An Online Appendix to:
Reinforcement Learning and Collusion

Clemens Possnig

University of Waterloo

July 12, 2025

1. The Algorithm Class

The following assumptions are sufficient for the results stated in Section 3 of the main text to go through, upon minor extensions to known results from stochastic approximation theory, to be found in Benaïm (1999), Borkar (2009), Michel Benaïm and Faure (2012). The setting considered here generalizes the one studied in the main text by allowing for a non-vanishing bias term, a robustification of the results discussed below.

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.

In keeping with the main text, we have $N > 1$ learning agents, and consider state variables taking $K > 0$ different values. Define $\overline{\mathbf{Y}} \subseteq \mathbb{R}^{NK}$ as the space of actions, stacked over i. The following constructs the family of bias functions that the results extend to:

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

$$\mathcal{B}^k_\gamma = \left\{ g : \overline{\mathbf{Y}} \to \mathbb{R}^{NK} \mid \sup_{x \in \overline{\mathbf{Y}}} \|g(x)\| + \sum_{j=1}^{k} \sup_{x \in \overline{\mathbf{Y}}} \|D^j g(x)\| \leq \gamma \right\}, \tag{1}$$

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

For notational ease, write $F(\rho) = F_S(\rho)$, as the stacking over i of $F_S^i(\rho) : \overline{\mathbf{Y}} \to \mathbb{R}^{NK}$, i's critic function. For all results to follow, state variables will be fixed. We identify time periods by n in order to distinguish the continuous timescale t used in the associated continuous time systems. The algorithms in most general form can jointly (stacked over i) be written as

$$\rho_{n+1} = \rho_n + \alpha_n \left[F(\rho_n) + g(\rho_n) + \delta_n + M_{n+1} \right], \tag{2}$$

where now $g(\rho_n) + \delta_n + M_{n+1} = \hat{F}_n(\rho_n) - F(\rho_n)$ is the representation of period-n errors in the critic estimation, and $g \in \mathcal{B}_\gamma^2$ for some $\gamma > 0$ represents a non-vanishing bias term. Fixing $g(\rho)$, we often write $F_g(\rho) = F(\rho) + g(\rho)$ to save space.

Assumption 1. *Let $\mathcal{F}_n$ be the σ-field generated by $\{\rho_n, \delta_n, M_n, \rho_{n-1}, \delta_{n-1}, M_{n-1} \ldots, \rho_0, \delta_0, M_0\}$, i.e. all the information available to the updating rule at a given period n.*

(i) Stepsizes α_n^i satisfy, for all i, to be square-summable, but not summable.

(ii) For all i, j, $\lim_{n\to\infty} \frac{\alpha_n^i}{\alpha_n^j}$ exists and lies in (c, ∞), for some $c > 0$.

(iii) F_g is Liptschitz continuous and grows sublinearly, i.e.

$$\limsup_{\|\rho\|\to\infty} \frac{\|F_g(\rho)\|}{\|\rho\|} < \infty.$$

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

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

(v) There exists a continuous function

$$\Omega : \overline{\mathbf{Y}} \to O(\overline{\mathbf{Y}}),$$

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

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

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

(vi)

$$E\left[\|\delta_n\|\right] = o(b_n),$$

where $b_n \to 0$ *satisfies* $\max_i \lim_{n\to\infty} \frac{b_n}{\alpha_n^i} = 0$, α_n^i *being* i*'s stepsize.*

(vii)

$$\sup_{n\geq 0} \mathbb{E}\left[\|\delta_n\|^2\right] < \infty,$$

(viii) *For all* $n' < n''$, $\delta_{n'}, \delta_{n''}$ *are uncorrelated conditional on* $\mathcal{F}_{n'}$.

(ix) *Iterates stay bounded almost surely:*

$$\sup_n \|\rho_n\| < \infty, \ a.s..$$

Point (i) is known as the Robbins-Monro condition (Robbins and Monro 1951) on stepsizes. It ensures that stepsizes converge slowly enough so that the whole real line can be mapped (as a continuous-time interval), while converging not too slowly in order for error terms to be averaged out. (ii) ensures that all stepsizes lie within the same order of magnitude. Point (iii) ensures global integrability and uniqueness of solutions to $\dot{\rho} = F(\rho)$. In the example of ACQ, it is an assumption on payoffs W^i, and that best responses can't grow too quickly. Point (iv) implies that given current information in period t, new errors due to $t+1$'s estimator of F are well-behaved. It is a common assumption in stochastic approximation theory. Point (v) ensures that some variance in error terms remains for all n; this is satisfied e.g. if the estimation of F involves exploratory noise, or stochasticity during the estimation as is true under randomized Bellman-iteration schemes. This assumption will be the main driver that pushes iterations away from unstable equilibria. Point (vi) ensures that the bias term vanishes faster than stepsizes. Points $(vii), (viii)$ are further regularity conditions on the bias term. Even though commonly made, point (ix) is often difficult to verify. It is common for results to be stated conditioning on the event that (ix) holds, see for example Michel Benaïm and Faure (2012). For a more general discussion of sufficient conditions for bounded iterates, see Borkar (2009), Chapter 2.

2. Proofs for the general Algorithm Class

Recall the following definition:

Definition 1. *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 $\lambda \in \Lambda$.*

Also define the limit set as

$$L_{S,g} = \bigcap_{t \geq 0} cl\left(\{\rho_\ell \,|\ell \geq t\}\right),$$

the set of limits of convergent subsequences ρ_{t_k}, keeping track of the existence of the bias function g.

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 $g \in \mathcal{B}^1_\gamma$ 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) $\mathbb{P}[L_{S,g} = \{\rho^g\}] > 0$.

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

Definition 2. *A profile ρ is an ε-equilibrium if for all players i all individual profiles $\rho' \in \overline{\mathbf{Y}}$ 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 $g \in \mathcal{B}_\gamma^1$ there is a $\bar{\varepsilon} > 0$ and a profile ρ^g such that*

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

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

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

Now to the proof: Taking F_g that satisfies Assumption 1, a solution to the differential equation $\dot{\rho} = F_g(\rho)$ can be defined as a flow $\phi : \mathbb{R} \times \overline{\mathbf{Y}} \to \overline{\mathbf{Y}}$. The following definition can be found in Mertikopoulos, Hsieh, and Cevher (2024, Section 4.2):

Definition 3. *Take a flow $\phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$ given some metric space $\mathcal{M}$, and a nonempty compact subset $\mathcal{S} \subseteq \mathcal{M}$. We say*

(1) $\mathcal{S}$ is invariant under ϕ if $\phi_t(\mathcal{S}) = \mathcal{S}$ for all $t \in \mathbb{R}$.

(2) $\mathcal{S}$ is an attractor of ϕ if it admits a neighborhood $\mathcal{W} \subseteq \mathcal{M}$ such that $d(\phi_t(w), \mathcal{S}) \to 0$ uniformly in $w \in \mathcal{W}$ as $t \to \infty$.

*(3) $\mathcal{S}$ is internally chain transitive (*ICT*) if it is invariant and $\phi\big|_{\mathcal{S}}$ has no attractors except $\mathcal{S}$.*

Point (3) is the main object of interest to algorithmic learners. Indeed, one can think of ICT sets as a generalization to periodic orbits of an ordinary differential equation, 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 1. *Impose Assumption 1. Almost surely, $L_{S,g}$ is an ICT set of the differential equation*

$$\dot{\rho} = AF_g(\rho(t)),$$

where $F_g(\rho(t)) = F(\rho(t)) + g(\rho(t))$, and A is a diagonal matrix where all diagonal entries are strictly positive, representing the limiting relative stepsizes of all algorithms.

Proof. The proof is a slight generalization of Borkar (2009, Theorem 2). The approach is to construct a linear interpolation of (2), and show that this will shadow solutions to $\dot{\rho} = AF_g(\rho(t))$ asymptotically, for large enough t. To deal with potentially asymptotically differing stepsizes, take e.g. 1's stepsize schedule α_n^1. By Assumption 1 (ii), we can multiply and divide each algorithm's iteration and write for each i

$$\begin{aligned}\rho_{n+1}^i &= \rho_n^i + \alpha_n^1 \frac{\alpha_n^i}{\alpha_n^1}\left[F(\rho_n) + g(\rho_n) + \delta_n + M_{n+1}\right] \\ &= \alpha_n^1 \bar{\alpha}^i \left[F(\rho_n) + g(\rho_n) + \delta_n + M_{n+1}\right] + \alpha_n^1 o_P(1),\end{aligned}$$

where the last term is due to the vanishing error $\left(\frac{\alpha_n^i}{\alpha_n^1} - \bar{\alpha}^i\right)$, which can be handled analogously to the error terms δ_n, which are discussed below. $\bar{\alpha}^i$ is then the ith diagonal element of scaling matrix A in the statement of this proposition. For notational ease, in the following we then write $\alpha_n = \alpha_n^1$, and proceed with the proof.

Following the notation in Borkar (2009), 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

$$X(\tau_n + s) = \rho_n + s\frac{\rho_{n+1} - \rho_n}{\alpha_{n+1}}, \quad s \in [0, \alpha_{n+1}]. \tag{3}$$

Following the proof of Borkar (2009, Theorem 2), we only need to take care of the additional term δ_n present in iteration (2). We will consider the accumulated δ_n, M_{n+1} error terms. First, note that

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

By Assumption 1, R_n is a standard error term in stochastic approximation theory, satisfying the usual assumptions of Robbins-Monro algorithms with martingale difference noise. It is

a standard result that R_n converges almost surely to zero.[1] We need to take care of the additional term δ_n present in iteration (2). It suffices to show that, for all $T > 0$

$$\sup_{n \leq k \leq m(\tau_n + T) - 1} \Psi_n^k \to 0, \tag{4}$$

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

$$\Psi_n^k \leq \sup_{n \leq k \leq m(\tau_n + T) - 1} \left\| \sum_{i=n}^{k} \alpha_{i+1} \Big(\|\delta_{i+1}\| - \mathbb{E}\big[\|\delta_{i+1}\| \mid \mathcal{F}_{i+1} \big] \Big) \right\| + \sum_{i=n}^{m(\tau_n + T) - 1} \alpha_{i+1} \mathbb{E} \|\delta_{i+1}\| \tag{5}$$

$$= R_{2,n} + K_n, \tag{6}$$

where $\mathcal{F}_i$ is the filtration defined in Assumption 1. Now, by Assumption 1 $(vi) - (viii)$, $R_{2,n}$ 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. As for K_n, recall from Assumption 1 (vi) that $\mathbb{E}\|\delta_n\| = o(b_n)$. Hence, 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 $\max_i \lim_{n \to \infty} \frac{b_n}{\alpha_n^i} = 0$. 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 (4) follows.

Hence, the arguments in Borkar (2009, Lemma 1, Theorem 2) extend to this case as well, which concludes the proof. □

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 1), 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

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

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

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, Melo, 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, Melo, et al. (1982, Proposition 2.18) 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, Theorem 2.15) . To prove convergence to an attractor $\{\rho^g\}$ with positive probability, a stronger result than Proposition 1 is first needed:

Assumption 2 (Condition 11, Faure and Roth (2010))**.** *There exists a map* $\omega : \mathbb{R}^3_+ \to \mathbb{R}_+$ *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 (2):

Lemma 2. *Suppose* δ_n, M_n *satisfy Assumption 1 (i),(ii),(iv). 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 1 (i) on α_n, and (iv) is satisfied for M_n. The remaining term Ψ_n^k can be treated analogously to the proof of Proposition 1. □

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. □

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 4. *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 3) subset of a neighborhood of ρ^g are candidates to being limiting points of the algorithm (2). 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 (2) 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 Michel 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 Michel 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 $\mathcal{C}^1$.

Michel Benaïm and Faure's result is concerned with time-interpolations of iterations such as (2). 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 **??** $(iv), (v)$, the result to be proved here holds true.

In fact, Michel Benaïm and Faure (2012, Hypothesis 2.2) is equivalent[2] to Assumption 2, 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

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

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

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

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

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

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

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. □

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, Melo, 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. □

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. □

3. Numerical Example and Simulations

I visualize the theory developed in the main text and here by considering a numerical example, under a Cournot-competition game. In this game, two algorithms choose quantities x and observe prices y as random signals of their opponent's choices.

The game is set up in line with Abreu, Pearce, and Stacchetti (1986)'s 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[3] 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 follows 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) then being interior.

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

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

conditional $c.d.f.$ and $\eta(y, X)$ of the stage game. η is defined as

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

the sensitivity of the price density to deviations in aggregate quantity.

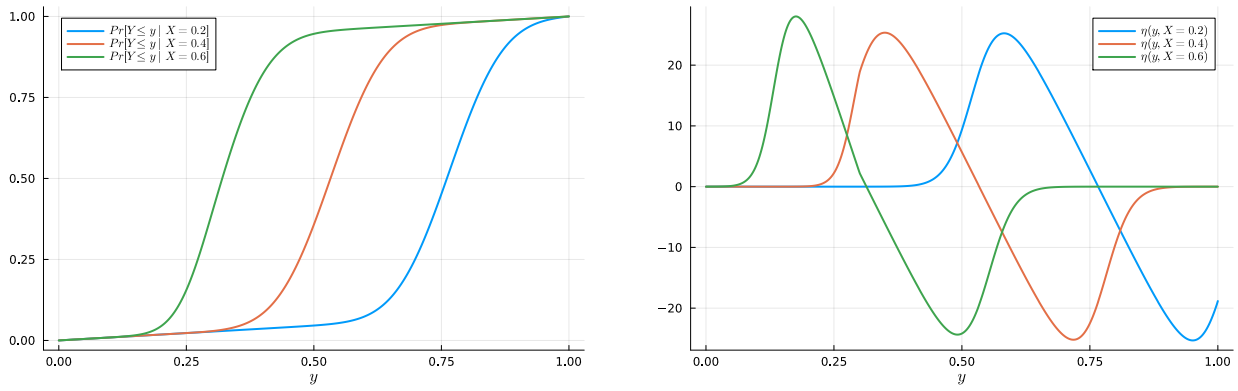

FIGURE 1. Left: C.d.f. conditional on different aggregate quantities. Right: $\eta(y, X)$ for different aggregate quantities X.

One can verify numerically that here, the best-response derivative tf the stage game, $BR_0'(x_N) = -0.39$, implying static stability of x_N. To computationally find the best payoff under bang-bang strategies, V, note first that $g(y; X)$ does not satisfy an MRLP. While for this p.d.f., $\eta(y, X)$ has a unique interior[4] zero, the function 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 4.2 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 $s \neq s'$. The approach of Section 4.2 in the main text (see equation (4) there) can be written as 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 $z \in \mathbf{Y}^4$, as equilibrium sets given threshold tuples z, as those thresholds fully pin down

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

transition probabilities. It is quick to check that Proposition 1 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 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.

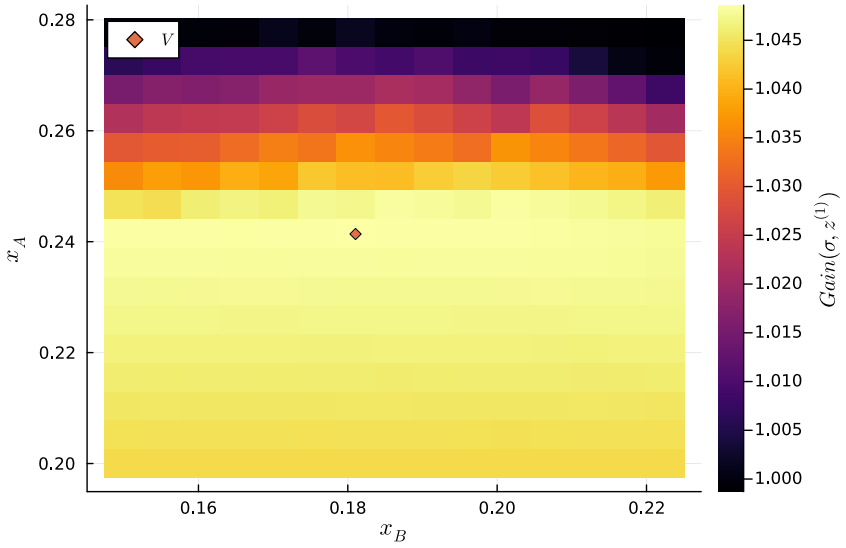

FIGURE 2. $Gain(\sigma, z)$ over varying thresholds $z = (y_a, y_B)$. The orange diamond indicates the location of the overall best equilibrium, V.

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. Based on notation from the main text, we can write for an ACQ learning rule: $\hat{F}_{S,t}(\rho_t)_s = \arg\max_{a' \in \mathbf{X}} Q^i_{t+1}(s, a') - \rho^i_t(s)$, where Q^i_{t+1} is the current period's Q-value function estimate.

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

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 critic estimation part of Q_t of the algorithm given above, I then simulate the reduced form algorithm below:

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 $\alpha_t = t^{-0.6}$, and $M^i_{t+1} \sim N(0, .1)$ is an i.i.d mean-zero Normal noise variable with variance 0.25, so that Assumption 1 holds.

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 take a binary state variable under which ρ_N, the repetition of static Nash, is globally attracting. 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 $\rho(A), \rho(B)$.

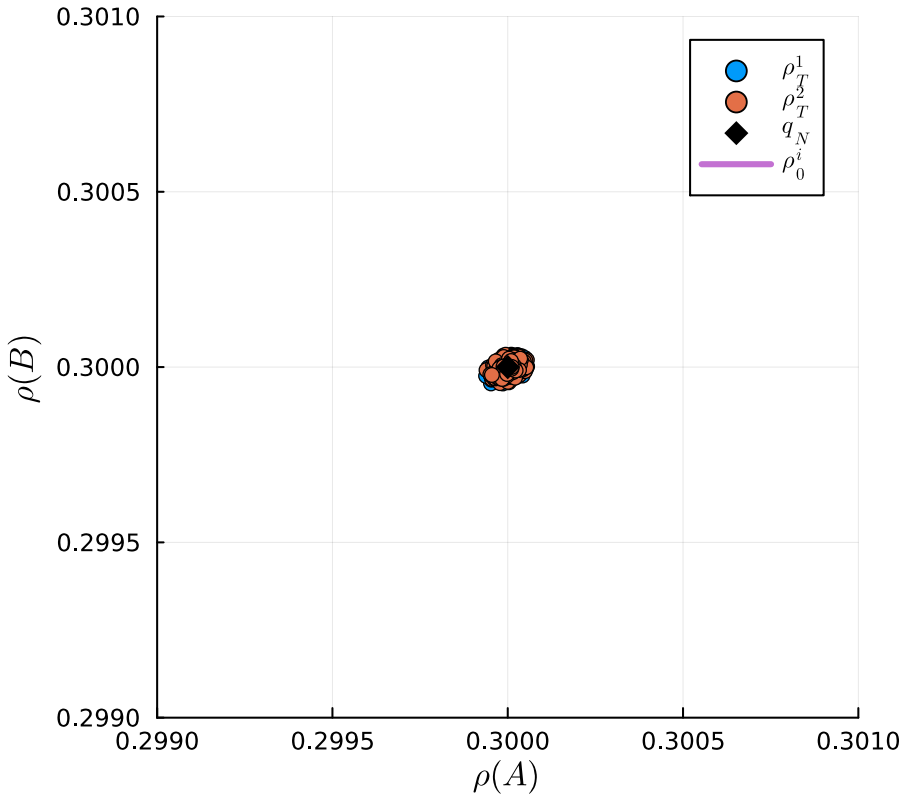

FIGURE 3. Final policies as dots ρ_T^i, for $i = 1, 2$ of 960 simulation runs, with $T = 10^6$. These runs were initialized globally, with ρ_0^i drawn uniformly from $\mathbf{X}^2$ for $i = 1, 2$. All runs converged to a close neighborhood of x_N. Note that the presence of shocks M_{t+1} pushes the process 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^*, the state variable supporting V. Even though the neighborhoods of starting values used in this scenario is the same as under the previous state variable, 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.

The existence of the third symmetric equilibrium is not surprising, as can be seen from the construction of this binary collusive equilibrium in a previous version of the main text, available upon request. In short, one may search for zeros of the payoff gradient along all symmetric quantities in states A and B. When ρ_N is unstable, this gradient must be

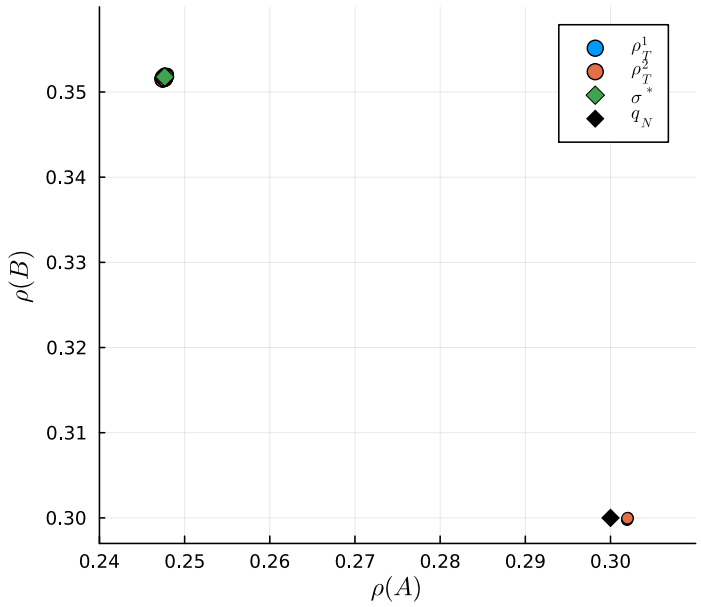

FIGURE 4. Final policies as dots ρ_T^i after a global initialization, with ρ_0^i drawn uniformly from $\mathbf{X}^2$ for $i = 1, 2$, with 960 simulation runs, with $T = 10^6$. 99.8% of runs converged to a neighborhood of the best equilibrium σ^* from these initial values. The remainder, 0.2%, converged 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.

increasing at ρ_N; it must however also be positive for small quantities, and negative for large ones, which leads to the existence of three (symmetric) zeros.

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 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, Holy, and Aschan (2022),Noack et al. (2023),Baran, Foster, et al. (2024),Pal et al. (2024),Mogensen, Villemot, et al. (2020),Isensee, Kornblith, et al. (2024),S. G. 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. Their focus is on finite action games, played for example by Q-learning algorithms. While a main example treated in the main text is an extension of Q-learning , it cannot accommodate Q-learning as a special case, nor is it a special case of Q-learning (ACQ, see discussion below). 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 are not accommodated in 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 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.

J. 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. 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, 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 oligopoly 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, 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.

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.

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

Baran, Mateusz, Claire Foster, et al. (2024). *JuliaArrays/StaticArrays.jl: v1.9.7*.

Benaïm, M (1999). "Dynamics of Stochastic Approximation, Le Seminaire de Probabilite', Springer Lecture Notes in Mathematics". In.

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

Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1, pp. 65–98. URL: https://doi.org/10.1137/141000671.

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.

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.

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.

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

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

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

Isensee, Jonas, Simon Kornblith, et al. (2024). *JuliaIO/JLD2.jl: v0.5.2*.

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

Johnson, Steven G. et al. (2023). *JuliaStrings/LaTeXStrigns.jl: v1.3.1*.

Kittisopikul, Mark, Timothy E. Holy, and Tomas Aschan (2022). *JuliaMath/Interpolations.jl: v0.14.7*.

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

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

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

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

Mertikopoulos, Panayotis, Ya-Ping Hsieh, and Volkan Cevher (2024). "A unified stochastic approximation framework for learning in games". In: *Mathematical Programming* 203.1, pp. 559–609.

Mogensen, Patrick K., Sebastien Villemot, et al. (2020). *JuliaNLSolver/NLsolve.jl: v4.5.1*.

Noack, Andreas et al. (2023). *JuliaParallel/DistributedArrays.jl: v0.6.7*.

Pal, Avik et al. (2024). *SciML/NonlinearSolve.jl: v3.14.0*.

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.

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.

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.