Submitted to *Econometrica*

AN ONLINE APPENDIX TO: REINFORCEMENT LEARNING AND COLLUSION

CLEMENS POSSNIG
Department of Economics, University of Waterloo

1. THE ALGORITHM CLASS

In this section, I provide the general reinforcement learning family the analysis of sections 2-3 in the main text applies to. 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 1: Each agent updates their policy according to the following adaptive procedure:

$$\rho_{n+1}^i \in \rho_n^i + \alpha_t \left[F^i(\rho_n) + B_n^i \right],$$

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

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

$$\rho_{n+1} \in \rho_n + \alpha_n \left[F(\rho_n) + B_n \right]. \tag{1}$$

For stepsizes, we assume:

ASSUMPTION 1—Robbins-Monro Condition: $\alpha_n \to 0$ with

$$\sum_{n=0}^{\infty} \alpha_n = \infty; \quad \sum_{n=0}^{\infty} \alpha_n^2 < \infty.$$

The iteration is generalized to an inclusion, as can be the case when F^i represents an argmax, which corresponds to the ACQ-algorithm presented in the main text. The class of

Clemens Possnig: cpossnig@uwaterloo.ca

RL algorithms studied here is determined by restrictions on $F(\rho)$ and B_n^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.

REMARK 1: The following are two important examples of what behavior B_t can be allowed to take:

1. $B_n = 0$ for all n 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_n is a martingale-difference noise with respect to some filtration $\mathcal{F}_n$, 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 (1), 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 (or criterion function) 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. The family of functions used to approximate $F(\rho)$ 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. The following constructs the family of bias functions that the results extend to:

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

$$\mathcal{B}_\gamma^k = \left\{ g : \overline{\mathbf{X}} \mapsto \mathbb{R}^{nL} \mid \sup_{x \in \overline{\mathbf{X}}} \|g(x)\| + \sum_{j=1}^{k} \sup_{x \in \overline{\mathbf{X}}} \|D^j g(x)\| \leq \gamma \right\}, \tag{2}$$

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

DEFINITION 2—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.

(i) $G(x) \subset R$ is convex, compact valued.

(ii) 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.

(iii) 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, $\mathcal{C}^1 \subset \mathcal{M}^1$. We assume that $\mathcal{U}_S$ contains hyperbolic rest points, which can then be treated in the main results. This would e.g. be true under ACQ-learning, but also under gradient schemes as outlined in the end of this section. Define the distance between points x and sets A as

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

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

DEFINITION 3—$\mathcal{C}^1$ Approximation:
Let Y be some space of observations (datasets) D_n to be used to approximate a mapping. Given $\gamma > 0$, bias function $g \in \mathcal{B}^1_\gamma$, 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 an integer $N > 0$ such that one can write for all $n \geq N$:

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

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

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

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

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

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

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

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

(iii)

$$\mathbb{E}\big[\|\delta_n\|\big] = o(b_n),$$

where $b_n \to 0$ is a sequence satisfying the following: there exists b'_n with $\frac{b'_n}{b_n} \to 0$, and b'_n satisfies Assumption 1.

(iv)

$$\sup_{n \geq 0} \mathbb{E}\big[\|\delta_n\|^2\big] < \infty.$$

One can interpret $g(\rho)$ as representing the bias part of the function approximation, and δ_n as a random variable such that $\mathbb{E}[\|\delta_n\|^2]$ represents the variance part. Points (iii) and (iv) bound the speed of convergence and variance of the error term δ_n to ensure that our characterization technique used in main results to come goes through. In fact, (iii) is useful as long as we have that the stepsize α_n in Definition 1 satisfies $\lim_{n\to\infty} \frac{\alpha_n}{b_n} = 0$. Thus, given a performance-criterion estimator satisfying Definition 3, choose α_n so that this is satisfied.

In the case of classical model-free Q learning (Watkins (1989)), D_n only needs to consist of $(s_k, a_k, r_k, s_{k+1})_{k=1}^n$, i.e. past observations of states, actions, payoffs, state transitions, and the initial Q-matrix.

Generally, one can think of $\mathcal{A}_g[F, D_n](\cdot)$ 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 2 and 3 characterizes the algorithm class that can be studied here.

ASSUMPTION 2:

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

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

(iii)

$$B_n^i = \mathcal{A}_{g^i}^i[F^i, D_{n+1}^i](\rho_t) - F_g^i(\rho_n) + M_{n+1}^i,$$

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

(iv) Stacked version of B_n^i:

$$B_n = \mathcal{A}_g[F, D_{n+1}](\rho_n) - F_g(\rho_n) + M_{n+1}.$$

(v) $\mathcal{F}_n$ is the σ-field generated by $\{\rho_n, D_n, M_n, \rho_{n-1}, D_{n-1}, M_{n-1}, \ldots, \rho_0, D_0, M_0\}$, i.e. all the information available to the updating rule at a given period n.

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

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

(vii) There exists a continuous function

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

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

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

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

(viii) Write $\delta_n = \mathcal{A}_g[F, D_{n+1}](\rho_n) - F_g(\rho_n)$. Then for all $n' < n''$, $\|\delta_{n'}\|, \|\delta_{n''}\|$ are uncorrelated conditional on $\mathcal{F}_{n'}$.

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_{n+1} generate enough noise so that any direction within a small open ball around ρ_n will be visited by ρ_{n+1} with positive probability, given $\mathcal{F}_n$. 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_{n+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 2: Note that Assumptions 1-5 in the main text are sufficient for the ACQ algorithm defined there to be within the above family. Most points are immediate. Based on Assumption 3 in the main text, define $\|\delta_n^i\| = C\left(\chi_n^i\right)^{\frac{1}{\beta}}$. Then Assumption **??** (ii) is sufficient for Definition 3 (iii) to be satisfied, by Jensen's inequality and since $\beta > 1$. Assumption 3 (iv) in the main text implies Assumption 2 $(viii)$.

1.1. *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:

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

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 when assuming that the state variable is irreducible, as in the main text. 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 $\rho \in \overline{\mathbf{X}}$

$$F_S^D(\rho) = \bigtriangledown W(\rho), \tag{4}$$

as the state dependent gradient dynamics. Take an iteration ρ_n and its respective function estimation target F as denoted in (1). If $F = F_D^S$, one can call the RL iteration 'Gradient Equivalent'.

For Gradient Equivalent iterations, if there is no asymptotic bias in the estimation of the gradient ($g(\rho) = 0$), the results here match the results in Mazumdar et al. (2020), but note that we study the possibility of repeated game strategies, which is not explicitly done there. Further, as noted in the introduction, the results extend Mazumdar et al. (2020) to the more commonly observed situation of non-vanishing biased function estimators.

2. PROOFS FOR THE GENERAL ALGORITHM CLASS

Recall the following definition:

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 $\lambda \in \Lambda$.

Theorem 1

The first result extends Theorem 1 in the main text:

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

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

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

DEFINITION 5: 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 as e.g. outlined in section 4 of the main text would then be:

COROLLARY 1: *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: Throughout we pick a stepsize sequence α_n s.t. Assumption 1 holds and $\lim_{n\to\infty} \frac{\alpha_n}{b_n} = 0$ for b_n given in Definition 3 (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 et al. (2005, Section 3.3):

DEFINITION 6:

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 $\dot{x} \in G(x)$[1], 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.

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

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

PROPOSITION 1: *Almost surely, $L_{S,g}$ is an ICT set of the differential inclusion*

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

where $F_g(\rho(t)) \in \mathcal{M}^1$ satisfies Definition 2 $(i), (ii)$.

PROOF: The algorithm (1) can be written as

$$\rho_{n+1} \in \rho_n + \alpha_n \left[F_g(\rho_n) + \delta_n + M_{n+1} \right], \tag{5}$$

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

We can now show first that iteration 5 is a perturbed solution to $\dot{\rho} \in F_g(\rho(t))$ as defined in Benaïm et al. (2005, Definition II). The approach is to construct a linear interpolation of (5), 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

$$X(\tau_n + s) = \rho_n + s \frac{\rho_{n+1} - \rho_n}{\alpha_{n+1}}, \quad s \in [0, \alpha_{n+1}]. \tag{6}$$

Following the proof of Hofbauer and Sandholm (2002, Proposition 1.3), we only need to take care of the additional term δ_n present in iteration 5 - but this can be done analogously to the proof of Proposition 4 of the main text. The result follows. *Q.E.D.*

Now, since payoffs are differentiable around ρ^*, point (1) of Theorem 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 1: *Suppose ρ^* is hyperbolic. Let $DF(\rho), DF_g(\rho)$ be the Jacobian of F, F_g, respectively. Then the eigenvalues of $DF_g(\rho^g)$ converge to the eigenvalues of $DF(\rho^*)$ uniformly over $g \in \mathcal{B}^1_\gamma$ as $\gamma \to 0$. Thus, for small enough γ, ρ^g has the same stability properties as ρ^*.*

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

Palis Jr et al. (1982, Proposition 2.18) 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 Palis Jr et al. (1982, Proposition 2.18) finishes the result. *Q.E.D.*

Since we know that all ρ^g must be asymptotically stable for γ small enough, one can apply Faure and Roth (2010, Theorem 2.15) . To prove convergence to an attractor $\{\rho^g\}$ with positive probability, a stronger result than Proposition 1 is first needed:

ASSUMPTION 3—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. $\lim_{n\to\infty} \omega(n, \varepsilon, T) = 0$.

Faure and Roth (2010, Proposition 2.16) states that Condition 11 above is satisfied for our M_{n+1} martingale difference sequence (i.e. if $\delta_n = 0$ for all n). I show next that this result extends to our case of (5):

LEMMA 2: *Suppose δ_n, M_n satisfy Definition 3 and Assumption 2. Then condition 11 is satisfied.*

PROOF: Note first that

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

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 2 (vi) is satisfied for M_n. The remaining term Ψ_n^k can be treated analogously to the proof of Theorem 1 in the main text. *Q.E.D.*

Finally Faure and Roth (2010, Theorem 2.15) 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. This can be

verified analogously to the approach in the proof of Theorem 1 of the main text. Thus, Faure and Roth (2010, Theorem 2.15) applies, concluding this proof. *Q.E.D.*

Theorem 2

The following generalizes Theorem 2 of the main text:

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

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

PROOF: 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 a neighborhood of rest point ρ^*, and we can allow $F(\rho)$ to be multivalued otherwise. Call this neighborhood U_{ρ^*}.

First, define invariant sets for given differential equations:

DEFINITION 7: Let $z(t, z_0)$ be the solution to some given differential equation $\dot{z} = f(z)$ with initial value z_0. Then a set S

- is invariant for f, if $z(t, z_0) \in \mathbf{S}$ 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 1 that only ICT sets (recall Definition 6) subset of a neighborhood of ρ^g are candidates to being limiting points of the algorithm (1). 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 of (5) with probability zero.

Note that by Lemma 1, there are $\gamma > 0$ small enough such that all ρ^g for $g \in \mathcal{B}^1_\gamma$ 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 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 $\mathcal{C}^1$.

Benaïm and Faure's result is concerned with time-interpolations of stochastic differential inclusions $F(\rho) \in \mathcal{M}^1$, such as (6). Their Theorem 3.12 states, translated in terms of this paper, that under an Assumption the authors refer to as Hypothesis 2.2, and Assumption 2 $(vi), (vii)$, the result to be proved here holds true.

In fact, Benaïm and Faure (2012, Hypothesis 2.2) is equivalent[2] to Assumption 3, which was shown to hold for our algorithm in Lemma 2. Thus, the result applies, concluding the proof.

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

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

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

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

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

where $\phi(t, \cdot)$ is a solution (flow) to the differential inclusion $\dot{\rho} \in F(\rho)$, 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

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

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

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 3: *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 1. 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. *Q.E.D.*

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

LEMMA 4: *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 Palis Jr et al. (1982, Lemmas 4.3 and 4.4), $B_{r_\varepsilon}(\rho^g) \subseteq N_g$, 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 3: 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 3 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. *Q.E.D.*

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. *Q.E.D.*

3. NUMERICAL EXAMPLE AND SIMULATIONS

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, under this conditional p.d.f., Proposition 2 of the main text 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 1 shows conditional $c.d.f.$ and $\eta(y, X)$ of the stage game.

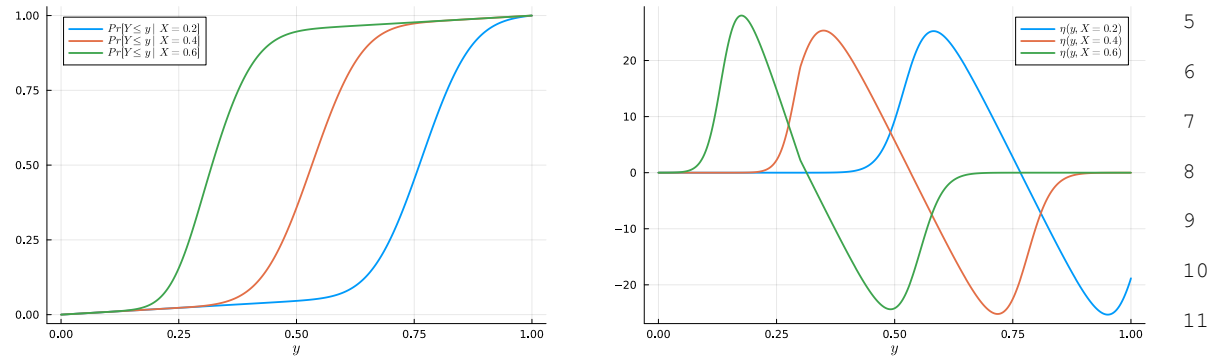

FIGURE 1.—Left: C.d.f. conditional on different aggregate quantities. Right: $\eta(y, X)$ for different aggregate quantities X.

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[3] zero as in the definition of $\mathcal{G}$ (see Definition 9 in the main text), $\eta(y, X)$ 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 Section 4.3 of the main text 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 the approach of Section 4.3 in the main text (see equation (9) there) 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 of the main text extends to this case.

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

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

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 2 shows a heatmap of $Gain\big(\sigma, z^{(1)}\big)$ for varying $z^{(1)} = \big(z_A^{(1)}, z_B^{(1)}\big)$. All eigenvalues of the associated linearized $F_S(\rho)$ at these equilibria are less that 1 in absolute value, hence stable. Thus, each equilibrium profile will be learned with positive probability under their respective state variable generated from the thresholds given.

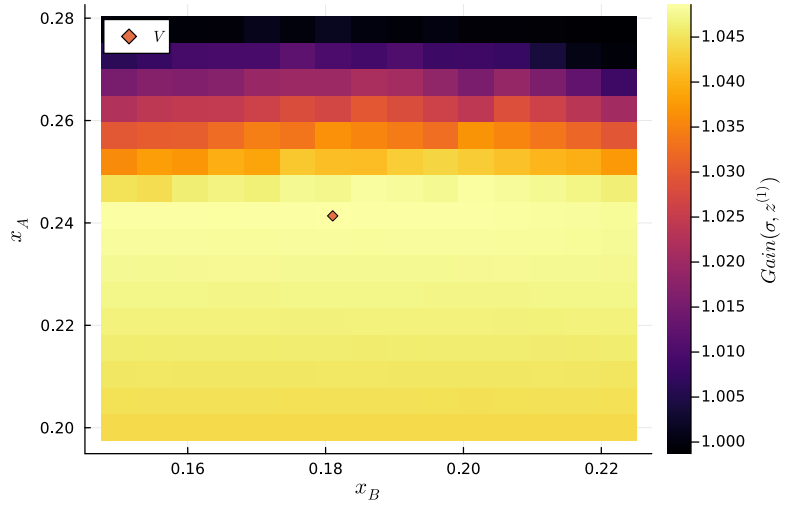

FIGURE 2.—$Gain(\sigma, z)$ over varying thresholds $z = (y_a, y_B)$. The orange diamond indicates the location of the overall best equilibrium, V.

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 (6) in the main text), 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. Recall from the main text the defintion of the ACQ learning rule:

$$\rho^i_{t+1}(s) \in \rho^i_t(s) + \alpha_t \Big[\arg\max_{a' \in \mathbf{X}} Q^i_{t+1}(s, a') - \rho^i_t(s) + M^i_{t+1} \Big], \tag{7}$$

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 Section 1 is used here: instead of simulating the estimation part of Q_t of the algorithm given above, I take Assumption 2 seriously, and simulate iteration (7) in the following way:

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

$$\rho^i_{t+1}(s) \in \rho^i_t(s) + \alpha_t \Big[\arg\max_{q' \in \mathbf{X}} Q^{i*}(s, x', \rho^{-i}_t) - \rho^i_t(s) + M^i_{t+1} \Big], \tag{8}$$

where $\alpha_t = t^{-0.6}$ satisfies the Robbins-Monro Assumption 1, 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 (8) replaces Q_t given in (7) by its estimation target Q^*. Thus, this iteration represents a noisy discretization of F rather than a simulation of a feasible model-free algorithm. As the results in Section 1 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 (7) 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 of the main text. 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 3. 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 $\rho(A), \rho(B)$.

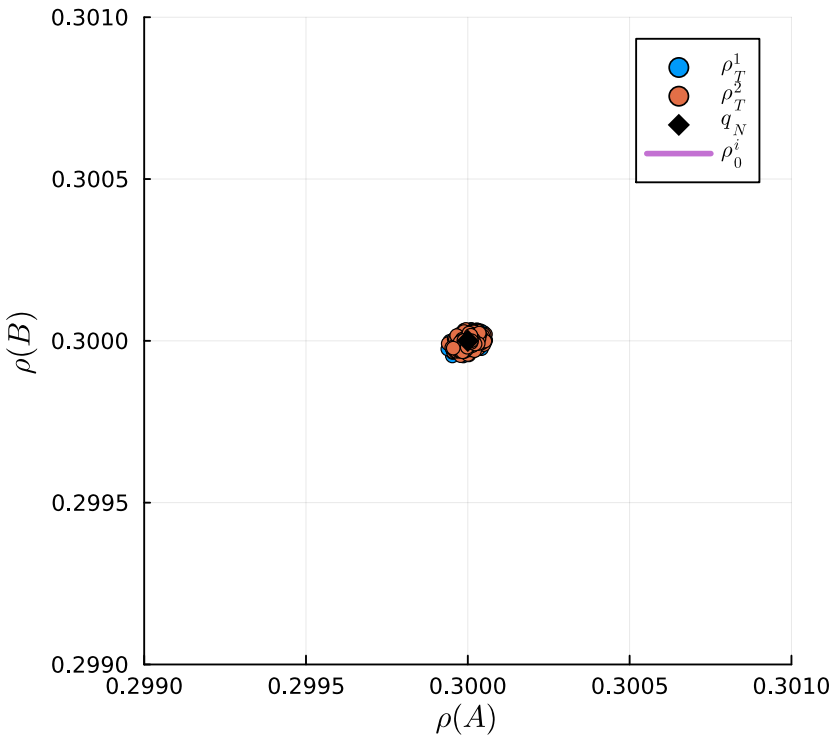

FIGURE 3.—Final policies as dots ρ_T^i, for $i = 1, 2$ of 960 simulation runs, with $T = 10^6$. These runs were initialized globally, with ρ_0^i drawn uniformly from $\mathbf{X}^2$ for $i = 1, 2$. All runs converged to a close neighborhood of x_N. Note that the presence of shocks M_{t+1} pushes the process 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.

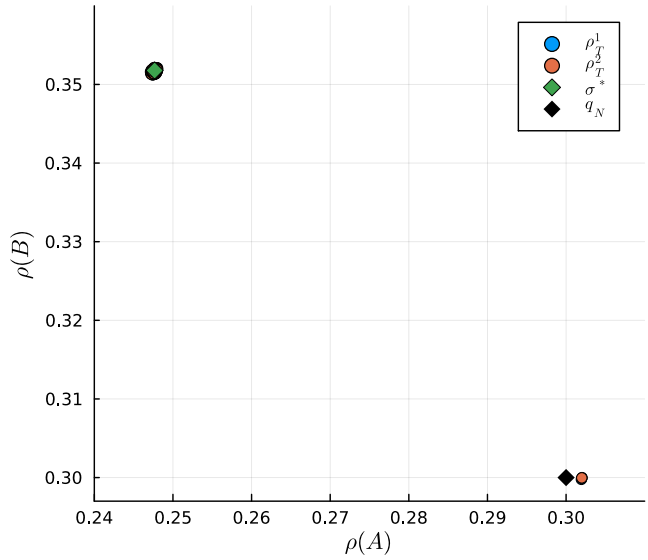

FIGURE 4.—Final policies as dots ρ_T^i after a global initialization, with ρ_0^i drawn uniformly from $\mathbf{X}^2$ for $i = 1, 2$, with 960 simulation runs, with $T = 10^6$. 99.8% of runs converged to a neighborhood of the best equilibrium σ^* from these initial values. The remainder, 0.2%, converged 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 in the main text. 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 2. However, in the case of this example, it is not only true that the iteration is pushed away, but also that it is pulled towards the collusive equilibrium σ. This 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 (Bezanson et al. (2017)). The following non-base packages where used in this example: Kittisopikul et al. (2022),Noack et al. (2023),Baran et al. (2024),Pal et al. (2024),Mogensen et al. (2020),Isensee et al. (2024),Johnson et al. (2023).

4. 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 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. The main text to this online appendix 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 et al. (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. At this point, the state of the art for general characterisations of long-run behavior of algorithms stops at showing whether a given outcome will be learned with positive, or zero probability. 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 et al. (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 con-

tinuous time fictitious play (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 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.

Furthermore, 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 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 et al. (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 et al. (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 et al. (2020)'s results are robust to the type of bias in the gradient estimation that my RL class allows.

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

BANCHIO, MARTINO AND GIACOMO MANTEGAZZA (2022): "Games of Artificial Intelligence: A Continuous-Time Approach," *arXiv preprint arXiv:2202.05946*. [21]

BARAN, MATEUSZ, CLAIRE FOSTER, ET AL. (2024): "JuliaArrays/StaticArrays.jl: v1.9.7," . [21]

BENAÏM, MICHEL AND MATHIEU FAURE (2012): "Stochastic approximation, cooperative dynamics and supermodular games," *The Annals of Applied Probability*, 22 (5), 2133–2164. [6, 12, 13]

BENAÏM, MICHEL, JOSEF HOFBAUER, AND SYLVAIN SORIN (2005): "Stochastic approximations and differential inclusions," *SIAM Journal on Control and Optimization*, 44 (1), 328–348. [8, 9]

BEZANSON, JEFF, ALAN EDELMAN, STEFAN KARPINSKI, AND VIRAL B SHAH (2017): "Julia: A fresh approach to numerical computing," *SIAM review*, 59 (1), 65–98. [21]

BORKAR, VIVEK S (2009): *Stochastic approximation: a dynamical systems viewpoint*, vol. 48, Springer. [2, 5]

BROWN, GEORGE W (1951): "Iterative solution of games by fictitious play," *Act. Anal. Prod Allocation*, 13 (1), 374. [23]

BROWN, ZACH Y AND ALEXANDER MACKAY (2021): "Competition in pricing algorithms," Tech. rep., National Bureau of Economic Research. [21]

CHICONE, CARMEN (2006): *Ordinary differential equations with applications*, vol. 34, Springer Science & Business Media. [12]

DOLGOPOLOV, ARTHUR (2024): "Reinforcement learning in a prisoner's dilemma," *Games and Economic Behavior*, 144, 84–103. [21]

FAURE, MATHIEU AND GREGORY ROTH (2010): "Stochastic approximations of set-valued dynamical systems: Convergence with positive probability to an attractor," *Mathematics of Operations Research*, 35 (3), 624–640. [11, 12, 13]

FILIPOV, ALEKSEI FEDOROVICH (1988): "Differential equations with discontinuous right-hand side," in *Amer. Math. Soc*, 191–231. [2]

FUDENBERG, DREW AND DAVID K LEVINE (2009): "Learning and equilibrium," *Annu. Rev. Econ.*, 1 (1), 385–420. [23]

HOFBAUER, JOSEF AND WILLIAM H SANDHOLM (2002): "On the global convergence of stochastic fictitious play," *Econometrica*, 70 (6), 2265–2294. [9, 23]

ISENSEE, JONAS, SIMON KORNBLITH, ET AL. (2024): "JuliaIO/JLD2.jl: v0.5.2," . [21]

JOHNSON, JUSTIN, ANDREW RHODES, AND MATTHIJS R WILDENBEEST (2020): "Platform design when sellers use pricing algorithms," *Available at SSRN 3753903*. [22]

JOHNSON, STEVEN G. ET AL. (2023): "JuliaStrings/LaTeXStrigns.jl: v1.3.1," . [21]

KITTISOPIKUL, MARK, TIMOTHY E. HOLY, AND TOMAS ASCHAN (2022): "JuliaMath/Interpolations.jl: v0.14.7," . [21]

LAMBA, ROHIT AND SERGEY ZHUK (2022): "Pricing with algorithms," *arXiv preprint arXiv:2205.04661*. [21]

LESLIE, DAVID S AND EDMUND J COLLINS (2006): "Generalised weakened fictitious play," *Games and Economic Behavior*, 56 (2), 285–298. [22, 23]

LESLIE, DAVID S, STEVEN PERKINS, AND ZIBO XU (2020): "Best-response dynamics in zero-sum stochastic games," *Journal of Economic Theory*, 189, 105095. [22]

MAZUMDAR, ERIC, LILLIAN J RATLIFF, AND S SHANKAR SASTRY (2020): "On gradient-based learning in continuous games," *SIAM Journal on Mathematics of Data Science*, 2 (1), 103–131. [7, 23, 24]

MOGENSEN, PATRICK K., SEBASTIEN VILLEMOT, ET AL. (2020): "JuliaNLSolver/NLsolve.jl: v4.5.1," . [21]

NOACK, ANDREAS, AMIT MURTHY, JAKE BOLEWSKI, VALENTIN CHURAVY, ET AL. (2023): "JuliaParallel/DistributedArrays.jl: v0.6.7," . [21]

PAL, AVIK ET AL. (2024): "SciML/NonlinearSolve.jl: v3.14.0," . [21]

PALIS JR, J, W DE MELO, ET AL. (1982): "Geometric Theory of Dynamical Systems," . [10, 15]

PAPADIMITRIOU, CHRISTOS AND GEORGIOS PILIOURAS (2018): "From nash equilibria to chain recurrent sets: An algorithmic solution concept for game theory," *Entropy*, 20 (10), 782. [9]

POSSNIG, CLEMENS (2022): "Learning to Best Reply: On the Consistency of Multi-Agent Batch Reinforcement Learning," . [24]

RAMASWAMY, ARUNSELVAN AND EYKE HULLERMEIER (2021): "Deep Q-Learning: Theoretical Insights from an Asymptotic Analysis," *IEEE Transactions on Artificial Intelligence*. [24]

SALCEDO, BRUNO (2015): "Pricing algorithms and tacit collusion," *Manuscript, Pennsylvania State University*. [21]

SAYIN, MUHAMMED, KAIQING ZHANG, DAVID LESLIE, TAMER BASAR, AND ASUMAN OZDAGLAR (2021): "Decentralized Q-learning in zero-sum Markov games," *Advances in Neural Information Processing Systems*, 34. [24]

WATKINS, CHRISTOPHER JOHN CORNISH HELLABY (1989): "Learning from delayed rewards," . [4]