Reinforcement Learning and Collusion

Clemens Possnig
University of Waterloo
September 18, 2024
Link to current version

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

JEL classification. C73, D43, D83.

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

I thank my committee members Li Hao, Vitor Farinha Luz, and Michael Peters for years of guidance and conversations. I am grateful to Alexander Frankel, Kevin Leyton-Brown, Wei Li, Vadim Marmer, Jesse Perla, Chris Ryan, and Kevin Song for many helpful discussions. I thank the participants at EC 22, GTA22, CORS/INFORMS 22, and CETC 22 for insightful comments. I also thank participants of the theory lunches at VSE for their extensive feedback and patience.

1. Introduction

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

Algorithms can help firms adapt to rapidly changing market environments, and potentially better serve their markets. However, recent empirical[1] and simulation-based[2] studies show that algorithms may learn to collude. This is a concern for consumer welfare. Moreover, legal systems are not currently adapted to deal with the kind of tacit collusion that might result from algorithmic competition. An antitrust regulator would need to understand how competing algorithms might lead to inefficient outcomes, depending on the market conditions, the nature of the competitive environment, and the details of the algorithms themselves. While the folk theorem tells us the set of possible payoffs rational players may achieve in a repeated interaction, the outcomes to algorithmic learning may not even constitute a subset of these. This paper is concerned with the analytical properties of these outcomes for a common family of reinforcement learning (RL) algorithms.

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

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

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

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

[2]Klein (2021), Calvano, Calzolari, Denicoló, et al. (2021) show that algorithms may learn to play repeated game strategies akin to typical carrot-and-stick type strategies studied in the economic theory literature.

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

Characterization of long-run behavior. The focus of this paper is actor-critic reinforcement learning. These algorithms keep track of an estimated performance criterion (the "critic", essentially a value function) and a policy function (the "actor") that is updated towards the maximizer of the performance criterion. The policy is a mapping from observables (states), such as past prices or other market data, to actions, e.g. prices or quantities. As a result, the class of algorithms studied here have the ability to learn repeated game strategies (e.g., take the observable to be a summary of the history of the interaction), in contrast to the stage game (myopic) strategies more commonly studied in the literature on learning in games.

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

Suppose that after a large enough number of iterations, the performance criterion estimator differs from the true criterion at most by a bounded, smooth bias term[5]. I then show that attractors of the underlying differential equation will be learned with positive probability, while unstable points will not be learned. In the case of ACQ, if the long run behavior of algorithms converges to a point, that point must be a Markov-perfect equilibrium (MPE) of the repeated stage game. The implication of this characterization is that it becomes necessary to understand what it means for a given MPE to be attracting.

Learning Nash. I consider the repeated Cournot game studied in Abreu, Pearce, and Stacchetti (1986)'s (henceforth APS) seminal work: There is a continuum of prices, which are realized randomly conditional on aggregate supply. Prices exhibit no time-dependence beyond dependence on actions. Instead of each other's actions, agents only observe public information (a state variable) that is correlated to actions, e.g. price realizations. In such a game, the repetition of the stage game Nash equilibrium is an MPE under any state variable. This equilibrium is useful in generating intuitions about stability for more general MPEs. I show that in order to learn the stage game Nash equilibrium, the details of the state variables that algorithms keep track of are

[3]Stepsizes signify the impact innovations have on policies in the algorithmic updating rule. They must satisfy the Robbins-Monro condition commonly invoked in the computer science literature. See Borkar (2009), Chapter 2.
[4]See Dutta and Upreti (2022), Grondman et al. (2012) for relevant surveys.
[5]The bias term is a modelling decision inspired by the fact that for many real-world performance criterion estimators, bias is unavoidable due to function approximation (Fujimoto, Hoof, and Meger (2018)). Also, often convergence proofs are lacking, in which case the fitness of an algorithm is shown by it doing better at benchmark tasks than previous algorithms. C.f. the discussion in Chapter 9 of François-Lavet et al. (2018). My results imply that the analysis of long run behavior is robust to well-behaved, non-vanishing bias.

essential. I provide the first step at a novel categorization of the coordinative ability granted to learning algorithms by their state variables. This is done by studying what kinds of state variables allow for algorithms to learn the stage game Nash equilibrium. To the best of my knowledge, this paper is the first to uncover this aspect of coordinative ability.

State variables are characterised by the space of their realizations and a transition function that pins down how the states evolve over time as a function of current state and actions taken by the algorithms. State transitions are state-independent if the transition function does not vary with the current state. This is the case in the past-periods price example: today's price realization does not depend on past period's prices after conditioning on actions.

As is commonly the case, stability properties of MPEs come down to how sensitive equilibrium constraints are to action deviations. It turns out that the conditions for stability of stage game equilibria reduce to a comparison between the slope of the (static) stage game best response and a growth rate of transition probabilities with respect to action deviations. When transition probabilities are more sensitive than a cutoff defined through the static best response slope, the stage game equilibrium will be rendered unstable, and therefore not learned.

Given state-independent transitions, I prove that stability of the stage game Nash equilibrium depends entirely on the stage game itself. In the Cournot game, stability is then determined by the well-known condition on static best response slopes. In other words, sensitivity of transition probability does not matter in this case, but only the slope of the stage game best response.[6] If the above bound on myopic best responses is satisfied for a given stage game Nash equilibrium, I call it *statically stable*. If for a given state variable the Nash equilibrium is stable under the resulting state-dependent best-response dynamics (and therefore learned with positive probability), I call it *dynamically stable*.

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

Learning to collude. Next, I apply my characterization to study collusion in the Cournot application. I provide conditions on payoffs and observables under which collusive equilibria exist that are attracting, and therefore will be learned with positive probability. I show that the payoff to the best among these collusive equilibria is lower bounded by optimal imperfect monitoring equilibria of the game restricted to finitely many actions, as analysed in APS.

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

To make the approach more concrete, I show for a class of payoff functions that there exists a simple binary state variable (falling into the class of direction-switching states) for which there exists a collusive equilibrium that is attracting. In addition, for this class of payoff functions, it is true that the unique stage game Nash equilibrium is statically stable and at the same time dynamically unstable. The result then ties in with my second contribution discussed above: when Nash is not learned, then what is learned can likely be a collusive outcome.

Finally, I provide a numerical example featuring the above properties: there is an attracting, collusive equilibrium as well as a unique stage game Nash equilibrium that is unstable. I verify in a simulation that ACQ learners will frequently converge to the collusive equilibrium while always avoiding static Nash. This continues to be true even under initializations close to the static Nash equilibrium, as suggested by the dynamic instability of that profile.

The results presented here indicate that a potentially useful regulatory approach to algorithmic competition works through observables of the algorithms, represented here as state variables. Such an approach would be common in spirit with approaches already in place to curb algorithmic bias through the control of covariates used by decision-making algorithms.[7]

Related Literature

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

Firstly, the literature on algorithmic collusion has received increasing attention in recent years. Assad et al. (2020) provide an empirical study supporting the hypothesis that algorithms may learn to play collusively, while there are many simulation studies suggesting the same, of which Calvano, Calzolari, Denicolo, et al. (2020), Calvano, Calzolari, Denicoló, et al. (2021), and Klein (2021) are important examples. A paper close in spirit to this study is Banchio and Mantegazza (2022). They consider a fluid approximation technique related to the stochastic approximation approach applied here, and recover interesting phenomena regarding the learning of cooperation for a class of RL algorithms. Meylahn and V. den Boer (2022), Loots and denBoer (2023) use ODE methods related to the ones used in this paper to prove that specific algorithms can learn to collude in a pricing game. Further important recent work in the area of algorithmic collusion includes Lamba and Zhuk (2022), Z. Y. Brown and MacKay (2021), Johnson, Rhodes, and Wildenbeest (2020), and Salcedo (2015). These papers feature stylized models of algorithmic competition, abstracting away from issues of learning and estimation, which are an important aspect of my analysis. Their relation to this work is discussed more thoroughly in Section 6.

Secondly, this paper connects to a long history of the theory of learning in games. Classically, this literature has been concerned with the ability of agents to learn a Nash equilibrium of the stage game when following a given learning rule (e.g. Milgrom and Roberts (1991), Fudenberg

[7]Lee, Resnick, Barton: "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms", May 22, 2019. Accessed June 12 2024.

and Kreps (1993)). More recent results concern learning in stochastic games (e.g. Leslie, Perkins, and Xu (2020)), where the state variable is taken as an exogenous object. The class of algorithms studied here has the ability to learn *repeated game strategies*, i.e. strategies that condition on summaries of the history of the game, implemented as automaton strategies. The games that can be studied here therefore contain stochastic games as a special case, but also allow for the case where the state that agents observe represents a finite history of the repeated interaction.

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

Thirdly, this paper makes use of an extensive body of research related to stochastic approximation theory (see for example Borkar (2009)) and hyperbolic theory (Palis Jr, Melo, et al. (1982)). There is a growing strand of the computer science literature devoted to establishing convergence proofs in multi-agent algorithmic environments. The paper in that area closest to this one is Mazumdar, Ratliff, and Sastry (2020).

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

This paper is structured as follows: In section 2 I give a brief introduction to RL, via the classical example of single-agent Markov-decision problems (MDPs). In section 3 I define the general economic environment our algorithms will play on, as well as ACQ learning, an important element of my RL class that will serve as the running example of the paper. I provide general limiting results in section 4. In section 5, I apply the results of the previous section to a repeated Cournot game, and give a numerical example with simulations in the end of the section. Section 6 concludes and discusses related literature more thoroughly.

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

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

Since this paper relies on technical methods that would overwhelm the main body, some sections are moved to the appendix. Appendix A characterizes 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 $a \in X$ repeatedly. There is a state variable S taking values in $\mathbf{S}$ so that in every $s \in \mathbf{S}$, the agent may find it best to choose different a. Given s, the agent's expected payoff from choosing a is denoted $u(a, s)$. The agent discounts the future with $\delta \in (0, 1)$, and aims to find a policy $\rho : S \mapsto X$ that maximizes future expected discounted payoffs

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

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

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

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

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

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

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

This Q- function is related to the value function by $V(s) = \max_{a \in X} Q^*(s,a)$. 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^*. This algorithm is celebrated due to its simplicity as well as minimal information requirement. One can use the algorithm without any knowledge of a payoff function and transition function, thus falling into the class of 'model-free' algorithms.

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

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

where $\beta_t \geq 0$ is a (possibly stochastic) sequence of numbers converging to zero. Importantly, notice that Q-learning does not specify a policy, just a performance criterion. Convergence results on Q_t require actions to be selected sufficiently often, but are generally agnostic about how actions a_t are sampled in every period. A common, basic sampling method is known as ε-greedy:

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

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

3. The Multi-Agent Setting

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

This section introduces actor-critic Q-learning (ACQ), a common evolution of Q-learning as introduced previously. In this section and the main body of the paper, I focus on ACQ for clarity. The general class of algorithms is broader and defined in Appendix A. A main advantage of ACQ over Q-learning is that it directly applies to continuous-action problems, which are the

focus of this paper.[8]. In that case, estimating a value function becomes a difficult task even when stationarity of the environment is satisfied. Commonly in such situations, algorithms use some form of parametric function approximation to generate an estimate, which can introduce bias. Often this involves deep neural networks due to their flexibility and scaleability. I refer to François-Lavet et al. (2018) for a thorough introduction to state-of-the-art RL techniques. This paper remains agnostic about the specificities of the value function approximation part of the algorithms. The goal is to make statements about what can be learned as long as the function approximation step is reasonably well behaved, a property to be defined in Assumption 3. While for clarity this section assumes unbiased function estimation, the results in Appendix A are more general.

There are n algorithms indexed by i, each having as action space a compact interval $\mathbf{X}_i$, with profile space $\mathbf{X} = \times_i \mathbf{X}_i$. A state variable S taking values in space $\mathbf{S}$ with $|\mathbf{S}| = L$[9] comes with a transition probability function $T : \mathbf{S}^2 \times \mathbf{X} \mapsto [0, 1]$ where I will maintain throughout the paper that each state space considered is irreducible, as specified below. Furthermore, after defining its transition probability function, I will refer to a state space S keeping implicitly in mind that it comes with its own transition probability. Each algorithm has a payoff function $u^i : \mathbf{X} \times \mathbf{S} \mapsto \mathbb{R}$, $\mathcal{C}^2$[10] in $\mathbf{X}$, and common discount factor $\delta \in (0, 1)$.

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

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

Assumption 1. *For all $\rho \in \overline{\mathbf{X}}$, the Markov chain induced by $T_{ss'}[\rho(s)]$ is irreducible and aperiodic.*[11]

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

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

[9]While it may be possible to carry out an analogous characterization 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.

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

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

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

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

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

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

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

where due to our assumption on irreducibility of the state space the optimal policy does not depend on the initial state s_0. The optimal policy is indeed optimal over all possible history-dependend policies since given a Markov stationary opponent profile ρ^{-i} there must be a Markov stationary best response.

Definition 1. *Define*

(i) $E_S \subset \overline{\mathbf{X}}$ *to be the set of Nash equilibria in policy profiles based on payoff functions* W^i*. In other words,* E_S *is the set of profiles* ρ^* *s.t.* $\rho^{*i} \in B_S^i(\rho^{*-i})$ *for all* i*.*

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

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

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

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

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

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

Now to state the running example of RL studied here. Assume that each algorithm uses ACQ to update their policy:[12]

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

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

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

where $\alpha_t > 0$ is a sequence of stepsizes converging to zero and M_{t+1}^i is an i.i.d, zero-mean, bounded variance noise generated as a means of exploring the policy space, commonly referred to as 'parameter noise exploration'[13][14]*.*

$Q_t^i(s, a)$ is an estimator[15] of

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

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

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

In what follows, whenever it is clear from context, write $Q_t^{i*} = Q^{i*}(s, a, \rho_t^{-i})$. Q_t^i is motivated from stationary MDPs as introduced in subsection 2.

Remark 1. *It is important to note that the use of this estimator in the multi-agent case faced here imposes an implicit behavioral assumption on each algorithm. Suppose that $Q_t^i(s, a) = Q_t^{i*}$, i.e. the estimator is 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_t^{i*} can be seen as an approximation to the true future expected value that would take into account evolving ρ_t^{-i}. Secondly, as stated before, this section is not concerned with a normative theory of how an optimal algorithm should behave. Rather, the interest is in developing a model that is realistic enough while staying analyzable, and forms an informational lower benchmark on algorithms in the sense that they can be allowed to be model-free.*

The following assumption ensures that maximizers of Q_t^i track the maximizers of the correct function Q_t^{i*} well when t is large enough. The classical Q- estimator (2) defined to motivate Q-learning will not be enough for this to be true, as it requires discretization of the continuous action space and may run into issues due to the underlying non-stationarity of the problem. However,

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

[14] Notice that (5) is generalized to an inclusion, since I allow for the possibility of the argmax having multiple values. If that is indeed the case, the algorithm picks arbitrarily, which will not affect the limiting characterization in ways that matter, as will be seen in section 4.

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

more involved estimation schemes exist for which Q_t^i can be shown to track Q_t^{i*}, as shown e.g. in Possnig (2022).

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

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

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

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

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

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

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

Assumption 3. *Assume for each i:*

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

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

(ii)

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

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

(iii)

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

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

In words, estimators Q_t^i converge uniformly in mean to Q_t^{i*}, with maximizers of Q_t^i converging at a related rate. Point (i) imposes a direct relationship between the uniform convergence of Q_t, and its maximizers. In the case of parametric function spaces such as $\mathcal{Q}$, this relationship commonly holds. Here in that case, χ_t^i can be written in terms of distance between estimated and 'true' parameter, and upper hemicontinuity of $A(Q)$ implies that the upper bound in (i) is

of the same order of magnitude as χ_t^i. More broadly, (i) is inspired by set-valued estimation results, e.g. conditions C.1 and C.2 in Chernozhukov, Hong, and Tamer (2007), adapted to this setting of maximization under time-dependent target. Point (ii) bounds the convergence speed in mean, and point (iii) ensures that large errors have negligible mass, which is important in the approximation results established in the next section. Point (iv) ensures that one can bound the variance of tail-sums of χ_t^i, which one can think of as the accumulated estimation error. $\mathcal{F}_t$ can be thought of as the information available to the algorithmic updating rule at a given period t. Such increasing sequences of σ-algebras are commonly employed in the analysis of stochastic difference equations such as (5).

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

For the stepsizes α_t I maintain the following:

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

$\alpha_t \to 0$ *with*

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

This assumption takes its name from the celebrated Robbins-Monro algorithm representation (Robbins and Monro (1951)). The assumption constrains the speed of convergence of α_t, needing to balance the averaging out of errors (i.e. 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 Benaïm and Faure (2012). For a more general discussion of sufficient conditions for bounded iterates, see Borkar (2009), Chapter 2.

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

4. Long Run Behavior: Main Results

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

$$L_S = \bigcap_{t \geq 0} \overline{\{\rho_\ell \,|\ell \geq t\}},$$

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

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

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

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

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

$$F_S(\rho) = \bar{B}_S(\rho) - \rho, \tag{6}$$

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

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

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

Proof Sketch of Theorem 1

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

Firstly, I make a general connection between the recursion in (5) and the differential inclusion F_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

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

Since the best-response may be multivalued, solutions to this inclusion are not guaranteed. However, assumptions on the regularity of F_S (which comes down to a linear growth condition, see Definition 11 in appendix A) 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. 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.

Notice that the stability property of an equilibrium depends on the performance criterion F_S used by the underlying algorithm. Further, this stability property depends on the state variable observed by algorithms, even if the same perfomance criterion is used.

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

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

Proof Sketch of Theorem 2

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

Theorems 1 and 2 show the full potential of the characterization. Asymptotically stable equilibria are equilibria that can be limiting points of the RL learning procedure, while unstable equilibria are not. The intuition is related to how RL learn to play: since such agents make errors due to estimation and also to explore their action space, opponent's strategy profiles are constantly perturbed. In other words, out of the view of a fixed agent i, the other agents are frequently deviating to policies nearby in the policy space. Now suppose the current profile ρ_t is close to an equilibrium ρ^*. Since i's updating rule tracks F_S, their policy will only stay close to ρ^* if the dynamics of F_S are somehow robust to deviations. This robustness is implied by asymptotic stability, and broken by unstable equilibria.

There is a caveat here, however: Theorem 1 does not state that all limiting points in $L_{S,g}$ will be equilibria of the underlying repeated game as played by rational players. Depending on details of the 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 Theorem 1 to the case of attracting cycles as in Faure and Roth (2010), and there exist results considering linearly unstable cycles (Benaïm and Faure (2012)) that suggest one may extend Theorem 2 to such linearly unstable cycles also.[16] Notice that this observation implies that the Folk theorem is neither necessary nor sufficient in describing the possible payoffs achieveable by learning algorithms.

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

5. Learning to Collude

In this section, I exemplify the potential of my characterisation via a repeated Cournot game played by RL algorithms falling into the family of ACQ learners.

Recall that in section 4, I established a link between state dependent best-response dynamics F_S as defined in (6), and long-run behavior of ACQ learners. This section will consider the existence and stability of equilibria under F_S in a repeated Cournot game more closely. This game can be shown to satisfy Assumptions 1 and 2. It follows that whenever an ACQ algorithm satisfies Assumptions 3 and 4, the long run characterizations of section 4 apply.

The game is set up in line with Abreu, Pearce, and Stacchetti (1986)'s (APS) oligopoly game: There are two agents, $i \in \{1, 2\}$. Actions are chosen as quantities $x \in \mathbf{X} = [0, M]$ for some large $M > 0$, with aggregate quantity X. I will sometimes write $X \in \mathbf{X}$, in the understanding that the actual space of aggregate quantites is $[0, 2M]$. The price outcome is stochastic, $y \in \mathbf{Y} = [0, \overline{Y}]$, continuously distributed conditional on X. The conditional price density is denoted $g(y; X)$ with full support on $\mathbf{Y}$, $\mathcal{C}^2$ in X for almost all X. Let the expected price conditional on X be

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

Stage game payoffs are symmetric[17] for $i \in \{1, 2\}$:

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

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

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

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

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

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

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

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

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

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

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

Let S_0 with $|\mathbf{S}_0| = 1$ be the trivial state variable. F_{S_0} then simplifies to the classical stage game strategy based best response dynamics, which I sometimes refer to as 'myopic' best response dynamics, given the repeated nature of the interaction at hand. Under F_{S_0}, it is well known that under general conditions on $u^i(\cdot, \cdot)$, there is a unique Nash equilibrium that is globally attracting

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

(Milgrom and Roberts (1990)). We will see how this Nash equilibrium may not be learned even if it is globally attracting under myopic best response dynamics.

5.1. Binary State Variables

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

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

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

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

for all s, x_i, i.e. transition probabilities only depend on aggregate quantities. I will therefore commonly write $P_{ss'}(x_1, x_2) = P_{ss'}(X)$ with $X = x_1 + x_2$. Let $\rho^i : \mathbf{S} \mapsto \mathbf{X}$ be each player's policy, and recalling the definition of W^i in (3), note that in the binary case one can derive

$$\begin{aligned} W^i(\rho, A) &= \omega^{-1}(\rho)\Big[(1 - \delta P_{BB}(\rho))u^i\big(\rho^i(A), \rho^{-i}(A)\big) + \delta P_{AB}(\rho)u^i\big(\rho^i(B), \rho^{-i}(B)\big)\Big], \\ W^i(\rho, B) &= \omega^{-1}(\rho)\Big[\delta(1 - P_{BB}(\rho))u^i\big(\rho^i(A), \rho^{-i}(A)\big) + (1 - \delta(1 - P_{AB}))u^i\big(\rho^i(B), \rho^{-i}(B)\big)\Big], \end{aligned} \tag{7}$$

where

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

Thus, W^i is a convex combination of stage game payoffs u^i over the two states, with weights being a function of transition probabilities. Notably, as $\delta \to 1$, these weights will converge to the unique stationary distribution over states given the policy profile ρ.[18]

In what follows I will focus on symmetric equilibria, and therefore drop the i- superscript for all objects, fixing our attention on player 1's payoffs. In the following, to further ease notation, I will adopt the following conventions:

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

As stated before, when u is regular, the unique static Nash equilibrium is globally attracting under myopic best-response dynamics (F_{S_0}), and therefore if ACQ-learners (and many other agents) played on the trivial state space S_0, they would converge to x_N with probability 1. I show next that even though that is true, a larger family of binary state variables exist so that when

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

they are used, ACQ learners will not learn this Nash equilibrium. This inspires the following definition:

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

To see how a statically stable equilibrium may not be dynamically stable, focus on a binary state variable. First, consider the incentives faced by a rational agent playing a binary Markov strategy at the stage game Nash equilibrium, $\rho^i(s) = x_N$ for both i, s. I call this policy ρ_N. Given regular u, note that for any twice differentiable interior transition probabilities, the Hessian of each player's optimization problem of maximizing (7) at ρ_N is negative definite. Thus, we can derive a player's best-response derivative at ρ_N. Letting $b(\rho)$ be the best response to ρ, consider player 1's best-response derivative in state A to an incremental change in opponent's policy in state A:

$$\frac{\partial b(\rho_N)(A)}{\partial \rho_N(A)} = BR'_N + \frac{\delta P'_{AB}(\rho_N)}{\omega_N} \frac{u_2^N}{u_{11}^N},$$

where $b(\rho_N) = \rho_N$ is the best response according to long-run payoffs as derived in (7), $\omega_N = \omega(\rho_N)$ signifies evaluation of $\omega(\rho)$ at ρ_N, and I use that $BR'_N = -\frac{u_{12}^N}{u_{11}^N}$. The terminology BR'_N is used since indeed, $-\frac{u_{12}^N}{u_{11}^N}$ is the slope of the stage-game best response function evaluated at x_N. The agent has a tradeoff between following incentives about payoffs today (static incentives), represented by BR'_N, and dynamic incentives considering effects on continuation payoffs, represented by the second term. The factor multiplying $\frac{u_2^N}{u_{11}^N}$ can be interpreted as the sensitivity of the sum of transition probabilities $P_{AB}(\rho_N) + P_{BA}(\rho_N)$ with respect to policy ρ, which I now denote as ζ_N.

Dynamic stability is affected by discounting. As noted before, as $\delta \to 1$, $\omega_N^{-1}(1 - \delta P_{BB}) \to \mu_A(\rho_N) = \frac{P_{BA}(\rho_N)}{P_{AB}(\rho_N)+P_{BA}(\rho_N)}$, the stationary probability of visiting A when ρ_N is played forever. On the other side, dynamic incentives are null when $\delta = 0$. Dynamic stability of the static Nash equilibrium is impacted by dynamic incentives in a straightforward manner:

Proposition 1. *Let u be regular and consider arbitrary transition probabilities $P_{ss'}$ for a binary state variable. Then ρ_N is dynamically unstable (i.e. unstable w.r.t. F_S) if and only if*

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

Proposition 1[19] uncovers the channels through which the static Nash equilibrium can be destabilized, and eventually through which algorithms in my class will learn to avoid this Nash equilibrium. On the one side, market conditions matter through the size of the slope of the static

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

best response BR'_N today and the weighted effect that an opponent's deviation has on stage game payoffs u in the future. On the other side, fixing the market conditions, state variables come into play through ζ_N, the total sensitivity of transition probabilities with respect to policy ρ. This quantity represents the aggregate effect of a marginal change in policy ρ on transition probabilities, which in turn control the correlation structure over states. For algorithms to avoid the static Nash equilibrium, only the magnitude of this sensitivity matters: For any payoff function u of bounded derivatives, there is a threshold so that once ζ_N surpasses that threshold, static Nash will not be learned. The set of state variables that can render a static Nash equilibrium unstable is therefore quite large. This intuition then allows to separate two factors that determine whether the RL will learn to play static Nash: properties of stage game payoffs u, and properties of the state variable's distribution, governed by $P_{ss'}$.

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

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

where

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

Note that regularity of u implies $z_1^* < 0 < z_2^*$, and $BR'_N \in (-1, 0)$.

5.2. Two Fundamental Families of Binary State Variables

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

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

$$f_{1R}(s_{t-1}, Y_{t-1}, (y_A, y_B)) = \begin{cases} A & \text{if } s_{t-1} = A \text{ and } Y_{t-1} \leq y_A \\ A & \text{if } s_{t-1} = B \text{ and } Y_{t-1} \leq y_B \\ B & \text{if } s_{t-1} = B \text{ and } Y_{t-1} > y_B \\ B & \text{if } s_{t-1} = A \text{ and } Y_{t-1} > y_A. \end{cases} \tag{8}$$

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

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

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

Corollary 2. *Let* ρ_N^{1R} *be a consistent* 1R*-policy that plays stage game Nash quantity* x_N *in every state. Then,* ρ_N^{1R} *is dynamically stable if and only if* x_N *is statically stable.*[20]

This result follows from Corollary 1, since under consistent 1R-policies, we must have that $\zeta_N = 0$ whenever $y_A = y_B$. Note that under consistent 1R-policies, $P_{AB} = 1 - P_{BA} = \mathbb{P}[Y \leq y_A]$. Thus, $P'_{AB}(\rho_N) + P'_{BA}(\rho_N)$=0. In general, this comes from the fact that, for every given current state, conditional distributions over future states are the same. I call this quality of a state variable 'state-independent transitions' (SIT).

The transition function of a state variable can be depicted in a transition diagram. Figure 1 shows the underlying transition diagram under a state following f_{1R}, for some y_A, y_B.

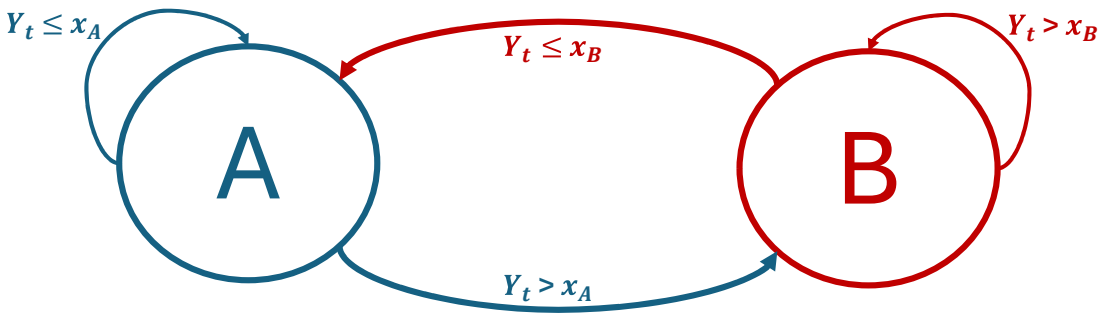

FIGURE 1. Transition Diagram: SIT

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

$$f_{DS}(s_{t-1}, Y_{t-1}, (y_A, y_B)) = \begin{cases} A & \text{if } s_{t-1} = A \text{ and } Y_{t-1} > y_A \\ A & \text{if } s_{t-1} = B \text{ and } Y_{t-1} \leq y_B \\ B & \text{if } s_{t-1} = B \text{ and } Y_{t-1} > y_B \\ B & \text{if } s_{t-1} = A \text{ and } Y_{t-1} \leq y_A. \end{cases} \tag{9}$$

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

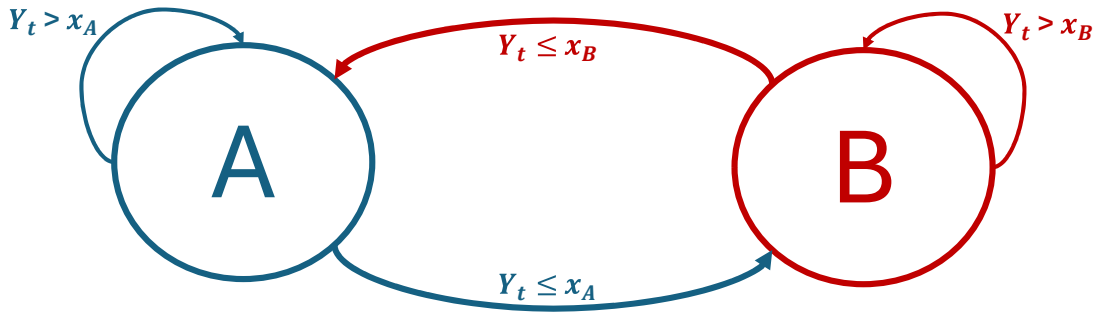

FIGURE 2. Transition Diagram: DS

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

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

Note that under consistent DS-policies, the probability of reaching any state s conditional on being in A is complementary to the probability of reaching s conditional on being in B. For a given quantity X, marginal deviations affect expected continuations in the opposite direction, depending on the current state. This fact introduces an essential difference in how states A, B are interpreted, even when $\rho(A) = \rho(B)$ is played. Let ρ_N^{DS} be a consistent DS-policy playing x_N in all states. It follows from Corollary 1 that there should exist conditional price distributions $g(y; X)$ so that ρ_N^{DS} will be dynamically unstable and therefore not learned by our RL. However, the issue is not immediate, since $g(y; X)$ plays a role both in construction of u and transition probabilities $P_{ss'}(X)$. In the following, I show that there is a set of regular payoff functions such that ρ_N^{DS} is indeed statically stable, but dynamically unstable. Moreover, this family of regular payoff functions will also allow for the existence of collusive equilibria.

For the sake of analytical tractability, I introduce a global shape restriction on $g(y; X)$ so that the resulting payoffs are regular, but will allow for a collusive equilibrium to exist under DS-policies.

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

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

$$\eta(y, X) \equiv \frac{\partial \log \big(g(y; X)\big)}{\partial X} \tag{10}$$

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

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

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

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

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

(iv)

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

MLRP ensures that lower prices are more sensitive to changes in quantities than higher prices. Points (ii) and (iii) will allow to construct collusive equilibria from first order conditions. Point (iv) ensures that a strict version of Definition 5 (iv) can be satisfied given this set of density functions. Note that point (iv) only becomes binding for large quantities x[21]. By carefully constructing $g(y; X)$ so as to have it loose sensitivity under large x, this can be made to hold. The numerical example at the end of this section verifies this.

Proposition 2. *For all $g \in \mathcal{G}$ there exist convex $c(x)$ such that the resulting u is regular. For a generic subset of $\mathcal{G}$*[22]*, there exists a state variable in S^* such that for all $\delta \in (0, 1)$ large enough, ρ_N^{DS} is dynamically unstable and there exists a symmetric equilibrium σ with $0 < \sigma_A < x_N < \sigma_B$.*

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

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

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

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

[22] A growth rate condition local to X_N.

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

Whether collusion will be learned, if it is a Nash equilibrium, still comes down to stability. This stability is determined by local conditions at the equilibrium. The dynamic instability of the Nash equilibrium is sufficient for the existence of collusion, but neither necessary for the existence, nor is it sufficient for stability of the collusive equilibrium. However, stability depends on quantities related to growth rates of transition probabilities and the stage game in manners analogous to the stability analysis of the static Nash equilibrium.

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

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

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

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

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

5.3. Relationship to the Best Equilibrium

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

First, let $\Gamma = \left(u^1, u^2\right)$ be the stage game as defined in the beginning of the section. Then one can define $\Gamma^\infty(\delta)$ as the infinite repetition of Γ where players discount expected long term payoffs by $\delta \in (0, 1)$. For any $0 < t$, define $b_t = \{Y_s\}_{0<s<t}$ to be a public history of the game, with $B_t = \mathbf{Y}^t$ the set of possible public histories up to time t. Then let $b_t^i = \{x_i^s\}_{0<s<t}$ be the private memory of a player's own actions, and define $B_t^i = \mathbf{X}^t$ as the set of those at period t. Now, a strategy of player i at period t can be written as map $\sigma_t^i : B_t \times B_t^i \mapsto \mathbf{X}$. A strategy is then a sequence $\sigma^i = \{\sigma_t^i\}_{t>0}$, with the set of such sequences denoted Σ. In keeping with APS, we can define strongly symmetric sequential equilibria (SSE) of Γ^∞ as profiles $\sigma = \{\sigma_t^1, \sigma_t^2)\}_{t>0}$ with $\sigma_t^1(b_t, b_t^1) = \sigma_t^2(b_t, b_t^2)$ whenever $b_t^1 = b_t^2$ (i.e. strategies that are 'public'), that are individually unimprovable for each player, with respect to their expected future discounted payoffs:

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

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

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

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

Under Γ^∞_K, APS' result applies: the payoff-maximizing SSE of Γ^∞_K can be achieved by a symmetric bang-bang profile $\sigma_K \in \mathbf{X}^4_K$, with V_K defined as its value. Notice further that under $g \in \mathcal{G}$, MLRP gives us that σ_K can be implemented as a DS-policy for some thresholds $(y_A, y_B) \in \mathbf{Y}^2$. This follows from the fact that optimal punishment regions of the price space, as characterised in APS, are monotone (binary) partitions under the MLRP. Payoffs under the 'good' state can be increased when the probability of punishment state is decreased; this can be done as long as incentives are preserved. Thus, starting from a price threshold $x \sim 0$, one can find the 'punishment set' of prices in the good state by considering all prices below x. By the MLRP, sensitivity of the conditional p.d.f. is maximal for the lowest prices, and decreases over prices, implying that this leads to the most efficient choice of punishment set in terms of probability of punishment.[23]

Define $W(\sigma, z)$ for symmetric profiles $\sigma \in \mathbf{X}^4$, thresholds $z \in \mathbf{Y}^2$ as long run binary state payoffs given those thresholds, abusing notation slightly from (7). Stretching notation slightly further, I will write $W(\sigma, \sigma', z)$ for profiles $(\sigma, \sigma') \in \mathbf{X}^4$ that are not necessarily symmetric.

Define

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

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

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

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

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

Analogously, define

$$V = \sup_{\substack{\sigma \in E^*(z) \\ z \in \mathbf{Y}^2}} W(\sigma, z). \tag{11}$$

Proposition 3.

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

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

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

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

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

5.4. Numerical Example and Simulations

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

One can verify numerically that here, $BR_0'(x_N) = -0.39$, implying static stability of x_N. To computationally find V, note that $g(y; X)$ does not satisfy MRLP. While for this p.d.f., $\eta(y, X)$ has a unique interior[24] zero as in the definition of $\mathcal{G}$ (9), $\eta(y, X)$ is not everywhere decreasing in y. However, $\eta(y, X)$ is single-peaked on the subsets $[0, \overline{y}(X))$, $(\overline{y}(X), \overline{Y}]$. An optimal assignment of punishment and reward regions as discussed in 5.3 is therefore still a binary partition of the price space - only that each state's punishment region is now described by two thresholds, instead of the previous single one. Let $\Omega_s = \left[z_s^{(1)}, z_s^{(2)}\right] \subset \mathbf{Y}$ for $s \in \{A, B\}$ be the 'switching' region in each state. Thus, when a price realizes in this region, states switch. It follows that here, $P_{ss'}(X) = \mathbb{P}[p \in \Omega_s \mid X]$, whenever $s \neq s'$. One can generalize (11) to an optimization program where V is maximized over $(z_s)_{s\in\{A,B\}}$, and $E^*(z), E_K(z)$ are re-defined accordingly, for $z \in \mathbf{Y}^4$. It is quick to check that Proposition 3 extends to this case.

[24]Interior isolated zero, which is sufficient here.

Thus, to numerically find V, I conduct a symmetric equilibrium search to determine $E^*(z)$ for a range of $z \in \mathbf{Y}^4$. Since each agent's value function $W(\sigma, \sigma', z)$ is concave in their policy σ, I conduct a search of symmetric zeros of the gradient of $W(\sigma, \sigma', z)$ with respect to σ. I consider a symmetric equilibrium to be found if $\max \left[\left| \nabla W(\sigma, \sigma', z) \Big|_{\sigma'=\sigma} \right| \right] \leq 10^{-14}$.

To visualize the possible values of best equilibria over a range of thresholds on a heatmap, define

$$V\big(z^{(1)}\big) = \max_{\substack{\sigma \in E^*(z) \\ (z_A^{(2)}, z_B^{(2)}) \in \mathbf{Y}^2}} W(\sigma, \sigma, z),$$

$$Gain\big(\sigma, z^{(1)}\big) = 100\Big(\frac{V\big(z^{(1)}\big)}{u_N} - 1\Big),$$

where $z^{(1)} = (z_A^{(1)}, z_B^{(1)})$. Thus, $V\big(z^{(1)}\big)$ is the best equilibrium given fixed lower bounds $z^{(1)}$ of Ω_A, Ω_B. $Gain\big(\sigma, z^{(1)}\big)$ is the percentage gain in long run payoffs of $V\big(z^{(1)}\big)$ versus the repetition of the static Nash payoff u_N.

Figure 4 shows a heatmap of $Gain\big(\sigma, z^{(1)}\big)$ for varying $z^{(1)} = \big(z_A^{(1)}, z_B^{(1)}\big)$, and an associated heatmap giving the maximal absolute eigenvalue of the linearized underlying dynamics F_S. Here, a value less than 1 indicates stability, hence that the given equilibrium profile will be learned with positive probability.

I construct a conditional p.d.f. $g(y; X)$, and convex cost resulting in a regular payoff function. For this game, the unique stage game Nash equilibrium x_N is statically stable, but dynamically unstable under a range of DS-policies. Furthermore, as argued in the previous section, under this conditional p.d.f., Proposition 3 applies.

Fix a discount factor $\delta = 0.98$. All numbers given in the example are rounded to two decimal points. Given domain $\mathbf{X} = [0, 1]$, and price support $\mathbf{Y} = [0, 1]$, Figure 3 shows conditional *c.d.f.* and $\eta(y, X)$ of the stage game.

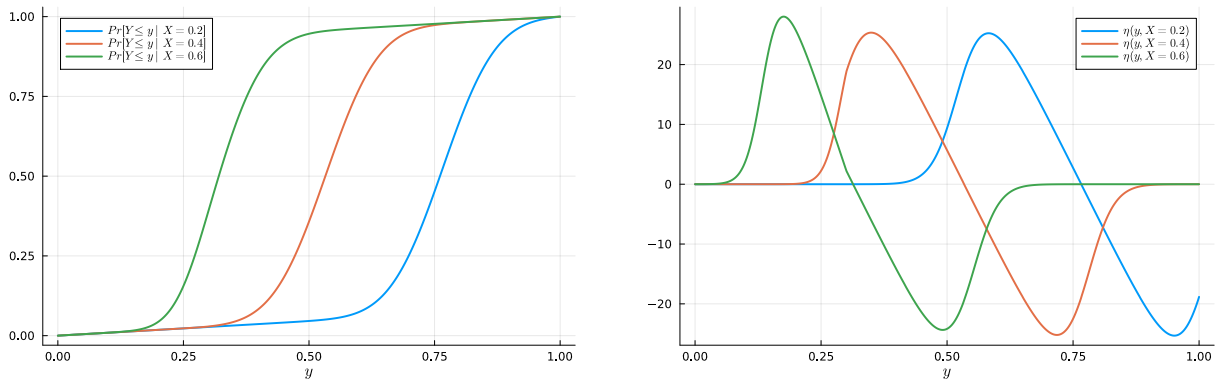

FIGURE 3. Left: C.d.f. conditional on different aggregate quantities. Right: $\eta(y, X)$ for different aggregate quantities X.

Thus, to numerically find V, I conduct a symmetric equilibrium search to determine $E^*(z)$ for a range of $z \in \mathbf{Y}^4$. Since each agent's value function $W(\sigma, \sigma', z)$ is concave in their policy σ, I conduct a search of symmetric zeros of the gradient of $W(\sigma, \sigma', z)$ with respect to σ. I consider a symmetric equilibrium to be found if $\max\left[\left|\nabla W(\sigma, \sigma', z)\Big|_{\sigma'=\sigma}\right|\right] \leq 10^{-14}$.

To visualize the possible values of best equilibria over a range of thresholds on a heatmap, define

$$V\big(z^{(1)}\big) = \max_{\substack{\sigma \in E^*(z) \\ (z_A^{(2)}, z_B^{(2)}) \in \mathbf{Y}^2}} W(\sigma, \sigma, z),$$

$$Gain\big(\sigma, z^{(1)}\big) = 100\Big(\frac{V\big(z^{(1)}\big)}{u_N} - 1\Big),$$

where $z^{(1)} = (z_A^{(1)}, z_B^{(1)})$. Thus, $V\big(z^{(1)}\big)$ is the best equilibrium given fixed lower bounds $z^{(1)}$ of Ω_A, Ω_B. $Gain\big(\sigma, z^{(1)}\big)$ is the percentage gain in long run payoffs of $V\big(z^{(1)}\big)$ versus the repetition of the static Nash payoff u_N.

Figure 4 shows a heatmap of $Gain\big(\sigma, z^{(1)}\big)$ for varying $z^{(1)} = \big(z_A^{(1)}, z_B^{(1)}\big)$, and an associated heatmap giving the maximal absolute eigenvalue of the linearized underlying dynamics F_S. Here, a value less than 1 indicates stability, hence that the given equilibrium profile will be learned with positive probability.

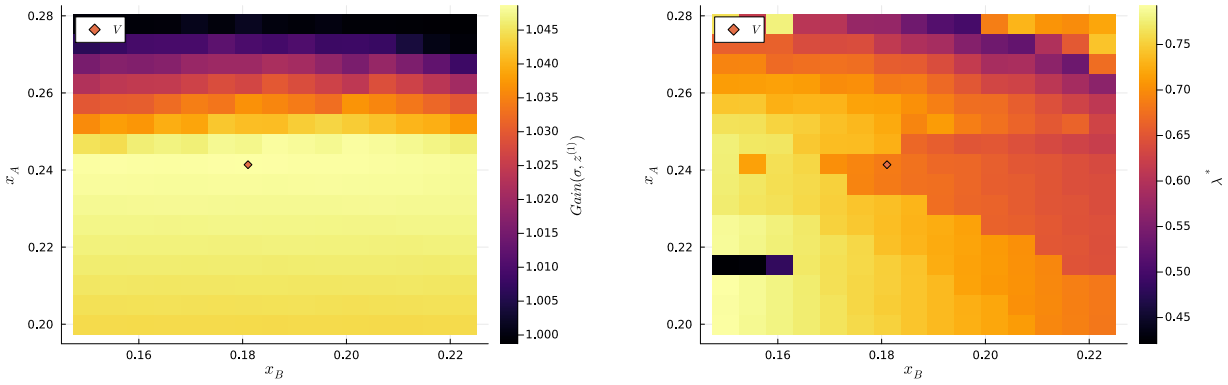

FIGURE 4. The left panel indicates $Gain(\sigma, z)$ over varying thresholds $z = (y_a, y_B)$. The right panel shows values of the maximal absolute eigenvalue of the equilibrium profiles supporting $V(z)$. All best equilibria $V(z)$ are attracting. The orange diamond indicates the location of the overall best equilibrium, V.

I finish by providing a simulation study of ACQ learners playing this game. Let z^* be the thresholds that support V, the computationally best equilibrium. Fix S^* to be the DS-state variable with transition function using Ω_A, Ω_B as switching regions pinned down by z^*. On the other side, fix S_{1R} to be the 1R-state variable (transitions as in (8)), under some threshold pair $z_{1R} = (z_A, z_B)$. The simulation study can now be used to see what ACQ-learners will learn if they observe either state variable. This simulation should be seen as a device to get intuitions about

the system dynamics after many iterations of the algorithm have passed. The characterization of long-run behavior given in subsection 4 is used here: instead of simulating the estimation part of Q_t of the algorithm given in Definition 2, I take Assumption 3 seriously, and simulate iteration (5) in the following way:

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

$$\rho^i_{t+1}(s) \in \rho^i_t(s) + \alpha_t \Big[\arg\max_{q' \in \mathbf{X}} Q^{i*}(s, x', \rho^{-i}_t) - \rho^i_t(s) + M^i_{t+1} \Big], \tag{12}$$

where $\alpha_t = t^{-0.6}$ satisfies the Robbins-Monro Assumption 4, and $M^i_{t+1} \sim N(0, .1)$ is an i.i.d mean-zero Normal noise variable with variance 0.25. Notice that (12) replaces Q_t given in (5) by its estimation target Q^*. Thus, this iteration represents a noisy discretization of F_{S^*} rather than a simulation of a feasible model-free algorithm. As the results in subsection 4 tell us, for algorithms in the class studied in this section this simulation will give us an equivalent representation of long-run trajectories of ρ_t to a full simulation of (5) when t is large.

In each simulation exercise, I run 960 separate simulations, and each for 10^6 periods. As will be seen, depending on the state variables of the algorithms involved, iterations move closer to the equilibrium in the neighborhood of which they started at, or move away from it, confirming the theory developed in this paper.

First, I consider the result given in Corollary 2. Since in this example, the Nash equilibrium is statically stable, its repetition under 1R-policies ρ_N is also stable. Thus, one would expect that once algorithms using 1R-state variables come close to the Nash equilibrium, they should stay close to it forever, and in the long run converge to it. This is what is evidenced by Figure 5. Since the state space is binary, the two algorithms' policies can be represented as points in the $\mathbf{X}^2$-plane. I now plot simulation outcomes in this plane, so that each simulation run is represented by two points in the plane spanned by $\rho(A), \rho(B)$.

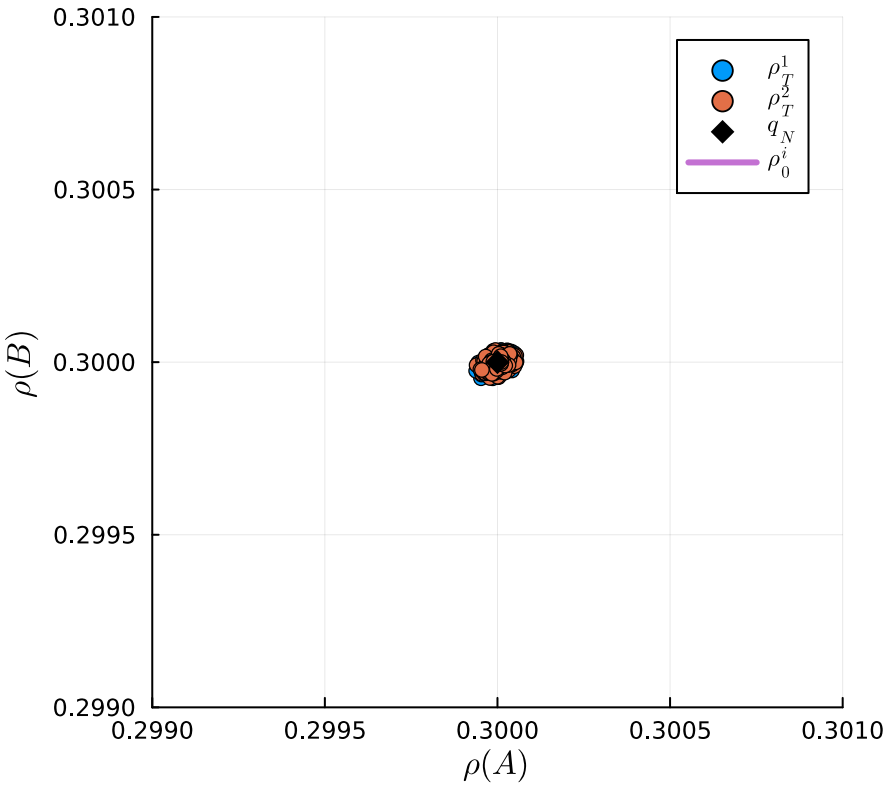

FIGURE 5. Final policies as dots ρ_T^i, for $i = 1, 2$ of 960 simulation runs, with $T = 10^6$. These runs were initialized globally, with ρ_0^i drawn uniformly from $\mathbf{X}^2$ for $i = 1, 2$. All runs converged to a close neighborhood of x_N. Note that the presence of shocks M_{t+1} pushes the process to continually move around the equilibrium, albeit in close proximity. The picture is analogous under a local initialization, with ρ_0^i drawn from a ball centered at x_N, at radius $0.01\|x_N\|$..

Now contrast this result with an analogous study given S^*. Even though the neighborhoods of starting values used in this scenario is the same as under 1R-policies, the picture is starkly different: none of the simulation runs converge to static Nash, which under the new state variable ceases to be dynamically stable.

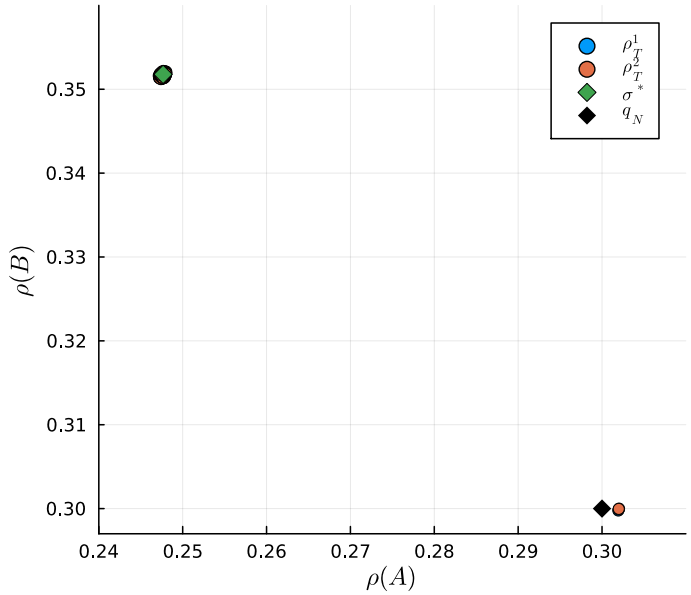

FIGURE 6. Final policies as dots ρ_T^i after a global initialization, with ρ_0^i drawn uniformly from $\mathbf{X}^2$ for $i = 1, 2$, with 960 simulation runs, with $T = 10^6$. 99.8% of runs converged to a neighborhood of the best equilibrium σ^* from these initial values. The remainder, 0.2%, converged to a neighborhood of the third symmetric equilibrium $\sim (0.3033, 0.2998)$, which is also stable. In both experiments, none of the simulations approached x_N in the long run.

The existence of the third symmetric equilibrium is not surprising, as can be seen from the construction of Ψ in the proof of Proposition 2. The outcome of a simulation with initialization centered at x_N, at radius $0.01\|x_N\|$, are similar: 98.1% of runs converged to a neighborhood of σ^*. Since x_N is dynamically unstable given state variable S^*, no matter how close the starting values of the iteration are, the iteration must be pushed away from ρ_N^{DS} as shown in the proof of Theorem 6. 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 together with the results of the global initialization 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 state variables used by algorithms. Even if one forced algorithms to initialize very close to a Cournot equilibrium, they can, given the right state variable, approach a collusive equilibrium instead.

The example was generated using the julia language (bezanson2017julia). The following non-base packages where used in this example: jlinterpolations,jldistribarrays,jlstaticarrays,jlnonlinsolve,jlnlsol,jld2,jlstr

6. Conclusion

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

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

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

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

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.

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

6.1. Discussion of related Literature

Firstly, Banchio and Mantegazza (2022) also consider a characterization of competing RL algorithms and apply it to games of economic interest. The class of algorithms they study intersects with the class studied in this paper, but there are important differences. It is unclear that their approach can accommodate actor-critic approaches that are featured here, as such approaches require a separate estimation technique that can introduce dependence of policy parameters on histories of past observations. This is important, since the actor-critic feature allows us to consider

closely the learning of repeated game strategies, which is not featured in the focus of Banchio and Mantegazza (2022).

Relatedly, Dolgopolov (2024) considers Q-learning in the prisoner's dilemma game. The author provides a novel long-run characterisation using Markov chains on a discretized approximation of the range of Q-values, and shows how tuning of stepsize and experimentation rate can decide whether cooperation emerges in the learning setting.

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

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

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

Johnson, Rhodes, and Wildenbeest (2020) look into platform design under algorithmic sellers. They investigate differing policies implemented by a platform designer wishing to promote competition or raise their own profits. They include a simulation study of Q-learning algorithms under different policy designs; clearly, results in this paper can be applied to study related RL algorithms under any given platform policy. Once a more tight characterization of the distribution over outcomes supported by a profile of algorithms is in place, one can go a step further and attempt to find the optimal platform policy for any given algorithm profile in my class.

There is now a growing area of research lying on the intersection of the theory of learning in games from the economics point of view, and the asymptotic theory of algorithmic learning from the computer science side. Leslie, Perkins, and Xu (2020)'s paper is an example of a paper intended more for economists, while applying language also common to the computer science literature. They consider zero-sum Markov games and construct an updating scheme related

to best response dynamics that converges to equilibria of the game. As they also keep track of separate policy and value function updates, their scheme falls into the class of actor-critic learning rules generally, while not falling into the class considered in this paper due to important assumptions on the updating speed differential between policy and performance criterion used there.

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

One can interpret results in this paper as showing that a subclass (ACQ) of the RL I consider can be seen as a GWFP process. Therefore, one can apply Leslie and Collins (2006) to conclude the limiting behavior of that process in games with the fictitious play property. However, there are many repeated games of interest that do not have this property; notably standard repeated oligopoly (Cournot) games where agents learn repeated game strategies. I analyze 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 (automaton) strategies, which to the best of my knowledge has so far only been considered under the restriction of Markov strategies for stochastic games.

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

Other papers related to asymptotic analysis of multi-agent systems commonly focus on developing a specific algorithm that behaves well in some metric, allow communication across algorithms, require information on the primitives of the game, or do not ask about the nature of the limiting

points. Notably, Ramaswamy and Hullermeier (2021) give a thorough analysis of deep learning techniques for Q-functions using gradient updates, without considering stability properties of rest points. Others focus on specific classes of games, for example zero sum games (Sayin et al. (2021)) and show convergence of multi-agent learning there.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Dolgopolov, Arthur (2024). "Reinforcement learning in a prisoner's dilemma". In: *Games and Economic Behavior* 144, pp. 84–103.

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

Faure, Mathieu and Gregory Roth (2010). "Stochastic approximations of set-valued dynamical systems: Convergence with positive probability to an attractor". In: *Mathematics of Operations Research* 35.3, pp. 624–640.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Leslie, David S, Steven Perkins, and Zibo Xu (2020). "Best-response dynamics in zero-sum stochastic games". In: *Journal of Economic Theory* 189, p. 105095.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A. The Algorithm Class

In this section, I provide the general reinforcement learning family the analysis of sections 3-4 applies to. Assume there are N algorithmic agents. Agents observe states on some fixed, finite state space S with $|S| = L$, and make per period choices (actions) in a compact interval $\mathbf{X}_i$. Let

$\overline{\mathbf{X}}_i = \mathbf{X}_i^L$, with policy profile space $\overline{\mathbf{X}} = \times_{i \in I} \overline{\mathbf{X}}_i$. Agents then follow a fixed rule (algorithm) to update their strategy profiles over time.

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

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

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

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

$$\rho_{t+1} \in \rho_t + \alpha_t \left[F(\rho_t) + B_t \right]. \tag{13}$$

The iteration is generalized to an inclusion, as can be the case when F^i represents an argmax, which corresponds to Q-iterations as in Definition 2. The class of RL algorithms studied here is determined by restrictions on $F(\rho)$ and B_t^i. Whenever there is multivaluedness, I allow the algorithm to pick arbitrarily. In our limiting characterization, this will show up as the possibility of multiple solutions (see Filipov (1988)), which will not affect the limiting statements as shown in section 4. Throughout, impose Assumptions 4 and 5.

Remark 2. *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 Borkar (2009)).

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

Considering the iteration (13), we can see that $F(\rho_t)$ features importantly as a mapping that provides the reinforcement of the iteration profile ρ_t. In many scenarios, $F(\rho)$ represents a performance criterion based on market and opponent conditions that are not known to the algorithm designer and must be estimated. $F(\rho)$ thus becomes an estimation target, and B_t can then be seen as the resulting estimation error.

First, I introduce the class of performance criteria $F(\rho)$, and what kinds of approximation methods can be considered here. As stated in section 3, this family can be allowed to not contain the estimation target, leading to asymptotically biased criterion function estimators. The long run characterisation result will later be shown to be robust to a certain family of biases. This robustification means it is sufficient for researchers to verify smoothness and bound a possible asymptotic bias, without needing to know the specific functional form of the bias.

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

$$\mathcal{B}_\gamma^k = \left\{ g : \overline{\mathbf{X}} \mapsto \mathbb{R}^{nL} \mid \sup_{x \in \overline{\mathbf{X}}} \|g(x)\| + \sum_{j=1}^{k} \sup_{x \in \overline{\mathbf{X}}} \|D^j g(x)\| \leq \gamma \right\}, \tag{14}$$

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

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

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

Note that, with some abuse of notation, $\mathcal{C}^1 \subset \mathcal{M}^1$. Now, define the distance between points x and sets A as

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

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

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

(i) For all $\rho \in \mathbf{X}$,

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

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

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

and $\delta_t \in \mathbb{R}^k$ a noise term,

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

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

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

(iii)

$$\mathbb{E}\big[\|\delta_t\|\big] = o(b_t),$$

where $b_t \to 0$ is a sequence satisfying the following: there exists b'_t with $\frac{b'_t}{b_t} \to 0$, and b'_t satisfies Assumption 4.

(iv)

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

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

the speed of convergence and variance of the error term δ_t to ensure that our characterization technique used in Theorems 1 and 2 goes through. In fact, (iii) is useful as long as we have that the stepsize α_t in Definition 10 satisfies $\lim_{t\to\infty} \frac{\alpha_t}{b_t} = 0$. Thus, given a performance-criterion estimator satisfying Definition 12, choose α_t so that this is satisfied.

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

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

Assumption 6.

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

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

(iii)

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

where $\mathcal{A}_g[F, D_t]$ *is a* $\mathcal{C}^1$ *Approximation of performance criterion* $F(\rho) \in \mathcal{M}^1$.

(iv) *Stacked version of* B^i_t*:*

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

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

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

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

(vii) *There exists a continuous function*

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

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

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

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

(viii) *Write* $\delta_t = \mathcal{A}_g[F, D_{t+1}](\rho_t) - F_g(\rho_t)$. *Then for all* $t' < t''$, $\|\delta_{t'}\|, \|\delta_{t''}\|$ *are uncorrelated conditional on* $\mathcal{F}_{t'}$.

We have discussed points $(i) - (iv)$. Point (v) constructs an increasing sequence of σ-fields, which in turn are used in the construction of the bounded Martingale-error term of (vi). This construction and assumption (vi) are common in the stochastic approximation literature concerned with the limiting behavior of stochastic difference equations (e.g. see Borkar (2009)). Point (vii) ensures that at minimum, errors M_{t+1} generate enough noise so that any direction within a small open ball around ρ_t will be visited by ρ_{t+1} with positive probability, given $\mathcal{F}_t$. This fact will prove useful in deterring the process from converging to unstable equilibria, and has been used e.g. in Benaïm and Faure (2012). Point $(viii)$ is analogous to the martingale-property of M_{t+1}, in that it ensures that the variance of sums of $\|\delta_t\|$ be bounded. These sums represent the accumulated estimation error, which need to vanish probabilistically in order for the ODE approximation to have any bite.

Remark 3. *Note that assumptions 1-5 are sufficient for the ACQ algorithm defined in section 3 to be within the above family. Most points are immediate. Based on Assumption 3, define* $\|\delta_t^i\| = C\big(\chi_t^i\big)^{\frac{1}{\beta}}$. *Then Assumption 3 (ii) is sufficient for Definition 12 (iii) to be satisfied, by Jensen's inequality and since* $\beta > 1$. *Assumption 3 (iv) implies Assumption 6 (viii).*

A.1. Gradient-type Algorithms

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

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

$$W^i(\rho^i, \rho^{-i}, s_0) = \mathbb{E} \sum_{t=0}^{\infty} \delta^t u^i(\rho(s_t), s_t), \tag{15}$$

where the expectation is made over the state transitions.

Then define

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

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

Now define for $\rho \in \overline{\mathbf{X}}$

$$F_S^D(\rho) = \bigtriangledown W(\rho), \tag{16}$$

as the state dependent gradient dynamics. Take an iteration ρ_t and its respective function estimation target F as denoted in (13). 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 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.

Appendix B. Best Response Dynamics: Stability

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

Let $S = \{A, B\}$ be any binary state space. For a given policy-profile $\alpha, \beta \in \mathbf{X}^2$, write best replies as $b_1(\beta) = (b_1^A, b_1^B)^\top \in BR_S^1(\beta)$, and $b_2(\alpha) = (b_2^A, b_2^B)^\top \in BR_S^2(\alpha)$. I consider the stability of rest points for the state-dependent best response dynamics under S, F_S (see (6)), given the stacked policies $\sigma \in \mathbf{X}^4$:

$$\dot{\sigma}_t = F_S(\sigma_t) = \begin{bmatrix} b_1(\sigma_{t,3}, \sigma_{t,4}) \\ b_2(\sigma_{t,1}, \sigma_{t,2}) \end{bmatrix} - \sigma_t. \tag{17}$$

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

$$X(\sigma^*) = \begin{bmatrix} -I_{2x2} & J_1(\sigma^*) \\ J_2(\sigma^*) & -I_{2x2}, \end{bmatrix} \tag{18}$$

where I_2 is the 2-dimensional identity matrix and

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

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

This linearization has a special structure one 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 characterized 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 $X(\sigma^*)$.

When considering symmetric equilibria, one 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. Given square matrix A, define Λ as the set of eigenvalues of the A. Then define

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

as the spectral radius of A.

We can then apply Remark 5 to our system and arrive at the following conclusion:

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

Proof. Using Remark 5, we get that

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

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

Appendix C. Proofs

Proof of Theorem 1

I prove the following more general result:

Proposition 4. *Let $\rho^* \in \mathcal{U}_S$ be asymptotically stable for F_S. Then for all γ small enough and all $g \in \mathcal{B}^1_\gamma$ there is a profile ρ^g such that*

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

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

Notice that accordingly, rest point ρ^g may not be an exact Nash equilibrium of the underlying game, but an ε-equilibrium:

Definition 13. *A profile ρ is an ε-equilibrium if for all players i all individual profiles $\rho' \in \overline{\mathbf{X}}$ and states $s \in \mathbf{S}$*

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

The implied statement in a game such as in section 5 would then be:

Corollary 4. *Let $\rho^* \in E$ be asymptotically stable for F_S. Then for all γ small enough and all $g \in \mathcal{B}^1_\gamma$ 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) $\mathbb{P}[L_{S,g} = \{\rho^g\}] > 0$.

Now to the proof: The proofs are given in terms of the general algorithm class treated in this paper, introduced in Appendix A. Throughout we pick a stepsize sequence α_t s.t. Assumption 4 holds and $\lim_{t\to\infty} \frac{\alpha_t}{b_t} = 0$ for b_t given in Definition 12 (iii). First, we prove the following result that employs known techniques from stochastic approximation theory.

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

Definition 14.

(1) Given a set $A \in \mathbf{X}$ and $x, y \in A$, we write $x \hookrightarrow_A y$ if for every $\varepsilon > 0$ and $T > 0$, there exists an integer $n \in \mathbb{N}$, solutions $x_1, ..., x_n$ to $\dot{x} \in G(x)$[25], and real numbers $t_1, ..., t_n$ greater than T such that:
a) $x_i(s) \in A$ for all $0 \leq x \leq t_i$, and for all $i = 1, ..., n$,
b) $\|x_i(t_i) - x_{i+1}(0)\| \leq \varepsilon$ for all $i = 1, ..., n-1$,
c) $\|x_1(0) - x\| \leq \varepsilon$ and $\|x_n(t_n) - y\| \leq \varepsilon$.
(2) A set $A \in \mathbf{X}$ is said to be internally chain transitive (ICT) if A is compact and $x \hookrightarrow_A y$ holds for all $x, y \in A$.

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

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

Proposition 5. *Almost surely, $L_{S,g}$ is an ICT set of the differential inclusion*

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

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

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

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

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

Proof. The algorithm (13) can be written as

$$\rho_{n+1} = \rho_n + \alpha_n \left[F_g(\rho_n) + \delta_n + M_{n+1} \right], \tag{19}$$

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

We can now show first that iteration 19 is a perturbed solution to $\dot{\rho} \in F_g(\rho(t))$ as defined in Definition II in Benaïm, Hofbauer, and Sorin (2005). The approach is to construct a linear interpolation of (19), and show that this will shadow solutions to $\dot{\rho} \in F_g(\rho(t))$ asymptotically, for large enough t. Following the notation in Hofbauer and Sandholm (2002), introduce:

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

Then, construct the interpolation as

$$X(\tau_n + s) = \rho_n + s \frac{\rho_{n+1} - \rho_n}{\alpha_{n+1}}, \quad s \in [0, \alpha_{n+1}]. \tag{20}$$

Following the proof of their Proposition 1.3, we only need to take care of the additional term δ_n present in iteration 19.

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

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

By Assumption 6, R_n is a standard error term in stochastic approximation theory, satisfying the usual assumptions of Robbins-Monro algorithms with martingale difference noise. The sufficient conditions of Benaïm, Hofbauer, and Sorin (2005) are satisfied here, so it is known that R_n converges almost surely to zero.[26] We need to take care of the additional term δ_n present in iteration 19. It suffices to show that, for all $T > 0$

$$\sup_{n \leq k \leq m(\tau_n+T)-1} \Psi_n^k \to 0, \tag{21}$$

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

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

$$\Psi_n^k \leq \sup_{n \leq k \leq m(\tau_n+T)-1} \left\| \sum_{i=n}^{k} \alpha_{i+1} \Big(\|\delta_{i+1}\| - \mathbb{E}\big[\|\delta_{i+1}\| \mid \mathcal{F}_{i+1}\big] \Big) \right\| + \sum_{i=n}^{m(\tau_n+T)-1} \alpha_{i+1} \mathbb{E}\|\delta_{i+1}\| \tag{22}$$

$$= R_{2,n} + K_n, \tag{23}$$

where $\mathcal{F}_i$ is the filtration defined in Assumption 6 (v). Now, by Assumption 6 $(viii)$ and square integrability of $\|\delta_n\|$, $R_{2,n}$ is the supremum on another martingale difference noise term with bounded variance, just as R_n. Thus, again for $R_{2,n}$ we have almost sure convergence to zero here. As for K_n, recall from Definition 12 that $\mathbb{E}\|\delta_n\| = o(b_n)$. We thus have that there exists some $C_K > 0$ such that for all n large enough,

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

by assumption that $\lim_{t\to\infty} \frac{\alpha_t}{b_t} = 0$. Thus, by square summability of α_i, the sum above must converge to zero in n, and therefore $K_n \to 0$ as well, and the result (21) follows.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Since we know that all ρ^g must be asymptotically stable for γ small enough, one can apply Faure and Roth (2010) (Thm 2.15). To prove convergence to an attractor $\{\rho^g\}$ with positive probability, a stronger result than Proposition 5 is first needed:

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

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

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

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

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

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

Lemma 5. *Suppose δ_n, M_n satisfy Definition 12 and Assumption 6. Then condition* 11 *is satisfied.*

Proof. Note first that

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

similarly as stated in the proof above. For R_n, Proposition 2.16 in Faure and Roth (2010) immediately applies, as it only requires the Robbins-Monro condition on α_n, and that Assumption

6 (vi) is satisfied for M_n. For Ψ_n^k, recall (22) from the previous proof. As noted there, $R_{2,n}$ is another bounded martingale-difference noise, so Proposition 2.16 applies again. Finally, K_n is a determinstic sequence converging to zero as shown above, so that the probability of the term being larger than any fixed $\varepsilon > 0$ is always zero for large enough n. The result follows. □

Finally, theorem 2.15 in Faure and Roth (2010) states that if condition 11 is satisfied, $\mathbb{P}[L_{S,g} = \{\rho^g\}] > 0$ holds as long as $\{\rho^g\}$ is *attainable* by the process ρ_n:

Definition 15. *A point p is attainable if, for any $n > 0$ and any neighborhood U of p*

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

Let $Att(X)$ be the set of attainable points for algorithm (13). Then we need that the basin of attraction of an attractor has nonempty intersection with $Att(X)$. This can be verified:

Lemma 6. *Let B be a basin of attraction of an attractor A for F_g. Suppose $\rho_t \in \overline{\mathbf{X}} \setminus B$. Then there exists $s > n$ such that $\rho_s \in B$ with positive probability.*

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

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

since F_g might be multivalued.

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

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

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

Thus, theorem 2.15 in Faure and Roth (2010) applies, concluding this proof. ■

Proof of Theorem 2

I prove the following more general statement:

Proposition 6. *Let $\rho^* \in \mathcal{U}_S$ be linearly unstable for F_S. Then for all γ small enough and all $g \in \mathcal{B}_\gamma^1$ there is an open neighborhood U_γ with $\rho^* \in U_\gamma$ such that*

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

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

The proof will use the Hartman-Grobman Theorem (c.f.Chicone (2006), Theorem 4.8), which connects the flow of a nonlinear ODE in the neighborhood of a hyperbolic rest point to the flow

of a linearized ODE. Since it works fully locally, our analysis only requires that $F(\rho)$ be single valued and $\mathcal{C}^1$ in U_{ρ^*}, and we can allow $F(\rho)$ to be multivalued otherwise.

First, define invariant sets for given differential equations:

Definition 16. *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 $z(t, z_0) \in \mathbf{S}$ holds for all $t \in \mathbb{R}$ and all $z_0 \in \mathbf{S}$.*
- *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 5 that only ICT sets (recall Definition 14) subset of a neighborhood of ρ^g are candidates to being limiting points of the algorithm (13). The singleton $\{\rho^g\}$ is an ICT set, and we show first that this is a limiting set of the algorithm with probability zero. Then we go on to show that for small enough γ, no other ICT sets can exist in a neighborhood around ρ^*, which finishes the proof.

1) $\{\rho^g\}$ is a limiting set with probability zero.

Note that by Lemma 4, there are $\gamma > 0$ small enough such that all ρ^g are linearly unstable just as ρ^*. We can thus apply Benaïm and Faure (2012), Theorem 3.12 to prove $\mathbb{P}[L_{S,g} = \rho^g] = 0$ in the following. Importantly, note that the conditions and analysis sufficient for the proof of Benaïm and Faure (2012)'s Theorem 3.12 are local with respect to ρ^g. Thus, the fact that F_g is globally potentially multivalued is of no importance, since in a small enough neighborhood around ρ^g it must be single-valued and $\mathcal{C}^1$.

I show here that the sufficient conditions for Benaïm and Faure (2012)'s Theorem 3.12 hold by definition of our algorithm under Assumption 6. First, we need that Hypothesis 2.2 of Benaïm and Faure (2012) holds. This hypothesis is equivalent[27] to condition 11 in Faure and Roth (2010), which was shown to hold for our algorithm in Lemma 5.

Finally, theorem 3.12 in Benaïm and Faure (2012) requires specifically that our Assumption 6 $(vi), (vii)$ hold, and thus the theorem holds true, concluding the proof of point 1.

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

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

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

[27] See Remark 2.14 in Faure and Roth (2010)

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 $\{\rho^*\}$ is an 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 7. *For any $\delta > 0$ with $a_0 < 1 - \delta$ there exists $\overline{\gamma} > 0$ such that for all $\gamma \in (0, \overline{\gamma}]$, there is a set of $\{a_g\}_{g \in \mathcal{B}^1_\gamma}$ as defined above with*

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

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

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

Lemma 8. *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 7: 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 7 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 Theorem 1 point 2, $\rho^g \to \rho^*$ uniformly over $g \in \mathcal{B}^1_\gamma$ as $\gamma \to 0$. Thus, there is γ small enough for which $\sup_{g \in \mathcal{B}^1_\gamma} |\rho^g - \rho^*| < \underline{z}$ and therefore $\rho^* \in N_g$ for all $g \in \mathcal{B}^1_\gamma$. Let $U_\gamma = \cap_{g \in \mathcal{B}^1_\gamma} N_g$. Since ρ^g for $g \in \mathcal{B}^1_\gamma$ are isolated invariant in U_γ by construction, the result follows. ■

Proof of Proposition 1

As discussed in Appendix B, one needs to linearize best responses at ρ_N to determine the stability of that profile. We need to characterize the eigenvalues $\lambda_{1,2}$ of the matrix of best-response derivatives of player 1 at symmetric Nash policies ρ_N:

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

We have that

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

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

Proof of Proposition 2

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

Lemma 9. *Suppose $g \in \mathcal{G}$. Then there exist a convex cost function $c(x)$ such that the resulting stage game payoffs $u(x_1, x_2)$ are regular.*

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

Recall the following conventions:

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

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

Lemma 10.

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

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

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

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

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

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

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

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

For the second claim, note that convexity of c and $u_{12}(x, x') < 0$ implies that $u_1(x, x)$ is strictly decreasing for all $x \in \mathbf{X}$, crossing 0 at x_N. By construction of y^*, note that by Definition 9 (ii),

$G_2(y^*; X) > 0$ for all $X \in int(\mathbf{X})$, with peak at X_N. Thus, the fraction $\frac{u_1(x,x)}{h'(2x)} \in (-\infty, \infty)$ is strictly decreasing over $x \in \mathbf{X}$, and the claim follows.

For the third claim, consider two cases:

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

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

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

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

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

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

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

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

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

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

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

Notice that for all such α^*, we also have $W_{12}(\alpha^*,\beta)=0$. This follows under irreducibility, since then initial states do not affect the optimal policy choice. If a policy is optimal, it must be optimal given any starting state s, and therefore one can characterize it through FOCs equivalently for any starting s.

Now for the proof of the Proposition:

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

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

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

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

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

$$z'(x_A)=-\frac{h'_B}{h'_A}\frac{u_{11}^A+u_{12}^A-2h''_A\frac{u_1^A}{h'_A}}{u_{11}^B+u_{12}^B+2h''_B\frac{u_1^B}{h'_B}}. \tag{25}$$

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

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

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

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

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

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

Finally to prove that $\Psi'(x_N) > 0$, note that

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

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

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

Thus we can write

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

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

Lemma 11. *For any $g \in \mathcal{G}$ there exists $\tilde{g}$ differing from g only on a neighborhood of (y^*, X_N), so that under this $\tilde{g}$, $\Psi'(x_N) > 0$ holds.*

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

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

We will perturb $g(y; X)$ so as to flip the inequality in our benefit. To this end, we will use that the left hand side depends directly on the c.d.f. evaluated at a point y^*, while the right hand side

depends only on an integral over all $p \in \mathbf{Y}$ of the c.d.f.. Note that

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

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

$$\tilde{g}_2(y; X) = g_2(y; X) + \Delta \mathbf{1}\{p \in N_1,\ Q \in N_2\} - \Delta \frac{\mu}{\mu_C} \mathbf{1}\{p \geq \sup N_1,\ Q \in N_2\}, \tag{26}$$

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

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

I will show that this perturbation is feasible for N_1 small enough relative to Δ, ensuring that $\tilde{g}(y; X) > 0$ remains true; the construction (26) ensures that $\tilde{g}_2$ integrates to zero. Define $[\underline{y}, \overline{y}] = N_1$, $[\underline{X}, \overline{X}] = N_2$. It follows that

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

and

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

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

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

Now for the expected price:

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

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

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

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

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

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

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

and thus, for all $Q \in N_2$, the decrease in $\tilde{g}(y;X)$ can be controlled via $\Delta\mu_2$ and $\Delta\mu\mu_2$. We can always find three sequences $\Delta_j, \mu_{1,j}, \mu_{2,j} > 0$ for all j such that $\Delta_j\mu_{1,j}$ increases, $\Delta_j\mu_{1,j}^2$ decreases, $\Delta_j\mu_{2,j}$ is weakly increasing, and $\Delta_j\mu_{1,j}\mu_{2,j}$ decreases.[28]

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

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

We are left to show that this zero is a global maximizer. Firstly, we note that the Hessian at σ must be negative definite: we see from (24) that $W_{12} = 0$, so the Hessian must be diagonal at σ. A sufficient condition for negative definiteness then is $h''_A > 0 > h''_B$ and $u^A > u^B$. The first one follows given Definition 9 (ii) and since $x_A < x_N < x_B$, the second one follows from the first order conditions:

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

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

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

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

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

satisfied so that the maximizer can only be interior. Taking derivatives, we get

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

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

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

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

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

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

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

Proof of Lemma 1

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

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

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

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

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

In the proof of Lemma 10 I show that we have that $\frac{u_1(x,x)}{h'(2x)}$ is strictly decreasing for all $x \in [0, M]$. At the same time, $u(x, x)$ is strictly decreasing for all x, which is necessary for $u_1(x, x) < 0$. Thus, for case (i) it must be that $x_A > x_B$, but since $\frac{u_1(x,x)}{h'(2x)}$ is strictly decreasing, there exists no such pair x_A, x_B to set $W_1 = W_2$. It follows that no such pair can be an equilibrium.

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

Proof of Lemma 2

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

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

$$\begin{aligned}
\frac{\partial \rho_A^{*1}}{\partial \rho_A^2} &= -1 + \phi_A^{-1}\Big[\Pi'_A - \omega^{-1}\delta P'_{AB}\Pi_A\Big],\\
\frac{\partial \rho_A^{*1}}{\partial \rho_B^2} &= \phi_A^{-1}\omega^{-1}\delta P'_{AB}\Pi_B,\\
\frac{\partial \rho_B^{*1}}{\partial \rho_A^2} &= \phi_B^{-1}\omega^{-1}\delta P'_{BA}\Pi_A,\\
\frac{\partial \rho_B^{*1}}{\partial \rho_B^2} &= -1 + \phi_B^{-1}\Big[\Pi'_B - \omega^{-1}\delta P'_{BA}\Pi_B\Big],
\end{aligned} \tag{27}$$

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

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

Some tedious algebra then allows re-writing determinant and trace of $J(\sigma, \sigma)$ using (27), Then:

$$\begin{aligned}
tr\big(J^*\big) &= -2 + \frac{\Pi'_A}{\phi_A}\big[1 - R_A\big] + \frac{\Pi'_B}{\phi_B}\big[1 - R_B\big];\\
det\big(J^*\big) &= \left[1 - \frac{\Pi'_A}{\phi_A}\frac{\Pi'_B}{\phi_B}\right] - \frac{\Pi'_A}{\phi_A}\big[1 - R_A\big]\left[1 - \frac{\Pi'_B}{\phi_B}\right] - \frac{\Pi'_B}{\phi_B}\big[1 - R_B\big]\left[1 - \frac{\Pi'_A}{\phi_A}\right].
\end{aligned} \tag{28}$$

Notice that for the stage game as constructed in Proposition 2, $\phi_s < u^s_{11}$ holds, and therefore $\frac{\Pi'_s}{\phi_s} \in (0, 1)$ can be guaranteed as long as $u^s_{12} \leq 0$, since $\Pi'_s = u^s_{11} - u^s_{12}$. Sign and magnitude of R_s depend on local conditions of both transition probabilities and the stage game quantity $\Pi(x_1, x_2)$. It is clear from (28) that both trace and determinant depend crucially on the quantities R_s. Indeed, if R_A, R_B are not too negative, stability of σ follows:

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

$$\big|tr\big(J^*\big)\big| - det\big(J^*\big) < 1. \tag{29}$$

Then, note from (28) that by the condition of the Proposition,

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

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

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

Thus, for R_A, R_B bigger than 0, the trace drops out in the condition in (29). The last equation then determines stability through the term $\left[1 - R_A - R_B\right]$. ■

Proof of Proposition 3

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

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

where

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

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

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

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

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

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

Lemma 12. *For all sequences* $\{\sigma_K\}$ *with* $\sigma_K \in E_K(z)$,

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

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

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

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

$$\left|W(\sigma_z, \sigma_K, z) - W(\sigma_z, \bar{\sigma}, z)\right| \leq \frac{\varepsilon}{3}. \tag{30}$$

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

$$\left|W(\sigma_K, \sigma_K, z) - W(\bar{\sigma}, \bar{\sigma}, z)\right| \leq \frac{\varepsilon}{3}. \tag{31}$$

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

$$\left|W(f^K(\sigma_z), \sigma_K, z) - W(\sigma_z, \sigma_K, z)\right| \leq \frac{\varepsilon}{3}. \tag{32}$$

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

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

Now we have

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

where $\beta_{1,K}$ corresponds to the projection error (32), and $\beta_{2,K}, \beta_{3,K}$ correspond to (30),(31) respectively. We have that $|\beta_{i,K}| \leq \frac{\varepsilon}{3}$, and thus

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

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

To finish the proof, note that Lemma 12 implies that for any feasible α_K, $\lim_{K\to\infty} F_z(\alpha_K) \subseteq E^*(z)$, with $E^*(z)$ being the continuous-action version of the equilibrium set for fixed z. We get that

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

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

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

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

Since $\Delta_z > 0$, we can find large enough discretizations s.t. σ can be approximated arbitrarily closely, in which case incentives must also align, by continuity of W. The contradiction follows. Hence, together with Lemma 12, we get that

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

■