Regret Matching with Finite Memory*

Rene Saran†
Maastricht University, Maastricht, The Netherlands

Roberto Serrano‡
Brown University, Providence, U.S.A.
IMDEA Social Sciences Institute, Madrid, Spain

This version: June 2010

Abstract

We consider the regret matching process with finite memory. For general games in normal form, it is shown that any recurrent class of the dynamics must be such that the action profiles that appear in it constitute a closed set under the "same or better reply" correspondence (CUSOBR set) that does not contain a smaller product set that is closed under "same or better replies," i.e., a smaller PCUSOBR set. Two characterizations of the recurrent classes are offered. First, for the class of weakly acyclic games under better replies, each recurrent class is monomorphic and corresponds to each pure Nash equilibrium. Second, for a modified process with random sampling, if the sample size is sufficiently small with respect to the memory bound, the recurrent classes consist of action profiles that are minimal PCUSOBR sets. Our results are used in a robust example that shows that the limiting empirical distribution of play can be arbitrarily far from correlated equilibria for any large but finite choice of the memory bound.

Keywords: Regret Matching; Nash Equilibria; Closed Sets under Same or Better Replies; Correlated Equilibria.

JEL: C72; C73; D83.

*We thank Antonio Cabrales for helpful comments.

†*Email address:* r.saran@maastrichtuniversity.nl; *Tel:* +31-43-3883763; *Fax:* +31-43-3884878

‡*Email address:* roberto_serrano@brown.edu

1 Introduction

We consider the regret matching process based on finite memory. Each player remembers the last m action profiles used in the game. Regret is calculated with respect to the average payoff obtained in those m periods. With respect to the last action chosen, the player calculates his or her regret from not having used other actions, when those actions replace the last action each time it was used in the m periods that the player recalls. The player switches with positive probability to those actions associated with positive regret, but continues to play the same action also with positive probability. This process corresponds exactly to the regret matching of Hart and Mas-Colell (2000), except that our players' memory is not unbounded.

A typical state of the m-period memory regret learning process is a list of m action profiles. It is shown that any recurrent class of the dynamics must be such that the action profiles that appear in it constitute a closed set under the "same or better reply" correspondence (CUSOBR set) that does not contain a smaller product set that is closed under "same or better replies," i.e., a smaller PCUSOBR set. Since this is only a necessary condition, in general games we are not able to offer a characterization of the recurrent classes of the m-period memory regret learning process. However, we offer two possible ways out that yield a characterization. First, for the class of weakly acyclic games under better replies, each recurrent class is monomorphic and corresponds to each pure Nash equilibrium of the game. Second, for a modified process in which agents sample at random from their bounded memory, if the sample size is sufficiently small with respect to the memory bound, the recurrent classes consist of action profiles that are minimal PCUSOBR sets.

Our findings turn out to shed interesting new light on the results in Hart and Mas-Colell (2000). These authors prove that the empirical distribution of play of their unbounded-memory regret matching process converges almost surely to the set of correlated equilibrium distributions. But as Hart and Mas-Colell (2000) themselves point out, little more is known about additional convergence properties of the empirical play distribution. In contrast, our analysis offers clear pointwise convergence conclusions. In weakly acyclic games, per period play in our process will almost surely in finite time be a pure Nash equilibrium that will be played for ever into the future. In terms of the empirical distribution of play, given the randomness in choosing the arbitrary initial conditions, we know that it is a correlated equilibrium in the convex hull of the pure-strategy Nash equilibria. In general, for any game, pointwise convergence of per period play can only happen to pure Nash

equilibria. But in addition, we know that play in finite time will enter one of the CUSOBR sets of the game not containing a smaller PCUSOBR set. This in turn implies that the empirical distribution converges to a point, which lies in the convex hull of distributions whose supports are CUSOBR sets that do not contain smaller PCUSOBR sets. As it turns out, this may give a long-run prediction far from the set of correlated equilibria, in the following sense. We offer a robust example, in which there is a unique correlated equilibrium distribution, and such that for any finite m –the memory bound– the limiting empirical play distribution concentrates with probability arbitrarily close to 1 on an action profile that is not in the support of the correlated equilibrium. The main result in Hart and Mas-Colell (2000) seems to depend crucially on the unbounded memory assumption.

1.1 Related Literature

Our process is part of the no-regret learning literature (e.g., Hannan (1957), Fudenberg and Levine (1995), Foster and Vohra (1998), Hart and Mas-Colell (2000)).[1] In related processes, Young (1993) shows that if players have bounded recall and play a myopic best reply to a sample drawn from their memory, where the sample is sufficiently small compared to the memory, then in games that are weakly acyclic (under "single best reply"), per period play converges to a pure Nash equilibrium.[2] Young (1998) proves that this learning dynamics converges to a minimal curb –closed under rational behavior– set for generic finite N-player games. In Hurkens (1995), players have bounded recall of m periods and play myopic best replies to their beliefs, where the belief of player i about player j is any distribution with its support in the set of actions played by player j during the last m periods. It is shown that this learning dynamics converges to a minimal curb set in all finite N-player games. Instead of myopic best reply, players in Josephson and Matros (2004) use an imitation dynamics. That is, players have bounded recall, and out of her memory, each player samples all past actions and the corresponding payoffs. She then plays the action that had the highest average payoff in her sample. The recurrent classes of this dynamics in all finite N-player games are monomorphic states and the main result is that the set of stochastically stable monomorphic states is a union of sets that are minimal closed sets under single better replies. Ritzberger and Weibull (1995) prove that the face of a product set (set of all mixed strategies with support in the set) is asymptotically

[1] See Fudenberg and Levine (1998), Hart (2005), Young (2004) or Sandholm (2009) for surveys of learning and related areas.

[2] "Single" means that only one player is allowed to change his or her action at a time.

stable under any sign-preserving selection dynamics (in continuous time) if and only if the set is closed under better replies. Each such set always contains an essential component of Nash equilibria that is strategically stable.

In Zapechelnyuk (2008), an agent is playing against nature. The agent has recall m and her adaptive behavior is a function of her *unconditional* regrets over the last m periods. He assumes that the agent plays according to a better-reply rule, which is defined by the following weak requirement: whenever there exists an action with positive unconditional regret, the agent does not play any action with non-positive unconditional regrets. Unconditional regret matching of Hart and Mas-Colell (2000) is a particular better-reply rule. He provides a 2×3 game example, where the agent is the row player and nature is the column player. He assumes that nature plays according to fictitious play with recall m, i.e., in every period, it plays a best reply to the agent's average play over the last m periods. Under this assumption, he proves that for any better-reply rule and for any large enough recall m, there exists an initial history and period T such that for all $t \geq T$, the probability that the agent's maximum unconditional regret over the last m periods is bounded away from 0 is bounded below by a positive constant. That is, any better-reply rule of the agent with large enough bounded recall is not universally consistent with nature's strategy. Apart from adaptive play being a function of unconditional regrets, the difference with respect to our example below is that nature's adaptive behavior (although a better-reply rule) is not the same as the agent's. In related work to Zapechelnyuk's, Lehrer and Solan (2009) find an adaptive rule with bounded recall that converges to the set of correlated equilibria by "restarting the memory."

Marden et al (2007) consider a regret based dynamics with fading memory and inertia. That is, with a positive probability each player repeats her last period's action and with the rest of the probability she updates her action as a function of her *unconditional* regrets where past regrets are exponentially discounted. Their result is that if players use this learning rule, then in games that are weakly acyclic under single better replies and in which no player is indifferent between distinct strategies, per period play converges to pure Nash equilibrium almost surely.

The rest of the paper is organized as follows. Section 2 describes regret matching with finite memory. Section 3 defines CUSOBR and PCUSOBR sets. We provide the results in Section 4. In Section 5, we discuss the connections with Hart and Mas-Colell (2000). Finally, Section 6 collects the proofs.

2 Regret Matching with Finite Memory

Consider a N-person game in normal form G, with a finite set of actions A_i for each player $i \in N$. Call $A = \prod_{i \in N} A_i$. Let $\pi_i(a_i, a_{-i})$ be her payoff when she chooses a_i and the other players choose a_{-i}.

Suppose that the players remember the last $m \geq 1$ action profiles. At the beginning of period $t+1$, let $(a^{t-m+1}, \ldots, a^t)$ be the history of action profiles played during the last m periods. Player i's average payoff over these m periods is given by $\Pi_i = \frac{1}{m} \sum_{k=t-m+1}^{t} \pi_i(a^k)$. Let a_i^t be the action played by player i in period t. For all $a_i' \neq a_i^t$, let $\Pi_i(a_i')$ be the average payoff over the last m periods that player i would have obtained had she played action a_i' every time she played action a_i^t during the last m periods. That is, $\Pi_i(a_i') = \frac{1}{m} \sum_{k=t-m+1}^{t} \upsilon_i^k(a_i')$, where

$$\upsilon_i^k(a_i') = \begin{cases} \pi_i(a_i', a_{-i}^k) & \text{if } a_i^k = a_i^t \\ \pi_i(a_i^k, a_{-i}^k) & \text{if } a_i^k \neq a_i^t. \end{cases} \tag{1}$$

Define $R_i(a_i') = \Pi_i(a_i') - \Pi_i$. Then, player i switches to action a_i' in period $t+1$ with probability $q(R_i(a_i')) > 0$ if and only if $R_i(a_i') > 0$, whereas she does not switch with the rest of probability, which we assume is positive, i.e., $\sum_{a_i' \neq a_i^t} q(R_i(a_i')) < 1$. This adaptive behavior is regret matching à la Hart and Mas-Colell (2000) but with finite recall.

Define a state of the matched players in a period to be the history of last m action profiles. Hence, the set of states is $H = A^m$.

Given G, for fixed $q(\cdot)$, regret matching with bounded recall describes an aperiodic Markov process $\bar{\mathcal{M}}_G(q)$ on the state space H. We identify its recurrent classes next. A *recurrent class* is a set of states such that if the process reaches one of them, it will never leave the set, and such that it does not admit a proper subset of states with the same property.

3 CUSOBR and PCUSOBR Sets

For any $(a_i, a_{-i}) \in A$, the *set of same-or-better replies* for player i is

$$R_i(a_i, a_{-i}) = \{a_i' \in A_i | \text{either } a_i' = a_i \text{ or } \pi_i(a_i', a_{-i}) > \pi_i(a_i, a_{-i})\}.$$

Let $R_G : A \to A$ be the *same-or-better-reply correspondence* of the game G, i.e.,

$$R_G(a_1, \ldots, a_N) = \prod_{i \in N} R_i(a_i, a_{-i}).$$

Definition 3.1. A set of action profiles $\hat{A} \subseteq A$ in G is *closed under same-or-better replies* (CUSOBR set) if for all $(a_1, \ldots, a_N) \in \hat{A}$, we have $R_G(a_1, \ldots, a_N) \subseteq \hat{A}$. A *minimal CUSOBR set* is a CUSOBR set that does not contain a proper subset that is a CUSOBR set.[3]

For any nonempty $\hat{A} \subseteq A$, define

$$\tilde{R}_G(\hat{A}) = \bigcup_{(a_1,\ldots,a_N) \in \hat{A}} \left(\prod_{i \in N} R_i(a_i, a_{-i}) \right).$$

Equivalently, $\hat{A}$ is a CUSOBR set if and only if $\hat{A}$ is a fixed point of $\tilde{R}_G$, i.e., $\tilde{R}_G(\hat{A}) = \hat{A}$.

It is easy to see that $(a_1, \ldots, a_N)$ is a pure Nash equilibrium of G if and only if $\{(a_1, \ldots, a_N)\}$ is a singleton minimal CUSOBR set. Furthermore, since G has a finite number of action profiles, there exists a minimal CUSOBR set.

Definition 3.2. $\hat{A} \subseteq A$ is a *product set of action profiles that is closed under same-or-better replies* (PCUSOBR set) if $\hat{A}$ is a product set and for all $(a_1, \ldots, a_N) \in \hat{A}$, we have $R_G(a_1, \ldots, a_N) \subseteq \hat{A}$. A *minimal PCUSOBR set* is a PCUSOBR set that does not contain a proper subset that is a PCUSOBR set.

For any nonempty $\hat{A} \subseteq A$, define

$$\hat{R}_G(\hat{A}) = \prod_{i \in N} \left(\bigcup_{(a_i, a_{-i}) \in \hat{A}} R_i(a_i, a_{-i}) \right).$$

Note that a product set $\hat{A}$ is a PCUSOBR set if and only if $\hat{A}$ is a fixed point of $\hat{R}_G$, i.e., $\hat{R}_G(\hat{A}) = \hat{A}$.

Remark: Every minimal CUSOBR set that is a product set is a minimal PCUSOBR set. Thus, in particular, a pure Nash equilibrium is both a singleton minimal CUSOBR set and a singleton minimal PCUSOBR set. Moreover, every minimal PCUSOBR set contains a minimal CUSOBR set. Hence, the set of minimal CUSOBR

[3]See Saran and Serrano (2010), where we study regret matching with one-period memory under fixed and random matching, for a comparison of CUSOBR sets with other set-valued concepts.

	L	R
U	$1,0$	$0,0$
M	$0,1$	$2,0$
D	$2,0$	$0,1$

(a)

	L	C	R
$\hat{U}$	$0,0$	$0,0$	$1,1$
U	$1,0$	$0,-1$	$0,0$
M	$0,1$	$2,0$	$0,0$
D	$2,0$	$0,1$	$0,0$

(b)

Figure 1

sets and minimal PCUSOBR sets coincide in games where all minimal CUSOBR sets are product sets. However, in some games, the set of minimal CUSOBR sets is a refinement of the set of minimal PCUSOBR sets. Game (a) in Figure 1 has a unique minimal CUSOBR set $\{(U,L),(M,L),(M,R),(D,L),(D,R)\}$, which is a refinement of its unique minimal PCUSOBR set $\{(U,L),(U,R),(M,L),(M,R),(D,L),(D,R)\}$. On the other hand, it is also possible that there exists a minimal CUSOBR set that is not a subset of any minimal PCUSOBR set of the game. For example, Game (b) in Figure 1 has a unique minimal PCUSOBR set $\{(\hat{U},R)\}$ but it has two minimal CUSOBR sets, $\{(\hat{U},R)\}$ and $\{(U,L),(M,L),(M,C),(D,L),(D,C)\}$.

4 Results

For any set of states $\hat{H} \subseteq H$, let $A(\hat{H}) \subseteq A$ be the set of all action profiles that are played in some state in $\hat{H}$.

Proposition 4.1. (a) *If $\hat{A}$ is a minimal PCUSOBR set of G, then there exists a recurrent class $\hat{H}$ of $\bar{\mathcal{M}}_G(q)$ such that $A(\hat{H}) \subseteq \hat{A}$.*
(b) *$\hat{H}$ is a recurrent class of $\bar{\mathcal{M}}_G(q)$ only if $A(\hat{H})$ is a CUSOBR set of G that does not contain a smaller PCUSOBR set.*

Remark: Due to inertia in the dynamics, for any $(a_1,\ldots,a_N) \in A(\hat{H})$, there exists a monomorphic state $(a^1,\ldots,a^m) \in \hat{H}$ such that $a^k = (a_1,\ldots,a_N)$ for all $k = 1,\ldots,m$. Hence, if $\hat{H}$ and $\hat{H}'$ are two recurrent classes of $\bar{\mathcal{M}}_G(q)$, then $A(\hat{H}) \bigcap A(\hat{H}') = \emptyset$.

A stronger result can be established if G is weakly acyclic under better replies.[4] A *better-reply graph* is defined as follows: each action profile of G is a vertex of the graph and there exists a directed edge from vertex $(a_1,\ldots,a_N)$ to vertex $(a'_1,\ldots,a'_N)$

[4]Young (2004) defines a smaller class of games that are in fact weakly acyclic under *single* better replies. See Saran and Serrano (2010) for a detailed comparison.

if and only if $(a_1, \ldots, a_N) \neq (a'_1, \ldots, a'_N)$ and $(a'_1, \ldots, a'_N) \in R_G(a_1, \ldots, a_N)$. A *sink* is a vertex with no outgoing edges. A *better-reply path* is a sequence of vertices $(a_1^1, \ldots, a_N^1), \ldots, (a_1^L, \ldots, a_N^L)$ such that there exists a directed edge from each $(a_1^l, \ldots, a_N^l)$ to $(a_1^{l+1}, \ldots, a_N^{l+1})$. The game G is *weakly acyclic under better replies* if from any action profile, there exists at least one better-reply path to a sink. Clearly, an action profile is a sink if and only if it is a pure Nash equilibrium of G. Thus, the game G is weakly acyclic under better replies if from any action profile there exists at least one better-reply path to a pure Nash equilibrium.

If G is weakly acyclic under better replies, then every CUSOBR set contains a pure Nash equilibrium, which is a singleton PCUSOBR set. Hence, we obtain the following corollary:

Corollary 4.2. *Suppose G is weakly acyclic under better replies. Then, $\hat{H}$ is a recurrent class of $\bar{\mathcal{M}}_G(q)$ if and only if $A(\hat{H})$ is a pure Nash equilibrium of G.*

We are, however, not able to strengthen Proposition 4.1 to an "if and only if" statement for games that are not weakly acyclic under better replies, which is the reason to turn to a random sampling version of the process next. That is, we thus far have assumed that players consider all the past periods in the m-period history. Instead, suppose that each player i independently draws a random sample of s periods $(a^1, \ldots, a^s)$ from the m-period history $(a^{t-m+1}, \ldots, a^t)$ and calculates her regrets relative to the latest action in her sample, a_i^s (unlike earlier, where the regrets are calculated relative to the latest action a_i^t).

Formally, let $\Pi_i^s = \dfrac{1}{s}\displaystyle\sum_{k=1}^{s} \pi_i(a^k)$ be player i's average payoff over her s-period sample. For all $a'_i \neq a_i^s$, let $\Pi_i^s(a'_i)$ be the average payoff over these s periods that player i would have obtained had she played action a'_i every time she played action a_i^s during these s periods. That is, $\Pi_i^s(a'_i) = \dfrac{1}{s}\displaystyle\sum_{k=1}^{s} \upsilon_i^k(a'_i)$, where

$$\upsilon_i^k(a'_i) = \begin{cases} \pi_i(a'_i, a_{-i}^k) & \text{if } a_i^k = a_i^s \\ \pi_i(a_i^k, a_{-i}^k) & \text{if } a_i^k \neq a_i^s. \end{cases}$$

Define $R_i^s(a'_i) = \Pi_i^s(a'_i) - \Pi_i^s$. Then, player i plays action a'_i in period $t+1$ with probability $q(R_i^s(a'_i)) > 0$ if and only if $R_i^s(a'_i) > 0$, whereas she does not switch with probability $1 - \sum_{a'_i \neq a_i^s} q(R_i^s(a'_i)) > 0$. This adaptive behavior is regret matching with bounded recall and random sampling.

As before, a state of the matched players in a period is the history of last m

action profiles. Hence, the set of states is still H. Given G, for fixed $q(\cdot)$, regret matching with bounded recall and random sampling describes an aperiodic Markov process $\tilde{\mathcal{M}}_G(q)$ on the state space H.

Proposition 4.3. *If s/m is sufficiently small, then $\hat{H}$ is a recurrent class of $\tilde{\mathcal{M}}_G(q)$ if and only if $A(\hat{H})$ is a minimal PCUSOBR set of G.*

5 Connections with Hart and Mas-Colell's Regret Matching

Hart and Mas-Colell (2000) study the long-run behavior when the players use regret matching but, in contrast to our model, have unbounded memory. Regret matching with unbounded recall is defined as follows: at the beginning of period $t+1$, let $(a^1, \ldots, a^t)$ be the history of action profiles played. The average payoff of player i over this history is given by $\Pi_i^t = \frac{1}{t}\sum_{k=1}^{t} \pi_i(a^k)$. Let a_i^t be the action played by player i in period t. For all $a_i' \neq a_i^t$, let $\Pi_i^t(a_i')$ be the average payoff that player i would have obtained had she played action a_i' every time she played action a_i^t in the history. That is, $\Pi_i^t(a_i') = \frac{1}{t}\sum_{k=1}^{t} v_i^k(a_i')$, where $v_i^k(a_i')$ is as in (1). Define $R_i^t(a_i') = \Pi_i^t(a_i') - \Pi_i^t$. Then, player i switches to action a_i' in period $t+1$ with probability

$$\frac{1}{c}\max\{R_i^t(a_i'), 0\},$$

whereas she does not switch with probability

$$1 - \frac{1}{c}\sum_{a_i' \neq a_i^t} \max\{R_i^t(a_i'), 0\},$$

which is positive for a sufficiently large constant c.

Let μ^t be the empirical distribution of play up to period t, i.e., for every $(a_1, \ldots, a_N)$,

$$\mu^t(a_1, \ldots, a_N) = \frac{1}{t}|\{1 \leq k \leq t | a^k = (a_1, \ldots, a_N)\}|.$$

The main theorem in Hart and Mas-Colell (2000) states the following: If the players use regret matching with unbounded recall, then the empirical distribution of play μ^t converges almost surely as $t \to \infty$ to the set of correlated equilibrium distributions of G. Nevertheless, as Hart and Mas-Colell (2000, p. 1132) themselves

point out, we know little about additional convergence properties of μ^t under regret matching with unbounded recall. In particular, it is not known whether μ^t converges to a "point", i.e., a distribution over the set of action profiles. We know that if there exists a finite time T such that for all $t > T$, μ^t lies in the set of correlated equilibria, then μ^t must be a pure Nash equilibrium for all $t > T$ (because the action profile does not change whenever μ^t is a correlated equilibrium as all regrets are zero). Hence, if μ^t does not converge to a pure Nash equilibrium, then the sequence $\{\mu^t\}_{t\geq 1}$ must lie infinitely often outside the set of correlated equilibria. Therefore, if μ^t converges to a point, then it can either converge to a pure Nash equilibrium or a correlated equilibrium on the boundary of the set of correlated equilibria.

To facilitate the comparison with regret matching with bounded recall (but no sampling), let's fix $q(\cdot)$ to be such that player i switches to action a_i' in period $t+1$ with probability $\frac{1}{c}\max\{R_i(a_i'), 0\}$, whereas she does not switch with probability $1 - \frac{1}{c}\sum_{a_i' \neq a_i^t} \max\{R_i(a_i'), 0\}$, where c is sufficiently large to ensure that the latter is positive.

In contrast to Hart and Mas-Colell (2000), we have precise results about the pointwise convergence of per period play under regret matching with bounded recall. If G is weakly acyclic under better replies, Corollary 4.2 tells us that per period play a^t will almost surely in finite time be a pure Nash equilibrium – the particular equilibrium depends on the initial history. In this case, μ^t converges as $t \to \infty$ to a correlated equilibrium distribution that lies in the convex hull of the pure Nash equilibrium distributions.[5]

More generally, Proposition 4.1 tells us that under regret matching with bounded recall, per period play a^t will almost surely in finite time enter some CUSOBR set that does not contain a smaller PCUSOBR set – again, the particular set depends on the initial history – and after that time, each of the action profiles that belong to this set, and only this set, will be played infinitely often. This in turn implies that μ^t converges as $t \to \infty$ to a point, which lies in the convex hull of distributions whose supports are CUSOBR sets that do not contain smaller PCUSOBR sets. However, as the following example illustrates, the empirical distribution of play need not converge to the set of correlated equilibrium distributions when players have bounded recall.

[5]This follows from the properties of the invariant distributions of the process. Since the recurrent classes of the process coincide with monomorphic states that are pure Nash equilibria, any invariant distribution of the process is a convex combination of the distributions whose supports are such monomorphic states. A similar remark applies to the convergence of μ^t when the game is not weakly acyclic under better replies.

Example 5.1. Suppose there are two players who repeatedly play the game in Figure 2, where $\epsilon \geq 0$.

	L	C	R
U	$0, 20$	$50, 15$	$60, 20$
D	$10, 30$	$40, 35$	$60 + \epsilon, 25$

Figure 2

Fix $m \geq 1$, and let $M(m, \epsilon)$ be the transition matrix of the Markov process when the players use regret matching with bounded recall of m. Let $M_{hh'}(m, \epsilon)$ be the hh' entry in this matrix, i.e., the probability of transition from state $h = (a^1, \ldots, a^m)$ to state $h' = (a'^1, \ldots, a'^m)$ in one period. Note that $a'^k = a^{k+1}$ for all $k = 1, \ldots, m-1$. Let i and j denote, respectively, the row and column players. Since the players choose their actions independently, $M_{hh'}(m, \epsilon) = i_{hh'}(m, \epsilon) j_{hh'}(m, \epsilon)$, where $i_{hh'}(m, \epsilon)$ and $j_{hh'}(m, \epsilon)$ are the probabilities that, respectively, the row player plays action $a_i'^m$ and the column player plays action $a_j'^m$ during the next period conditional on state h. Since $R_j(\cdot)$ does not depend on ϵ, $j_{hh'}(m, \epsilon)$ does not depend on ϵ. Thus, $j_{hh'}(m, \epsilon) = j_{hh'}(m, 0)$ for all ϵ. Similarly, if h is such that $a_j^k \neq R$ for all $k = 1, \ldots, m$, then $i_{hh'}(m, \epsilon) = i_{hh'}(m, 0)$ for all ϵ. So suppose h is such that out of all the periods in which player i played a_i^m, player j played L and R in, respectively, l and r periods.

First, let $a_i^m = U$. Then, conditional on state h, $R_i(D) = \frac{10}{m}(2l + r - m) + \epsilon \frac{r}{m}$. Therefore, if $\epsilon < \min\{\frac{10}{m}, c - 10\}$ (note that $c > 10$ to ensure positive probability of inertia in the process when $\epsilon = 0$), then the probability that the row player switches to D the next period is

$$\begin{array}{cc} \frac{10}{cm}(2l + r - m) + \epsilon \frac{r}{cm} < 1, & \text{if } \frac{10}{m}(2l + r - m) \geq 0 \\ 0, & \text{if } \frac{10}{m}(2l + r - m) < 0. \end{array}$$

Hence, for all $\epsilon < \min\{\frac{10}{m}, c - 10\}$, we have:

- if $a_i'^m = D$, then

$$i_{hh'}(m, \epsilon) = \begin{cases} i_{hh'}(m, 0) + \epsilon \frac{r}{cm} < 1, & \text{if } \frac{10}{m}(2l + r - m) \geq 0 \\ i_{hh'}(m, 0), & \text{otherwise.} \end{cases}$$

- if $a_i'^m = U$, then

$$i_{hh'}(m,\epsilon) = \begin{cases} i_{hh'}(m,0) - \epsilon\frac{r}{cm} > 0, & \text{if } \frac{10}{m}(2l+r-m) \geq 0 \\ i_{hh'}(m,0), & \text{otherwise.} \end{cases}$$

Next, let $a_i^m = D$. Then, conditional on state h, $R_i(U) = \frac{10}{m}(m-2l-r) - \epsilon\frac{r}{m}$. Therefore, if $\epsilon < \frac{10}{m}$ and $c > 10$, then the probability that the row player switches to U the next period is

$$\begin{array}{cc} \frac{10}{cm}(m-2l-r) - \epsilon\frac{r}{cm} \in (0,1), & \text{if } \frac{10}{m}(m-2l-r) > 0 \\ 0, & \text{if } \frac{10}{m}(m-2l-r) \leq 0. \end{array}$$

Hence, for all $\epsilon < \frac{10}{m}$, we have:

- if $a_i'^m = U$, then

$$i_{hh'}(m,\epsilon) = \begin{cases} i_{hh'}(m,0) - \epsilon\frac{r}{cm} > 0, & \text{if } \frac{10}{m}(m-2l-r) > 0 \\ i_{hh'}(m,0), & \text{otherwise.} \end{cases}$$

- if $a_i'^m = D$, then

$$i_{hh'}(m,\epsilon) = \begin{cases} i_{hh'}(m,0) + \epsilon\frac{r}{cm} < 1, & \text{if } \frac{10}{m}(m-2l-r) > 0 \\ i_{hh'}(m,0), & \text{otherwise.} \end{cases}$$

Thus, whenever $\epsilon < \min\{\frac{10}{m}, c-10\}$, there exists a $Q(m)$ such that $M(m,\epsilon) = M(m,0) + \epsilon Q(m)$.

If $\epsilon > 0$, then the set of all action profiles is the game's unique CUSOBR set that does not contain a smaller PCUSOBR set. Hence, it follows from Proposition 4.1 that the Markov process defined by $M(m,\epsilon)$ has a unique recurrent class and hence, a unique invariant distribution, $\mu(m,\epsilon)$. Then,

$$\mu(m,\epsilon) = \mu(m,\epsilon)M(m,\epsilon) = \mu(m,\epsilon)M(m,0) + \epsilon\mu(m,\epsilon)Q(m).$$

There exists a subsequence where $\mu(m,\epsilon)$ converges pointwise to say $\mu(m)$ as $\epsilon \to 0$. Hence, along this subsequence, we have

$$\mu(m) = \lim_{\epsilon\to 0}\mu(m,\epsilon) = \left(\lim_{\epsilon\to 0}\mu(m,\epsilon)\right) M(m,0) = \mu(m)M(m,0).$$

That is, $\mu(m)$ is an invariant distribution of the Markov process defined by $M(m,0)$.

But if $\epsilon = 0$, then $\{(U, R)\}$ is the game's unique CUSOBR set that does not contain a smaller PCUSOBR set; the only other CUSOBR sets are the set A and the set $A \setminus \{(D, R)\}$. Therefore, the Markov process defined by $M(m, 0)$ has a unique invariant distribution, with support on the monomorphic state in which (U, R) is played. Hence, we conclude that for any memory m, there exists an $\epsilon_m > 0$ such that the unique invariant distribution of the Markov process defined by $M(m, \epsilon_m)$ puts probability close to 1 on the monomorphic state in which (U, R) is played.

For any $\epsilon > 0$, the game has a unique correlated equilibrium, in which each of the action profiles (U, L), (U, C), (D, L) and (D, C) has probability equal to 0.25. Thus, fixing a finite m as large as one wishes, we have argued that there exists an ϵ small enough such that the empirical distribution of play of the game as $t \to \infty$ is concentrated on the outcome (U, R) a proportion of time close to 1: this is very "far" from the unique correlated equilibrium distribution of the game.

On the other hand, for any $\epsilon > 0$ and taking $m = \infty$, it follows from the result in Hart and Mas-Colell (2000) that the limiting empirical distribution of play must approximate the unique correlated equilibrium. Our analysis shows that, in obtaining this result, the infinite tail of memory is crucial.

6 Proofs

Proof of Proposition 4.1: Suppose $\hat{A}$ is a PCUSOBR set. Let $P_i(\hat{A})$ be the projection of $\hat{A}$ on A_i. Pick any action profile $(a_1, \ldots, a_N) \in \hat{A}$. Suppose that in period t, the dynamics is in state $(a^{t-m+1}, \ldots, a^t) \in H$ such that $a^k = (a_1, \ldots, a_N), \forall k = t - m + 1, \ldots, t$. We argue by induction that for all $t' \geq t$, the state in period t', $(a^{t'-m+1}, \ldots, a^{t'})$ is such that $a^k \in \hat{A}, \forall k = t' - m + 1, \ldots, t'$. This is clearly true for $t' = t$. Now, suppose this is true for $t'' \geq t$. Consider $a^{t''+1}$. It must be that for all i, either $a_i^{t''+1} = a_i^{t''}$ or there exists a a^k, where $t'' - m + 1 \leq k \leq t''$, such that $a_i^k = a_i^{t''}$ and $\pi_i(a_i^{t''+1}, a_{-i}^k) > \pi_i(a_i^k, a_{-i}^k)$. If $a_i^{t''+1} = a_i^{t''}$, then obviously $a_i^{t''+1} \in P_i(\hat{A})$. On the other hand, since $a^k \in \hat{A}$ (follows from the induction hypothesis) and $\hat{A}$ is a PCUSOBR set, we again have $a_i^{t''+1} \in P_i(\hat{A})$. Since this is true for all i, $a^{t''+1} \in \prod_{i \in N} P_i(\hat{A}) = \hat{A}$, where the equality follows since $\hat{A}$ is a product set, which completes the induction argument.

This implies that starting from period t, any action profile that does not belong to $\hat{A}$ is played with zero probability. Hence, there exists a recurrent class $\hat{H}$ such that $A(\hat{H}) \subseteq \hat{A}$. The first statement in the proposition follows from this fact.

Next, suppose $\hat{H}$ is a recurrent class of $\bar{\mathcal{M}}_G(q)$. We first argue that $A(\hat{H})$

is a CUSOBR set. Pick any action profile $(a_1, \ldots, a_N) \in A(\hat{H})$. There exists a $(a^1, \ldots, a^m) \in \hat{H}$ such that $a^k = (a_1, \ldots, a_N), \forall k = 1, \ldots, m$ (because there is inertia in the dynamics and $\hat{H}$ is a recurrent class). Let $(a'_1, \ldots, a'_N) \in R_G(a_1, \ldots, a_N)$. From state $(a^1, \ldots, a^m)$, there is a positive probability that the dynamics will move to the new state $(a^2, \ldots, a^m, a^{m+1})$, where $a^{m+1} = (a'_1, \ldots, a'_N)$. Since $\hat{H}$ is a recurrent class, it must be that $(a^2, \ldots, a^m, a^{m+1}) \in \hat{H}$ and hence, $(a'_1, \ldots, a'_N) \in A(\hat{H})$.

Now, suppose $A(\hat{H})$ is a CUSOBR set that contains a smaller PCUSOBR set $\hat{A}$. Then there exists a recurrent class of $\bar{\mathcal{M}}_G(q)$, H' such that $A(H') \subseteq \hat{A} \subset A(\hat{H})$, a contradiction. This completes the proof of the second statement in the proposition. □

Proof of Proposition 4.3: As in the previous proof, we can argue that if $\hat{A}$ is a PCUSOBR set, then there exists a recurrent class $\hat{H}$ of $\tilde{\mathcal{M}}_G(q)$ such that $A(\hat{H}) \subseteq \hat{A}$.

We argue that if $\hat{H}$ is a recurrent class of $\tilde{\mathcal{M}}_G(q)$, then $A(\hat{H})$ contains a PCUSOBR set. Pick any action profile $(a_1, \ldots, a_N) \in A(\hat{H})$. Recall the definition of $\hat{R}_G$ and to simplify notation, we instead write $\hat{R}$. Consider the iteration

$$\hat{R}(\{(a_1, \ldots, a_N)\}) \subseteq \hat{R}^2(\{(a_1, \ldots, a_N)\}) \subseteq \ldots \subseteq \hat{R}^l(\{(a_1, \ldots, a_N)\}) \ldots$$

Since the set of action profiles is finite, there exists a finite l' such that for all $l \geq l'$, $\hat{R}^l(\{(a_1, \ldots, a_N)\}) = \hat{R}^{l+1}(\{(a_1, \ldots, a_N)\}) = \tilde{A}$. By construction, $\tilde{A}$ is a PCUSOBR set.

Let $s|A| < m$. Since $\hat{H}$ is a recurrent class, starting at any state in $\hat{H}$, the action profile $(a_1, \ldots, a_N)$ will be played after finite time. Then each player can repeatedly draw a sample in which $a^s = (a_1, \ldots, a_N)$ and therefore, this action profile will be played for the next m periods due to inertia, i.e., there exists a $(a^1, \ldots, a^m) \in \hat{H}$ such that $a^k = (a_1, \ldots, a_N), \forall k = 1, \ldots, m$. Let $(a'_1, \ldots, a'_N) \in \hat{R}(\{(a_1, \ldots, a_N)\}) \setminus \{(a_1, \ldots, a_N)\}$. Starting with state $(a^1, \ldots, a^m)$ in period t, there is a positive probability that $(a'_1, \ldots, a'_N)$ is played for the next s periods. This is because in each $t + k$ period, where $1 \leq k \leq s$, each player can draw a s-period sample in which only $(a_1, \ldots, a_N)$ is played. Let $(a''_1, \ldots, a''_N) \in \hat{R}(\{(a_1, \ldots, a_N)\}) \setminus \{(a_1, \ldots, a_N), (a'_1, \ldots, a'_N)\}$. Starting with period $t+s$, there is a positive probability that $(a''_1, \ldots, a''_N)$ is played for the next s periods. This is because in each $t + s + k$ period, where $1 \leq k \leq s$, each player can again draw a s-period sample in which only $(a_1, \ldots, a_N)$ is played. It is clear that in finite time, we will obtain a history h in which each action profile in $\hat{R}(\{(a_1, \ldots, a_N)\}$ is played for at least s periods. Let

$(\tilde{a}_1, \ldots, \tilde{a}_N) \in \hat{R}^2(\{(a_1, \ldots, a_N)\}) \setminus \hat{R}(\{(a_1, \ldots, a_N)\})$. Hence, for all i, there exists a $(\tilde{a}'_i, \tilde{a}'_{-i}) \in \hat{R}(\{(a_1, \ldots, a_N)\})$ such that $\tilde{a}_i \in R_i(\tilde{a}'_i, \tilde{a}'_{-i})$. In each of the s periods following history h, there is a positive probability that player i will draw a s-period sample in which only $(\tilde{a}'_i, \tilde{a}'_{-i})$ is played. Hence, there is a positive probability that $(\tilde{a}_1, \ldots, \tilde{a}_N)$ will be played during these s periods. Continuing the argument, we see that we will obtain a history $\tilde{h}$ in which all action profiles in $\tilde{A}$ are played at least s times. Since $\hat{H}$ is a recurrent class, history $\tilde{h} \in \hat{H}$. Hence, $\tilde{A} \subseteq A(\hat{H})$.

So far we have argued that: (i) if $\hat{A}$ is a PCUSOBR set, then there exists a recurrent class $\hat{H}$ such that $A(\hat{H}) \subseteq \hat{A}$, and (ii) if $\hat{H}$ is a recurrent class, then $A(\hat{H})$ contains a PCUSOBR set. It follows from these statements that a minimal PCUSOBR set $\hat{A}$ contains a $A(\hat{H})$, where $\hat{H}$ is a recurrent class, which in turn contains a PCUSOBR set $\tilde{A}$. Since $\hat{A}$ is a minimal PCUSOBR set, it must be that $\hat{A} = A(\hat{H})$. On the other hand, if $\hat{H}$ is a recurrent class, then $A(\hat{H})$ contains a PCUSOBR set and hence a minimal PCUSOBR set $\tilde{A}$, which in turn contains a $A(\tilde{H})$, where $\tilde{H}$ is a recurrent class. But $\tilde{H} = \hat{H}$ and hence, $A(\hat{H}) = \tilde{A}$. Thus, the proposition is established. □

References

Foster, D. P., Vohra, R. V., 1998. Asymptotic Calibration. Biometrika 85, 379-390.

Fudenberg, D., Levine, D. K., 1995. Universal Consistency and Cautious Fictitious Play. Journal of Economic Dynamics and Control 19, 1065-1089.

Fudenberg, D., Levine, D. K., 1998. The Theory of Learning in Games. Cambridge: MIT Press.

Hannan, J., 1957. Approximation to Bayes Risk in Repeated Play. In: Dresher, M., et al. (Eds.). Contributions to the Theory of Games III. Princeton: Princeton University Press, 97-139.

Hart, S., 2005. Adaptive Heuristics. Econometrica 73, 1401-1430.

Hart, S., Mas-Colell, A., 2000. A Simple Adaptive Procedure Leading to Correlated Equilibrium. Econometrica 68, 1127-1150.

Hurkens, S., 1995. Learning by Forgetful Players. Games and Economic Behavior 11, 304-329.

Josephson, J., Matros, A., 2004. Stochastic Imitation in Finite Games. Games and Economic Behavior 49, 244-259.

Lehrer, E., Solan, E., 2009. Approachability with Bounded Memory. Games and Economic Behavior 66, 995-1004.

Marden, J. R., Arslan, G., Shamma, J. S., 2007. Regret Based Dynamics: Convergence in Weakly Acyclic Games. AAMAS '07: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. New York: ACM, 194-201.

Ritzberger, K., Weibull, J. W., 1995. Evolutionary Selection in Normal-Form Games. Econometrica 63, 1371-1399.

Sandholm, W. H., 2009. Evolutionary Game Theory. In: Meyers, R. (Ed.). Encyclopedia of Complexity and Systems Science. New York: Springer, 3176-3205.

Saran, R., Serrano, R., 2010. Ex-Post Regret Learning in Games with Fixed and Random Matching: The Case of Private Values. Mimeo, Department of Economics, Brown University.

Young, H. P., 1993. The Evolution of Conventions. Econometrica 61, 57-84.

Young, H. P., 1998. Individual Strategy and Social Structure. Princeton: Princeton University Press.

Young, H. P., 2004. Strategic Learning and its Limits. Oxford: Oxford University Press.

Zapechelnyuk, A., 2008. Better-Reply Dynamics with Bounded Recall. Mathematics of Operations Research 33, 869-879.