Reinforcement Learning and Collusion

Clemens Possnig
Vancouver School of Economics, University of British Columbia
September 28, 2022
Link to current version

Abstract. This paper presents an analytical characterization of long run outcomes arising from learning algorithms playing a repeated game. I show that these outcomes correspond to attracting Markov-perfect equilibria. Whether an equilibrium is attracting is determined by details of a tractable differential equation. I give necessary and sufficient conditions on the game and on the algorithms for the stage-game Nash equilibrium not to be a long run outcome. Applying the approach in a Cournot game, I give conditions under which algorithms learn to collude 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 thank Alexander Frankel, Kevin Leyton-Brown, Vadim Marmer, 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

A growing fraction of economic interactions[1] is carried out by algorithms which learn from observing real-time data. A learning algorithm is an updating rule for a policy. The policy is a mapping from observables, such as market data and information about competitors, to actions, e.g. prices to set or quantities to sell. The updating rule takes real-time observations, such as payoffs, and uses those to adapt the policy to achieve higher profits. What market outcomes can we expect when algorithms compete?

Recent studies have alerted economists to the possibility that algorithms may learn to collude. 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%. Using numerical simulations, (Klein (2021), Calvano, Calzolari, Denicolo, et al. (2020),Calvano, Calzolari, Denicoló, et al. (2021)) show that after many periods of interaction, algorithms may learn to play collusive policies. Their evidence suggests that algorithms in the long-run may not only be able to sustain profits above the competitive level, but also learn to play repeated game strategies akin to typical carrot-and-stick type strategies studied in the economic theory literature.

These observations are problematic not only due to their implied welfare consequences, but also because most legal systems are currently not adapted to appropriately deal with tacit collusion resulting from algorithmic pricing (Calvano, Calzolari, Denicolò, et al. (2019)). Firstly, it is difficult to establish collusion absent proof of explicit agreements to collude. Secondly, the outcome of collusion may lack intent, since firms may not even realize the potential of their algorithms to learn to play collusively, or claim so. Thus, in order to develop better informed antitrust procedures there is a need to better understand the channels through which market conditions and details of competing algorithms affect outcomes.

This paper contributes to this problem by providing a characterisation of the long-run learning behavior of algorithms playing a repeated game. Algorithmic learning is represented by a stochastic difference equation, finding the limits of which is commonly considered a complex problem. I show that one can solve this problem by studying the stability of rest points of a deterministic differential equation, greatly simplifying the required analysis.

[1]Real-world applications in strategic environments range from autonomous driving to algorithmic pricing (Gronauer and Diepold (2022)), while the number of games algorithms beat the best human players is ever growing (c.f. Walker, B. (Mar 14. 2020). *The Games That AI Won.*towardsdatascience. https://towardsdatascience.com/the-games-that-ai-won-ff8fd4a71efc).

To the best of my knowledge, this is the first analytical study providing characterisations of long-run behavior of learning algorithms in the context of oligopolistic competition.[2]

For an important subclass of the algorithms I consider, the differential equation characterising long-run behavior is a state-dependent best-response dynamic. Only Markov-perfect equilibria that are attracting under this dynamic can be outcomes when the policy profiles of the algorithms converge. The implication of this characterisation is that it becomes necessary to understand what it means for a given equilibrium to be attracting.

As a useful benchmark outcome, I study when the repetition of the stage-game Nash equilibrium is attracting. This is important since such an outcome may be the best one can hope for in terms of maximizing consumer welfare, and also since fixing this outcome allows for interesting comparisons among possible observables algorithms may condition their policies on. One can see these observables as the information-states conditioned on by a policy. As a result, I provide the first step at a novel categorization of the coordinative ability granted to learning algorithms by a given type of information. This is done by studying what kinds of observables allow for algorithms to learn the stage-game Nash equilibrium.

Applying my characterisation 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.

The focus of this paper is on long-run behavior of actor-critic reinforcement learning (RL) algorithms. Reinforcement learning is a fundamental machine learning paradigm concerned with algorithms that update policies over time towards actions that have performed well in the past. Actor-critic algorithms keep track of two main objects: an estimate of a performance measure (e.g. a value function), and a policy function that is being updated using the performance measure. The name "actor-critic" was introduced in the computer science literature to allude to the fact that the performance measure, "the critic", gives feedback about the performance of the policy, "the actor". As a useful benchmark throughout the paper, I consider model-free RL. Such RL maintain estimates of their performance measure without explicitly modeling their own or their opponent behavior and payoffs. Thus, such algorithms fall into the class of adaptive, uncoupled learning rules. This serves

[2]A notable exception is Banchio and Mantegazza (2022), a paper in development concurrently with this project, having similar aims but using different methods discussed in the literature review.

as a minimal-information benchmark, which will be shown to be sufficient to lead to the emergence of collusion.[3]

A well-studied special case in my class is actor-critic Q-learning[4] (ACQ). For these algorithms, the performance measure is a value function (referred to as "Q-function" in the literature). Policies are updated towards to optimal policy given the current estimate of the value function. To fall into my class, two main sufficient conditions must be satisfied: firstly, after a long enough amount of play, the value-function estimator differs from the true value function at most by a bounded, smooth bias term (which can be zero)[5]. Secondly, the step-sizes at which the policies are updated decrease over time within a given spectrum of speeds[6]. If these two conditions are satisfied, I show that the resulting policy iteration is a noisy discretization of a differential equation. Since under ACQ policies are updated towards an optimal policy, it then makes sense that trajectories of the policies will approach trajectories of a state-dependent best-response dynamic. As a result, when the process of policies converges to a point, that point must be a Markov-perfect equilibrium which is attracting (also called "stable") under the best response dynamic.

In general, I show that once an RL falls into my class, the connection to an underlying deterministic differential equation can be made to characterise the long-run behavior using a method known as stochastic approximation (V. S. Borkar (2009)). I then show that it is enough to consider attractors of that differential equation, while unstable points will never be learned. An equilibrium of a differential equation is attracting if once trajectories of the differential equation reach a neighborhood of that equilibrium, they converge to it. Conversely, an equilibrium is repelling (also called "unstable") if no matter how close a trajectory is to that equilibrium, it can be repelled from it and will not converge to it.

While my analysis applies to general repeated continuous action games, I apply my methodology to study ACQ learners in the workhorse model of oligopolistic Cournot competition. The benchmark outcome is repeating the stage-game Nash equilibrium in every period. This allows for a better understanding of the coordinative behaviors a given policy

[3]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.

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

[5]The bias term is a modelling decision inspired by the fact that for many real-world performance-measure 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). While all results in the paper go through if the underlying performance measure has no asymptotic bias, I show that the results are robust to biased estimates.

[6]Stepsizes must satisfy the Robbins-Monro condition commonly invoked in the computer science literature. See V. S. Borkar (2009), Chapter 2.

space can support. A policy space consists of the definition of observables policies condition on, and the actions they can play. A definition of observables comes with the space of their realizations, and a transition function that pins down how the observables evolve over time as a function of actions taken by the algorithms. A simple example would be the observation of the previous period's price as an observable, where the price is realized as a function of quantities sold by each algorithm. Different policy spaces may support different equilibria, but stage-game equilibria are equilibria under any policy space, making tight comparisons of policy spaces possible. To the best of my knowledge, this is the first such comparative statics exercise.

I first prove that a simple condition on market fundamentals is necessary and sufficient for a given stage-game equilibrium to be learned under a class of policy spaces I call "1-recall policy space". The condition is a bound on the slope of the myopic best response function evaluated at that stage-game Nash equilibrium implying the stability of that equilibrium under myopic best-response dynamics[7]. In common textbook-versions of the Cournot game[8] there is a unique, interior, and symmetric stage-game Nash equilibrium that satisfies this condition. 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)). If the stage-game equilibrium is stable under myopic best-response dynamics, we will call it *statically attracting*. In contrast, I then characterise a policy space I call *direction-switching* with the following property: For any value the above payoff condition takes, there exists an element in the class of direction-switching policy spaces so that the stage-game Nash equilibrium will never be learned. Once that is true for a given equilibrium and policy-space, we will call the Nash equilibrium *dynamically repelling*.

To understand this comparative statics result with respect to the policy space, it is necessary to first discuss policy spaces more thoroughly. The restriction in the analysis in this paper is that the domain (state-space) of a policy must be finite.[9] This domain is the space of realizations of a commonly observable state-variable s. I show that what is crucial in determining whether a given stage-game equilibrium can be learned is the evolution of

[7]This refers to classical best-response dynamics that consider the learning of stage-game strategies.

[8]This holds for a large class of payoff functions including linear demand and convex cost, under some boundary conditions preventing the monopoly equilibrium to exist.

[9]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.

states, and not the cardinality of the state-space. The state variable transitions from state to state according to some probability law that can depend on actions of the algorithms, but is assumed Markov stationary given a fixed set of actions. In the Cournot example, 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. As mentioned before, an example of a state is then the previous period's price realization. As a result, a policy would be a mapping from price realizations to actions, and more commonly referred to as a public 1-recall policy, where the word "public" refers to the fact that the price publicly observed.

In the example of public 1-recall policies, states transition to new states *independently* of the current state. If given two states s, s', the same aggregate quantity is played, then the transition probability to any new state s^* (in this example, the probability of observing a given price s^*) must be the same in s as in s'. As mentioned before, when the common policy space is 1-recall, it is necessary and sufficient to consider the condition on the slope of the static best-response to determine whether the stage-game equilibrium is learnable (can be learned with positive probability), and hence is dynamically attracting. In other words, under 1-recall policies, information on the evolution of the state is irrelevant when determining whether a stage-game equilibrium will be learned or not.

Next I show that in general, details of the state evolution can be crucial in determining whether the stage-game equilibrium will be learned. I introduce direction-switching states, which 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 moves with the aggregate quantity Q in the same way as does the probability of transitioning from 2 to 2. As an example, let $P_{ss'}(Q)$ be the likelihood of moving from s to s' given aggregate quantity Q. The state-space is direction switching if $P_{11}(Q) = P_{22}(Q)$. Then the intuition best comes to mind if in addition the probability of transition from 1 to 1 is monotone increasing in aggregate quantity Q. This implies that the probability of transition from 2 to 1 must be monotone decreasing in Q, and thus the direction of the effect of adjusting one's quantity *switches* as we consider differing starting states. I show that under this state-space, for any value the before-mentioned condition on market fundamentals takes, there is a transition probability such that the stage-game equilibrium will be dynamically repelling and therefore not learned. These observations indicate an important difference between 1-recall and direction switching in the *coordinative ability* these policy spaces can grant to learning algorithms. In addition, I will discuss later on

that state-spaces with this property support simple two-quantity carrot-stick style collusive equilibria.

The intuition for how a statically attracting Nash equilibrium can become dynamically repelling under state-dependent policies becomes apparent when considering deviations from the equilibrium. Once we consider state-dependent policies, we introduce dynamic incentives to the updating equation of a learning algorithm. When computing best responses at the Nash equilibrium, there is a static incentive that considers myopically optimal behavior in the given period, and a dynamic incentive, that considers how a change in actions affects continuation payoffs. The static incentive given a small deviation of the opponent is to move in the opposite direction from the perturbation, but by an amount smaller than the original perturbation of the opponent. This is what makes the equilibrium statically attracting, and what generates the typical cobweb-style adjustment process in discrete time. However, the dynamic incentive concerns itself with the likelihood of visiting a given state. Suppose that the opponent's policy is decreased marginally in state s^*, while staying equal to the original static Nash equilibrium quantity in all other states. The standard Cournot model tells us that this state now has become a desirable state, since payoffs must be uniformly higher (which can be seen from the demand being decreasing in quantities). Dynamically then there is an incentive to visit this state more often in the future.

This shows that optimal play depends on static and dynamic incentives. To understand whether an equilibrium is dynamically attracting or not, it is necessary to consider the higher-dimensional differential system resulting from having a binary state-space. Recall that when an equilibrium is attracting, policy profiles that are slightly perturbed from the equilibrium will converge back to it under the dynamics. I show that since transitions are state-independent under 1-recall policies, any perturbation of the equilibrium collapses to a perturbation along the 45-degree line. In other words, any perturbation must approach a perturbation equal in all states, which renders any dynamic consideration in determining optimal behavior mute (since all states are affected equally). As a result, only myopic incentives matter, and so follows the property of 1-recall policies discussed before. On the other hand, under direction-switching policies I show that the dynamic effect gets reinforced when perturbing the equilibrium. If the dynamic effect counteracts the static effect, and overpowers it given parameters of the game and transition probabilities, this renders the static Nash equilibrium dynamically repelling even if it may be statically attracting.

Calvano, Calzolari, Denicoló, et al. (2021) show in their simulation studies that RL can learn to play collusive strategies in a Cournot game. Collusion in that context is a policy that for some states plays low quantities (correspoding to high prices) and high quantities (punishment quantities) for others.

My characterisation implies that once one knows the policy space of the algorithms, and the payoff environment of the game, one can determine whether an equilibrium will be learned with positive probability by checking its stability property, which comes down to an eigenvalue-condition[10]. This condition can serve as an empirical test. Using data on the firm's payoff structure, market demand, and the policy spaces of the algorithms competing, one can device a one-sided test to see whether a given equilibrium can be learned with positive probability or not.

In the Cournot application, I show for a class of payoff functions that there exists a simple binary policy space (the above-mentioned direction switching policy space) 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. I then go on to study a numerical example where such an attracting, collusive equilibrium exists and verify in a simulation that ACQ learners initialized in a neighborhood of that equilibrium will indeed converge to it. I also simulate learners 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.

Relation to the 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. As mentioned above, Assad et al. (2020) give an important 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. Their class intersects with the class studied here, but there are important differences. It is unclear that their approach can accommodate actor-critic

[10]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.

approaches that are importantly 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). Furthermore, the results studied in their paper are different in flavor than the results here. It appears that the result they uncover is more an example of an inability of algorithms to learn the dominant strategy equilibrium, rather than an ability of the algorithms to learn collusive behavior as shown in this paper. The collusive outcomes considered in this paper are outcomes that can be recovered from rational agents playing repeated games, which are not featured in Banchio and Mantegazza (2022).

Secondly, this paper connects to a long history of the theory of learning in games. My RL class contains algorithms that impose little informational assumptions, commonly called "model-free" as defined earlier. Thus, the RL class considered here can be seen as examples 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), Milgrom and Roberts (1991), Fudenberg and Kreps (1993), Fudenberg and Levine (2009), Gaunersdorfer and Hofbauer (1995) and many more. The fact that in my paper players learn to play repeated game policies is in contrast to the classical game theoretic learning literature, that generally considers the process of learning to play static game Nash equilibria. In that sense, this project sheds 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. At the same time, 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 have to live in that policy space. 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.

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. Perhaps more due to notation and framing than due to content, Leslie, Perkins, and Xu (2020)'s paper is considered as research in the theory of learning in games much more so than algorithmic learning theory.

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 (Brown (1951)), or stochastic ficitious play (c.f. Hofbauer and Sandholm (2002)), depending on details of the process. Their framework allows to conclude asymptotic behavior of learning processes once one has shown that the process is a GWFP process. They show that GWFP converges in games that have the fictitious play property. Notably that class includes zero-sum games, submodular games, and potential games.

One can interpret results in this paper as showing that a subclass (ACQ) of the RL I consider can be seen as a GWFP process. Therefore, one can apply Leslie and Collins (2006) to conclude the limiting behavior of that process in games with the fictitious play property. However, there are many repeated games of interest that do not have this property; notably standard repeated oligopoly (Cournot) games. 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.

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 by 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. I give lower level sufficient conditions algorithms must satisfy in order to fall into the class I consider, and also provide a full example of an algorithm that satisfies all those conditions. 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.
Finally, this paper can be interpreted as casting RL competition as an equilibrium selection mechanism. It is a common observation in the learning literature that when players follow heuristic learning rules, they may not be able to learn to play all possible equilibria of the game, or not even converge to an equilibrium. This paper is no different, and also allows for the possibility of games for which players may converge to cycles. However, the classical literature was developed as 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, which shows that the selection ability of competing RL is something novel. 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.

The 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.
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 ways of the example of an agent solving a multi-armed bandit problem. For a thorough introduction,

consider Sutton and Barto (2018).

Consider an agent choosing actions $q \in A$ 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

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

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

RL algorithms are updating rules meant for the learning of optimal policies or value functions for a given problem. Such algorithms are commonly used to solve Markov decision problems (MDPs). In Generally, 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

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

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 ρ^* simply by maximizing Q^* in every state.

C. J. C. H. Watkins (1989) then proposed a RL algorithm that estimates Q^*. In the language we introduced earlier, the algorithm takes estimates of Q^* as the relevant performance measure. 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:

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

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 measure. 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 ε-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 ε, she samples uniformly from A.

For a suitable sequence β_t, one can show that Q_t converges in probability to Q^* if states form a Markov chain controlled by q_t and actions are sampled ε-greedily (c.f. C. J. Watkins and Dayan (1992))[11]. Stationarity of the state-transitions conditional on a fixed policy ρ is an important ingredient of the standard convergence proof for Q- learning. If stationarity fails, one can imagine that learning of the correct Q^* may fail also.

3. 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 payoff function, state transitions, and opponents, the player may resort to Q-learning again. What if all players in the game apply this method to learn their optimal policies?

One would immediately have to do away with any stationarity assumptions. If each player acted ε-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. As mentioned in the Introduction, 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

[11] 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.

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 ρ_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 as defined in Section 2. 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 measure, and introduce a class of algorithms that perform reasonably well in the function approximation step, up to a well-behaved asymptotic bias term. I believe that allowing for an asymptotic bias significantly increases the number of learning algorithms that fall into our class of RL agents, due to the inherent problems these agents face while learning, 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[12]. In that case unbiasedly estimating a value function becomes a daunting task even when stationarity of the environment is satisfied. Very commonly in such situations algorithms use some form of parametric function approximation to generate an estimate, which can introduce bias. Often this involves deep neural networks due to their flexibility and scaleability. I refer to François-Lavet et al. (2018) for a thorough introduction to state of the art RL techniques and a deeper dive into issues of biased estimation of value functions and their gradients. I will show that the bias I allow in this class does not affect the main results developed in the next section.

[12]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.

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 $\gamma > 0$, let $\mathcal{B}^i_\gamma$ be the set of C^i functions with bounded derivatives :

$$\mathcal{B}^i_\gamma = \big\{g : E \mapsto E; \mid \sup_{x\in\bar{A}} \|g(x)\| + \sum_{j=1}^{i} \sup_{x\in\bar{A}} \|D^j g(x)\| \leq \gamma\big\}, \tag{3}$$

  where $D^j g$ represents the jth derivative.
- For any set B, define $conv[B]$ 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 $|S| = L$ 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 it's transition probability function, I will refer to a state space S keeping implicitly in mind that it comes with it's 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 $\delta \in (0, 1)$.
Each RL updates a policy function $\rho^i_t : S \mapsto A_i$ the long-run behavior of which is the 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.*[13]

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 strategy profiles $[\rho^i, \rho^{-i}] \in \bar{A}$:

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E}\sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \tag{4}$$

where the expectation is taken over the randomness in the stage-game payoffs and state transitions.
Then define $B^i_S(\rho^{-i})$ as the optimal strategy given a profile $\rho^{-i} \in \bar{A}_{-i}$, chosen from the constraint set of stationary, S-state strategies:

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

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

where due to our assumption on irreducibility of the state space the optimal strategy does not depend on the initial state s_0, and it is indeed optimal over all possible dynamic policies since given a Markov stationary opponent profile ρ^{-i} there must be a Markov stationary best response.
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 ρ^* 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 ρ^* and the Hessian of each agent's optimization problem at ρ^* is negative definite.*

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

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

- *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.*

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):[14][15]

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

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

where $\alpha_t > 0$ is a sequence of stepsizes converging to zero and M_{t+1}^i is an i.i.d, zero-mean, bounded variance noise generated as a means of exploring the policy space[16][17].

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

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

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

$$\Big\{ \max_{q' \in A} Q^{i*}(s, q', \rho_t^{-i}) \Big\}_{s \in S} = BR_S^i(\rho_t^{-i}).$$

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 ρ_t^{-i} fixed forever into the future.* Having read that every algorithm uses (5) to update their policies, this supposition is clearly incorrect. However, firstly as stepsizes for ρ_t^i decrease, computing Q^* can be seen as an approximation to the true future expected value that would take into account evolving ρ_t^{-i}. Secondly, as stated before, this paper is not looking into a positive theory of how an optimal algorithm should behave. Rather, the interest

[14] 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.

[15] 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.

[16] 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 measure Q^* is being estimated. For continuous action problems, this is a common means to aid exploration (Plappert et al. (2017)).

[17] 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.

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. *Suppose there is a small $\gamma > 0$ and bias functions $g^i \in \mathcal{B}_\gamma^2$ such that for each i,*

$$Q_t^i(s,q) - Q^{i*}(s,q,\rho_t^{-i}) = g^i(s,q,\rho_t^{-i}) + o_P(1).$$

Note that Assumption 3 allows for an asymptotic bias term in the Q_t estimation, which one may assume to be uniformly small but not necessarily vanishing. As stated in the beginning of this section, this allowance 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 α_t I maintain the following:

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

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

This assumption takes its name from the celebrated Robbins-Monro algorithm represenation (Robbins and Monro (1951)). The assumption constrains the speed of convergence of α_t, 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 ρ_t:

Assumption 5. *Iterates stay bounded almost surely:*

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

Even though commonly made, Assumption 5 is often difficult to verify. It is common for authors to give all their results conditioning on the event that 5 holds, see for example Benaım and Faure (2012). For a more general discussion of sufficient conditions for bounded iterates, see V. S. Borkar (2009), Chapter 2.

With Assumptions 3 and 4 in place, I will show that one can apply results from stochastic approximation theory (see e.g. V. S. Borkar (2009)) to connect the long-run behavior of ρ_t to limiting sets of solutions to an underlying differential equation. Given Assumption 3, one can convince oneself that this differential equation will have to do with the computation of a best response. This is indeed the case, as will become clear shortly.

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*

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

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

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

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

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

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

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

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

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 ρ_t to solutions of the differential inclusion

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

Since the best-response may be multi-valued, solutions to this inclusion are not guaranteed. However, assumptions on the regularity of F_B^S (which comes down to a linear growth condition) allow us to show that there is a global solution in the sense of Filipov (1988).

When considering that the updating rate α_t converges to zero, one may convince oneself that the recursion in (5) looks similar to a discrete time approximation to a time-derivative. The idea then is to show that the time-interpolated version of ρ_t indeed must stay close, with probability one, to solutions of an underlying differential inclusion. The limiting behavior of ρ_t can then be deduced from a subset of the limiting behaviors of the differential inclusion above.

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

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

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

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

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

Notice that the stability property of an equilibrium depends on the performance measure 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. $g \in \mathcal{B}^1_\gamma$. The stability of ρ^* itself depends further on the state space observed by the algorithms. I therefore emphasize this dependence by writing F_B^S as the best response dynamics defined on state space S.

Proposition 2. *Let* $\rho^* \in \mathcal{U}_S$ *be linearly unstable for* F_B^S. *Then for all* γ *small enough and all* $g \in \mathcal{B}^1_\gamma$ *there is an open neighborhood* U_γ *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 ρ^* and the rest points ρ^g. ρ^g being unstable hyperbolic implies that there exists an unstable manifold that ρ^g lies on, which acts as a repeller to the differential inclusion F_g. I go on to show that due to the instability of ρ^g and nonvanishing variance of M_{t+1}, no matter how close the algorithm updates come to ρ^g, and no matter how large t is, there is always a high probability that ρ_t lands on the unstable manifold and therefore must move away from ρ^g. Finally I show the existence of a neighborhood U_γ. I show that due to the hyperbolicity of ρ^*, ρ^g, there is a neighborhood U around ρ^g with $\rho^* \in U$ such that ρ^g is the only internally chain transitive set within U. Recall that ρ^* is not internally chain transitive for the perturbed system F_g, and the result follows.

Corollary 1 and Proposition 2 show the full potential of our characterisation. Asymptotically stable equilibria are equilibria that can be limiting points of the RL learning procedure, while unstable equilibria are not. The intuition is related to how RL learn to play: since such agents make errors due to estimation and also to explore their action space, opponent's strategy profiles are constantly perturbed. In other words, out of the view of a fixed agent i, the other agents are frequently deviating to policies nearby in the policy space. Now suppose the current profile ρ_t is close to an equilibrium ρ^*. Since i's updating rule tracks F_B^S, their policy will only stay close to ρ^* if the dynamics of F_B^S are somehow robust to deviations. This robustness is implied by asymptotic stability, and broken by unstable equilibria.
There is a caveat here however: Corollary 1 does not state that all limiting points in $L_{S,g}$ will be equilibria of the game. Depending on details of the environment, one may or may

not be able to rule out the case where algorithm updates get trapped in a cycle, or other more complex behavior not involving rest points (see Papadimitriou and Piliouras (2018)). I do not include cycles in the above definition, however it is straightforward to extend Proposition 1 to the case of attracting cycles as in Faure and Roth (2010), and there exist results considering linearly unstable cycles (Benaım and Faure (2012)) that suggest one may extend Proposition 2 to such linearly unstable cycles also.[18]
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. These are testable conditions on the algorithms and underlying market conditions, respectively. Thus, these represent viable avenues of further empirical research.
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 $Y \in \{P_L, P_H\}$.

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

- 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:

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

- 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 $q_i \in I$:

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

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

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

for all s, q_i, i.e. transition probabilities only depend on aggregate quantities. I will therefore sometimes write $P_{ss'}(q_1, q_2) = P_{ss'}(Q)$ with $Q = q_1 + q_2$.
Throughout, let $\rho^i : S \mapsto I$ be each players policy, and recalling the definition of W^i in (4), note that in the binary case one can derive

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

where

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

Thus, W^i is a convex combination of stage-game payoffs u^i over the two states, which weights being a function of transition probabilities. Notably, as $\delta \to 1$, these weights will converge to the unique stationary distribution over states given the policy profile ρ.[19]
In what follows I will eventually focus on symmetric equilibria, and therefore drop the i- superscript for all objects, fixing our attention on player 1's payoffs. Suppose ρ^{*1} is an interior best response to ρ^2 for which local optimality conditions hold with a negative definite Hessian. One can then use the implicit function theorem to find the derivative of 1's best response with respect to ρ^2, which will be an essential building block in finding stability conditions of an equilibrium. Since policies are vectors in I^2, from now on I will use the conventions $\rho^i(s) = \rho^i_s \in I$ for all i, s.

$$J(\rho^{*1}, \rho^2) = \begin{bmatrix} \frac{\partial \rho^{*1}_A}{\partial \rho^2_A} & \frac{\partial \rho^{*1}_A}{\partial \rho^2_B} \\ \frac{\partial \rho^{*1}_B}{\partial \rho^2_A} & \frac{\partial \rho^{*1}_B}{\partial \rho^2_B} \end{bmatrix}. \tag{7}$$

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

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

[19][Uniqueness is implied by our irreducibility assumption 1.]

More explicitly, one can then write

$$\begin{aligned}
\frac{\partial \rho_A^{*1}}{\partial \rho_A^2} &= -1 + \frac{\omega^{-1}\delta P'_{AB}(u_2^A - u_1^A) + u_{11}^A - u_{12}^A}{\omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A}, \\
\frac{\partial \rho_A^{*1}}{\partial \rho_B^2} &= \frac{\omega^{-1}\delta P'_{AB}(u_1^B - u_2^B)}{\omega^{-1}\delta P''_{AB}(u^B - u^A) + u_{11}^A}, \\
\frac{\partial \rho_B^{*1}}{\partial \rho_A^2} &= \frac{\omega^{-1}\delta P'_{BB}(u_2^A - u_1^A)}{\omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B}, \\
\frac{\partial \rho_B^{*1}}{\partial \rho_B^2} &= -1 + \frac{\omega^{-1}\delta P'_{BB}(u_1^B - u_2^B) + u_{11}^B - u_{12}^B}{\omega^{-1}\delta P''_{BB}(u^B - u^A) + u_{11}^B}.
\end{aligned} \tag{8}$$

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) $P'(2q) < 0$ *for all* $q < M$.
(iv) There exists $K \in (0, M)$ *with* $u_1(0, 2K) < 0$ *and such that*

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

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 "$P'(Q) + qP''(Q) \leq 0$" 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 ζ_N be the DS-policy that plays q_N in every state. Then ζ_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 ζ_N to determine the stability of that profile. Taking (8), 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 ζ_N dynamically unstable. Thus, for any SD-policy with large enough $P'_{AB}(Q_N)$, the static equilibrium will be dynamically unstable no matter if it is statically stable. □

Proposition 3[20] 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

[20]Proposition 3 holds under more general state-and price spaces, as established in Appendix B.2 .

incentives. As discussed in the proof above, stability of ζ_N comes down to considering the eigenvalues of the matrix of best-response derivatives at ζ_N, which I re-state here:

$$J(\zeta_N, \zeta_N) = \begin{bmatrix} -\frac{u_{12}^N}{u_{11}^N} + D_N & -D_N \\ -D_N & -\frac{u_{12}^N}{u_{11}^N} + D_N \end{bmatrix}, \tag{9}$$

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 ζ_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, ζ_N is dynamically unstable if

$$-\frac{u_{12}^N}{u_{11}^N} + 2D_N > 1. \tag{10}$$

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. Recall that by definition of regularity, q_N must be statically stable. The second term in (10) 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 ζ_N, and call player 2's DS-policies β. 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-subsitute-nature of the stage game, player 1's static incentive is to increase their quantity, by an amount less than ε. This is also what enforces the static stability of q_N. In (10), 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 (10), which I call the *dynamic incentive*. It features the derivative in the likelihood of staying in A, $-\frac{P'_A B}{\omega} < 0$, 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 (10), 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 SD-policies as allowing for more coordinative ability to the RL when it comes to learning a stage-game Nash equilibrium[21] than 1R-policies as defined below[22].

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* $P \in \mathbf{P}$*.*

Proposition 4. *Let* ρ_N *be the 1R-policy that plays stage-game Nash quantity* q_N *in every state. Then* ρ_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. □

This result is in stark contrast to Proposition 5, as discussed using (10). 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. Note that for DS-policies, one can write

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

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

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

[21] Here coordinative ability refers to the fact that RL can learn to coordinate to move *away* from the Nash equilibrium.
[22] 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.

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 ζ_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 $v_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. As further explained below, this fact implies that any perturbation around ρ_N will collapse to a perturbation along v_0. The following thought experiment explains why this is important: Suppose at ρ_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 ρ_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 $J(\cdot,\cdot)$.[23]
First, let 2-dimensional matrix $\mathbf{M}$ be diagonalizeable, and consider the problem of finding solutions to linear ODE

$$\dot{x} = \mathbf{M}x. \tag{13}$$

Let v_1, v_2 and λ_1, λ_2 be eigenvectors and eigenvalues of $\mathbf{M}$. Then solutions to (13) can be written as linear combinations

$$x(t) = c_1 e^{\lambda_1 t} v_1 + c_2 e^{\lambda_2 t} v_2, \tag{14}$$

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 ζ_N, which 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 $v_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Solutions to linear 2-dimensional ODEs based on a defective matrix take a different form from (14). 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

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

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.

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 RL?"
Given a subset of regular u, I will now 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

[23]For a more detailed discussion of stability analysis under F_B^S and related issues see Appendix B.

there exists a symmetric, collusive equilibrium σ 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 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:

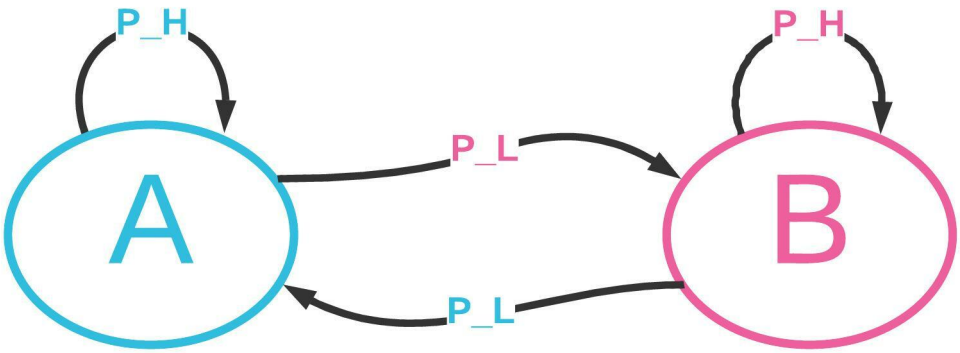

FIGURE 1. State Transition Diagram

In words, the 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

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

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

Since the goal of this exposition is analytical tractability[24], I will introduce a global shape restriction on h that 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*

[24]Tractability goes a long way when it comes to proving statements about eigenvalues of matrices that are of dimension ≥ 4, which is the smallest system one can face when considering repeated game strategy profiles for ≥ 2 players.

functions with lower bounded second derivative as

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

Let ζ_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, ζ_N is dynamically unstable and there exists a symmetric equilibrium σ 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 σ', 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

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

where $\underline{h} \in [0, 1)$ is strictly increasing, with $\underline{h}'(0) = 0$, $\underline{h}'(x_1) = h'_A$, $\underline{h}'(x_4) = h'_B$ and $\underline{h}''(Q) > 0$ for all $Q \in [0, x_1]$, and $\overline{h} \in [0, 1)$ is strictly increasing, with $\overline{h}'(M) = 0$, $\overline{h}''(x_4) = 0$ and $\overline{h}''(Q) < 0$ for all $Q \in (x_4, M)$. Elements of $\hat{\mathcal{G}}$ are therefore piecewise-linear versions of elements of $\mathcal{G}$. The idea is that this construction facilitates a numerical example, while still allowing for existence of a collusive equilibrium using similar intuition as in Proposition 5.

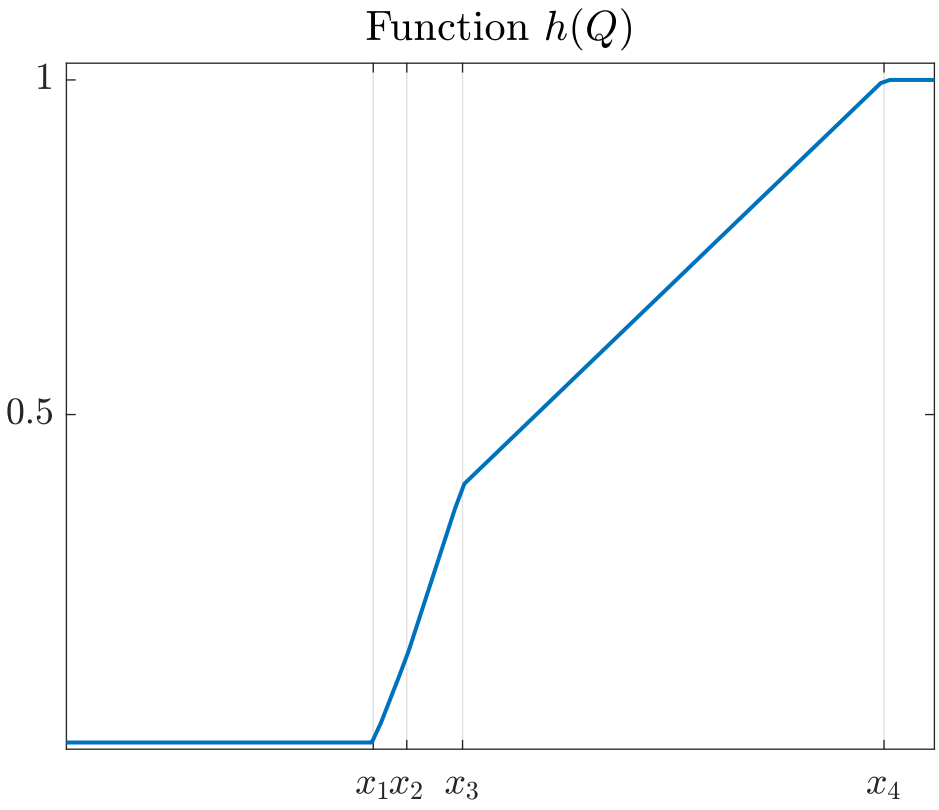

FIGURE 2. The plot indicates the location of the cutoffs x_i for piecewise linear $h(Q)$.

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:

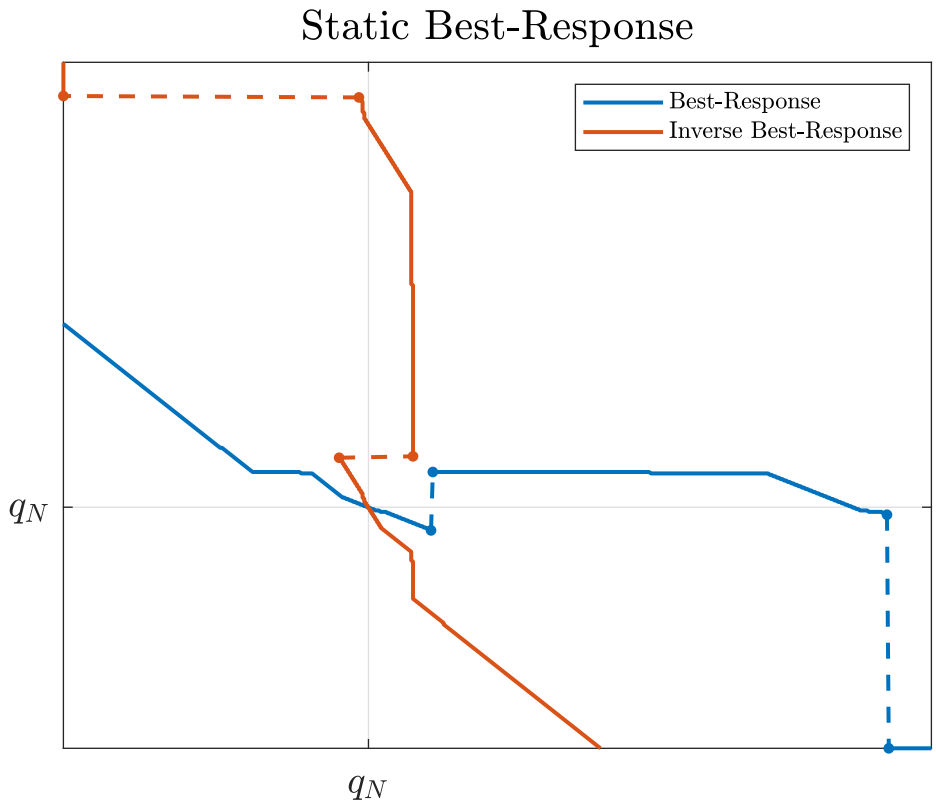

FIGURE 3. The plot shows the unique intersection of static best-response functions in the numerical example, where $q_N = 1.58$. Best responses jump upwards at a value larger than q_N. This is due to the non-concavity introduced by the requirement $h'_B < h'_N$, which enforces the S-shape of the piecewise-linear $h(Q)$ 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 (10) that the Nash equilibirum is statically stable but indeed dynamically unstable. At the same time, this numerical example supports a pair of symmetric, collusive equilibria σ, σ' with $\sigma_A = 1.52 < \sigma_B = 1.94$ (rounded to two decimal points) and the quantities flipped for σ'. (see the discussion after Proposition 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,

$$\rho^i_{t+1}(s) \in \rho^i_t(s) + \alpha_t \Big[\arg\max_{q' \in I} Q^{i*}(s, q', \rho^{-i}_t) - \rho^i_t(s) + M^i_{t+1} \Big], \tag{15}$$

where $\alpha_t = t^{-0.6}$ satisfies the Robbins-Monro Assumption 4, and $M^i_{t+1} \sim N(0, .25)$ is an i.i.d mean-zero Normal noise variable with variance 0.25. Notice that (15) 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 ρ_t to a full simulation of (5) when t is large. Running a more in-depth simulation experiment including the estimation part of Q_t will be an insightful object of further investigation.

Note that the long-run characterisations given in Section 4 are local in nature: if the iteration ρ_t at some point t enters a basin of attraction for a given stable equilibrium ρ^*, then the iteration will converge to that equilibrium with large probability.[25] 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

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

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 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 ρ_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 $\rho(A), \rho(B)$.

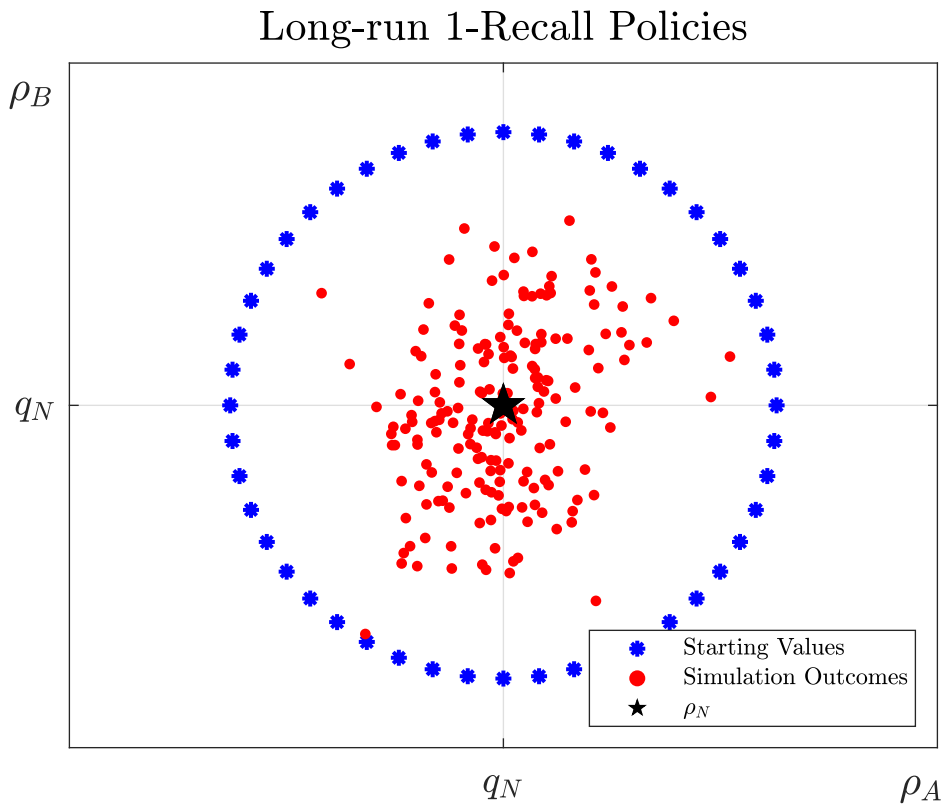

FIGURE 4. The final policy profiles of 96 simulation runs of 25000 iterations each are shown. Simulations are started in a circular neighborhood of ρ_N, with a radius of $.025\|\rho_N\|$. All simulations remained in the neighborhood, as should be expected given the stability of ρ_N.

Now contrast this result with an analogous study given DS-policies 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 ζ_N. Since q_N is dynamically unstable under S^*-policies, no matter how close the starting values of the iteration are, the iteration must be pushed away from ζ_N as shown in the proof of Proposition 2. However, in the case of this example, it is not only true that the iteration is pushed away, but also that it is pulled towards the collusive equilibrium σ. This indicates that the basin of attraction for the collusive equilibrium in this example is not confined to a small neighborhood of the equilibrium but in fact quite large. This scenario also underlines the weight of consideration that should be given to the analysis of policy spaces given two competing algorithms. Even if one forced algorithms to initialize very close to a Cournot equilibrium, they can, given the right 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.

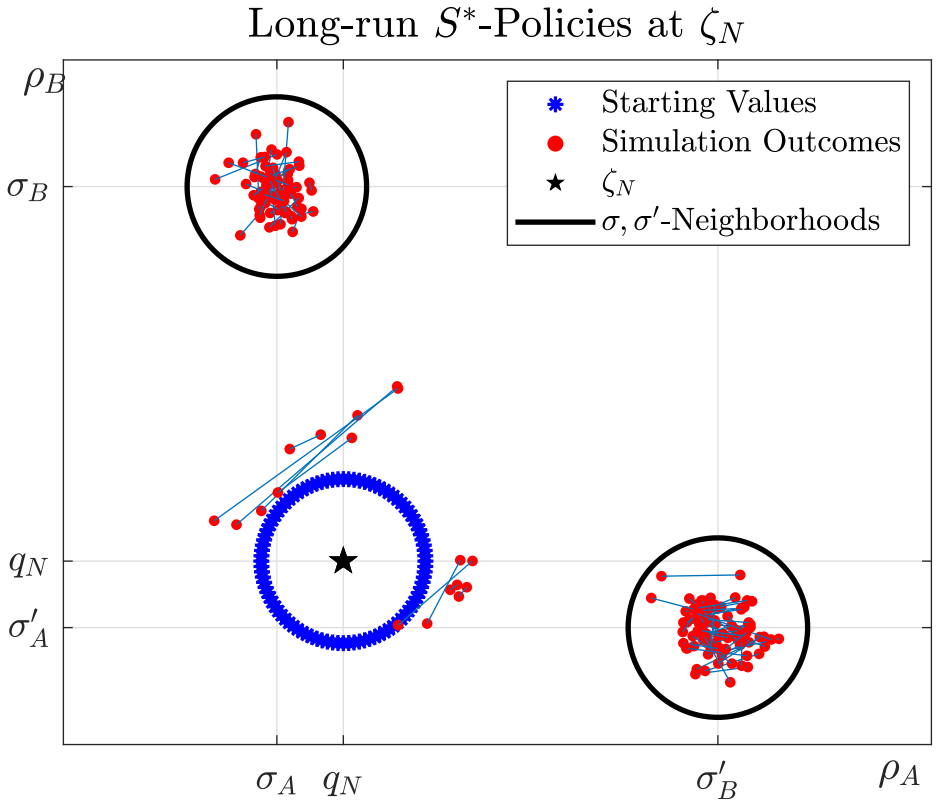

FIGURE 5. The final policy profiles of 96 simulation runs of 25000 iterations each are shown. Simulations are started in a circular neighborhood of ζ_N, with a radius of $.025\|\zeta_N\|$. All simulations left the neighborhood of starting values, with most conglomerating at one of the two collusive equilibria σ, σ'. To see that outcomes indeed approached σ, σ', two red dots connected by a line represent a single simulation outcome. The black circles represent neighborhoods of σ, σ' with radius $.25\|\sigma\|$.

With a similar exercise it can be seen that the collusive equilibria indeed attract the algorithm iterations if starting values are analgously defined as for the two above discussed simulations:
Figure 5.1 shows how after starting in a neighborhood of the collusive equilibrium σ, iterations stayed there for the course of the simulation. An analogous picture can be generated when initializing in a neighborhood of σ'.

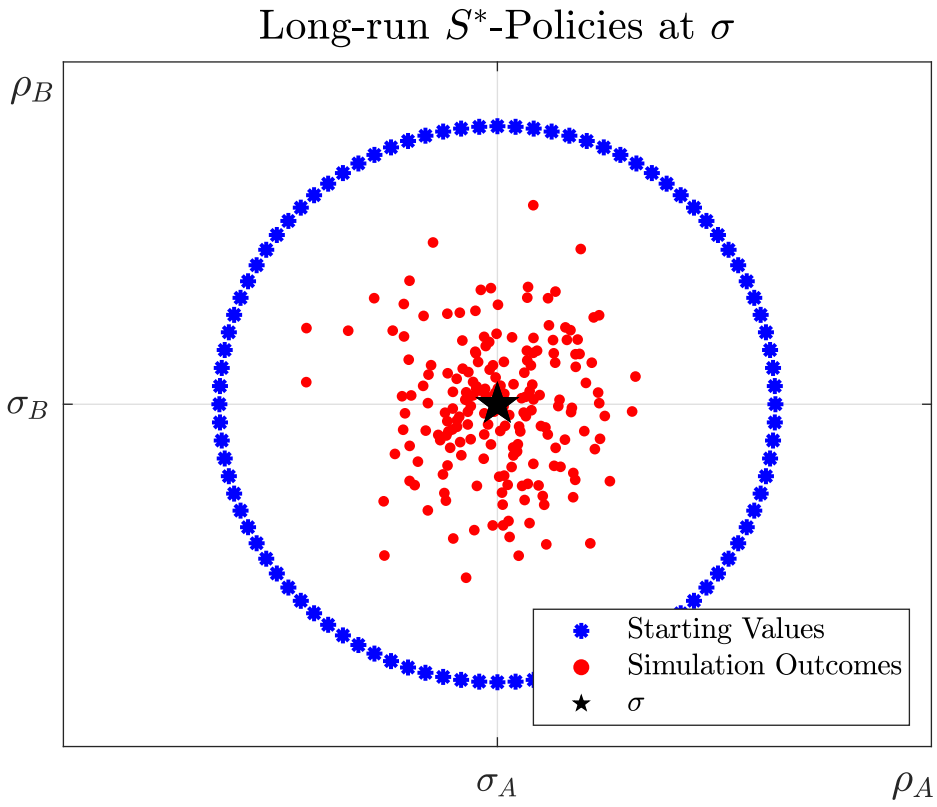

FIGURE 6. The final policy profiles of 96 simulation runs of 25000 iterations each are shown. Simulations are started in a circular neighborhood of σ, with a radius of $.025\|\sigma\|$. All simulations remained in the neighborhood, as should be expected given the stability of σ.

6. Conclusion

This paper considers the long-run behavior of a class of RL algorithms and shows that one can interpret said long-run behavior by considering the stability of repeated game equilibria according to an underlying differential equation. By ways of the application of collusion in repeated games, I observe the usefulness of this framework: it allows one to consider comparative statics exercises on the long-run learning behavior of RL with respect to details of the game and algorithms.

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 related 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.

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*.

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

Benaım, 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.

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.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, et al. (2019). "Algorithmic pricing what implications for competition policy?" In: *Review of industrial organization* 55.1, pp. 155–171.

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.

Gronauer, Sven and Klaus Diepold (2022). "Multi-agent deep reinforcement learning: a survey". In: *Artificial Intelligence Review* 55.2, pp. 895–943.

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.

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

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.

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.

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

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

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

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

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

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:*

$$\rho^i_{t+1} \in \rho^i_t + \alpha_t \left[F^i(\rho_t) + B^i_t \right],$$

where $\alpha_t > 0$ is a decreasing stepsize sequence, $F(\rho_t)$ is a (possibly multi-valued) mapping, and B^i_t represents a (possibly multi-valued) error term.
I stack the above iteration over $i \in I$ to get to the representation I study:

$$\rho_{t+1} \in \rho_t + \alpha_t \left[F(\rho_t) + B_t \right]. \tag{16}$$

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 one can allow B_t^i to behave. 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 (16) 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 (16), we can see that $F(\rho_t)$ features importantly as a mapping that provides the reinforcement of the iteration profile ρ_t. In many scenarios, $F(\rho)$ represents a performance measure 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 measures $F(\rho)$, and what kinds of approximation methods we can consider:
First the class of performance measures to be approximated:

Definition 11 (Candidate Performance Measures)**.** *We define the set $\mathcal{M}^1$ of (possibly multivalued) maps G with compact domain $X \subset \mathbb{R}^k$ and range $\mathcal{P}[R]$ for compact set $R \subset \mathbb{R}^B$ 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,* $\mathcal{U}_G = \bigcup C$*, such that* $G(x)$ *is single-valued and* C^1 *for* $x \in \mathcal{U}_G$*.*

Remark 2. *We allow for multi-valuedness to be able to handle to common learning scheme of actor-critic Q-learning, which maintains estimates of the argmax of a value function as introduced in Section 3. Note however that* $\mathcal{C}^1 \subset \mathcal{M}^1$*.*

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 measure* $F \in \mathcal{M}^1$ *if there is an increasing sequence of* σ*-fields* $\mathcal{F}_t$ *generated by datasets* $D_t \in Y$*, a bias function* $g \in \mathcal{B}^1_\gamma$ *and an integer* $N > 0$ *such that we can write for all* $n \geq N$*:*

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

$$\sup_{z \in \mathcal{A}_g[F, D_t](\rho)} d\big(z, F(\rho)\big) < \gamma + \delta(\rho, D_t),$$

where $\delta(\rho, D_t) \geq 0$ *is such that* $\sup_{\rho \in X} \delta(\rho, D_t) \to_p 0$ *as* $n \to \infty$*. "*$\to_p$*" denotes convergence in probability.*

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

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

with $g \in \mathcal{B}^1_\gamma$*, and* $R(\rho, D_t)$ *is a (possibly singleton) set such that*

$$\sup_{\rho \in X} \sup_{\delta_t(\rho) \in R(\rho, D_t)} \|\delta_t(\rho)\| \to_p 0,$$

as $n \to \infty$*.*

One can interpret $g(\rho)$ *as representing the bias part of the function approximation, and* $\delta(\rho, D_t)$ *as a random variable such that* $\mathbb{E}[\|\delta(\rho, D_t)\|^2 \,|\, \mathcal{F}_t]$ *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 measure 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 $g \in \mathcal{B}^1_\gamma$.

(2) Let $D_t \in Y$ be a sequence of datasets.

(3)

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

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

(4) Stacked version of B^i_t:

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

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

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

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

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

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

where Ω_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}, ε_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:

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \tag{17}$$

where the expectation is made over the state transitions.
Then we can define

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

as the gradient of agent i's long term payoff evaluated at $[\rho^i, \rho^{-i}]$. By abuse of notation, write $\bigtriangledown W(\rho)$ 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 $\rho \in \bar{A}$

$$F_D^S(\rho) = \bigtriangledown W(\rho), \tag{18}$$

as the state dependent gradient vector field. Take an iteration ρ_t and its respective function estimation target F as denoted in (16). 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 ρ_0^i that would allow for the process ρ_t to converge to an unstable restpoint ρ^*. By definition of instability, the paths that could attract ρ_t to ρ^*, if they exist, must be

a lower-dimensional subspace of profile space $\bar{A}$. The statement we can then make is that the probability that a randomly chosen starting profile ρ_0 will converge to unstable ρ^* must equal 0, since any lower dimensional subspace of 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 $\sigma \in A^4$:

$$\dot{\sigma}_t = F_S(\sigma_t) = \begin{bmatrix} b_1(\sigma_{t,3}, \sigma_{t,4}) \\ b_2(\sigma_{t,1}, \sigma_{t,2}) \end{bmatrix} - \sigma_t. \tag{19}$$

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

$$X(\sigma^*) = \begin{bmatrix} -I_{2x2} & J_1(\sigma^*) \\ J_2(\sigma^*) & -I_{2x2}, \end{bmatrix} \tag{20}$$

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 $X(\sigma^*)$:

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

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

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

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

$$b_1(b_2(\sigma_1, \sigma_2)) - (\sigma_1, \sigma_2)^T \tag{21}$$

with respect to σ. We can then interpret stability graphically as a scenario in which (21) 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 σ^*, we have $b_1(\sigma^*) = b_{2(}\sigma^*)$. Further, since we have symmetric payoff functions, we have $J_1 = J_2 = J$ as the matrix of derivatives of the best reply function. We can then apply Remark 5 to our system and arrive at the conclusion. Firstly, given square matrix A, define Λ as the set of eigenvalues of the A. Then define

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

as the spectral radius of A.

Lemma 2. *Suppose $\alpha^* = \beta^* = \sigma^*$ is an interior, symmetric equilibrium. Let $\bar{\kappa}$ be the real part of the spectral radius of M.*
Then σ^* is asymptotically stable if $\bar{\kappa} < 1$, and unstable if $\bar{\kappa} > 1$.

Proof. Using Remark 5, we get that

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

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

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 (8) we can write

$$det(J) = -tr(J) - 1 + \frac{(u_{11}^A - u_{12}^A)(u_{11}^B - u_{12}^B)}{\phi_A \phi_B} + J_{12}\frac{P'_{BB}}{P'_{AB}}\frac{(u_{11}^A - u_{12}^A)}{\phi_B} + J_{21}\frac{P'_{AB}}{P'_{BB}}\frac{(u_{11}^B - u_{12}^B)}{\phi_A}, \tag{22}$$

where $tr(J)$ 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

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

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

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

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

$$|tr(J)| - det(J) < 1.$$

Using (22), we get

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

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 σ^* 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

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

Finally, note from (8) that for any such equilibrium, $J_{12}<0$ and $J_{21}<0$. Thus, if $tr(J)\leq 0$ (i.e. the condition of this Proposition holds), we have that

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

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 $A\in\mathbb{R}^{nxn}$ 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 λ 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. □

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 $S=\{s_1,...,s_K\}$, 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 $s_i\in S$:

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

the discounted expected payoff from playing ρ if the opponent plays γ.

Fixing the profile ρ,γ, we can also use the vector form

$$W = \Big[W(\rho,\gamma,s_1),...,W(\rho,\gamma,s_K)\Big]^T,$$
$$U = \Big[u(\rho(s_1),\gamma(s_1)),...,u(\rho(s_K),\gamma(s_K))\Big]^T,$$

to write

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

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

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 $i \leq K < j$:*

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

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

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

$$
\begin{aligned}
\tilde{W}_j &= [I_K - \delta T]^{-1}(1-\delta)\Big[\frac{\partial U}{\partial \gamma_{j-K}} - \frac{\partial U}{\partial \rho_{j-K}}\Big] \\
\tilde{W}_{ii} &= [I_K - \delta T]^{-1}\delta\frac{\partial^2 T}{(\partial \rho^i)^2}W + [I_K - \delta T]^{-1}(1-\delta)\frac{\partial^2 U}{(\partial \rho^i)^2} \\
\tilde{W}_{ij} &= [I_K - \delta T]^{-1}\delta\frac{\partial^2 T}{\partial \rho^i \partial \gamma_{j-K}}W + [I_K - \delta T]^{-1}(1-\delta)\frac{\partial^2 U}{\partial \rho^i \partial \gamma_{j-K}} \\
&\quad + [I_K - \delta T]^{-1}\delta\frac{\partial T}{\partial \rho^i}\tilde{W}_j.
\end{aligned}
\tag{23}
$$

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*

$$B = \Big[b_1\iota_n, ..., b_n\iota_n\Big],$$

where ι_n is a vector of n ones, so that B is a matrix of columns of constants $b_1, ..., b_n$. 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 $b^T = (b_1, ..., b_n)$, with superscript T denoting the transpose. By the matrix determinant Lemma, for any $\lambda \in \mathbb{R}$ we can write

$$det\Big(A - \lambda I_n\Big) = det\Big((a-\lambda)I_n\Big) + v^T adj\Big((a-\lambda)I_n\Big)\iota_n,$$

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

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

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, $eig(M) = \{\lambda_1, ..., \lambda_K\}$, is s.t.

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

where

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

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

$$det[M - \lambda I_n] = det[M_4 - \lambda I_k] det\left[M_1 - \lambda I_m - M_2(M_4 - \lambda I_k)^{-1} M_3\right]. \tag{24}$$

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

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

Then we can apply the Sherman-Morrison formula to get

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

So we get

$$M_2(M_4-\lambda I_k)^{-1}M_3 = (d_1-\lambda)^{-1}\Big[\iota_m b^T\iota_k c^T - M_2\iota_k d_2^T M_3\big(d_1-\lambda+d_2^T\iota_k\big)^{-1}\Big]$$
$$= \frac{b^T\iota_k}{d_1-\lambda+d_2^T\iota_k}\iota_m c^T.$$

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

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

Plugging these results into (24), we get

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

The result then follows from finding all roots λ of the above.

□

In our enlarged Cournot game Γ_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 Γ_E and let ρ_N be the 1R-policy that plays stage-game Nash quantity q_N in every state. Then ρ_N is asymptotically stable if and only if q_N is.*

Proof. To determine the stability of ρ_N, we need to compute the eigenvalues of the linearized P-state-dependent best-response dynamics $(F_B^{\mathbf{P}})$ at ρ_N. As shown in Lemma 2, we can equivalently consider the eigenvalues of the gradient of the best response in $\mathbf{P}$-statespace at ρ_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 (23) can then be used to find this gradient of the best response.
First, write $W_{ij} = \tilde{W}_{ij}(s_1)$ for all $i \leq K$, $\jmath \leq 2K$. Then we can write the Hessian as $H = diag\big(W_{ii}\big)$, 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

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

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

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

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 ρ_N, we can make multiple observations that will greatly simplify the structure of M:

Firstly, long term payoffs $= \mathbf{U^N}$, since ρ_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 ρ_N, we can simplify the elements of H, R to

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

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 ρ_N is constant for all states, we can write this down in the more simple form

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

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 ρ_N is constant for all states, we can write $T = \iota_K p^T$, where $p^T = (Pr_1, ..., Pr_K)$ is a column-vector of the probability of observing each price $k = 1, .., K$ 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 $H = u_{11}^N [(1-\delta) e_1 + diag(\delta Pr_i)]$. 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 $B = [b_1, ..., b_K]$ 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. □

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 restriciton.

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, $P \in \mathbf{P} = \{P_1, ..., P_K\}$, quantity choice q in interval I. Let the cost function c be the same as in Γ^*. 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 $S_E = \{s_1, ..., s_L\}$ 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 $T : S_E \times \mathbf{P} \mapsto S_E$:

$$T(s,P) \in \begin{cases} S_E^A & \text{if } s \in S_E^A \wedge P \in P_H^A \\ S_E^A & \text{if } s \in S_E^B \wedge P \in P_L^A \\ S_E^B & \text{if } s \in S_E^A \wedge P \in P_L^B \\ S_E^B & \text{if } s \in S_E^B \wedge P \in P_H^B \end{cases}, \tag{25}$$

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 ρ^* 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 $s \in S_E^G$ 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 $s \in S_E^G$.

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 Γ^* and Γ^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 σ^* 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 ρ^*. Then if $F(Q, s), H$ and P, P_E are matching, ρ^* 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 ρ^* 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 σ^* is an S^*-equilibrium.

(3) σ^* is asymptotically stable if ρ^* 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 ρ^* is asymptotically stable if σ^* is asymptotically stable.

Corollary 3. *Take a pair Γ^*, Γ_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 Γ^* if and only if it is a static Nash equilibrium of Γ_E.

(2) q_N is statically asymptotically stable for Γ^* if and only if it is statically asymptotically stable for Γ_E.

(3) Let σ_N, ρ_N be the S^*, S_E strategies respectively playing q_N in all states. σ_N is asymptotically stable if and only if ρ_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). □

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 σ, σ' and S_E-policies ρ, γ, recursively define

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

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

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

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

Lemma 5. *Let σ^*, ρ^* be policies as in the statement of Proposition 8. Then*

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

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

$$W(\rho^*,\rho^*,s_i) - W(\rho^*,\rho^*,s_j)$$
$$= \delta\sum_{k=1}^{K}\Big(Pr\big[s_k \,|\, \rho^*(s_i)+\rho^*(s_i),\ s_i\big] - Pr\big[s_k \,|\, \rho^*(s_j)+\rho^*(s_j),\ s_j\big]\Big)W(\rho^*,\rho^*,s_k) = 0,$$

since by construction $\rho^*(s_i) = \rho^*(s_j)$ if $i,j \le \bar{l}$, 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

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

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. □

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

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

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

$$\begin{aligned}
W(\rho^*, \rho^*, s_i) \geq & (1-\delta)u\big(q, \rho^*(s_i)\big) && (26)\\
& + \delta Pr\big[A \,|\, q + \sigma^*(A),\ A\big] W(\rho^*, \rho^*, s_1) && (27)\\
& + Pr\big[B \,|\, q + \sigma^*(A),\ A\big] W(\rho^*, \rho^*, s_K), && (28)\\
\iff \tilde{W}(\sigma^*, \sigma^*, A) \geq & (1-\delta)u\big(q, \sigma^*(A)\big) && (29)\\
& + \delta Pr\big[A \,|\, q + \sigma^*(A),\ A\big] \tilde{W}(\sigma^*, \sigma^*, A) && (30)\\
& + Pr\big[B \,|\, q + \sigma^*(A),\ A\big] \tilde{W}(\sigma^*, \sigma^*, B), && (31)
\end{aligned}$$

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

For the other direction, suppose ρ^* is an S_E-equilibrium, and take σ^* 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 (26) and the claim follows: σ^* 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 $Pr\big[s' \,|\, Q,\, s\big] > 0$ for all Q, s, s'. As in the proof of Proposition 4, we can use (23) 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 $H = diag(\{W_{ii}\}_{i \leq K})$ is the Hessian.
We now show that we can apply Lemma 4 to M. Define

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

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

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

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

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

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

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 safe 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 $\big[I_K - \delta T\big]^{-1}$ by using the Sherman-Morrison formula twice.

First, let

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

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,

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

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

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

Re-writing things multiple times allows us to get to

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

where

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

where

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

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

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

where we safe notation to write $Pr_l^{s\prime} = \frac{\partial Pr[s_j|q+q_s]}{\partial q}\Big|_{q=q_s}$ for $s \in \{A, B\}$. Using our information on transitions for state space S_E (25), we can further re-write this as

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

Using the transition information, we can also write

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

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

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

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

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

Now following the notation in Lemma 4 , let

$$a_1 = -\Big[\delta h''(2q_A)\big(V_B - V_A\big) + (1-\delta)u_{11}^A\Big]^{-1}\Big[\delta h''(2q_A)\big(V_B - V_A\big) + (1-\delta)u_{12}^A\Big],$$
$$d_1 = -\Big[\delta h''(2q_B)\big(V_A - V_B\big) + (1-\delta)u_{11}^B\Big]^{-1}\Big[\delta h''(2q_B)\big(V_A - V_B\big) + (1-\delta)u_{12}^B\Big].$$

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 $s \in \{A, B\}$.

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

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

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

In the notation of Lemma 4, we have

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

Finally notice that

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

Recall $J_1(\sigma^*)$, the matrix of best response-derivatives for the S^*-collusive equilibrium σ^*. 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

$$p = tr\big(J_1(\sigma^*)\big), \quad q = det\big(J_1(\sigma^*)\big),$$

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

Thus, if σ^* is unstable, ρ^* 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 $d_1 < 1$:

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

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 ρ^*.

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 $P_H > 0 > \big(P_L - P_H\big)$ such that

$$g(0) < \frac{P_H}{-\big(P_L - P_H\big)} < h(2D). \tag{32}$$

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

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

and therefore the above implies that there exists a $c(q)$ such that $u_1(0, D) < 0 < u_1(0,0)$ (pinned down only by $c'(0)$). Then since $u_1(0, \hat{q})$ strictly decreases in $\hat{q} \in [0, M]$, there exists $M^* \in (0, 2D)$ such that $u_1(0, M^*) = 0$. This corresponds to K^* in Lemma 1. Finally to check whether Definition 6 (iv), note that

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

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

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

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

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

$$
\begin{aligned}
W_1 &= \omega^{-1}(1-\delta P_{BB})\Big[\omega^{-1}\delta P'_{AB}(u^B-u^A)+u_1^A\Big],\\
W_2 &= \omega^{-1}(\delta P_{GB})\Big[\omega^{-1}\delta P'_{BB}(u^B-u^A)+u_1^B\Big],\\
W_{11} &= -2\omega^{-1}\delta P'_{AB}W_1+\omega^{-1}(1-\delta P_{BB})\Big[\omega^{-1}\delta P''_{AB}(u^B-u^A)+u_{11}^A\Big],\\
W_{22} &= 2\omega^{-1}\delta P'_{BB}W_2+\omega^{-1}(\delta P_{GB})\Big[\omega^{-1}\delta P''_{BB}(u^B-u^A)+u_{11}^B\Big],\\
W_{12} &= \omega^{-1}\delta\Big[P'_{AB}\frac{1-\delta P_{BB}}{\delta P_{GB}}W_2-P'_{BB}\frac{\delta P_{GB}}{1-\delta P_{BB}}W_1\Big],\\
W_{13} &= W_{11}+\omega^{-1}(1-\delta P_{BB})\Big[\omega^{-1}\delta P'_{AB}(u_1^A-u_2^A)+u_{12}^A-u_{11}^A\Big],\\
W_{24} &= W_{22}+\omega^{-1}(\delta P_{GB})\Big[\omega^{-1}\delta P'_{BB}(u_2^B-u_1^B)+u_{12}^B-u_{11}^B\Big],\\
W_{14} &= -\omega^{-1}\delta P'_{BB}\frac{\delta P_{GB}}{1-\delta P_{BB}}W_1+\omega^{-1}(1-\delta P_{BB})\omega^{-1}\delta P'_{AB}\Big[\omega^{-1}\delta P'_{BB}(u^B-u^A)+u_2^B\Big]\\
&= W_{12}+\omega^{-1}(1-\delta P_{BB})\omega^{-1}\delta P'_{AB}(u_2^B-u_1^B),\\
W_{23} &= \omega^{-1}\delta P'_{AB}\frac{1-\delta P_{BB}}{\delta P_{GB}}W_2-\omega^{-1}(\delta P_{GB})\omega^{-1}\delta P'_{BB}\Big[\omega^{-1}\delta P'_{AB}(u^B-u^A)+u_2^A\Big]\\
&= W_{12}+\omega^{-1}(\delta P_{GB})\omega^{-1}\delta P'_{BB}(u_1^A-u_2^A).
\end{aligned}
\tag{33}
$$

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

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

Notice that for all such α^*, we also have $W_{12}(\alpha^*,\beta)=0$. This follows for any optimal Markov policy by the identity in (**??**) by irreducibility. 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 τ there exist $P_H > 0 > \Delta$ and a convex $c(q)$ such that $q_N \in \mathcal{N}$. In particular, there exist $P_H > 0 > \Delta$ and a convex $c(q)$ such that $q_N = \tau$.

Proof. As argued in Lemma 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. □

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 τ.

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 τ. First let us fix $P_H, \Delta, c'(\tau)$ so that $q_N = \tau$. We also have that $h'(2q)$ is strictly increasing for $q \in [0, q_N)$, and strictly decreasing for $q \in (q_N, M]$. Finally, we have that $u_1(q_N, q_N) = 0$, so that the fraction must be strictly decreasing for $q \in [0, D]$. Note that we have so far only imposed two point conditions on convex $c(q)$, for $c'(0), c'(\tau)$. 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

$$\frac{u_1(q,\hat{q})}{h'(q+\hat{q})} = \frac{P_H + \Delta h(q+\hat{q}) - c'(q)}{h'(q+\hat{q})} + \Delta q. \tag{34}$$

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

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

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

Case 3: $\hat{q} \in (M^*, M]$.
Taking equation (34) and the above argument together implies that again, $\frac{u_1(q,\hat{q})}{h'(q+\hat{q})}$ is strictly decreasing, for any $q \in [0, M]$, since $M^* > Q_N$ holds (i.e. $h''(q+\hat{q}) \leq 0$). The claim follows.

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

Now for the proof of the Proposition:

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

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

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

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

By Lemma 8 we have that for any $q_G \in (0, q_N]$ there exists a unique $q_B \in [q_N, M)$ such that

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

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

$$z'(q_G) = -\frac{h'_B}{h'_G}\frac{u^A_{11} + u^A_{12} - 2h''_G\frac{u^G_1}{h'_G}}{u^B_{11} + u^B_{12} + 2h''_B\frac{u^B_1}{h'_B}}. \tag{35}$$

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

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

Firstly, recall that as in Lemma 7 we have that for $q > 0$ small enough, $u_1(q, q) > 0$ must hold. Now consider $\Psi(q_G)$:

$$\Psi(q_G) > 0 \;\Leftrightarrow\; \omega^{-1}\delta h'(2q_G)(u^B - u^A) + u^A_1 > 0.$$

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

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

$$\begin{aligned}
\Psi'(q_N) &= W^N_{11} + W^N_{13} + W^N_{14}z'(q_N) = W^N_{11} + W^N_{13} - W^N_{14} \\
&= \omega^{-1}(1 - \delta(1 - g_N))\Big[u^N_{11} + u^N_{12} - \omega^{-1}\delta h'_N u^N_2 + \omega^{-1}\delta h'_N u^N_2 z'(q_N)\Big] \\
&= \omega^{-1}(1 - \delta(1 - g_N))\Big[u^N_{11} + u^N_{12} - 2\omega^{-1}\delta h'_N u^N_2\Big] \\
&= \omega^{-1}(1 - \delta(1 - g_N))u^N_{11}\Big[1 + \frac{u^N_{12} - 2\omega^{-1}\delta h'_N u^N_2}{u^N_{11}}\Big].
\end{aligned}$$

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

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

Thus we can write

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

where $R_1 = 2\Delta g_N q_N h''_N$ vanishes for q_N close enough to τ, and $R_2 = \frac{1-\delta}{2\delta}(u_{12}^N + u_{11}^N)$ vanishes as $\delta \to 1$. Then for $\delta < 1$ close enough to 1 and q_N close enough to τ, the condition stated in the Proposition is sufficient for $\Psi'(q_N) > 0$.

This together with $\Psi(q) > 0$ for q small, $\Psi(q) < 0$ for q large, allows us to use the intermediate value theorem. It gives us that there exists $q_G < q_N < q_B$ such that $W_1(\sigma) = W_2(\sigma) = 0$ for $\sigma = (q_G, q_B, q_G, q_B)$.

We are left to show that this zero is a global maximizer. Firstly we note that the Hessian at σ must be negative definite: we see from (33) that $W_{12} = 0$, so the Hessian must be diagonal at σ. A sufficient condition for negative definitess then is $h''_G > 0 > h''_B$ and $u^A > u^B$. The first one follows since $q_G < q_N < q_B$, the second one follows from the first order conditions:

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

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

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

holds for all $q \in S_g$. Equivalently, we can show that $q = q_G$ 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

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

By construction, $H^G(q_G, q_G) = 0$. Since the Hessian is negative definite at q_G, q_G, $H_1^G(q_G, q_G) = \frac{\partial H^G(q,q_G)}{\partial q}\Big|_{q=q_G} < 0$. Recall that in the proof of Lemma 8 we we showed that q_G is the only solution to $H^G(q, q_G) = 0$, but also that $\frac{u_1(q,q_G)}{h'(q+q_G)}$ is strictly decreasing over $q \in [0, q^*(q_G)]$. Thus, $H^G(0, q_G) > 0$ and $H^G(M/2, q_G) < 0$ must hold and q_G is globally optimal.
Now, in state D we do the analogous argument, take derivatives to get

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

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 $\frac{u_1(q,q_G)}{h'(q+q_G)}$ 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 $\sigma = (q_G, q_D)$ is the unique best reply to an opponent playing σ, and thus σ is a symmetric equilibrium as required.
■

C.2. Proof of Proposition 1

First we prove the following result that employs known techniques from stochastic approximation theory:

Proposition 9. *With probabilitiy one, $L_{S,g}$ is an internally chain transitive (ICT) set[26] of the differential inclusion*

$$\dot{\rho} \in F_g(\rho(t)) \equiv conv[F(\rho(t))] + g(\rho(t)).$$

Proof. Write $\bar{M}_{t+1} = M_{t+1} + \varepsilon_t - g(\rho_t)$, then the algorithm (16) can be written as

$$\rho_{t+1} = \rho_t + \alpha_t \left[F_g(\rho_t) + \bar{M}_{t+1} + \delta_t\right],$$

were all assumption for the set valued convergence Theorem 3.6 in Benaım, Hofbauer, and Sorin (2005) hold. □

Since payoffs are differentiable around ρ^*, point 1 follows as long as ρ^g and ρ^* are close. For point 2, we will prove something more general: as long as ρ^* is hyperbolic, point 2 holds.
This follows because when ρ^* is hyperbolic, there is a neighborhood U around 0 such that F has a differentiable inverse on U. Next, note that ρ^g solves

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

[26]Importantly, these sets include rest points and limit cycles (if they exist). We refer to Benaım, Hofbauer, and Sorin (2005) Definition 6 for a definition, and Papadimitriou and Piliouras (2018) for an intuitive discussion.

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

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

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

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

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

Proof. 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 $\rho^g \to \rho^*$ uniformly for $g \in \mathcal{B}^1_\gamma$ as $\gamma \to 0$ (see above proof of point 2), we get that

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

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

Now that we know that all ρ^g must be asymptotically stable for γ small enough, we can apply Faure and Roth (2010) (Thm 2.8).

We only need to verify that our game satisfies their attainability condition:

Definition 13. *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 (16). 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 $z \in int(E)$ 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.) □

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 $\mathcal{U}$ is also where F, F_g are single valued. Solution curves are unique whenever they intersect $\mathcal{U}$.

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_{ρ^*}, and we can allow $F(\rho)$ to be multivalued otherwise.

First, we define invariant sets for given differential equations:

Definition 14. *Let $z(t, z_0)$ be the solution so some given differential equation $\dot{z} = f(z)$ with initial value z_0. Then a set S*

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

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

Given a $g \in \mathcal{B}^1_\gamma$, we know from Proposition 9 that only ICT sets subset of a neighborhood of ρ^g are candidates to being limiting points of the algorithm (16). The singleton $\{\rho^g\}$ is

an ICT set, and we show first that this cannot be a limiting set of the algorithm. Then we go on to show that for small engough γ, no other ICT sets can exists in a neighborhood around ρ^*, which finishes the proof.

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

Note that by Lemma 9, there are $\gamma > 0$ small enough such that all ρ^g are linearly unstable just as ρ^*. We can thus apply Benaım 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 Benaım and Faure (2012), Hypothesis 2.2 exists given our assumptions on ε_t, M_t. Benaım 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 ρ^* and ρ^g.

We will prove that there are no other invariant sets in such a neighborhood. Since ICT sets are subsets of invariant sets, this will complete the proof.
We can use Hartman-Grobman to show that there are open neighborhoods N_g, N_0 with $\rho^* \in N_0, \rho^g \in N_g$ such that ρ^*, ρ^g are isolated invariant sets in their respective neighborhoods. These neighborhoods are nontrivial for all γ small enough, which follows from both ρ^*, ρ^g being hyperbolic:
By Hartman-Grobman and hyperbolicity there exists a homeomorphism H on a neighborhood $N \subseteq U_{\rho^*}$ of ρ^* with $H(\rho^*) = \rho^*$ such that

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

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

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

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

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

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

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

and ρ^* is isolated invariant set for $\phi(t,\rho)$.
Since ρ^g are hyperbolic for γ small enough, an analogous argument gives us that ρ^g are isolated invariant also. Let N_g be the neighborhood on which the homeomorphism is defined that connects flows of F_g to flows of the linearized system $DF_g(\rho^g)$. By definition, $\rho^g \in N_g$, and we know that ρ^g is isolated invariant in N_g. We are left to show that for γ small enough, for all $g \in \mathcal{B}^1_\gamma$, $\rho^* \in N_g$:
To prove this, we will argue that each N_g contains a ball $B^g_z(\rho^g)$, for which the radius $z > 0$ can be lower bounded by a number that depends only on the eigenvalues of $DF(\rho^*)$ and γ. First we need an auxiliary Lemma to show how eigenvalues of $DF_g(\rho^g)$ vary continuously in γ. First some more notation:
For small enough γ, all ρ^g are hyperbolic when $g \in \mathcal{B}^1_\gamma$. Fix such a g. Define $\rho_l > 0$ to be the smallest positive eigenvalue of $DF_g(\rho^g)$, and $\rho_u < 0$ be the largest negative eigenvalue of $DF_g(\rho^g)$. Now let $a_g \in (0,1)$ be any number such that

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

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

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

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

Proof. Apply Lemma 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. □

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

Lemma 12. *Suppose ρ^* is hyperbolic for F. Fix a small $\underline{z} > 0$. Then there is $\bar{\gamma}$ such that for all $\gamma \leq \bar{\gamma}$, and all $g \in \mathcal{B}^1_\gamma$, there is $B^g_z(\rho^g) \subseteq N_g$ with $z \geq \underline{z}$.*

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

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

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

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

By Lemmas 4.3 and 4.4 of Palis Jr, Melo, et al. (1982), $B_{r_\varepsilon}(\rho^g) \subseteq N_g$, if $\varepsilon < \overline{\varepsilon}_g$.
We are left to show that r_ε can be made to depend only on the eigenvalues of $DF(\rho^*)$ and γ.
Notice that small enough $\underline{z} > 0$ pins down the $\delta > 0$ referred to in Lemma 11: Let

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

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

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

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.

■