to find a policy  $\rho : S \mapsto A$  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

$$V(s) = \max_{q \in A} \left\{ u(q, s) + \delta \mathbb{E}[V(s') | q, s] \right\}.$$

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

$$Q^*(s, q) = u(q, s) + \delta \mathbb{E} \left[ \max_{q' \in A} Q^*(s', q') | q, s \right]. \quad (1)$$



This  $Q$ -function is related to the value function by  $V(s) = \max_{q \in A} Q^*(s, q)$ . The  $Q$ -function can thus be seen as a function evaluating the expected payoff from selling  $q$  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  $\rho^*$  simply by maximizing  $Q^*$  in every state.

C. J. C. H. Watkins (1989) then proposed a RL algorithm that estimates  $Q^*$ . In the language we 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  $q_t$ , and the next state  $s_{t+1}$  are observed.

The algorithm takes some initial value  $Q_0$ , and then updates the following way:

$$Q_{t+1}(s, q) = \begin{cases} Q_t(s, q) + \beta_t \left[ u_t + \delta \max_{q' \in A} Q_t(s_{t+1}, q') - Q_t(s, q) \right] & \text{if } s_t = s, q_t = q \\ Q_t(s, q) & \text{otherwise} \end{cases}, \quad (2)$$

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  $q_t$  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  $\varepsilon$ -greedy:

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

For a suitable sequence  $\beta_t$ , one can show that  $Q_t$  converges in probability to  $Q^*$  if states form a Markov chain controlled by  $q_t$  and actions are sampled  $\varepsilon$ -greedily (c.f. C. J. Watkins and Dayan (1992))<sup>9</sup>. Stationarity of the state-transitions conditional on a fixed policy  $\rho$  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.

---

<sup>9</sup>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.

### 3. Multi-Agent Reinforcement Learning

Now imagine the player described in Section 2 in fact faces multiple competitors in a market, which transforms our MDP into a game. Without any knowledge about their pay-off 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?

One would immediately have to do away with any stationarity assumptions. If each player acted  $\varepsilon$ -greedily, their policy would evolve over time, meaning that each player on their own faces a non-stationary environment, and convergence of  $Q$ -learning can fail. Calvano, Calzolari, Denicoló, et al. (2021) simulate this situation and show that still, one may see convergence to policies that resemble repeated-game strategies well-known to economists (for example,  $T$ -period punishment schemes as characterised in Green and Porter (1984)). The purpose of this paper is to characterize a class of algorithms for which it is possible to analytically describe the limiting policies resulting from algorithms 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  $\rho_t$  at the same time as estimating the  $Q$ -function, making it an actor-critic  $Q$ -learning (ACQ) rule as defined in the Introduction. In this section and the main body of the paper, 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.<sup>10</sup> There are multiple reasons why this is a difficult situation for such agents when it comes to learning a good policy. The fact that there are multiple agents involved in updating independent policies implies that each agent learns in a nonstationary environment. This manifests itself via each algorithm’s performance criterion estimators having to follow a moving target, since their opponent’s profiles are moving over time. As a result, value function approximations can be inconsistent. For a more thorough discussion of the issues introduced in multi-agent learning, see 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

---

<sup>10</sup>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.

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, as outlined above.

Albeit not an issue unique to multi-agent learning, the results in this paper are concerned with algorithms that learn to play continuous action policies<sup>11</sup>. In that case unbiasedly estimating a value function becomes a daunting 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 scalability. 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.

To introduce the model, first some mathematical definitions are required: Let  $X, Y$  be two metric spaces.

- Let  $\mathcal{C}^i[X, Y]$  be the set of functions that is  $i$  times continuously differentiable, with domain  $X$  and range  $Y$ .
- For any set  $B$ , define  $\overline{B}$  as the closure and  $\text{conv}[B]$  as the convex closure.

There are  $n$  algorithms indexed by  $i$ , each having as action space a compact interval  $A_i$ , with profile space  $A = \times_i A_i$ . A finite state space  $S$  with  $|S| = L$ <sup>12</sup> comes with a transition probability function  $T : S^2 \times A \mapsto (0, 1)$  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  $u^i : A \times S \mapsto \mathbb{R}$ ,  $C^2$  in  $A$ , and common discount factor  $\delta \in (0, 1)$ .

Algorithms update a policy function  $\rho_t^i : S \mapsto A_i$  the long-run behavior of which is the

---

<sup>11</sup>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  $Pr[q|s] = \frac{\exp(Q(s,q))}{\sum_{q'} \exp(Q(s,q'))}$ . This again falls into our continuous control scenario.

<sup>12</sup>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 paper 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 paper applies.

object of our interest. Since states are discrete, policy profile  $\rho_t \in \bar{A} = A^{nL}$  can be represented as a vector in  $\mathbb{R}^{nL}$ .

**Assumption 1.** *For all  $\rho \in \bar{A}$ , the Markov chain induced by  $T_{ss'}[\rho(s)]$  is irreducible.*<sup>13</sup>

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

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

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

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  $\rho^{-i} \in \bar{A}_{-i}$ , chosen from the constraint set of stationary,  $S$ -state policies:

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

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  $\rho^{-i}$  there must be a Markov stationary best response.

Define  $E_S \subset \bar{A}$  to be the set of Nash equilibria in policy profiles based on payoff functions  $W^i$ , as the set of profiles  $\rho^*$  s.t.  $\rho^{i*} \in B_S^i(\rho^{*-i})$  for all  $i$ . Then

**Definition 1.** *A Nash equilibrium  $\rho^* \in E_S$  is called 'differential Nash equilibrium' if first order conditions hold for each agent at  $\rho^*$  and the Hessian of each agent's optimization problem at  $\rho^*$  is negative definite.*

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

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

- *Given state space  $S$ , stationary equilibrium profiles  $\rho^* \in \bar{A}$  exist. Call the set of such equilibria  $E_S$ .*
- *There exist  $\rho^* \in E_S$  that are differential Nash equilibria.*

<sup>13</sup>For Definitions see e.g. Appendix A in Puterman (2014)

A sufficient condition for both points in Assumption 2 to hold is the existence of an interior static Nash equilibrium given  $u(r, s)$  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 space  $S$  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 space and current state  $s$  is a common observable to all algorithms competing. Thus, the state space can be interpreted as a model of what kind of information the RL is allowed to condition their policy on. In section 5 I will show how details of the state transitions will factor importantly in whether RL will be able to learn to play a stage game Nash equilibrium.

Now we are ready to state the running example of RL I consider. Assume that each algorithm uses the following adaptive rule to update their policy, which is known as actor-critic Q learning (ACQ):<sup>1415</sup>

**Definition 2.** *Each algorithm  $i$  updates policies  $\rho_t^i$  according to*

$$\rho_{t+1}^i(s) \in \rho_t^i(s) + \alpha_t \left[ \arg \max_{q' \in A} Q_t^i(s, q') - \rho_t^i(s) + M_{t+1}^i \right], \quad (5)$$

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’<sup>1617</sup>.

<sup>14</sup>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 that is widely cited and 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.

<sup>15</sup>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. Our 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.

<sup>16</sup>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 exist, 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).

<sup>17</sup>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.

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

$$Q^{i*}(s, q, \rho_t^{-i}) = u(q, s) + \delta \mathbb{E} \left[ \max_{q' \in A} Q^{i*}(s', q', \rho_t^{-i}) \mid q, s \right],$$

the correct  $Q^*$ -function conditional on  $i$ 's opponents playing profile  $\rho_t^{-i}$  forever into the future. This  $Q^*$  is related to  $W$  through the equation

$$\max_{q' \in A} Q^{i*}(s, q', \rho^{-i}) = \max_{\rho \in A_i} W^i(\rho, \rho^{-i}, s).$$

$Q_t^i$  is motivated from stationary MDPs as introduced in Section 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  $Q_t^i(s, q) = Q^{i*}(s, q, \rho_t^{-i})$ , 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  $\rho_t^{-i}$  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  $\rho_t^i$  decrease, computing  $Q^*$  can be seen as an approximation to the true future expected value that would take into account evolving  $\rho_t^{-i}$ . Secondly, as stated before, this paper is not looking into a positive 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 the following section, the assumptions made here will be sufficient to allow for collusive and other interesting behavior to emerge.

I will assume that  $Q_t^i$  is an estimator that tracks the correct function  $Q^{i*}$  well when  $t$  is large enough. The basic  $Q$ -estimator (2) defined to motivate  $Q$ -learning will not be enough to make this happen, as it requires discretization of the continuous action space and may run into issues due to the underlying non-stationarity of the problem. As the focus of this paper is on the limiting behavior of algorithms that do their job well enough, I maintain the high-level assumption

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

$$P\text{-}\lim_{t \rightarrow \infty} \sup_{(s, q) \in S \times A} \|Q_t^i(s, q) - Q^{i*}(s, q, \rho_t^{-i}) - g^i(s, q, \rho_t^{-i})\| = 0.$$

This Assumption 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 with probability 1. Furthermore, this limiting difference, or asymptotic bias, is a function of  $t$  only through its dependence on current period's profile  $\rho_t$ . Thus, Assumption 3 allows for an asymptotic bias term in the  $Q_t$  estimation, and later

on I will assume this bias to be small enough so that limits of  $\rho_t$  can be analysed as if  $g$  did not exist. As stated in the beginning of this section, the allowance of  $g$  is made 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 at all trivial, since it sweeps away the issue of non-stationarity when it comes to estimating  $Q_t$  discussed before. 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 more involved algorithms that can deal with both of these issues. In Possnig (2022), I develop low-level sufficient conditions on hyperparameters of actor-critic algorithms for which this is satisfied. Their estimation procedure is technically more involved but the intuition remains similar as in (2). For the stepsizes  $\alpha_t$  I maintain the following:

**Assumption 4.** *Robbins-Monro Condition on stepsizes:*

$\alpha_t \rightarrow 0$  with

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

This assumption takes its name from the celebrated Robbins-Monro algorithm representation (Robbins and Monro (1951)). The assumption constrains the speed of convergence of  $\alpha_t$ , that needs to balance the averaging out of errors  $M_{t+1}$  (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  $\rho_t$ :

**Assumption 5.** *Iterates stay bounded almost surely:*

$$\sup_t \|\rho_t\| < \infty, \text{ a.s..}$$

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 Benaim 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  $\rho_t$  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.

Define for  $\rho \in \bar{A}$

$$F_B^S(\rho) = \bar{B}_S(\rho) - \rho, \tag{6}$$



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

## 4. Limiting Behavior

Throughout, maintain Assumptions 2 - 5.

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

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

the set of limits of convergent subsequences  $\rho_{t_k}$ .

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.

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

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

Now I will 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. This is helpful in that one may find it hard to exactly determine what  $g$  is in a given application, but one could make assumptions that the bias  $g$  relevant for the application is at least bounded. For  $\gamma > 0$ , let  $\mathcal{B}_\gamma^i$  be the set of  $C^i$  functions with bounded derivatives :

$$\mathcal{B}_\gamma^i = \left\{ g : E \mapsto E; \mid \sup_{x \in \bar{A}} \|g(x)\| + \sum_{j=1}^i \sup_{x \in \bar{A}} \|D^j g(x)\| \leq \gamma \right\}, \quad (7)$$

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

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

- (1)  $\sup_{g \in \mathcal{B}_\gamma^1} |\rho^g - \rho^*| \rightarrow 0$  as  $\gamma \rightarrow 0$ .
- (2)  $P[L_{S,g} = \{\rho^g\}] > 0$ .

### Proof Sketch of Proposition 1

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

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  $\rho_t$  to solutions of the differential inclusion

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

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  $\alpha_t$  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  $\rho_t$  indeed must stay close, with probability one, to solutions of an underlying differential inclusion. The limiting behavior of  $\rho_t$  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  $\rho^*$  and  $\rho^g$ . 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  $\rho^*$  there is a unique rest point  $\rho^g$ . Further, stability of  $\rho^*$  must carry over to stability of  $\rho^g$ . Once it is established that  $\rho_t$  tracks solutions to the above inclusion over time, it makes sense that attracting points of the differential system will also attract  $\rho_t$  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  $\rho_t$  one may only see  $\varepsilon$ -equilibria of the game, defined below:

**Definition 5.** A profile  $\rho$  is an  $\varepsilon$ -equilibrium if for all players  $i$  all individual profiles  $\rho' \in \bar{A}$  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  $\gamma$  small enough and all  $g \in \mathcal{B}_\gamma^1$  there is a  $\bar{\varepsilon} > 0$  and a profile  $\rho^g$  such that

- (1)  $\rho^g$  is an  $\varepsilon$ -equilibrium for all  $\varepsilon \geq \bar{\varepsilon}$ .
- (2)  $\sup_{g \in \mathcal{B}_\gamma^1} |\rho^g - \rho^*| \rightarrow 0$  as  $\gamma \rightarrow 0$ .
- (3)  $P[L_{S,g} = \{\rho^g\}] > 0$ .

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  $\rho^*$  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  $\gamma$  small enough and all  $g \in \mathcal{B}_\gamma^1$  there is an open neighborhood  $U_\gamma$  with  $\rho^* \in U_\gamma$  such that*

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

### 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  $\rho^*$  and the rest points  $\rho^g$ .  $\rho^g$  being unstable hyperbolic implies that there exists an unstable manifold that  $\rho^g$  lies on, which acts as a repeller to the differential inclusion  $F_g$ . I go on to show that due to the instability of  $\rho^g$  and nonvanishing variance of  $M_{t+1}$ , no matter how close the algorithm updates come to  $\rho^g$ , and no matter how large  $t$  is, there is always a high probability that  $\rho_t$  lands on the unstable manifold and therefore must move away from  $\rho^g$ . Finally I show the existence of a neighborhood  $U_\gamma$ . I show that due to the hyperbolicity of  $\rho^*, \rho^g$ , there is a neighborhood  $U$  around  $\rho^g$  with  $\rho^* \in U$  such that  $\rho^g$  is the only internally chain transitive set within  $U$ . Recall that  $\rho^*$  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  $\rho_t$  is close to an equilibrium  $\rho^*$ . Since  $i$ 's updating rule tracks  $F_B^S$ , their policy will only stay close to  $\rho^*$  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 (Benam and Faure (2012)) that suggest one may extend Proposition 2 to such linearly unstable cycles also.<sup>18</sup>

In the rest of the paper I restrict attention to the equilibrium limiting points of the algorithm learning process.

## 5. Learning to Collude in a Cournot Game

In this section, I apply my limiting characterisation to a game of repeated Cournot-competition played by ACQ algorithms as introduced in Definition 2.

Given the link between state dependent best-response dynamics  $F_B^S$  and long-run RL behavior I have established in the previous sections, this section will consider the existence and stability of equilibria under  $F_B^S$  in a repeated Cournot game more closely. First, I study the static Nash equilibrium, which under general conditions is a unique, symmetric stage game equilibrium of the Cournot game. I show how the state variables RL condition their policies on can affect their ability to learn or move away from the stage game Nash equilibrium. I will uncover necessary conditions on policy spaces so that the stage game Nash equilibrium will not be learned. In addition, I give sufficient conditions on the payoff environment so that this holds.

Next, I give a specific example of a type of state space that, when conditioned on by RL, not only renders the stage game Nash equilibrium unlearnable by RL, but also supports a collusive equilibrium that is learnable. This shows an additional warning sign when it comes to types of commonly observable states that may have problematic welfare consequences when RL condition their policies on them.

Finally, I give a fully developed numerical example and provide simulation studies visualizing the intuitions provided in the section. The game is set up as follows:

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

$$Pr[Y = P_L | Q] = h(Q),$$

with  $h'(Q) \geq 0$ .

- Expected price conditional on  $Q$ :

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

---

<sup>18</sup>The inclusion of an analysis of limit cycles is an interesting avenue of further research, but would be beyond the scope of this paper.

- 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  $Q = q_1 + q_2$ .

As shown in the previous section, determining the stability of equilibria is essential in determining whether they can be learned or not. Since stability of equilibria is defined with respect to state-dependent best response dynamics  $F_B^S$ , the definition of the state space is part of what can determine stability. Recall that I therefore emphasize this dependence by writing  $F_B^S$  as the best response dynamics defined on state space  $S$ .

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. Under  $F_B^{S_0}$ , it is well known that under general conditions on  $P(Q)$ , 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 space  $S = \{A, B\}$ . Define for any  $s \in S$ , and  $q_i \in I$ :

$$P_{sB}(q_1, q_2) = \Pr[s' = B \mid s; q_1, q_2],$$

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 sections. Also assume that

$$P_{sB}(q_1, q_2) = \Pr[s' = B \mid s; q_1 + q_2],$$

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  $\rho^i : S \mapsto I$  be each players policy, and recalling the definition of  $W^i$  in (3), note that in the binary case one can derive

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

where

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

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  $\delta \rightarrow 1$ , these weights will

converge to the unique stationary distribution over states given the policy profile  $\rho$ .<sup>19</sup>

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  $\rho^{*1}$  is an interior best response to  $\rho^2$  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  $\rho^2$ , which will be an essential building block in finding stability conditions of an equilibrium. Since policies are vectors in  $I^2$ , from now on I will use the conventions  $\rho^i(s) = \rho_s^i \in I$  for all  $i, s$ .

$$J(\rho^{*1}, \rho^2) = \begin{bmatrix} \frac{\partial \rho_A^{*1}}{\partial \rho_A^2} & \frac{\partial \rho_A^{*1}}{\partial \rho_B^2} \\ \frac{\partial \rho_B^{*1}}{\partial \rho_A^2} & \frac{\partial \rho_B^{*1}}{\partial \rho_B^2} \end{bmatrix}. \quad (8)$$

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

- $u^A = u(\rho_A, \rho_A)$ ,  $u^B = u(\rho_B, \rho_B)$ .
- $u_i^s = \frac{\partial u^s}{\partial q_i}$  and  $u_{ij}^s = \frac{\partial u_i^s}{\partial q_j}$ , 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.

More explicitly, one can then write

$$\begin{aligned} \frac{\partial \rho_A^{*1}}{\partial \rho_A^2} &= -1 + \frac{\omega^{-1} \delta P'_{AB}(u_2^A - u_1^A) + u_{11}^A - u_{12}^A}{\omega^{-1} \delta P''_{AB}(u^B - u^A) + u_{11}^A}, \\ \frac{\partial \rho_A^{*1}}{\partial \rho_B^2} &= \frac{\omega^{-1} \delta P'_{AB}(u_1^B - u_2^B)}{\omega^{-1} \delta P''_{AB}(u^B - u^A) + u_{11}^A}, \\ \frac{\partial \rho_B^{*1}}{\partial \rho_A^2} &= \frac{\omega^{-1} \delta P'_{BB}(u_2^A - u_1^A)}{\omega^{-1} \delta P''_{BB}(u^B - u^A) + u_{11}^B}, \\ \frac{\partial \rho_B^{*1}}{\partial \rho_B^2} &= -1 + \frac{\omega^{-1} \delta P'_{BB}(u_1^B - u_2^B) + u_{11}^B - u_{12}^B}{\omega^{-1} \delta P''_{BB}(u^B - u^A) + u_{11}^B}. \end{aligned} \quad (9)$$

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  $q \in I$ .
- (iii)  $Y'(2q) < 0$  for all  $q < M$ .

<sup>19</sup>[Uniqueness is implied by our irreducibility assumption 1.]

(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  $I$ , 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 spaces exist so that when RL condition their policies on them, they will not learn the statically stable Nash equilibrium:

First define a set of binary-state strategies that are special in the way they prescribe state transitions.

**Definition 7.** *Say a binary state policy is a DS-policy (‘DS’ for direction-switching) if the underlying state transitions are irreducible and  $P_{AB}(Q) = 1 - P_{BB}(Q)$  holds for all  $Q$ . Denote the DS- state space as  $S^{DS}$ .*

In words, the probability of reaching any state  $s$  conditional being in  $A$  is complementary to the probability of reaching  $s$  conditional on being in  $B$ . Notice that this can affect how players evaluate continuation payoffs in an important way: for a given quantity  $Q$ , marginal deviations affect expected continuations in the opposite direction. This fact introduces an essential difference in how states  $A, B$  are interpreted.

**Proposition 3.** *Let  $u$  be regular and  $\zeta_N$  be the DS-policy that plays  $q_N$  in every state. Then  $\zeta_N$  is dynamically unstable (i.e. unstable w.r.t.  $F_B^{S^{DS}}$ ) if*

$$-\frac{u_{12}^N}{u_{11}^N} + 2D_N > 1,$$



where

$$D_N = \delta \frac{P'_{AB}(Q_N)}{\omega} \frac{\delta u_2^N}{u_{11}^N}.$$

*Proof.* As discussed in Appendix B, one needs to linearize best responses at  $\zeta_N$  to determine the stability of that profile. Taking (9), this simplifies to

$$J(\zeta_N, \zeta_N) = \begin{bmatrix} -\frac{u_{12}^N}{u_{11}^N} + D_N & -D_N \\ -D_N & -\frac{u_{12}^N}{u_{11}^N} + D_N \end{bmatrix},$$

with  $D_N$  as defined in the statement of this Proposition. The resulting matrix has two eigenvalues  $\lambda_j \in \{-\frac{u_{12}^N}{u_{11}^N}, -\frac{u_{12}^N}{u_{11}^N} + 2D_N\}$  for  $j \in \{1, 2\}$ . The first one is the familiar eigenvalue representing the slope of the static best-response which determines static stability of  $q_N$ . The second eigenvalue is now of interest to us. By regularity,  $\frac{u_2^N}{\omega u_{11}^N} > 0$ . Thus, if  $P'_{AB}(Q_N) > 0$  and large enough, it is possible for this eigenvalue to cross the threshold of 1 and therefore render  $\zeta_N$  dynamically unstable. Thus, for any DS-policy with large enough  $P'_{AB}(Q_N)$ , the static equilibrium will be dynamically unstable no matter if it is statically stable.  $\square$

Proposition 3<sup>20</sup> reveals that a binary state policy playing the static Nash equilibrium repeatedly can be unstable under  $F_B^{S^{DS}}$  (i.e. dynamically unstable), and therefore cannot be a limiting point of the algorithm-process according to Proposition 2. The condition determining the dynamic stability of Nash reveals a trade-off between static and dynamic incentives. As discussed in the proof above, stability of  $\zeta_N$  comes down to considering the eigenvalues of the matrix of best-response derivatives at  $\zeta_N$ , which I re-state here:

$$J(\zeta_N, \zeta_N) = \begin{bmatrix} -\frac{u_{12}^N}{u_{11}^N} + D_N & -D_N \\ -D_N & -\frac{u_{12}^N}{u_{11}^N} + D_N \end{bmatrix}, \quad (10)$$

where  $D_N$  is defined in the statement of Proposition 3. The fact that  $D_N > 0$  is the same number for every derivative above is not surprising, since I start from  $\zeta_N$ , which is a profile at which every state has the same action and value. What is important is the sign multiplied to  $D_N$ . As stated in Proposition 3,  $\zeta_N$  is dynamically unstable if

$$-\frac{u_{12}^N}{u_{11}^N} + 2D_N > 1. \quad (11)$$

Notice that the first term represents the derivative of the best response in the *static* Cournot game, evaluated at  $q_N$ . It is well known that static stability of  $q_N$  is equivalent to  $-\frac{u_{12}^N}{u_{11}^N} \in (-1, 1)$ , so this term is directly related to the static stability property of  $q_N$ .

---

<sup>20</sup>Proposition 3 holds under more general state-and price spaces which are non-binary, as established in Appendix B.2.

Recall that by definition of regularity,  $q_N$  must be statically stable. The second term in (11) represents the dynamic incentive introduced when *DS*-policies are played. To understand it, it is best to consider the following thought experiment:

Suppose player 1 plays  $\zeta_N$ , and call player 2's *DS*-policies  $\beta$ . Start at  $\beta = \zeta_N$ . Now let player 2 deviate to  $\beta' = (q_N - \varepsilon, q_N)$ . In other words, 2 marginally decreases their quantity in state *A*, while keeping the quantity in state *B* constant. How will player 1 react?

By the strategic-substitute-nature of the stage game, player 1's static incentive is to increase their quantity, by an amount less than  $\varepsilon$ . This is also what enforces the static stability of  $q_N$ . In (11), the decrease of player 2's quantity in state *A* would imply a positive first term. The other effect of 2's deviation on 1's payoff however is that  $u(q_N, q_N - \varepsilon) > u^N$ . In other words, 1's payoff in state *A* increased. From a dynamic perspective, it is beneficial for 1 to increase the likelihood of being in the good state, which corresponds to *decreasing* 1's quantity *A*. This incentive is what is represented by the second term in (11), which I call the *dynamic incentive*. It features the derivative in the likelihood of staying in *A*,  $-\frac{P'_{AB}}{\omega} < 0$ , multiplied by the change in the *A*-state's payoff  $u_2^N$  due to the deviation of player 2, weighted by the Hessian  $u_{11}^N$ . 2's deviation then implies a negative second term in (11), which introduces the tension between static and dynamic incentive mentioned above.

Thus in the above game, the static and dynamic stability property of the Nash equilibrium differs. This fact is not an immediate implication from checking the dynamic stability based on any non-trivial state space. I show now that in fact, for there to be a difference between static and dynamic stability, it is necessary for the dynamic strategy to *not* be a one-recall strategy (1R-policy). That is, when the state space is represented by a one-period recall of a public observable, (in our example: the past period's price), then the static and dynamic stability property of the Nash equilibrium must coincide.

This insight can be seen as the first step in introducing a categorization of state-dependent strategies based on the evolution of the state. Categorize *DS*-policies as allowing for more coordinative ability to the RL when it comes to learning a stage game Nash equilibrium<sup>21</sup> than 1R-policies as defined below<sup>22</sup>.

**Definition 8.** A public 1R - policy can be defined as policy  $\rho : \mathbf{P} = \{P_L, P_H\} \mapsto I$ , so that states are price realizations representing last periods observed price. This can equivalently be defined as having a state space  $\mathbf{P}$  with transition function  $T(s, P) \in \mathbf{P}$  such that  $T(s, P) = P$  for all  $s \in \mathbf{P}$ , and all price observations  $P \in \mathbf{P}$ .

<sup>21</sup>Here coordinative ability refers to the fact that RL can learn to coordinate to move *away* from the Nash equilibrium.

<sup>22</sup>The restriction to a binary-price game is not necessary for this insight. A treatment of a game with arbitrarily many prices is given in Appendix B.1.

**Proposition 4.** *Let  $\rho_N$  be the 1R-policy that plays stage game Nash quantity  $q_N$  in every state. Then  $\rho_N$  is asymptotically stable if and only if  $q_N$  is.*

*Proof.* The proof is analogous to the proof of Proposition 7 in Appendix B.1.  $\square$

This result is in stark contrast to Proposition 5, as discussed using (11). It implies that in fact, for 1R-policies, information about transition probabilities and dynamic incentives is irrelevant when determining whether a Nash equilibrium can be learned. The intuition why 1R-strategies are not enough to make a distinction in the stability of a Nash equilibrium comes down to how dynamic incentives behave when 1R-strategies are played.

Intuitions are provided by comparing the best response derivatives in the case of  $DS$ -policies versus the case of 1R-strategies. For 1R-strategies, let state  $A$  correspond to  $P_L$  and state  $B$  correspond to  $P_H$ . Take the transition probability  $P_{sB}$  resulting in  $DS$ -policies and denote  $P^{1R}$  as the transition probability resulting in the 1R-policy, so that  $P_{AA}^{1R}(Q) = Pr[P_L \mid Q]$ . To make comparisons between these two state variables more direct, also assume that  $P_{sB}^{1R}(Q) = P_{ss}(Q)$  for all  $Q$ . This assumption is satisfied if a  $DS$ -state follows a transition function given observations of the price, as is studied more thoroughly later on (c.f. Figure 5). Now note that for  $DS$ -policies, one can write

$$J(\zeta_N, \zeta_N) = -\frac{u_{12}^N}{u_{11}^N} I_2 + D_N \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = -\frac{u_{12}^N}{u_{11}^N} I_2 + D_N K^*, \quad (12)$$

where  $I_2$  denotes the 2-dimensional identity matrix and  $K^* = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ . In comparison let  $\rho_N$  be the 1R-policy that plays  $q_N$  in every state. Then

$$J(\rho_N, \rho_N) = \begin{bmatrix} -\frac{u_{12}^N}{u_{11}^N} + D_N & -D_N \\ D_N & -\frac{u_{12}^N}{u_{11}^N} - D_N \end{bmatrix} = -\frac{u_{12}^N}{u_{11}^N} I_2 + D_N K^{1R}, \quad (13)$$

where  $K^{1R} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$ . The importance is in the difference between  $K^*$  and  $K^{1R}$ . One can read these matrices as encoding how a perturbation in an opponent's strategy dynamically affects the best response. Firstly, each column  $c$  represents how for a perturbation in opponent's strategy in state  $c$ , both state's best response quantities will be changed. Importantly, note that  $K^{1R}$  tells us that for a given perturbation in state  $c$ , one reacts optimally *in the same direction* no matter which state's optimal quantity is considered. This is precisely due to the fact that  $K^{1R}$  is constructed based on the 1R-transition function. No matter what state it is, the likelihood of moving to another state conditional on a fixed

quantity has to be the same. In contrast, the transitions due to  $DS$  strategy  $\zeta_N$  give a change in direction for each state: each column of  $K^*$  has alternating signs. Thus for a perturbation in a given state  $c$ ,  $K^*$  implies differential optimal responses across states. This is the essence of the difference between 1R and  $S^*$ -strategies. Due to its sign-structure,  $K^{1R}$  is a *nil-potent* matrix, meaning that  $(K^{1R})^2 = 0$ . As a result the dynamic effect  $D_N$  cancels out in the eigenvalue calculation of  $J(\rho_N, \rho_N)$ , while  $K^*$ 's sign structure is such that the dynamic effect is reinforced, leading to the second eigenvalue of  $J(\zeta_N, \zeta_N)$  being different from  $-\frac{u_{12}^N}{u_{11}^N}$ .

A further intuition can be gained by realizing that the nil-potency of  $K^{1R}$  renders  $J(\rho_N, \rho_N)$  defective; meaning that it has a repeat eigenvalue  $\lambda_1 = \lambda_2 = -\frac{u_{12}^N}{u_{11}^N}$ , but a unique eigenvector  $v_0 = (1, 1)'$ . As further explained below, this fact implies that any perturbation around  $\rho_N$  will collapse to a perturbation along  $v_0$ . The following thought experiment explains why this is important: Suppose at  $\rho_N$ , player 2 perturbs their policy along  $v_0$ . This represents the one perturbation in 2-dimensional policy space for which the optimization problem faced by player 1 is equivalent to the myopic problem, as they are starting from  $\rho_N$ , and seeing a perturbation equal in both states. It is then natural that when the equilibrium is statically stable, this type of perturbation can not break its stability, and players will not escape the Nash equilibrium.

To elaborate how this result emerges, one needs to consider the role of eigenvalues and eigenvectors in characterising the behavior of solutions to ordinary differential equations (ODEs) close to an equilibrium. The celebrated Hartman-Grobman Theorem (c.f. Chicone (2006), Theorem 4.8) connects the flow of a nonlinear ODE in the neighborhood of a hyperbolic equilibrium (see Definition 4) to the flow of a linearized ODE at the equilibrium. In our case, as discussed above, linearizing the relevant ODE comes down to finding the best-response derivatives at the equilibrium. Our relevant ODE is then a 4-dimensional system, as the policy profile is 4-dimensional. However, due to the symmetry of payoffs and the equilibrium, it is enough to get intuitions based on a single agent's 2-dimensional best response derivative matrix  $J(\cdot, \cdot)$ .<sup>23</sup>

First, let 2-dimensional matrix  $\mathbf{M}$  be diagonalizable, and consider the problem of finding solutions to linear ODE

$$\dot{x} = \mathbf{M}x. \tag{14}$$

---

<sup>23</sup>For a more detailed discussion of stability analysis under  $F_B^S$  and related issues see Appendix B.

Let  $v_1, v_2$  and  $\lambda_1, \lambda_2$  be eigenvectors and eigenvalues of  $\mathbf{M}$ . Then solutions to (14) can be written as linear combinations

$$x(t) = c_1 e^{\lambda_1 t} v_1 + c_2 e^{\lambda_2 t} v_2, \quad (15)$$

for any scalar  $c_1, c_2$ . I take the zero-vector as equilibrium without loss, as any linear ODE can be shifted by a constant vector. Then  $\lambda_j < 0$  for both  $j$  implies that 0 is a global attractor, while if  $\lambda_j > 0$  for at least one  $j$ , there is a direction that repels from 0. This is the image one should have in mind in case of DS-policy  $\zeta_N$ , since  $J(\zeta_N, \zeta_N)$  is not defective, having a 2-dimensional eigenspace that spans the whole plane. As a result,  $J(\zeta_N, \zeta_N)$  can have a repelling eigenvalue when dynamic effect  $D_N$  is large enough.

Now, recall that  $J(\rho_N, \rho_N)$  is a *defective* matrix, with unique eigenvector  $v_0 = (1, 1)'$ . Solutions to linear 2-dimensional ODEs based on a defective matrix take a different form from (15). It can be shown that there must exist a generalized eigenvector  $v_1$ , linearly independent from  $v_0$  (that is of course not an actual eigenvector to the defective matrix), so that one can write the general solution

$$x(t) = e^{\lambda t} [c_1 v_0 + c_2 (v_1 + t v_0)].$$

Now suppose (as in our case of  $J(\rho_N, \rho_N)$ ) that  $\lambda < 0$ . 0 is attracting as expected, however the solutions now can be split into a fast direction  $v_1$ , and a slow direction  $v_0$ , meaning that any solution will first approach the slow direction  $v_0$  and then approach 0. In other words, the unique eigenvector can be seen as the *essential* direction of flow close to the equilibrium, since for large  $t$  the fast direction represented by the generalized eigenvector will have collapsed. The intuition given in the beginning of this discussion then follows: all perturbations must collapse to trivial perturbations along  $v_0$ .<sup>24</sup>

Now that it is established that for policy spaces allowing for transitions to switch directions as specified for *DS*-policies one can have dynamically unstable  $q_N$ , one may ask: “what, then, could be learned by the RL?”

Given a subset of regular  $u$ , I construct a specific *DS*-policy for which not only is  $q_N$  dynamically unstable (while by definition of regular  $u$  being statically stable), but also there exists a symmetric, collusive equilibrium  $\sigma$  with  $\sigma_A < q_N < \sigma_B$ . In other words, an equilibrium in which in state  $A$ , collusive quantities are played, while in state  $B$ , punishment

---

<sup>24</sup>When considering the full 4-dimensional system, analogous intuitions can be derived, with the slight complication being that the 4-dimensional system has 2 unique eigenvectors; therefore still being defective. The two eigenvectors in that system are the unit vector corresponding to  $v_0$  above, and a vector  $v_1 = (1, 1, -1, -1)'$ , which correspond to perturbations in which both agents change their policy the same way in each state, but in opposite directions. Thus the intuition that only myopic incentives matter is not affected.

quantities are played.

I call this specific  $DS$ -state space  $S^*$ . It comes with the following transition which evolves based on observations of the current period's price outcome  $Y_t$ . State transitions can be represented in a diagram:

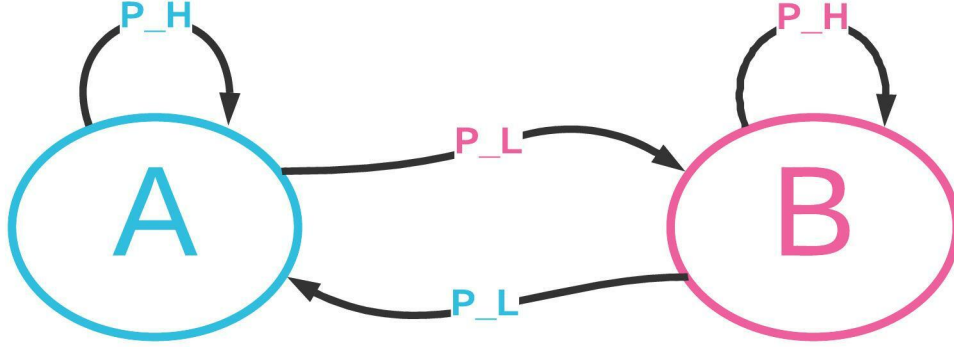


FIGURE 1. State Transition Diagram

In words, 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

$$P_{AB}(Q) = \Pr[P_L \mid Q] = h(Q); \quad P_{BB}(Q) = \Pr[P_H \mid Q] = 1 - h(Q),$$

from which one can directly verify that  $S^*$  is indeed an  $S^{DS}$ -state space.

Since the goal of this exposition is analytical tractability<sup>25</sup>, 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, I will assume for  $h$  (and therefore  $P(Q)$ ) to take an  $S$ -shaped form:

**Definition 9.** Fix  $M > 0$ . Let  $\mathcal{G}^o$  be the space of strictly monotone, twice differentiable functions  $h : [0, M] \mapsto [0, 1]$ . Let  $\underline{h}'' = \min_{Q \in (0, M)} h''(Q)$ . Define a space of  $S$ -shaped functions with lower bounded second derivative as

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

<sup>25</sup>Tractability goes a long way when it comes to proving statements about eigenvalues of matrices that are of dimension  $\geq 4$ , which is the smallest system one can face when considering repeated game strategy profiles for  $\geq 2$  players.

Let  $\zeta_N$  be the  $S^*$ -policy playing  $q_N$  in every state.

**Proposition 5.** *There exists  $h \in \mathcal{G}$ ,  $P_H > P_L \geq 0$  and a convex  $c(q)$  such that resulting  $u$  is regular,  $\zeta_N$  is dynamically unstable and there exists a symmetric equilibrium  $\sigma$  with  $0 < \sigma_A < q_N < \sigma_B$ .*

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 5 also implies that there exists a another symmetric collusive equilibrium  $\sigma'$ , satisfying  $\sigma'_A = \sigma_B, \sigma'_B = \sigma_A$ .

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.

The following subsection gives a numerical example that supports a stable collusive equilibrium.

## 5.1. Example: Stable Collusion

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  $[0, M]$ , 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

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

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  $\bar{h} \in [0, 1)$  is strictly increasing, with  $\bar{h}'(M) = 0$ ,



$\bar{h}''(x_4) = 0$  and  $\bar{h}''(Q) < 0$  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 5.

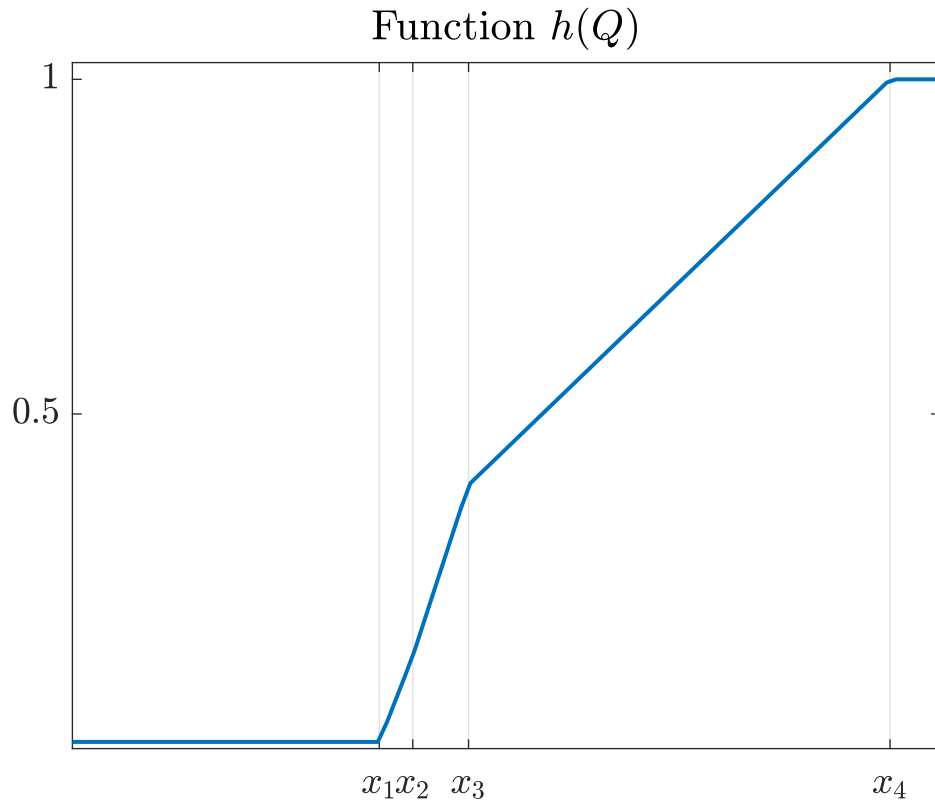


FIGURE 2. The plot indicates the location of the cutoffs  $x_i$  for piecewise linear  $h(Q)$ .

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:

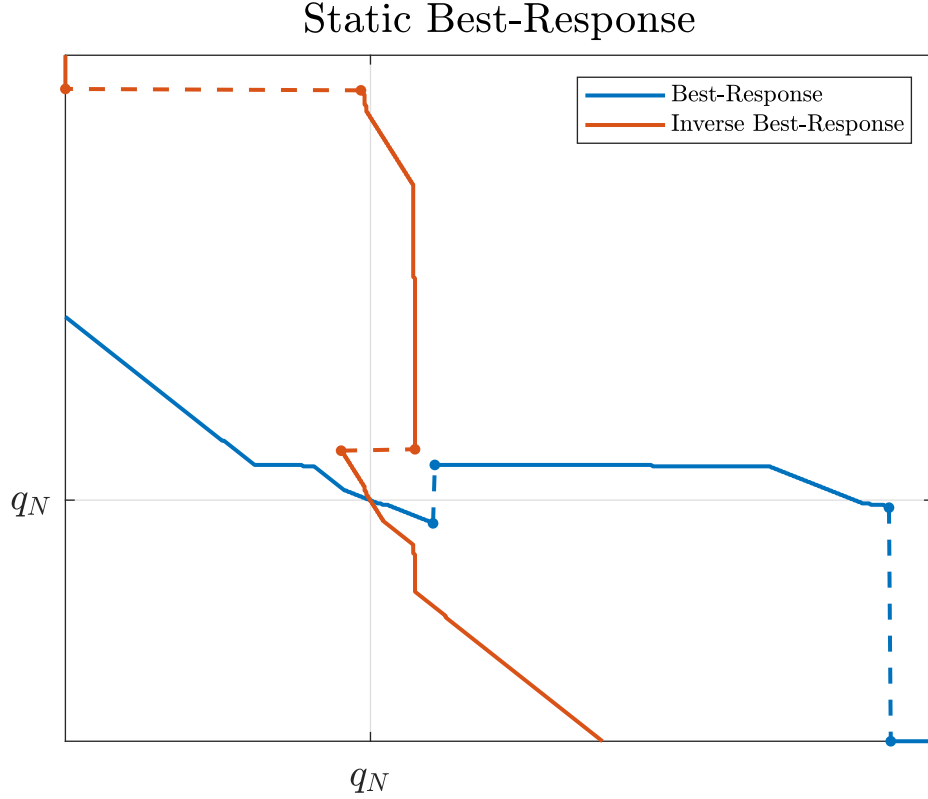


FIGURE 3. The plot shows the unique intersection of static best-response functions in the numerical example, where  $q_N = 1.58$ . Best responses jump upwards at a value larger than  $q_N$ . This is due to the non-concavity introduced by the requirement  $h'_B < h'_N$ , which enforces the S-shape of 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,

$$-\frac{u_{12}^N}{u_{11}^N} = -0.5; \quad D_N = 1.14,$$

which implies from (11) that the Nash equilibrium is statically stable but indeed dynamically unstable. At the same time, this numerical example supports a pair of symmetric, collusive equilibria  $\sigma, \sigma'$  with  $\sigma_A = 1.52 < \sigma_B = 1.94$  (rounded to two decimal points) and the quantities flipped for  $\sigma'$ . (see the discussion after Proposition 5. 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 Section 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  $s$ ,

$$\rho_{t+1}^i(s) \in \rho_t^i(s) + \alpha_t \left[ \arg \max_{q' \in I} Q^{i*}(s, q', \rho_t^{-i}) - \rho_t^i(s) + M_{t+1}^i \right], \quad (16)$$

where  $\alpha_t = t^{-0.6}$  satisfies the Robbins-Monro Assumption 4, and  $M_{t+1}^i \sim N(0, .25)$  is an i.i.d mean-zero Normal noise variable with variance 0.25. Notice that (16) 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 Section 4 tell us, for algorithms in the class studied in this paper this simulation will give us an equivalent representation of long-run trajectories of  $\rho_t$  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 Section 4 are local in nature: if the iteration  $\rho_t$  at some point  $t$  enters a basin of attraction for a given stable equilibrium  $\rho^*$ , then the iteration will converge to that equilibrium with large probability.<sup>26</sup> 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 25000 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 policy space of the algorithms involved, iterations move closer to the

---

<sup>26</sup>In fact, approximations of this probability can be made given the neighborhood of  $\rho^*$  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 developed 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 Proposition 4. Since in this example, the Nash equilibrium is statically stable, its repetition under 1R-policies  $\rho_N$  is also stable. Thus, one would expect that once algorithms running on the 1R-policy space 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 shown in Figure 5.1. Since the state space is binary, the two algorithms' policies can be represented as points in the  $I^2$ -plane. I now plot simulation outcomes in this plane, so that each simulation run is represented by two points in the plane spanned by  $\rho(A), \rho(B)$ .

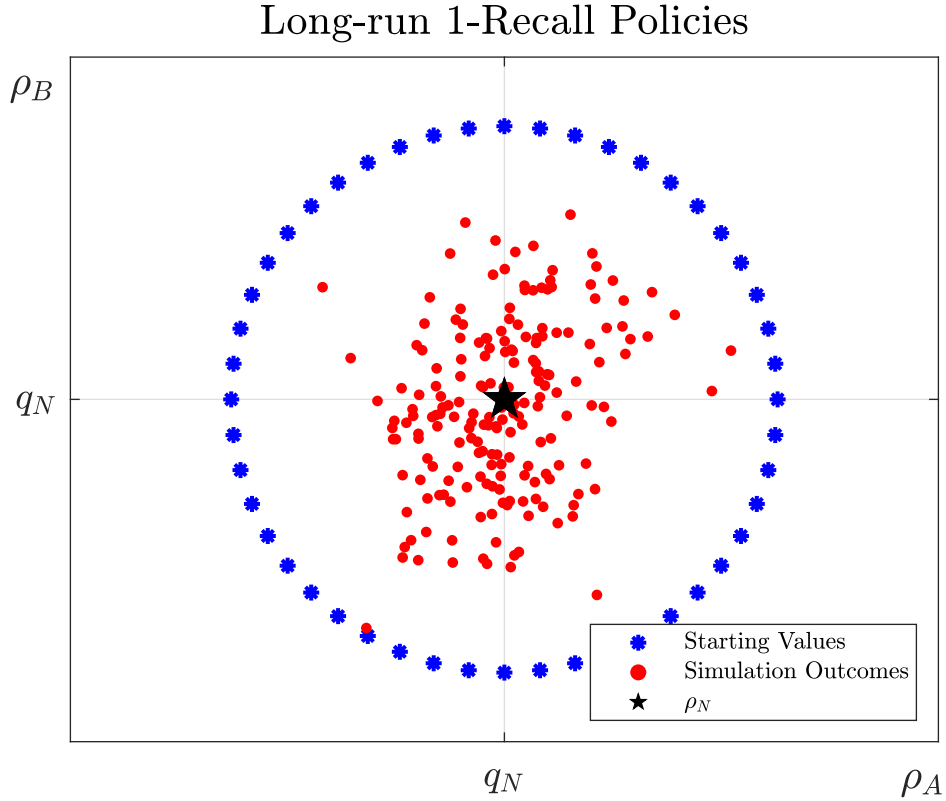


FIGURE 4. The final policy profiles of 96 simulation runs of 25000 iterations each are shown. Simulations are started in a circular neighborhood of  $\rho_N$ , with a radius of  $.025\|\rho_N\|$ . All simulations remained in the neighborhood, as should be expected given the stability of  $\rho_N$ .

Now contrast this result with an analogous study given  $DS$ -policies supported on the  $S^*$ -policy space shown in Figure 5.1. 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  $\zeta_N$ . 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  $\zeta_N$  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  $\sigma$ . 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 policy space (for example, a  $DS$ -policy space), 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.

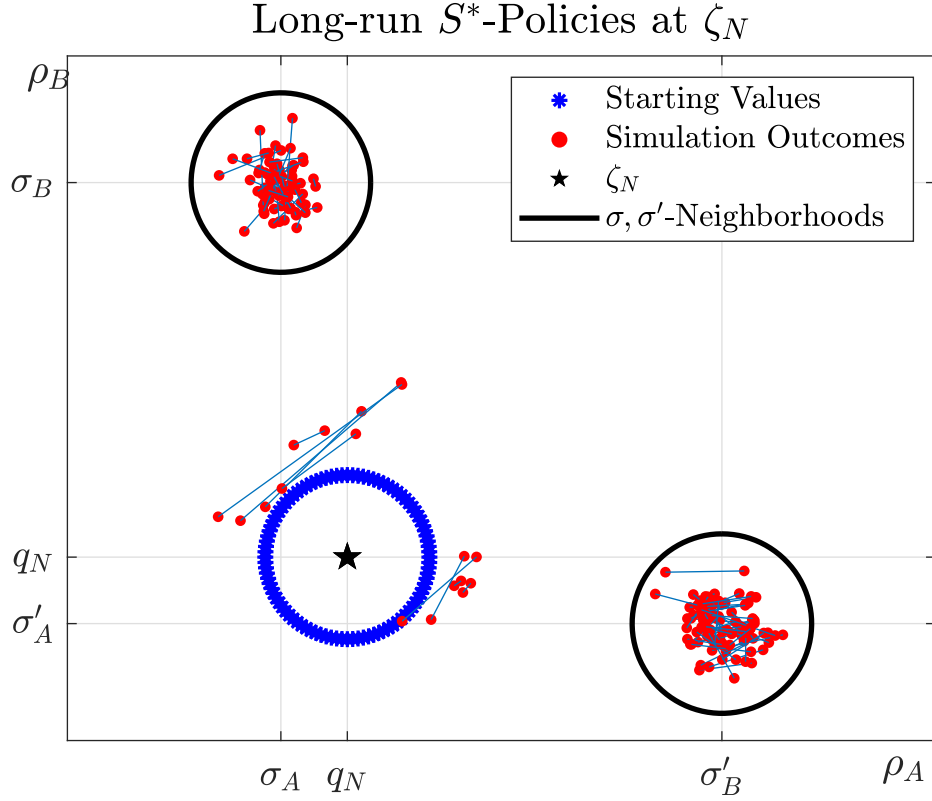


FIGURE 5. The final policy profiles of 96 simulation runs of 25000 iterations each are shown. Simulations are started in a circular neighborhood of  $\zeta_N$ , with a radius of  $.025\|\zeta_N\|$ . All simulations left the neighborhood of starting values, with most conglomerating at one of the two collusive equilibria  $\sigma, \sigma'$ . To see that outcomes indeed approached  $\sigma, \sigma'$ , two red dots connected by a line represent a single simulation outcome. The black circles represent neighborhoods of  $\sigma, \sigma'$  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 5.1 shows how after starting in a neighborhood of the collusive equilibrium  $\sigma$ , iterations stayed there for the course of the simulation. An analogous picture can be generated when initializing in a neighborhood of  $\sigma'$ .

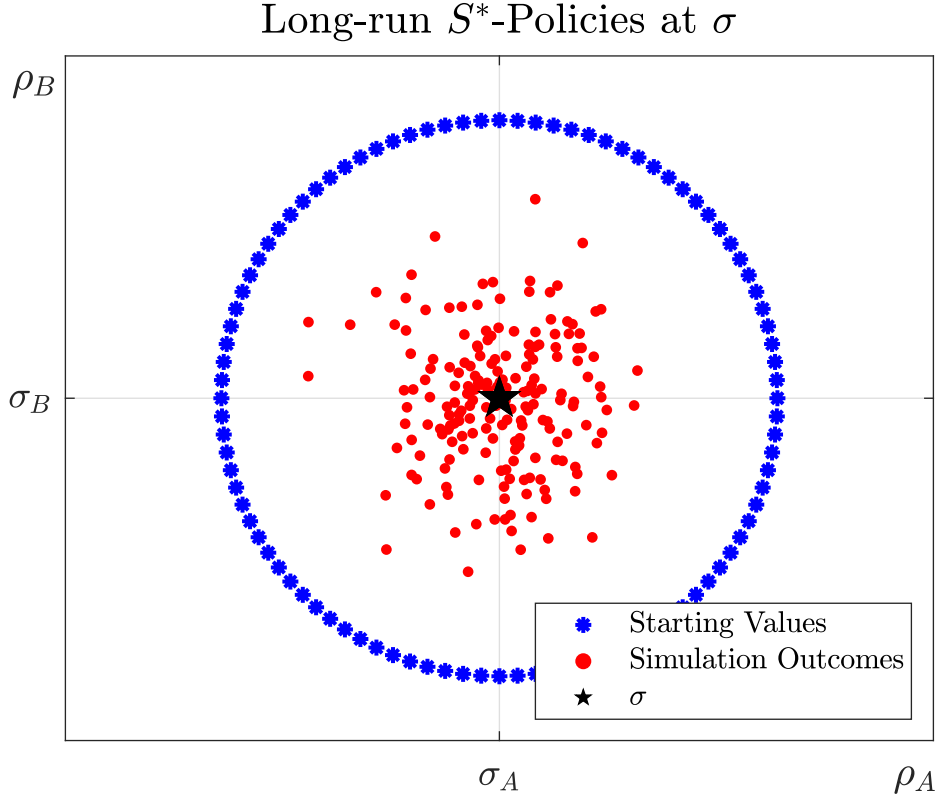


FIGURE 6. The final policy profiles of 96 simulation runs of 25000 iterations each are shown. Simulations are started in a circular neighborhood of  $\sigma$ , with a radius of  $.025\|\sigma\|$ . All simulations remained in the neighborhood, as should be expected given the stability of  $\sigma$ .

## 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 way 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.

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?

I show that if the dynamic policies considered are 1-recall policies, RL can learn to play the static Nash equilibrium if and only if it satisfies a payoff-condition that is independent from the state-transitions underlying the 1-recall policy. However, *DS*-policies can prevent RL from learning the static Nash equilibrium even if it is statically stable, and may allow them to learn to play collusively. This 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 an 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 fictitious 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 more general treatment on asymptotic analysis of RL 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

- 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*.
- Benaim, Michel and Mathieu Faure (2012). “Stochastic approximation, cooperative dynamics and supermodular games”. In: *The Annals of Applied Probability* 22.5, pp. 2133–2164.
- Benaim, 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 (Oct. 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.
- Green, Edward J and Robert H Porter (1984). “Noncooperative collusion under imperfect price information”. In: *Econometrica: Journal of the Econometric Society*, pp. 87–100.
- 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). “Consistency of Multi-Agent Batch Reinforcement Learning”. URL: [https://cjmpossnig.github.io/papers/marlbatchconv\\_CPossnig.pdf](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, Gagan 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. Algorithm Class Definitions

In this section I provide the general reinforcement learning procedure considered in this paper, abstracting away from underlying details of the game environment that is being played on. Assume there is a set  $I$  of algorithmic agents with  $|I| = N$ . Agents observe states on some fixed, finite state space  $S$  with  $|S| = L$ , and make per period choices (actions) in compact interval  $A_i$ , with policy profile space  $\bar{A} = \times_{i \in I} \bar{A}_i$ . 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:*

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

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  $i \in I$  to get to the representation I study:

$$\rho_{t+1} \in \rho_t + \alpha_t [F(\rho_t) + B_t]. \quad (17)$$

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 how 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 will be made clear in section 4.

Throughout, I will 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)).

The case where  $F^i(\rho)$  is a gradient to some payoff function is treated in Mazumdar, Ratliff, and Sastry (2020).

- (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.

**Remark 1.** Notice that Definition (17) does not exclude the case in which the function to be approximated is fully known, or there is no bias term. Our 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. In the case where  $F(X)$  is a gradient, this scenario is similarly treated in Mazumdar, Ratliff, and Sastry (2020). We also refer to Mazumdar, Ratliff, and Sastry (2020) for a list of classes of algorithms that are included in our definition.

Considering the iteration (17), we can see that  $F(\rho_t)$  features importantly as a mapping that provides the reinforcement of the iteration profile  $\rho_t$ . 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 approximated.  $F(\rho)$  thus becomes an approximation target, and  $B_t$  can then be seen as the resulting error term. If  $F(\rho)$  were fully known, as is allowed by definition, we can set  $B_t = 0$  for all  $t$ .

To introduce the definition of error term  $B_t^i$  that can be covered in this paper, we first need to define the class of performance criterions  $F(\rho)$ , and what kinds of approximation methods we can consider:

First the class of performance criterions to be approximated:

**Definition 11** (Candidate performance criterions). We 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  $C \subseteq X$  of positive measure,  $\mathcal{U}_S = \bigcup C$ , such that  $G(x)$  is single-valued and  $C^1$  for  $x \in \mathcal{U}_S$ .

**Remark 2.** We allow for multi-valuedness to be able to handle the 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  $\mathcal{C}^1 \subset \mathcal{M}^1$ .

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

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

Then, the definition of function approximators we allow:

**Definition 12** ( $C^1$  Approximation).

Let  $Y$  be some space of observations to be used to approximate a function. Given  $\gamma > 0$ , we say that a function approximation operator  $\mathcal{A}_g : \mathcal{M}^1 \times Y \mapsto \mathcal{M}^1$  is a  $C^1$  Approximation of a performance criterion  $F \in \mathcal{M}^1$  if there is an increasing sequence of  $\sigma$ -fields  $\mathcal{F}_t$  generated by datasets  $D_t \in Y$ , a bias function  $g \in \mathcal{B}_\gamma^1$  and an integer  $N > 0$  such that we 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

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

and  $\delta_t \rightarrow_p 0$  as  $n \rightarrow \infty$ . “ $\rightarrow_p$ ” denotes convergence in probability.

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

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

with  $g \in \mathcal{B}_\gamma^1$ , and  $\delta_t \rightarrow_p 0$  as  $n \rightarrow \infty$ .

One can interpret  $g(\rho)$  as representing the bias part of the function approximation, and  $\delta_t$  as a random variable such that  $\mathbb{E}[\|\delta_t\|^2 | \mathcal{F}_t]$  represents the variance part.

In the case of classical model-free  $Q$  learning as in Section 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  $C^1$  function after enough data (large  $n$ ) has been accumulated.

Fix small  $\gamma > 0$  and observation space  $Y$ . 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.**

- (1) Let the bias function  $g \in \mathcal{B}_\gamma^1$ .
- (2) Let  $D_t \in Y$  be a sequence of datasets.
- (3)

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

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

- (4) Stacked version of  $B_t^i$ .

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

- (5)  $\mathcal{F}_t$  is the  $\sigma$ -field generated by  $\{\rho_t, D_t, M_t, \rho_{t-1}, D_{t-1}, M_{t-1}, \dots, \rho_0, D_0, M_0\}$ , i.e. all the information available to the updating rule at a given period  $t$ .
- (6)  $M_{t+1}$  is a Martingale-difference noise. There is  $0 < \bar{M} < \infty, q \geq 2$  such that for all  $t$

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

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

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

where  $\Omega_t$  is symmetric positive definite for all  $t$ .

- (8) Write  $\varepsilon_t = \mathcal{A}_g[F, D_t](\rho_t) - F(\rho_t)$ . Then  $M_{t+1}, \varepsilon_t$  are independent conditional on  $\mathcal{F}_t$ .

## A.1. Gradient-type Algorithms

We can now give a brief overview of the kind of gradient-type algorithms that are included in our class: A few definitions are in order to properly understand this class:

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

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \quad (18)$$

where the expectation is made over the state transitions.

Then we can define

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

as the gradient of agent  $i$ 's long term payoff evaluated at  $[\rho^i, \rho^{-i}]$ . By abuse of notation, write  $\nabla W(\rho)$  as the stacked gradients of all agents, where without much loss we 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 we can define for  $\rho \in \bar{A}$

$$F_D^S(\rho) = \nabla W(\rho), \quad (19)$$

as the state dependent gradient vector field. Take an iteration  $\rho_t$  and its respective function estimation target  $F$  as denoted in (17). 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 3** (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  $\rho_0^i$  that would allow for the process  $\rho_t$  to converge to an unstable restpoint  $\rho^*$ . By definition of instability, the paths that could attract  $\rho_t$  to  $\rho^*$ , if they exist, must be a lower-dimensional subspace of profile space  $\bar{A}$ . The statement we can then make is that the probability that a randomly chosen starting profile  $\rho_0$  will converge to unstable  $\rho^*$  must equal 0, since any lower dimensional subspace of  $E$  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 Section 5.

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

$$\dot{\sigma}_t = F_S(\sigma_t) = \begin{bmatrix} b_1(\sigma_{t,3}, \sigma_{t,4}) \\ b_2(\sigma_{t,1}, \sigma_{t,2}) \end{bmatrix} - \sigma_t. \quad (20)$$

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

$$X(\sigma^*) = \begin{bmatrix} -I_{2 \times 2} & J_1(\sigma^*) \\ J_2(\sigma^*) & -I_{2 \times 2} \end{bmatrix} \quad (21)$$

where  $I_2$  is the 2-dimensional identity matrix and

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

for  $i \in \{1, 2\}$ .

This linearization has a special structure we can exploit:

**Remark 4.** 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  $X(\sigma^*)$ :

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

Then all eigenvalues are characterised as the zeros of  $ch(\lambda)$ . Remark 4 tells us that

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

That is, if  $\mu$  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:

$$b_1(b_2(\sigma_1, \sigma_2)) - (\sigma_1, \sigma_2)^T \quad (22)$$

with respect to  $\sigma$ . We can then interpret stability graphically as a scenario in which (22) 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, we can go even further:

**Remark 5.** 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  $\sigma^*$ , 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 5 to our system and arrive at the conclusion. Firstly, given square matrix  $A$ , define  $\Lambda$  as the set of eigenvalues of the  $A$ . Then define

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

as the spectral radius of  $A$ .

**Lemma 2.** Suppose  $\alpha^* = \beta^* = \sigma^*$  is an interior, symmetric equilibrium. Let  $\bar{\kappa}$  be the real part of the spectral radius of  $M$ .

Then  $\sigma^*$  is asymptotically stable if  $\bar{\kappa} < 1$ , and unstable if  $\bar{\kappa} > 1$ .

*Proof.* Using Remark 5, we get that

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

Thus, if  $\mu$  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.  $\square$

**Proposition 6.** Consider an interior, symmetric equilibrium under  $S^*$  domain,  $(q_A, q_B) = \sigma^*$  with  $q_A < q_B$  as constructed in Proposition 5. It is asymptotically stable if

$$\frac{\partial b_1^A(\sigma^*)}{\partial \beta_A} + \frac{\partial b_1^B(\sigma^*)}{\partial \beta_B} \leq 0.$$

*Proof.* Consider the matrix  $J$  as defined for  $X(\sigma^*)$  above. We will show that the sufficient condition in Lemma 2 is implied by this Proposition. Using (9) we can write

$$\det(J) = -\text{tr}(J) - 1 + \frac{(u_{11}^A - u_{12}^A)(u_{11}^B - u_{12}^B)}{\phi_A \phi_B} + J_{12} \frac{P'_{BB}}{P'_{AB}} \frac{(u_{11}^A - u_{12}^A)}{\phi_B} + J_{21} \frac{P'_{AB}}{P'_{BB}} \frac{(u_{11}^B - u_{12}^B)}{\phi_A}, \quad (23)$$

where  $\text{tr}(J)$  is the trace of  $J$ ,  $J_{ij}$  is the element of row  $i$ , column  $j$  of  $J$ , and

$$\phi_k = \omega^{-1} \delta f'_k(u^B - u^A) + u_{11}^k,$$

for  $k \in \{A, B\}$ .

Next, note that for a 2x2 matrix  $J$  we can write the eigenvalues as

$$\mu_{1,2} = \frac{1}{2} \text{tr}(J) \pm \sqrt{\frac{1}{4} \text{tr}(J)^2 - \det(J)}.$$

Assume for now that the eigenvalues are real. We will show at the end that this must indeed be the case. Then  $\bar{\kappa}$ , the spectral radius of  $J$ , equals

$$\bar{\kappa} = \frac{1}{2} |\text{tr}(J)| + \sqrt{\frac{1}{4} \text{tr}(J)^2 - \det(J)}.$$

Then  $\bar{\kappa} < 1$  is equivalent to

$$|\text{tr}(J)| - \det(J) < 1.$$

Using (23), we get

$$\begin{aligned} |\text{tr}(J)| - \det(J) &= |\text{tr}(J)| + \text{tr}(J) + 1 \\ &\quad - \frac{(u_{11}^A - u_{12}^A)(u_{11}^B - u_{12}^B)}{\phi_A \phi_B} - M_{12} \frac{P'_{BB}}{P'_{AB}} \frac{(u_{11}^A - u_{12}^A)}{\phi_B} - M_{21} \frac{P'_{AB}}{P'_{BB}} \frac{(u_{11}^B - u_{12}^B)}{\phi_A}. \end{aligned}$$

Finally, note that

$$\begin{aligned} u_1(q_1, q_2) &= P'(Q)q_1 + P(Q); & u_2(q_1, q_2) &= P'(Q)q_1, \\ u_{11}(q_1, q_2) &= 2P'(Q) + P''(Q)q_1; & u_{12}(q_1, q_2) &= P'(Q) + P''(Q)q_1, \end{aligned}$$

and therefore

$$u_1(q_1, q_2) - u_2(q_1, q_2) = P(Q) > 0; \quad u_{11}(q_1, q_2) - u_{12}(q_1, q_2) = P'(Q) < 0$$

by initial assumption. Further, local optimality and non-degeneracy of  $\sigma^*$  requires  $\phi_k < 0$  for  $k = A, B$ . The fact that we are considering a two-threshold equilibrium implies  $P'_{AB} > 0 > P'_{BB}$ . All of these facts together imply

$$\frac{(u_{11}^A - u_{12}^A)(u_{11}^B - u_{12}^B)}{\phi_A \phi_B} > 0; \quad \frac{P'_{BB}}{P'_{AB}} \frac{(u_{11}^A - u_{12}^A)}{\phi_B} < 0; \quad \frac{P'_{AB}}{P'_{BB}} \frac{(u_{11}^B - u_{12}^B)}{\phi_A} < 0.$$

Finally, note from (9) that for any such equilibrium,  $J_{12} < 0$  and  $J_{21} < 0$ . Thus, if  $\text{tr}(J) \leq 0$  (i.e. the condition of this Proposition holds), we have that

$$\bar{\kappa} < 1 \iff -\frac{(u_{11}^A - u_{12}^A)(u_{11}^B - u_{12}^B)}{\phi_A \phi_B} - J_{12} \frac{P'_{BB}}{P'_{AB}} \frac{(u_{11}^A - u_{12}^A)}{\phi_B} - J_{21} \frac{P'_{AB}}{P'_{BB}} \frac{(u_{11}^B - u_{12}^B)}{\phi_A} < 0,$$

and plugging in all our signs shows that this must hold.

To see that the eigenvalues are real, I will use the fact that  $J$  is a strict  $Z$ -matrix. Strict  $Z$ -matrices are defined as all matrices  $A \in \mathbb{R}^{n \times n}$  s.t.  $a_{ij} < 0$  if  $i \neq j$ . One can then write

$$J = xI - B;$$

where  $B_{ij} > 0$  for all  $i, j$  and  $x > 0$ . Thus, if  $\lambda$  is an eigenvalue of  $B$ ,  $x - \lambda$  must be an eigenvalue of  $J$ . The Perron-Frobenius Theorem tells us that the spectral radius of  $B$ ,  $\kappa(B)$ , is real and corresponds to a positive eigenvalue of  $B$ . But then,  $x - \kappa(B)$  is a real eigenvalue of  $J$ , and since  $J$  has only two eigenvalues, both must be real.  $\square$

## B.1. General State Space: Payoff Characterisations and Stability

In this section, we will show how to connect our insights on stability of the static Nash equilibrium and collusive equilibrium defined for the binary price game to games with more general state spaces and many possible price outcomes.

Consider the general situation in which  $S = \{s_1, \dots, s_K\}$ , and given aggregate quantities  $Q$  we define transition probabilities between states  $s_k, s_{k'}$  as  $P_{kk'}(Q)$ , assuming throughout that  $P_{kk'}(Q) > 0$  for all  $Q, s_k, s_{k'} \in S$ . Let  $\rho : S \mapsto A$  be a policy with  $S$ -domain.

For long term payoffs we define recursively, for all  $s_i \in S$ :

$$W(\rho, \gamma, s_i) = (1 - \delta)u(\rho(s_i), \gamma(s_i)) + \delta \sum_{k=1}^K P_{kk'}(\rho(s_i) + \gamma(s_i))W(\rho, \gamma, s_k),$$

the discounted expected payoff from playing  $\rho$  if the opponent plays  $\gamma$ .

Fixing the profile  $\rho, \gamma$ , we can also use the vector form

$$\begin{aligned} W &= \left[ W(\rho, \gamma, s_1), \dots, W(\rho, \gamma, s_K) \right]^T, \\ U &= \left[ u(\rho(s_1), \gamma(s_1)), \dots, u(\rho(s_K), \gamma(s_K)) \right]^T, \end{aligned}$$

to write

$$W = (1 - \delta)U + \delta TW = [I_K - \delta T]^{-1} (1 - \delta)U,$$

where  $T = (P_{kk'})_{k,k' \in \{1, \dots, K\}}$  is the Markov transition matrix given a fixed profile  $\rho, \gamma$ .

**Remark 6.** *The inverse of  $I_K - \delta T$  exists and has all elements non-negative for all  $\delta < 1$ , since it is an  $M$ -matrix.*

**Corollary 2.** *The derivatives of vector  $W$  can be written as, for  $i \leq K < j$ :*

$$\begin{aligned} W_i &= [I_K - \delta T]^{-1} \delta \frac{\partial T}{\partial \rho^i} W + [I_K - \delta T]^{-1} (1 - \delta) \frac{\partial U}{\partial \rho^i} \\ W_j &= [I_K - \delta T]^{-1} \delta \frac{\partial T}{\partial \gamma_{j-K}} W + [I_K - \delta T]^{-1} (1 - \delta) \frac{\partial U}{\partial \gamma_{j-K}} \\ W_{ii} &= [I_K - \delta T]^{-1} \delta \frac{\partial^2 T}{(\partial \rho^i)^2} W + [I_K - \delta T]^{-1} (1 - \delta) \frac{\partial^2 U}{(\partial \rho^i)^2} - 2 [I_K - \delta T]^{-1} \delta \frac{\partial T}{\partial \rho^i} \tilde{\mathbf{W}}_i \\ W_{ij} &= [I_K - \delta T]^{-1} \delta \frac{\partial^2 T}{\partial \rho^i \partial \gamma_{j-K}} W + [I_K - \delta T]^{-1} (1 - \delta) \frac{\partial^2 U}{\partial \rho^i \partial \gamma_{j-K}} \\ &\quad + [I_K - \delta T]^{-1} \delta \frac{\partial T}{\partial \gamma_{j-K}} \tilde{\mathbf{W}}_i + [I_K - \delta T]^{-1} \delta \frac{\partial T}{\partial \rho^i} \tilde{\mathbf{W}}_j. \end{aligned}$$

*Proof.* This follows from some matrix algebra, importantly using the following fact:

For a matrix function  $X$  of variable  $y$ , let  $\partial X$  be the partial derivative of  $X$  with respect to  $y$ . Then  $\partial(X^{-1}) = -(X^{-1})(\partial X)(X^{-1})$ .  $\square$

If  $W_i = 0$ , we can further simplify this, using the fact that  $\mathbf{P}$  depends only on aggregate quantities in each state.

$$\begin{aligned} \tilde{W}_j &= [I_K - \delta T]^{-1} (1 - \delta) \left[ \frac{\partial U}{\partial \gamma_{j-K}} - \frac{\partial U}{\partial \rho_{j-K}} \right] \\ \tilde{W}_{ii} &= [I_K - \delta T]^{-1} \delta \frac{\partial^2 T}{(\partial \rho^i)^2} W + [I_K - \delta T]^{-1} (1 - \delta) \frac{\partial^2 U}{(\partial \rho^i)^2} \\ \tilde{W}_{ij} &= [I_K - \delta T]^{-1} \delta \frac{\partial^2 T}{\partial \rho^i \partial \gamma_{j-K}} W + [I_K - \delta T]^{-1} (1 - \delta) \frac{\partial^2 U}{\partial \rho^i \partial \gamma_{j-K}} \\ &\quad + [I_K - \delta T]^{-1} \delta \frac{\partial T}{\partial \rho^i} \tilde{W}_j. \end{aligned} \tag{24}$$

We can now prove some results for the asymptotic stability of specific types of strategy profiles, for specific types of general state spaces. To do this, we need auxiliary results on the characterisation of eigenvalues for general state spaces. The following Lemmas will help with this:



**Lemma 3.** Suppose we have a square matrix  $A \in \mathbb{R}^{n \times n}$  that can be written as  $A = aI_n + B$ , for some  $a \in \mathbb{R}$ , where

$$B = \begin{bmatrix} b_1 \iota_n, \dots, b_n \iota_n \end{bmatrix},$$

where  $\iota_n$  is a vector of  $n$  ones, so that  $B$  is a matrix of columns of constants  $b_1, \dots, b_n$ . Then the eigenvalues of  $A$  are

$$\begin{aligned} & a \text{ with algebraic multiplicity } n - 1; \\ & a + \sum_{k=1}^n b_k \text{ with algebraic multiplicity } 1. \end{aligned}$$

*Proof.* Firstly, we can write  $A = aI_n + \iota_n b^T$ , where  $b^T = (b_1, \dots, b_n)$ , with superscript  $T$  denoting the transpose. By the matrix determinant Lemma, for any  $\lambda \in \mathbb{R}$  we can write

$$\det(A - \lambda I_n) = \det((a - \lambda)I_n) + \iota_n^T \text{adj}((a - \lambda)I_n) \iota_n,$$

where  $\text{adj}(A)$  is the adjoint of matrix  $A$ . Using that by definition,  $A \text{adj}(A) = \det(A)I_n$ , we get that

$$\begin{aligned} \det(A - \lambda I_n) &= (a - \lambda)^n + \iota_n^T (a - \lambda)^{n-1} \iota_n \\ &= (a - \lambda)^{n-1} \left[ a - \lambda + \sum_{k=1}^n b_k \right], \end{aligned}$$

and the result follows. □

**Lemma 4.** Suppose  $M \in \mathbb{R}^{n \times n}$  s.t.

$$M = \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix}$$

is a matrix consisting of blocks  $M_i$ . For  $m, k$  s.t.  $m + k = n$ , suppose we can write those blocks as

$$\begin{aligned} M_1 &= a_1 I_m + \iota_m a_2^T; & M_2 &= \iota_m b^T \\ M_3 &= \iota_k c^T; & M_4 &= d_1 I_k + \iota_k d_2^T. \end{aligned}$$

We assume  $a_1, d_1$  are nonzero scalars, and  $a_2, c$  vectors in  $\mathbb{R}^m$ , and  $d_2, b$  vectors in  $\mathbb{R}^k$ . Then the set of eigenvalues of  $M$ ,  $\text{eig}(M) = \{\lambda_1, \dots, \lambda_K\}$ , is s.t.

$$\begin{aligned}\lambda_1 &= a_1 \quad \text{with algebraic multiplicity: } m-1, \\ \lambda_2 &= d_1 \quad \text{with algebraic multiplicity: } k-1, \\ \lambda_{3,4} &= \frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q},\end{aligned}$$

where

$$\begin{aligned}p &= a_1 + d_1 + a_2^T \iota_m + d_2^T \iota_k; \\ q &= (b^T \iota_k)(c^T \iota_m).\end{aligned}$$

*Proof.* For  $\lambda \notin \text{eig}(M_4)$ , we can use the Schur-complement to write the characteristic equation of  $M$  as

$$\det[M - \lambda I_n] = \det[M_4 - \lambda I_k] \det[M_1 - \lambda I_m - M_2(M_4 - \lambda I_k)^{-1} M_3]. \quad (25)$$

Then we can use the special form of all these blocks. First, note that we can apply the matrix-determinant Lemma to get

$$\begin{aligned}\det(M_4 - \lambda I_k) &= (d_1 - \lambda)^k + d_2^T (d_1 - \lambda)^{k-1} I_k \iota_k \\ &= (d_1 - \lambda)^{k-1} [d_1 - \lambda + d_2^T \iota_k].\end{aligned}$$

Then we can apply the Sherman-Morrison formula to get

$$\begin{aligned}(M_4 - \lambda I_k)^{-1} &= (d_1 - \lambda)^{-1} I_k - \frac{(d_1 - \lambda)^{-2} \iota_k d_2^T}{1 + (d_1 - \lambda)^{-1} d_2^T \iota_k} \\ &= (d_1 - \lambda)^{-1} \left[ I_k - \iota_k d_2^T (d_1 - \lambda + d_2^T \iota_k)^{-1} \right].\end{aligned}$$

So we get

$$\begin{aligned}M_2(M_4 - \lambda I_k)^{-1} M_3 &= (d_1 - \lambda)^{-1} \left[ \iota_m b^T \iota_k c^T - M_2 \iota_k d_2^T M_3 (d_1 - \lambda + d_2^T \iota_k)^{-1} \right] \\ &= \frac{b^T \iota_k}{d_1 - \lambda + d_2^T \iota_k} \iota_m c^T.\end{aligned}$$

It follows that we can write, by the matrix determinant lemma,

$$\begin{aligned}\det[M_1 - \lambda I_m - M_2(M_4 - \lambda I_k)^{-1} M_3] &= (a_1 - \lambda)^m + \left[ a_2^T - \frac{b^T \iota_k}{d_1 - \lambda + d_2^T \iota_k} c^T \right] (a_1 - \lambda)^{m-1} I_m \iota_m \\ &= (a_1 - \lambda)^{m-1} \left[ a_1 - \lambda + a_2^T \iota_m - \frac{(b^T \iota_k)(c^T \iota_m)}{d_1 - \lambda + d_2^T \iota_k} \right].\end{aligned}$$

Plugging these results into (25), we get

$$\det[M - \lambda I_n] = (a_1 - \lambda)^{m-1} (d_1 - \lambda)^{k-1} \left[ (a_1 - \lambda + a_2^T \iota_m) (d_1 - \lambda + d_2^T \iota_k) - (b^T \iota_k) (c^T \iota_m) \right].$$

The result then follows from finding all roots  $\lambda$  of the above. □

In our enlarged Cournot game  $\Gamma_E$ , a public one-recall strategy (1R - strategy) can be defined as policy  $\rho : \mathbf{P} \mapsto A$ , so that states are price realizations representing last periods observed price. See Definition 8. We show in the following that the insight of Proposition 4 can be made in the enlarged game of arbitrarily many price outcomes.

**Proposition 7.** *Take  $\Gamma_E$  and let  $\rho_N$  be the 1R-policy that plays stage game Nash quantity  $q_N$  in every state. Then  $\rho_N$  is asymptotically stable if and only if  $q_N$  is.*

*Proof.* To determine the stability of  $\rho_N$ , we need to compute the eigenvalues of the linearized  $P$ -state-dependent best-response dynamics ( $F_B^{\mathbf{P}}$ ) at  $\rho_N$ . As shown in Lemma 2, we can equivalently consider the eigenvalues of the gradient of the best response in  $\mathbf{P}$ -statespace at  $\rho_N$ .

Since by assumption, states are irreducible (i.e. for every aggregate quantity  $Q$ , every price has positive probability of occurring after finite time), we can fix an arbitrary initial state  $s_1$  (which in the case of this 1R-policy corresponds to an arbitrary price in  $\mathbf{P}$ ) when computing first order conditions to compute best responses. The implicit function theorem and (24) can then be used to find this gradient of the best response.

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

$$W_i = [I_K - \delta T]^{-1} \left[ \delta \frac{\partial T}{\partial \rho^i} W + (1 - \delta) \frac{\partial U}{\partial \rho^i} \right].$$

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

$$\begin{aligned} W_{ij} &= \left( \partial [I_K - \delta T]^{-1} \right) [I_K - \delta T] W_i \\ &\quad + [I_K - \delta T]^{-1} \left[ \delta \frac{\partial^2 T}{\partial \rho^i \partial \rho_j} W + (1 - \delta) \frac{\partial^2 U}{\partial \rho^i \partial \rho_j} \right]. \end{aligned}$$

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

Now define  $R = [W_{ij}]_{i < j}$  as the matrix cross derivatives between an agents own strategy

$\rho(s_i)$  and an opponents strategy  $\gamma(s_{j-K})$ . Then we can define, using the implicit function theorem, the best response derivative matrix as

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

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

Firstly, long term payoffs =  $\mathbf{U}^N$ , since  $\rho_N$  prescribes the same action in each state.

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

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

$$\begin{aligned} W_{ii} &= [I_K - \delta T]_1^{-1} (1 - \delta) \frac{\partial^2 U}{(\partial \rho^i)^2}, \\ W_{ij} &= [I_K - \delta T]_1^{-1} (1 - \delta) \left[ \frac{\partial^2 U}{\partial \rho^i \partial \gamma_{j-K}} + \delta \frac{\partial T}{\partial \rho^i} [I_K - \delta T]^{-1} \frac{\partial U}{\partial \gamma_{j-K}} \right], \end{aligned}$$

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

$$\begin{aligned} W_{ii} &= [I_K - \delta T]_1^{-1} (1 - \delta) u_{11}^N e_i, \\ W_{ij} &= [I_K - \delta T]_1^{-1} (1 - \delta) \left[ u_{12}^N e_i + \delta \frac{\partial T}{\partial \rho^i} [I_K - \delta T]^{-1} u_2^N e_i \right], \end{aligned}$$

if  $i = j - K$ , and

$$W_{ij} = [I_K - \delta T]_1^{-1} (1 - \delta) \left[ \delta \frac{\partial T}{\partial \rho^i} [I_K - \delta T]^{-1} u_2^N e_{j-K} \right],$$

otherwise, following the notation used in section 5.

Now again since  $\rho_N$  is constant for all states, we can write  $T = \iota_K p^T$ , where  $p^T = (Pr_1, \dots, Pr_K)$  is a column-vector of the probability of observing each price  $k = 1, \dots, K$  given aggregate quantity choice  $2q_N$ . In that case, we can use the Sherman-Morrison formula to derive the inverse

$$[I_K - \delta T]^{-1} = I_K + \frac{\delta}{1 - \delta} \iota_K p^T,$$

implying that each of row of  $[I_K - \delta T]^{-1}$  sums to  $\frac{1}{1 - \delta}$ .

It follows that  $W_{ii} = \delta u_{11}^N Pr_i$  for all  $1 < i \leq K$ , and  $W_{11} = u_{11}^N [1 - \delta + \delta Pr_1]$ . The the

difference for index  $i$  reflects the fact that we fixed initial state 1. We can thus write the Hessian as  $H = u_{11}^N[(1 - \delta)e_1 + \text{diag}(\delta Pr_i)]$ . Going step-by-step, we get

$$[I_K - \delta T]^{-1} u_2^N e_{j-K} = u_2^N \left[ e_{j-K} + \frac{\delta}{1 - \delta} Pr_{j-K} \iota_K \right],$$

so that

$$\begin{aligned} & [I_K - \delta T]_1^{-1} (1 - \delta) \delta \frac{\partial T}{\partial \rho^i} [I_K - \delta T]^{-1} u_2^N e_{j-K} \\ &= \begin{cases} \delta^2 u_2^N Pr_i Pr'_j & \text{if } i > 1 \\ \delta[1 - \delta + \delta Pr_i] u_2^N Pr'_j & \text{if } i = 1 \text{ otherwise} \end{cases}, \end{aligned}$$

again using that  $(\partial p)^T \iota_K = 0$  by definition. Finally, when taking the ratio  $H^{-1}R$ , we can see that difference for row  $i = 1$  to other rows does not matter since the factors are cancelled correctly between  $H$  and  $R$ , so that we get

$$\frac{W_{ij}}{W_{ii}} = \begin{cases} \frac{\delta u_2^N Pr'_j}{u_{11}^N} & \text{if } j - K \neq i, \\ \frac{u_{12}^N}{u_{11}^N} + \frac{\delta u_2^N Pr'_j}{u_{11}^N} & \text{otherwise} \end{cases}.$$

We can use this to write

$$M = -\frac{u_{12}^N}{u_{11}^N} I_K + B,$$

where  $B = [b_1, \dots, b_K]$  is a matrix of column vectors of constants:  $b_j = -\frac{\delta u_2^N Pr'_j}{u_{11}^N} \iota_K$ .

Finally we can apply Lemma 3 to find the eigenvalues of  $M$ . We immediately get  $n - 1$  eigenvalues equal  $-\frac{u_{12}^N}{u_{11}^N}$ , which equals the eigenvalue of the static Nash equilibrium  $q_N$ . For the last eigenvalue, recall  $(\partial p)^T \iota_K = 0$ , so it collapses to  $-\frac{u_{12}^N}{u_{11}^N}$  as well, and we are done.  $\square$

## B.2. Generalizing $S^*$

We show here that the insights we can gain from a binary-price game with collusive equilibrium based on binary state  $S^*$  hold more generally, i.e. do not depend crucially on the binary restriction.

First recall the example game with binary prices and quantity choice  $q$  in interval  $I$ , with the usual expected Cournot payoffs, given some cost function  $c$ . For the resulting payoff

functions  $u^i(q_i, q_{-i})$  and set of players  $N$ , call this game  $\Gamma^* = \langle u^i, N \rangle$ . Define the binary statespace  $S^* = \{A, B\}$  and binary price outcome  $P \in \{P_L, P_H\}$  as in section 5.

We say that  $(q_A, q_B)$  is a binary state equilibrium if the policy  $\sigma^* : S^* \mapsto I$  is such that  $\sigma^*(A) = q_A, \sigma^*(B) = q_B$  (i.e. it is an  $S^*$ -policy) and is a best reply to itself, with the objective being long term discounted expected profits. We let  $H(Q_s) = Pr[P_L | Q_s]$  for  $s \in S^*$ .

Now consider an enlarged game with  $K \geq 2$  possible price outcomes,  $P \in \mathbf{P} = \{P_1, \dots, P_K\}$ , quantity choice  $q$  in interval  $I$ . Let the cost function  $c$  be the same as in  $\Gamma^*$ . For the resulting payoff functions  $u^i(q_i, q_{-i})$  and set of players  $N$ , call this the enlarged game  $\Gamma_E = \langle u^i, N \rangle$ , where enlargement happened only with respect to the possible set of price observations.

Throughout we assume  $h_k(Q) = Pr[P = P_k | Q] > 0$  for all  $k, Q$ .

Define two binary partitions of  $\mathbf{P}$ :

$$\mathbf{P}_A = \{P_L^A, P_H^A\}; \quad \mathbf{P}_B = \{P_L^B, P_H^B\}.$$

Notet that for the sake of our argument, these partitions don't have to be monotone.

Given  $\mathbf{P}_A, \mathbf{P}_B$ , we define the *enlarged* state space  $S_E = \{s_1, \dots, s_L\}$  for  $L \geq 2$  and give a binary partition of this state space as  $S_E = S_E^A \cup S_E^B$ . States in this enlarged state space evolve with the following state transition  $T : S_E \times \mathbf{P} \mapsto S_E$ :

$$T(s, P) \in \begin{cases} S_E^A & \text{if } s \in S_E^A \wedge P \in P_H^A \\ S_E^A & \text{if } s \in S_E^B \wedge P \in P_L^A \\ S_E^B & \text{if } s \in S_E^A \wedge P \in P_L^B \\ S_E^B & \text{if } s \in S_E^B \wedge P \in P_H^B \end{cases}, \quad (26)$$

where importantly, the transition *within*  $S_E^A, S_E^B$  will not matter for our result, so we can be agnostic about it.

Let  $F(Q, s) = Pr[s' \in S_E^A | Q, s]$  be the probability of transitioning into some state in  $S_E^A$ , conditional on aggregate quantity  $Q$  and current state  $s$ .

We define policy  $\rho^* : S_E \mapsto I$  as a policy that plays only two quantities given the enlarged state space:  $\rho^*(s) = q_A$  for all  $s \in S_E^A$ , and  $\rho^*(s) = q_B$  for all  $s \in S_E^B$ . Generally we call policies on  $S_E$ -domain  $S_E$ -policies, and call them binary  $S_E$ -policies if they are defined as  $\rho^*$  is. We call a profile of them  $S_E$ -equilibrium if they are best replies to each other, again under the objective of long term discounted expected payoffs.

Firstly, one should think of states  $s \in S_E^A$  as good states (low quantity), and others as punishment states. Going back to the original, binary state and price equilibrium, this is the intuitive connection to  $S^*$ .

It is interesting to note that for the following reduction result, it does not matter how exactly the state transition maps *within*  $S_E^A, S_E^B$ . All that matters is the probability, given a current state  $s$ , to map into any future state  $s \in S_E^A$ .

We say that aggregate transition probability  $F(Q, s)$  and binary probability function  $h$  are matching if

$$F(Q, s) = \begin{cases} 1 - h(Q) & \text{if } s \in S_E^A \\ h(Q) & \text{if } s \in S_E^B \end{cases}$$

$$F'(Q, s_i) = \begin{cases} -h'(Q) & \text{if } s \in S_E^A \\ h'(Q) & \text{if } s \in S_E^B \end{cases}$$

$$F''(Q, s_i) = \begin{cases} -h''(Q) & \text{if } s \in S_E^A \\ h''(Q) & \text{if } s \in S_E^B \end{cases}$$

for all  $Q$ , where for any function  $f(Q)$ ,  $f', f''$  represent first and second derivatives with respect to  $Q$ .

Finally, to connect payoff functions between  $\Gamma^*$  and  $\Gamma^E$ , we need to verify the following property:

Let  $P(Q) = h(Q)P_L + (1 - h(Q))P_H$  and  $P_E(Q) = \sum_{k=1}^K h_k(Q)P_k$  be the expected price functions for the two games respectively. Then if  $P(Q) = P_E(Q)$ ,  $P'(Q) = P'_E(Q)$ ,  $P''(Q) = P''_E(Q)$  holds for all  $Q$ , we say that  $P, P_E$  are matching. Notice that starting from a binary price game, we can always find a game with many prices so that  $P, P_E$  are matching by defining  $h_k(Q) = \frac{1}{K}h(Q)$  for all  $k < K$ , with  $g_K(Q)$  pinned down as the residual probability. However, there is a restriction from the other direction: given a game of many prices, there may not be a binary pair of prices  $P_L, P_H$  so that  $P, P_E$  can match.

**Proposition 8.**

- (1) Suppose  $S^*$ -policy  $\sigma^*$  is a binary state equilibrium. For any enlarged state space  $S_E$  and partitions  $\mathbf{P}_A, \mathbf{P}_B$  given set of prices  $\mathbf{P}$ , define the binary  $S_E$  policy  $\rho^*$ . Then if  $F(Q, s), H$  and  $P, P_E$  are matching,  $\rho^*$  is an  $S_E$ -equilibrium.
- (2) Conversely, take an enlarged state space  $S_E$  with set of prices  $\mathbf{P}$  and partitions  $\mathbf{P}_A, \mathbf{P}_B$ , and binary strategy  $\rho^*$  that is an  $S_E$ -equilibrium. If we can find  $h(Q), P_L, P_H$  such that  $F, H$  and  $P, P_E$  are matching, then  $S^*$ -policy  $\sigma^*$  is an  $S^*$ -equilibrium.

- (3)  $\sigma^*$  is asymptotically stable if  $\rho^*$  is asymptotically stable. If in addition  $u_{11}(q, q) + u_{12}(q, q) \leq 0$  for all  $q \in I$  (concavity along the 45 degree line), then  $\rho^*$  is asymptotically stable if  $\sigma^*$  is asymptotically stable.

**Corollary 3.** Take a pair  $\Gamma^*, \Gamma_E$  with partitions  $\mathbf{P}_A, \mathbf{P}_B$  and  $S_E$  so that  $F, H$  and  $P, P_E$  are matching.

- (1) Then  $q_N$  is a static Nash equilibrium of  $\Gamma^*$  if and only if it is a static Nash equilibrium of  $\Gamma_E$ .
- (2)  $q_N$  is statically asymptotically stable for  $\Gamma^*$  if and only if it is statically asymptotically stable for  $\Gamma_E$ .
- (3) Let  $\sigma_N, \rho_N$  be the  $S^*, S_E$  strategies respectively playing  $q_N$  in all states.  $\sigma_N$  is asymptotically stable if and only if  $\rho_N$  is.

*Proof.* Points (1, 2) follow since for determining static equilibria, players only care about expected prices  $P(Q)$ , and for stability, all we need is matching first and second derivatives at the equilibrium points.

Point (3) follows from the proof of Proposition 8, point (3).  $\square$

**Remark 7.** Proposition 8 includes the scenario where  $\mathbf{P} = \{P_L, P_H\}$ , and  $S_E = S^* \times \bar{S}$ , for any arbitrary finite  $\bar{S}$ . Then  $S_E^A, S_E^B$  can be constructed intuitively and resulting  $F(Q, s)$  must be matching with  $h$ , so the result from Proposition 8 applies.

### B.3. Proof of Proposition 8

The proof is done for the case of two agents, but it is readily generalizeable. For  $S^*$ -policies  $\sigma, \sigma'$  and  $S_E$ -policies  $\rho, \gamma$ , recursively define

$$W(\rho, \gamma, s_i) = (1 - \delta)u(\rho(s_i), \gamma(s_i)) + \delta \sum_{k=1}^K Pr[s_k | \rho(s_i) + \gamma(s_i), s_i] W(\rho, \gamma, s_k),$$

$$\tilde{W}(\sigma, \sigma', s) = (1 - \delta)u(\sigma(s), \sigma'(s)) + \delta \sum_{s' \in S^*} Pr[s' | \sigma(s) + \sigma'(s), s] \tilde{W}(\sigma, \sigma', s'),$$

as the long term expected payoffs for both repeated games  $\Gamma^E, \Gamma^*$  respectively.

Let  $\tilde{W}$  be the long term payoff in the repeated version of  $\Gamma^*$ , i.e. where  $S^*$ -strategies are chosen. We proceed by proving the following Lemmas:



**Lemma 5.** *Let  $\sigma^*, \rho^*$  be policies as in the statement of Proposition 8. Then*

$$\begin{aligned} W(\rho^*, \rho^*, s_i) &= \tilde{W}(\sigma^*, \sigma^*, A) \text{ if } i \leq \bar{l}, \\ W(\rho^*, \rho^*, s_i) &= \tilde{W}(\sigma^*, \sigma^*, B) \text{ if } i > \bar{l}. \end{aligned}$$

*Proof.* First we show that  $W(\rho^*, \rho^*, s_i) = W(\rho^*, \rho^*, s_j)$  if  $i, j \leq \bar{l}$ , and also if  $i, j > \bar{l}$ .  
By definition,

$$\begin{aligned} &W(\rho^*, \rho^*, s_i) - W(\rho^*, \rho^*, s_j) \\ &= \delta \sum_{k=1}^K \left( Pr[s_k | \rho^*(s_i) + \rho^*(s_i), s_i] - Pr[s_k | \rho^*(s_j) + \rho^*(s_j), s_j] \right) W(\rho^*, \rho^*, s_k) = 0, \end{aligned}$$

since by construction  $\rho^*(s_i) = \rho^*(s_j)$  if  $i, j \leq \bar{l}$ , and also if  $i, j > \bar{l}$ , and by definition of  $T(s, Q)$ .

Having shown this, we can use the fact that  $F, h$  are matching to write

$$\begin{aligned} &W(\rho^*, \rho^*, s_1) = (1 - \delta)u(\rho^*(s_1), \rho^*(s_1)) \\ &+ \delta \sum_{k=\bar{l}+1}^K Pr[s_k | \rho^*(s_1) + \rho^*(s_1), s_1] W(\rho^*, \rho^*, s_1) + \sum_{k=1}^{\bar{l}} Pr[s_k | \rho^*(s_1) + \rho^*(s_1), s_1] W(\rho^*, \rho^*, s_k) \\ &= (1 - \delta)u(\rho^*(s_1), \rho^*(s_1)) \\ &+ \delta Pr[A | \sigma^*(A) + \sigma^*(A), A] W(\rho^*, \rho^*, s_1) + \delta Pr[B | \sigma^*(A) + \sigma^*(A), A] W(\rho^*, \rho^*, s_K), \end{aligned}$$

where we chose  $s_1, s_K$  arbitrarily due to the first result of this proof. Thus,  $W(\rho^*, \rho^*, s_i)$  for  $i \leq \bar{l}$  is the result of the same recursion as  $\tilde{W}(\sigma, \sigma', A)$ , and the same holds for  $i > \bar{l}$  and  $B$ . The result follows.  $\square$

Now suppose that  $\sigma^*$  is an  $S^*$ -equilibrium and  $\rho^*$  is the corresponding  $S_E$ -policy. To show that  $\rho^*$  is an  $S_E$ -equilibrium, it is sufficient to show that  $\rho$  satisfies all one-shot deviations. In other words, we need to show that, for all  $i, q \in I$

$$W(\rho^*, \rho^*, s_i) \geq (1 - \delta)u(q, \rho^*(s_i)) + \delta \sum_{k=1}^K Pr[s_k | q + \rho^*(s_i), s_i] W(\rho^*, \rho^*, s_k).$$

By Lemma 5, if  $i \leq \bar{l}$  we can re-write the above as

$$W(\rho^*, \rho^*, s_i) \geq (1 - \delta)u(q, \rho^*(s_i)) \quad (27)$$

$$+ \delta Pr[A | q + \sigma^*(A), A] W(\rho^*, \rho^*, s_1) \quad (28)$$

$$+ Pr[B | q + \sigma^*(A), A] W(\rho^*, \rho^*, s_K), \quad (29)$$

$$\iff \tilde{W}(\sigma^*, \sigma^*, A) \geq (1 - \delta)u(q, \sigma^*(A)) \quad (30)$$

$$+ \delta Pr[A | q + \sigma^*(A), A] \tilde{W}(\sigma^*, \sigma^*, A) \quad (31)$$

$$+ Pr[B | q + \sigma^*(A), A] \tilde{W}(\sigma^*, \sigma^*, B), \quad (32)$$

which must be true since  $\sigma^*$  is an  $S^*$ -equilibrium. Similarly for  $i > \underline{l}$ . Thus,  $\rho^*$  must be an  $S_E$ -equilibrium.

For the other direction, suppose  $\rho^*$  is an  $S_E$ -equilibrium, and take  $\sigma^*$  as the corresponding  $S^*$ -policy as constructed in the statement of the Proposition. By an argument analogous to the proof of Lemma 5, we get that the statement of the Lemma still goes through. But then we can do an argument analogous to (27) and the claim follows:  $\sigma^*$  is an  $S^*$ -equilibrium.

Finally, we will prove the claim about stability of the equilibria.

We will continue in a similar manner as in the proof of Proposition 4. Recall that we assume throughout that  $Pr[s' | Q, s] > 0$  for all  $Q, s, s'$ . As in the proof of Proposition 4, we can use (24) and it is without loss to consider  $W_{ij} = W_{ij}(s_1)$  for all  $i \leq K, j \leq 2K$  when computing the matrix of best response derivatives. We follow the notation in the proof of Proposition 4:

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

$$M = -H^{-1}R,$$

where  $H = \text{diag}(\{W_{ii}\}_{i \leq K})$  is the Hessian.

We now show that we can apply Lemma 4 to  $M$ . Define

$$f^T = [I_L - \delta T]_1^{-1} \in \mathbb{R}^{1 \times K}.$$

Then note that for any matrix  $A(j)$  that is zero everywhere except in row  $j$ , we have that

$$f^T A(j) = f_j a^T,$$

where  $a^T \in \mathbb{R}^{1 \times K}$  is the  $j$ 'th row of  $A(j)$ . With some abuse of notation, let  $T_i, T_{ii}, T_{ij}$  be the  $i$ 'th rows of

$$\frac{\partial T}{\partial \rho^i}, \quad \frac{\partial^2 T}{\partial \rho^i \partial \rho^i}, \quad \frac{\partial^2 T}{\partial \rho^i \partial \rho_j},$$

respectively. It follows that we can write for all  $i \leq K$

$$W_{ii} = \delta f_i T_{ii} W + (1 - \delta) f_i u_{11}(\rho^i, \gamma_i),$$

And we can write for all  $i < j \leq 2K$

$$W_{ij} = \begin{cases} \delta f_i T_{ii} W + (1 - \delta) f_i u_{12}(\rho^i, \gamma_i) + \delta f_i T_i W_j, & \text{if } i = j - K; \\ \delta f_i T_i W_j, & \text{otherwise} \end{cases},$$

where we recall that

$$W_j = [I_L - \delta T]^{-1} (1 - \delta) \left[ \frac{\partial U}{\partial \gamma_{j-K}} - \frac{\partial U}{\partial \rho_{j-K}} \right] = (1 - \delta) x_{j-K} (u_2(\rho_{j-K}, \gamma_{j-K}) - u_1(\rho_{j-K}, \gamma_{j-K})),$$

where we define  $x_j \in \mathbb{R}^{K \times 1}$  to be the  $j$ 'th column of  $[I_L - \delta T]^{-1}$ .

Now we can write for all  $i \leq K < j \leq 2K$ :

$$\frac{W_{ij}}{W_{ii}} = \left[ \delta T_{ii} W + (1 - \delta) u_{11}(\rho^i, \gamma_i) \right]^{-1} \begin{cases} \delta T_{ii} W + (1 - \delta) u_{12}(\rho^i, \gamma_i) \\ + (1 - \delta) \delta T_i x_{j-K} (u_2(\rho_{j-K}, \gamma_{j-K}) - u_1(\rho_{j-K}, \gamma_{j-K})) & \text{if } i = j - K; \\ (1 - \delta) \delta T_i x_{j-K} (u_2(\rho_{j-K}, \gamma_{j-K}) \\ - u_1(\rho_{j-K}, \gamma_{j-K})) & \text{otherwise} \end{cases}.$$

Now, recall that by construction of the  $S_E$ -policy, we have that  $W(i) = \tilde{W}(\sigma^*, \sigma^*, A)$  for all  $i \leq \bar{l}$ , and  $W(i) = \tilde{W}(\sigma^*, \sigma^*, A)$  for all  $i > \bar{l}$ . To save notation, we write  $V_s = W(\sigma^*, \sigma^*, s)$  for  $s \in \{A, B\}$ .

Next let us consider  $x_j$  more closely. We can in fact compute the matrix  $[I_K - \delta T]^{-1}$  by using the Sherman-Morrison formula twice.

First, let

$$p_s = (Pr[s_1 | 2q_s], \dots, Pr[s_L | 2q_s]),$$

for  $s \in \{A, B\}$ . Then we can write

$$T = \begin{bmatrix} \iota_{\bar{l}} \\ \mathbf{0}_m \end{bmatrix} p_A^T + \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix} p_B^T,$$

where  $\mathbf{0}_k$  is a  $k$ - vector of zeros and  $m = L - \bar{l}$ . Now define

$$Q = I_L - \delta \begin{bmatrix} \iota_{\bar{l}} \\ \mathbf{0}_m \end{bmatrix} p_A^T.$$

So we can write

$$I_L - \delta T = Q - \delta \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix} p_B^T.$$

Then by the Sherman-Morrison formula,

$$Q^{-1} = I_L + \frac{\delta \begin{bmatrix} \iota_{\bar{l}} \\ \mathbf{0}_m \end{bmatrix} p_A^T}{1 - \delta(1 - h_A)},$$

where we follow the notation of Section 5 and write  $h_s = h(2q_s)$  for  $s \in \{A, B\}$ . and

$$[I_K - \delta T]^{-1} = Q^{-1} + \frac{\delta Q^{-1} \delta \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix} p_B^T Q^{-1}}{1 - \delta p_B^T Q^{-1} \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix}}.$$

Re-writing things multiple times allows us to get to

$$[I_K - \delta T]^{-1} = I_L + \begin{bmatrix} \iota_{\bar{l}} \\ \mathbf{0}_m \end{bmatrix} a_A^T + \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix} a_B^T,$$

where

$$a_A = \frac{\delta}{(1 - \delta)\omega} \left[ p_A (1 - \delta z + \frac{\delta^2 h_G h_B}{1 - \delta(1 - h_A)}) + \delta h_A p_B \right];$$

$$a_B = \frac{\delta}{(1 - \delta)\omega} \left[ p_A \delta h_B + p_B (1 - \delta(1 - h_A)) \right],$$

where

$$\omega = 1 + \delta(h_A + h_B - 1),$$

$$z = 1 - h_B + \frac{\delta h_B h_A}{1 - \delta(1 - h_A)}.$$

Importantly, we have  $a_A^T \iota_L = a_B^T \iota_L = \frac{\delta}{1-\delta}$ , so that  $f^T \iota_L = \frac{1}{1-\delta}$ .

Also notice that following the above representation,  $x_{j-K}(i) = a_{A,j} + \mathbf{1}\{i = j - K\} \in \mathbb{R}$  for all  $i \leq \bar{l}$ , and  $x_{j-K}(i) = a_{B,j} + \mathbf{1}\{i = j - K\} \in \mathbb{R}$  for all  $i > \bar{l}$ .

Thus

$$T_i x_j = \begin{cases} \sum_{l \leq \bar{l}} Pr_l^{A'} a_{A,j} + \sum_{\bar{l} < l \leq L} Pr_l^{A'} a_{B,j} + Pr_j^{A'} & \text{if } i \leq \bar{l} \\ \sum_{l \leq \bar{l}} Pr_l^{B'} a_{A,j} + \sum_{\bar{l} < l \leq L} Pr_l^{B'} a_{B,j} + Pr_j^{B'} & \text{if } i > \bar{l} \end{cases},$$

where we safe notation to write  $Pr_l^{s'} = \frac{\partial Pr[s_j | q + q_s]}{\partial q} \Big|_{q=q_s}$  for  $s \in \{A, B\}$ . Using our information on transitions for state space  $S_E$  (26), we can further re-write this as

$$T_i x_j = \begin{cases} h'(2q_A)(a_{B,j} - a_{A,j}) + Pr_j^{A'} & \text{if } i \leq \bar{l} \\ h'(2q_B)(a_{A,j} - a_{B,j}) + Pr_j^{B'} & \text{if } i > \bar{l} \end{cases}.$$

Using the transition information, we can also write

$$T_{ii} W = \begin{cases} h''(2q_A)(V_B - V_A) & \text{if } i \leq \bar{l} \\ -h''(2q_B)(V_B - V_A) & \text{if } i > \bar{l} \end{cases}.$$

Which then allows us to conclude, for  $i \leq \bar{l}$ :

$$\begin{aligned} \frac{W_{ij}}{W_{ii}} &= \left[ \delta h''(2q_A)(V_B - V_A) + (1 - \delta)u_{11}^A \right]^{-1} \\ &\begin{cases} \delta h''(2q_A)(V_B - V_A) + (1 - \delta)u_{12}^A \\ + (1 - \delta)\delta \left[ h'(2q_A)(a_{B,j-K} - a_{A,j-K}) + Pr_{j-K}^{A'} \right] (u_2^A - u_1^A) & \text{if } i = j - K; , \\ (1 - \delta)\delta \left[ h'(2q_A)(a_{B,j-K} - a_{A,j-K}) + Pr_{j-K}^{A'} \right] (u_2^A - u_1^A) & \text{otherwise} \end{cases} \end{aligned}$$

and if  $i > \bar{l}$ :

$$\begin{aligned} \frac{W_{ij}}{W_{ii}} &= \left[ \delta h''(2q_B)(V_A - V_B) + (1 - \delta)u_{11}^B \right]^{-1} \\ &\begin{cases} \delta h''(2q_B)(V_A - V_B) + (1 - \delta)u_{12}^B \\ + (1 - \delta)\delta \left[ h'(2q_B)(a_{A,j-K} - a_{B,j-K}) + Pr_{j-K}^{B'} \right] (u_2^B - u_1^B) & \text{if } i = j - K; . \\ (1 - \delta)\delta \left[ h'(2q_B)(a_{A,j-K} - a_{B,j-K}) + Pr_{j-K}^{B'} \right] (u_2^B - u_1^B) & \text{otherwise} \end{cases} \end{aligned}$$

Now following the notation in Lemma 4 , let

$$\begin{aligned} a_1 &= -\left[\delta h''(2q_A)(V_B - V_A) + (1 - \delta)u_{11}^A\right]^{-1} \left[\delta h''(2q_A)(V_B - V_A) + (1 - \delta)u_{12}^A\right], \\ d_1 &= -\left[\delta h''(2q_B)(V_A - V_B) + (1 - \delta)u_{11}^B\right]^{-1} \left[\delta h''(2q_B)(V_A - V_B) + (1 - \delta)u_{12}^B\right]. \end{aligned}$$

Then we can apply Lemma 4 to  $M$ . To conclude, we need to compute the quantities given in the Lemma that pin down  $\lambda_{3,4}$ .

To do this, we will compute  $a_s^T \iota_{\bar{l}}$ , and  $a_s^T \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix}$ , for  $s \in \{A, B\}$ .

$$\begin{aligned} b_A &= a_A^T \iota_{\bar{l}} = \frac{\delta}{(1 - \delta)\omega} \left[ (1 - h_A)(1 - \delta(1 - h_A)) + \delta h_A h_B \right] \\ &= \frac{\delta(\omega - h_A)}{(1 - \delta)\omega}; \\ b_B &= a_B^T \iota_{\bar{l}} = \frac{\delta}{(1 - \delta)\omega} \left[ (1 - h_A)\delta h_B + h_B(1 - \delta(1 - h_A)) \right] \\ &= \frac{\delta h_B}{(1 - \delta)\omega}. \end{aligned}$$

Recalling that  $a_A^T \iota_L = a_B^T \iota_L = \frac{\delta}{1 - \delta}$ , we get

$$\begin{aligned} a_A^T \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix} &= \frac{\delta}{1 - \delta} - b_A = \frac{\delta h_A}{(1 - \delta)\omega}; \\ a_B^T \begin{bmatrix} \mathbf{0}_{\bar{l}} \\ \iota_m \end{bmatrix} &= \frac{\delta}{1 - \delta} - b_B = \frac{\delta(\omega - h_B)}{(1 - \delta)\omega}. \end{aligned}$$

In the notation of Lemma 4, we have

$$\begin{aligned}
a_2^T \iota_{\bar{l}} &= - \left[ \delta h''(2q_A)(V_B - V_A) + (1 - \delta)u_{11}^A \right]^{-1} \\
&\quad (1 - \delta)\delta(u_2^A - u_1^A)h'(2q_A) \left[ b_B - b_A - 1 \right]; \\
c^T \iota_{\bar{l}} &= \left[ \delta h''(2q_B)(V_A - V_B) + (1 - \delta)u_{11}^B \right]^{-1} \\
&\quad (1 - \delta)\delta(u_2^A - u_1^A)h'(2q_B) \left[ b_B - b_A - 1 \right]; \\
b^T \iota_m &= \left[ \delta h''(2q_A)(V_B - V_A) + (1 - \delta)u_{11}^A \right]^{-1} \\
&\quad (1 - \delta)\delta(u_2^B - u_1^B)h'(2q_A) \left[ b_B - b_A - 1 \right]; \\
d_2^T \iota_m &= \left[ \delta h''(2q_B)(V_A - V_B) + (1 - \delta)u_{11}^B \right]^{-1} \\
&\quad (1 - \delta)\delta(u_2^B - u_1^B)h'(2q_B) \left[ b_B - b_A - 1 \right].
\end{aligned}$$

Finally notice that

$$\begin{aligned}
1 + b_A - b_B &= \frac{\delta}{(1 - \delta)\omega} (\omega - h_A - h_B) + 1 \\
&= \omega^{-1}.
\end{aligned}$$

Recall  $J_1(\sigma^*)$ , the matrix of best response-derivatives for the  $S^*$ -collusive equilibrium  $\sigma^*$ . Take  $p, q$  from the definition of  $\lambda_{3,4}$  in Lemma 4. Then putting all the derivation above together, some more re-writing reveals that

$$p = \text{tr}(J_1(\sigma^*)), \quad q = \det(J_1(\sigma^*)),$$

where  $\text{tr}, \det$  represent trace and determinant. It follows that  $\lambda_{3,4}$  equal the two eigenvalues of  $J_1(\sigma^*)$ .

Thus, if  $\sigma^*$  is unstable,  $\rho^*$  must be unstable also. Finally, note that  $a_1 \in (-1, 1)$  always holds by construction of  $q_A$  and definition of  $g$ . That is because  $q_A < q_N$ , and therefore  $0 \geq u_{12}^A > u_{11}^A$ .

Since  $u_{12}^B \geq u_{11}^B$  must hold, we always have  $d_1 > -1$ . However, by definition of  $g$  it is

possible that  $u_{12}^B \geq 0$ , which means we need an additional assumption to ensure  $d_1 < 1$ :

$$\begin{aligned} d_1 &< 1 \\ \Leftrightarrow u_{11}^B - u_{12}^B &> 2\left(\omega^{-1}\delta h_B''(u^A - u^B) + u_{11}^B\right) \\ \Leftrightarrow u_{11}^B + u_{12}^B &< -2\omega^{-1}\delta h_B''(u^A - u^B) \end{aligned}$$

which holds if  $u_{11}^B + u_{12}^B \leq 0$ . Thus, if as stated in the Proposition  $u_{11}(q, q) + u_{12}(q, q) \leq 0$  for all  $q$ , we are done:  $d_1 \in (-1, 1)$ , so it can never affect conclusions about stability of  $\rho^*$ .

## Appendix C. Proofs

### C.1. Proof of Proposition 5

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

**Lemma 6.** *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  $-\underline{h}'' = \frac{h'(2D)}{D}$ . That is equivalent to  $P''(2D)D + P''(2D) = 0$ . Now, since  $g$  is strictly increasing, there exist  $P_H > 0 > (P_L - P_H)$  such that

$$g(0) < \frac{P_H}{-(P_L - P_H)} < h(2D). \quad (33)$$

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

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

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  $M^* \in (0, 2D)$  such that  $u_1(0, M^*) = 0$ . This corresponds to  $K^*$  in Lemma 1. Finally to check whether Definition 6 (iv), note that

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

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

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

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



Now we need the following observations based on the definition of  $W$  in section 5:

$$\begin{aligned}
W_1 &= \omega^{-1}(1 - \delta P_{BB}) \left[ \omega^{-1} \delta P'_{AB}(u^B - u^A) + u_1^A \right], \\
W_2 &= \omega^{-1}(\delta P_{AB}) \left[ \omega^{-1} \delta P'_{BB}(u^B - u^A) + u_1^B \right], \\
W_{11} &= -2\omega^{-1} \delta P'_{AB} W_1 + \omega^{-1}(1 - \delta P_{BB}) \left[ \omega^{-1} \delta P''_{AB}(u^B - u^A) + u_{11}^A \right], \\
W_{22} &= 2\omega^{-1} \delta P'_{BB} W_2 + \omega^{-1}(\delta P_{AB}) \left[ \omega^{-1} \delta P''_{BB}(u^B - u^A) + u_{11}^B \right], \\
W_{12} &= \omega^{-1} \delta \left[ P'_{AB} \frac{1 - \delta P_{BB}}{\delta P_{AB}} W_2 - P'_{BB} \frac{\delta P_{AB}}{1 - \delta P_{BB}} W_1 \right], \\
W_{13} &= W_{11} + \omega^{-1}(1 - \delta P_{BB}) \left[ \omega^{-1} \delta P'_{AB}(u_1^A - u_2^A) + u_{12}^A - u_{11}^A \right], \\
W_{24} &= W_{22} + \omega^{-1}(\delta P_{AB}) \left[ \omega^{-1} \delta P'_{BB}(u_2^B - u_1^B) + u_{12}^B - u_{11}^B \right], \\
W_{14} &= -\omega^{-1} \delta P'_{BB} \frac{\delta P_{AB}}{1 - \delta P_{BB}} W_1 + \omega^{-1}(1 - \delta P_{BB}) \omega^{-1} \delta P'_{AB} \left[ \omega^{-1} \delta P'_{BB}(u^B - u^A) + u_2^B \right] \\
&= W_{12} + \omega^{-1}(1 - \delta P_{BB}) \omega^{-1} \delta P'_{AB}(u_2^B - u_1^B), \\
W_{23} &= \omega^{-1} \delta P'_{AB} \frac{1 - \delta P_{BB}}{\delta P_{AB}} W_2 - \omega^{-1}(\delta P_{AB}) \omega^{-1} \delta P'_{BB} \left[ \omega^{-1} \delta P'_{AB}(u^B - u^A) + u_2^A \right] \\
&= W_{12} + \omega^{-1}(\delta P_{AB}) \omega^{-1} \delta P'_{BB}(u_1^A - u_2^A).
\end{aligned} \tag{34}$$

Then, an optimal, non-degenerate, interior strategy  $\alpha^*$  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  $\alpha^*$ , 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 7.** *Suppose  $h \in \mathcal{G}$  and*

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

Then for all neighborhoods  $\mathcal{N}$  of  $\tau$  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 6, 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 6 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.  $\square$

We can use Lemma 7 to make our analysis cleaner. As long as  $g$  satisfies the condition of the Lemma,  $q_N$  can be treated as a primitive of the model, replacing  $\tau$ .

**Lemma 8.** Suppose  $h \in \mathcal{G}$  and  $P_H > 0 > \Delta$  and a convex cost function  $c(q)$  such that Lemma 6 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

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

then there exist  $P_H > 0 > \Delta$  and a convex  $c(q)$  that satisfy Lemma 6 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 6 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 6 that  $u_1(q, q)$  is strictly decreasing for all  $q \in [0, D]$ . From Lemma 7 we get that we can take  $q_N$  arbitrarily close to  $\tau$ . 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  $c'(0), c'(\tau)$ . If we now additionally impose, 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 6,  $u_{11}(q, \hat{q}) < 0$  for all such  $q, \hat{q}$  and the fraction is monotone still.

Finally, notice that we can write

$$\frac{u_1(q, \hat{q})}{h'(q + \hat{q})} = \frac{P_H + \Delta h(q + \hat{q}) - c'(q)}{h'(q + \hat{q})} + \Delta q. \quad (35)$$

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

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

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 (35) 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  $\tau$  such that for all  $P_H > 0 > \Delta$  and a convex  $c(q)$  that give  $q_N \in N'$ , the Lemma still goes through.  $\square$

Now for the proof of the Proposition:

Assume for  $h$ :  $h'(0) = h'(M) = 0$  and

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

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

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

By Lemma 8 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

$$z'(q_A) = -\frac{h'_B u_{11}^A + u_{12}^A - 2h''_G \frac{u_1^A}{h'_G}}{h'_G u_{11}^B + u_{12}^B + 2h''_B \frac{u_1^B}{h'_B}}. \quad (36)$$

It is then not surprising that at  $q_N$ ,  $z'(q_N) = -1$ . Now define  $\Psi(q_A) = W_1(q_c, z(q_A), q_A, z(q_A))$  as the first order condition of  $W$  with respect to  $q_A$ , substituting in  $z(q_A)$  so that at every

$q_A$ ,  $W_2(q_c, z(q_A), q_A, z(q_A)) = W_1(q_c, z(q_A), q_A, z(q_A))$  must hold. Thus, any zero of  $\Psi(q_A)$  must set both first order conditions to zero.

Since  $\sigma_N$  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 7 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

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

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

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

Thus we can write

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

where  $R_1 = 2\Delta g_N q_N h''_N$  vanishes for  $q_N$  close enough to  $\tau$ , and  $R_2 = \frac{1-\delta}{2\delta}(u_{12}^N + u_{11}^N)$  vanishes as  $\delta \rightarrow 1$ . Then for  $\delta < 1$  close enough to 1 and  $q_N$  close enough to  $\tau$ , 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  $\sigma$  must be negative definite: we see from (34) that  $W_{12} = 0$ , so the Hessian must be diagonal at  $\sigma$ . A sufficient condition for negative definiteness 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  $\sigma$  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

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

holds for all  $q \in S_g$ . 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

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

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 8 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

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

Where again by the negative definite Hessian, we have  $H_1^B(q, q_B) < 0$ . Then in the proof of Lemma 8 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  $\sigma = (q_A, q_D)$  is the unique best reply to an opponent playing  $\sigma$ , and thus  $\sigma$  is a symmetric equilibrium as required.

■

## C.2. Proof of Proposition 1

The proofs are given in terms of the general algorithm class treated in this paper, introduced in Appendix A. 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 Benaim, 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  $x_1, \dots, x_n$  to  $\dot{x} \in G(x)$ <sup>27</sup>, and real numbers  $t_1, \dots, t_n$  greater than  $T$  such that:
  - a)  $x_i(s) \in A$  for all  $0 \leq s \leq t_i$ , and for all  $i = 1, \dots, n$ ,
  - b)  $\|x_i(t_i) - x_{i+1}(0)\| \leq \varepsilon$  for all  $i = 1, \dots, n-1$ ,
  - 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  $x, y \in A$ .

Importantly, these sets include rest points and limit cycles (if they exist). We refer to Benaim, Hofbauer, and Sorin (2005) Definition 6 for a definition, and Papadimitriou and Piliouras (2018) for an intuitive discussion. The following result shows why these sets are of importance in our analysis:

**Proposition 9.** *With probability 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  $\rho$ , and particularly

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

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

*Proof.* The algorithm (17) can be written as

$$\rho_{t+1} = \rho_t + \alpha_t [F_g(\rho_t) + \delta_t + M_{t+1}],$$

where  $\delta_t = \mathcal{A}_g[F, D_t](\rho_t) - F_g(\rho_t)$ .

We can now show that a version of Corollary 4 in Chapter 5 of V. S. Borkar (2009) now

---

<sup>27</sup>Recall that  $G(x)$  is an inclusion, so uniqueness of solutions cannot be guaranteed.

goes through. First notice that by our assumptions on  $M_{t+1}$  in Assumption 6, we have that this Corollary 4 goes through if it were true that  $\delta_t = 0$  for all  $t$ . By Definition 12,  $\delta_t = o_P(1)$ , so we can apply extension 3 in Chapter 2 of V. S. Borkar (2009) and the result follows.  $\square$

Since payoffs are differentiable around  $\rho^*$ , point 1 follows as long as  $\rho^g$  and  $\rho^*$  are close. For point 2, we will prove something more general: as long as  $\rho^*$  is hyperbolic, point 2 holds.

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

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

Since  $\|g\|_1 \leq \gamma$ , for  $\gamma$  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  $\rho^g$  close enough to  $\rho^*$  must also be asymptotically stable. The next Lemma gives a more general result:

**Lemma 9.** *Suppose  $\rho^*$  is hyperbolic. Then the eigenvalues of  $DF_g(\rho^g)$  converge to the eigenvalues of  $DF(\rho^*)$  uniformly over  $g \in \mathcal{B}_\gamma^1$  as  $\gamma \rightarrow 0$ . Thus, for small enough  $\gamma$ ,  $\rho^g$  has the same stability properties as  $\rho^*$ .*

*Proof.* We will show that eigenvalues of a hyperbolic matrix  $DF(\rho^*)$  vary continuously in  $C^1$  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  $\rho^g \rightarrow \rho^*$  uniformly for  $g \in \mathcal{B}_\gamma^1$  as  $\gamma \rightarrow 0$  (see above proof of point 2), we get that

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

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

Since we know that all  $\rho^g$  must be asymptotically stable for  $\gamma$  small enough, we can apply Faure and Roth (2010) (Thm 2.8).

We only need 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  $p$*

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

We let  $Att(X)$  be the set of attainable points for algorithm (17). Then we need that the basin of attraction of an attractor has nonempty intersection with  $Att(X)$ . This should be true given our support condition on  $M_{t+1}$  and the assumption that equilibria must be interior:

**Lemma 10.** *Let  $B$  be a basin of attraction of an attractor  $A$  for  $F_g$ . Suppose  $\rho_t \in \bar{A} \setminus B$ . 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  $z \in B$ , we can pin down the necessary shock  $M_z$  to reach it:

$$M_z = \frac{z - \rho_t}{\alpha_t} - F_g(\rho_t).$$

Since  $z \in \text{int}(E)$  by definition,  $M_z$  is in the support of  $M_{t+1}$  for every  $t$ . For any ball  $B_z$  around  $z$ , we can 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.)  $\square$

All other conditions that are sufficient for the model-algorithm to converge to the attractor hold by assumption 6.

■

### C.3. 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), Thm 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  $C^1$  in  $U_{\rho^*}$ , and we can allow  $F(\rho)$  to be multivalued otherwise.

First, we define invariant sets for given differential equations:

**Definition 15.** *Let  $z(t, z_0)$  be the solution to 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  $g \in \mathcal{B}_\gamma^1$ , we know from Proposition 9 that only ICT sets (recall Definition 13) subset of a neighborhood of  $\rho^g$  are candidates to being limiting points of the algorithm (17). 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 enough  $\gamma$ , no other ICT sets can exist in a neighborhood around  $\rho^*$ , which finishes the proof.

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

Note that by Lemma 9, there are  $\gamma > 0$  small enough such that all  $\rho^g$  are linearly unstable just as  $\rho^*$ . We can thus apply Benaim and Faure (2012), Thm 3.12 to prove  $P[L_{S,g} = \rho^g] = 0$  first:

We can show that the sufficient conditions for this hold by definition of our algorithm under assumption 6. According to Faure and Roth (2010) Proposition 2.16, we have that the bounding function required in Benaim and Faure (2012), Hypothesis 2.2 exists given our assumptions on  $\varepsilon_t, M_t$ . Benaim and Faure (2012)'s Hypothesis 3.6 is then also satisfied, at least in a neighborhood of the rest point. As noted by their Remark 3.7, all conditions only need to hold in a neighborhood of the unstable point, so set-valued gradients outside the neighborhood are allowed.

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

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  $\rho^*, \rho^g$  are isolated invariant sets in their respective neighborhoods. These neighborhoods are nontrivial for all  $\gamma$  small enough, which follows from both

$\rho^*, \rho^g$  being hyperbolic:

By Hartman-Grobman and hyperbolicity there exists a homeomorphism  $H$  on a neighborhood  $N \subseteq U_{\rho^*}$  of  $\rho^*$  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  $\dot{\rho} \in \text{conv}[F(\rho)]$ , and  $\psi(t, \cdot)$  is the solution to the ODE  $\dot{y} = DF(\rho^*)(y - \rho^*)$ . Given a neighborhood  $U \subseteq N$  of  $\rho^*$ , define

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

We will show that  $\rho^* = \text{inv}(U)$ , and therefore it is isolated invariant.

Notice that  $\text{inv}(U)$  can be rewritten as

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

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

$$\text{inv}(U) = \rho^*,$$

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

Since  $\rho^g$  are hyperbolic for  $\gamma$  small enough, an analogous argument gives us that  $\rho^g$  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  $\rho^g$  is isolated invariant in  $N_g$ . We are left to show that for  $\gamma$  small enough, for all  $g \in \mathcal{B}_\gamma^1$ ,  $\rho^* \in N_g$ :

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

For small enough  $\gamma$ , all  $\rho^g$  are hyperbolic when  $g \in \mathcal{B}_\gamma^1$ . 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 11.** *For any  $\delta > 0$  with  $a_0 < 1 - \delta$  there exists  $\bar{\gamma} > 0$  such that for all  $\gamma \in (0, \bar{\gamma}]$ , there is a set of  $\{a_g\}_{g \in \mathcal{B}_\gamma^1}$  as defined above with*

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

*Proof.* Apply Lemma 9. Since there is a one-to-one mapping between eigenvalues and  $\{e^{\rho^u}, e^{-\rho^l}\}$ , we can find numbers  $a_g$ . The result follows.  $\square$

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

**Lemma 12.** *Suppose  $\rho^*$  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  $g \in \mathcal{B}_\gamma^1$ , there is  $B_z^g(\rho^g) \subseteq N_g$  with  $z \geq \underline{z}$ .*

*Proof.* For small enough  $\gamma$ , all  $\rho^g$  are hyperbolic when  $g \in \mathcal{B}_\gamma^1$ . Fix such a  $g$ . Given some  $\varepsilon > 0$ , let  $r_\varepsilon$  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

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

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

We are left to show that  $r_\varepsilon$  can be made to depend only on the eigenvalues of  $DF(\rho^*)$  and  $\gamma$ .

Notice that small enough  $\underline{z} > 0$  pins down the  $\delta > 0$  referred to in Lemma 11: Let

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

For  $\delta > 0$  small enough, choose  $\bar{\gamma} > 0$  such that Lemma 11 holds. It follows from the Lemma that  $\hat{z}(\bar{\gamma}) > 0$ . Then any  $\underline{z} < \hat{z}(\bar{\gamma})$  satisfies our conditions and the conclusion follows.  $\square$

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

■

#### C.4. Proof of Lemma 1

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

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

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  $q_2 \in I$  so that there exists unique  $K^* < 2K$  with  $u_1(0, K^*) = 0$ . 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  $q_2 < K^*$  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  $q_2 \in I$  there is a unique best response  $q^*(q_2)$  that is pinned down by first order conditions whenever  $q_2 \leq K^*$ , and equals zero otherwise.

Whenever  $q_2 \leq K^*$ , 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  $q_2 < K^*$  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:  $q^*(0) < K^*$ . 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  $q_2 < K^*$  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.

