Dyn Games Appl (2012) 2:160–175
DOI 10.1007/s13235-011-0021-8

Regret Matching with Finite Memory

Rene Saran · Roberto Serrano

Published online: 9 June 2011

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

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

R. Saran (✉)
Maastricht University, Maastricht, The Netherlands
e-mail: r.saran@maastrichtuniversity.nl

R. Serrano
Brown University, Providence, USA
e-mail: roberto_serrano@brown.edu

R. Serrano
IMDEA Social Sciences Institute, Madrid, Spain

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 to the regret matching of Hart and Mas-Colell [11], 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. We shall refer to such a set as an ω-set, for short. We discuss below the relationships between our CUSOBR sets, PCUSOBR sets, and ω-sets and previous concepts in the literature, such as closed sets under rational behavior (curb sets) or the set of rationalizable action profiles. The investigation of all the properties of ω-sets is beyond our scope here, but they seem to be sets that are distinct from the previous solution concepts mentioned.

Since our first finding 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.

Foster and Vohra [3] and Fudenberg and Levine [8] obtain learning procedures for which the empirical distribution of play converges to the set of correlated equilibria. These procedures are complex, requiring sophisticated updating of beliefs. Hart and Mas-Colell's regret matching was the first particularly simple and intuitive procedure leading to the set of correlated equilibria in all games, a striking result. It turns out, however, that the regret matching process with finite memory may give a long-run prediction far from the set of correlated equilibria, in the following sense. We use our first results to show that for any finite memory bound m, there exists a game such that the limiting empirical distribution of play concentrates with probability arbitrarily close to 1 on an action profile that is not in the support of the game's unique correlated equilibrium. Thus, the main result in Hart and Mas-Colell [11] seems to depend crucially on the unbounded memory assumption. Nevertheless, we do not know whether our finding will be robust to the opposite order of limits. That is, we do not know whether, for a given game, the limit as $m \to \infty$ of the empirical distribution of play in the corresponding limiting process will converge to the set of correlated equilibria. This remains an important open question. On the other hand, our findings are robust to the memory bounds being player dependent.

1.1 Related Literature

Our process is part of the no-regret learning literature (e.g., [4, 6, 9, 11]).[1] In related processes, Young [24] 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

[1] See Fudenberg and Levine [7], Hart [10], Young [26], or Sandholm [19] for surveys of learning and related areas.

memory, then in games that are weakly acyclic (under "single best reply"), per-period play converges almost surely to a pure Nash equilibrium.[2] Young [25] proves that this learning dynamics converges almost surely to a minimal curb set for generic finite N-player games. In contrast, we show that if the players calculate their regrets with respect to a randomly drawn sample from their memory and the sample size is small enough, then the regret matching dynamics converges almost surely in any finite N-player game to a minimal PCUSOBR set.

Similar connections between limits of various learning dynamics and set-valued solution concepts have been discovered in the literature. For instance, Hurkens [12] provides a learning dynamics that converges almost surely to a minimal curb set in all finite N-player games. In Hurkens' dynamics, 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. Instead of myopic best reply, players in Josephson and Matros [13] 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 [18] prove that the face of a product set (i.e., set of all mixed strategies with support in the set) is asymptotically stable under any sign-preserving selection dynamics (in continuous time) if and only if the set is closed under weakly better replies. Each such set always contains an essential component of Nash equilibria that is strategically stable.

There are other better-reply dynamics that yield convergence to pure Nash equilibria in some games. Young [26] considers a better-reply process with finite memory and inertia, in which each player repeats her last action with an exogenous probability, and with the rest of the probability, she chooses an action according to some distribution over the set of actions with positive *unconditional* regret over the finite number of periods she remembers. Young [26] proves that the better-reply process converges almost surely in per-period play to a pure Nash equilibrium in games that are weakly acyclic under *single*-better replies. Friedman and Mezzetti [5] obtain the same convergence result for their better-reply dynamics in which a randomly selected player randomly samples her other actions and switches to the sampled action if it is a better reply to the status quo action profile. Marden et al. [15] study a regret based dynamics with fading—instead of finite—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 a pure Nash equilibrium almost surely. We argue in Sect. 4 that any game that is weakly acyclic under single-better replies is also weakly acyclic under better replies but the converse is not true. Thus, for regret matching with bounded recall, we provide the result of convergence to pure Nash equilibria for a larger class of games.

A couple of other papers in the literature show that evolutionary dynamics need not converge to the set of correlated equilibrium distributions. Viossat [22] gives an example of a 4×4 game in which from an open set of initial conditions, the single action profile in

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

the support of the game's unique correlated equilibrium is eliminated under the replicator dynamics. Viossat [23] generalizes this result to an open set of games and many other deterministic evolutionary dynamics defined on continuous state spaces. However, the unique correlated equilibrium in these games is a strict Nash equilibrium and hence remains asymptotically stable under those evolutionary dynamics. In contrast, we have a stochastic dynamics on a discrete state space. As a result, we are able to show that for *any* initial condition, the limiting empirical distribution of play is concentrated on an action profile that is not in the support of the game's unique correlated equilibrium distribution.

In Zapechelnyuk [27], 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 nonpositive unconditional regrets. Unconditional regret matching of Hart and Mas-Colell [11] 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 that of the agent. In the work related to Zapechelnyuk [27], Lehrer and Solan [14] find an adaptive rule with bounded recall that converges to the set of correlated equilibria by "restarting the memory."

1.2 Plan of the Paper

The rest of the paper is organized as follows. Section 2 describes regret matching with finite memory. Section 3 defines CUSOBR, PCUSOBR, and ω-sets. We provide the results in Sect. 4. In Sect. 5, we discuss the connections with Hart and Mas-Colell [11]. Finally, Sect. 6 collects the proofs.

2 Regret Matching with Finite Memory

Consider an N-person game in normal form G, with a finite set of actions A_i for each player $i \in N$ (we use N to denote both the set and the number of players). Let $A = \prod_{i \in N} A_i$, and let $\pi_i(a_i, a_{-i})$ be the payoff of player i 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 $t+1$, where $t \geq m$, let $(a^{t-m+1}, \ldots, a^t)$ be the history of action profiles played during the last m periods (the initial history, when $t < m$, is formed arbitrarily). The average payoff of player i 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 to be positive, i.e., $\sum_{a_i' \neq a_i'} q(R_i(a_i')) < 1$.[3] This adaptive behavior is regret matching à la Hart and Mas-Colell [11] but with finite recall.[4]

Define a state 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. A *recurrent class* is a set of states such that if the process reaches one of them, it never leaves the set, and such that it does not admit a proper subset of states with the same property. The recurrent classes of $\bar{\mathcal{M}}_G(q)$ are closely related to the sets of action profiles in G that are closed under "same or better replies" of the players. We define these sets in the next section.

3 CUSOBR, PCUSOBR, and ω-Sets

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

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

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

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

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 $B_G(a_1, \ldots, a_N) \subseteq \hat{A}$. A *minimal CUSOBR set* is a CUSOBR set that does not contain a proper subset that is a CUSOBR set.

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

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

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

Sets that are closed under weakly better replies [18] are CUSOBR sets. However, the converse is not true, i.e., a CUSOBR set need not be closed under weakly better replies since a player can have weakly better replies that have the same payoff and hence, do not belong to the set of same-or-better replies. For the same reason, CUSOBR sets are not closed under rational behavior (curb) sets [1].

[3] Several different specifications of $q(\cdot)$ are possible. For instance, let A^* be the maximum number of actions that any player has and $\Delta^* = \max_i (\max_{(a_i, a_{-i})} (\max_{a_i'} \pi_i(a_i', a_{-i}) - \pi_i(a_i, a_{-i})))$. Then any $q(\cdot)$ such that $q(r) \in [0, \frac{1}{A^* \Delta^*})$ and $q(r) > 0 \iff r > 0$ will fulfill these properties. In our analysis, we fix $q(\cdot)$ to be one such function.

[4] Our specification of the switching probabilities is more general as it allows for several "sign-preserving" functional forms. It is only in Sect. 5, where we compare our results with Hart and Mas-Colell [11], that we consider the particular "proportional" functional representation used by Hart and Mas-Colell [11].

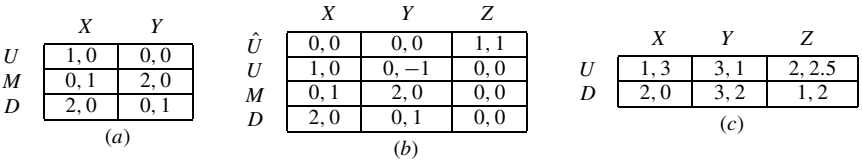

Fig. 1 Games illustrating relationships between different set-valued concepts

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, i.e., $\hat{A} = \prod_{i \in N} \hat{A}_i$, where $\emptyset \neq \hat{A}_i \subseteq A_i \ \forall i \in N$, and for all $(a_1, \ldots, a_N) \in \hat{A}$, we have $B_G(a_1, \ldots, a_N) \subseteq \hat{A}$. 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

$$\hat{B}_G(\hat{A}) = \prod_{i \in N} \Bigg(\bigcup_{(a_i, a_{-i}) \in \hat{A}} B_i(a_i, a_{-i}) \Bigg).$$

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

Since G has a finite number of action profiles, there exist both minimal CUSOBR and minimal PCUSOBR sets in G. It is also 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 and a singleton minimal PCUSOBR set.

Every minimal CUSOBR set that is a product set is a minimal PCUSOBR set. Moreover, every minimal PCUSOBR set contains a minimal CUSOBR set. Hence, the set of minimal CUSOBR 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 Fig. 1 has a unique minimal CUSOBR set $\{(U, X), (M, X), (M, Y), (D, X), (D, Y)\}$, which is a refinement of its unique minimal PCUSOBR set that is equal to the set of all action profiles. 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 Fig. 1 has a unique minimal PCUSOBR set $\{(\hat{U}, Z)\}$, but it has two minimal CUSOBR sets, $\{(\hat{U}, Z)\}$ and $\{(U, X), (M, X), (M, Y), (D, X), (D, Y)\}$.

We introduce one more class of sets before we present our results.

Definition 3.3 A set of action profiles $\hat{A} \subseteq A$ in G is an *ω-set* if $\hat{A}$ is a CUSOBR set that does not contain a proper subset that is a PCUSOBR set.

Clearly, every minimal CUSOBR set or minimal PCUSOBR set is an ω-set. Thus, every game has at least one ω-set. It is also possible that a game has an ω-set that is neither a minimal CUSOBR set nor a minimal PCUSOBR set (see Example 4.3).

Every singleton ω-set is a singleton minimal CUSOBR set and a singleton minimal PCUSOBR set. Hence, a pure Nash equilibrium is a singleton ω-set and vice versa. Thus, the set of pure Nash equilibria of any game G lies in the intersection of the set of its minimal CUSOBR sets and the set of its minimal PCUSOBR sets, and this intersection in turn lies in the set of its ω-sets. In contrast, a pure Nash equilibrium is a curb set or a set that is closed under weakly better replies only if it is strict.

Fig. 2 Fashion Game

	X	Y	Z
X	1, 0	0, 0	0, 1
Y	0, 1	1, 0	0, 0
Z	0, 0	0, 1	1, 0

Another solution concept that is weaker than Nash equilibrium is that of rationalizability [2, 17]. However, there is no logical relation between the sets of actions of a player that are supported in minimal CUSOBR sets, minimal PCUSOBR sets and ω-sets, and the set of her rationalizable actions. Consider Game (a) in Fig. 1. It has two ω-sets, one equal to its unique minimal CUSOBR set and the other equal to its unique minimal PCUSOBR set. Thus, action U of the row player is supported in the game's unique minimal CUSOBR set, unique minimal PCUSOBR set, and both ω-sets. However, U is not rationalizable for the row player since it is strictly dominated by the mixed strategy in which the row player plays M with probability 0.25 and D with probability 0.75. On the other hand, consider Game (c) in Fig. 1. This game has a unique ω-set (and hence, unique minimal CUSOBR set and unique minimal PCUSOBR set) equal to $\{(D, Y)\}$. However, all actions of both players are rationalizable in this game. Since action Y of the column player is weakly dominated by action Z, this example further shows that the sets of actions of a player that are supported in minimal CUSOBR sets, minimal PCUSOBR sets, and ω-sets need not intersect with the set of her undominated actions.

As a final example, consider the game in Fig. 2, which was originally given by Shapley [21] to prove that fictitious play need not converge to Nash equilibrium. Young [25] calls it the *Fashion Game*. To give his description, let X, Y, and Z stand for red, yellow, and blue, respectively. The row player is a fashion follower who likes to imitate the column player. On the other hand, the column player is a fashion leader who prefers to wear a color that contrasts with the other player's choice. Thus, the column player prefers blue when the row player wears red; the column player prefers red when the row player wears yellow, and the column player prefers yellow when the row player wears blue.

In this game, $\{(X, X), (X, Z), (Z, Z), (Z, Y), (Y, Y), (Y, X)\}$ is the unique minimal CUSOBR set and, therefore, an ω-set of the game. This set corresponds to a fashion cycle of red $\rightarrow$ blue $\rightarrow$ yellow $\rightarrow$ red in which the column player keeps defining new fashion and the row player keeps catching up to the former. This game has one more ω-set, which is the set of all action profiles (this is also the game's unique minimal PCUSOBR set). The larger ω-set coincides with the set of all rationalizable action profiles.

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 an ω-set.*

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 $A(\hat{H}) \cap A(\hat{H}') = \emptyset$. Therefore, according to the

Fig. 3 For long memory lengths, recurrent classes can support action profiles outside minimal CUSOBR sets

	X	Y	Z
U	$1, 50$	$0, 0$	$400, 0$
M	$0, 100$	$200, 0$	$0, 100$
D	$2, 0$	$0, 10$	$0, 0$

necessary condition in part (b) of Proposition 4.1, each recurrent class of the regret matching with bounded recall dynamics corresponds to a distinct ω-set of G.

Saran and Serrano [20] provide a complete characterization of the recurrent classes of the dynamics when players have only one-period memory.

Proposition 4.2 (Saran and Serrano [20]) *Suppose $m = 1$. $\hat{H}$ is a recurrent class of $\bar{\mathcal{M}}_G(q)$ if and only if $A(\hat{H})$ is a minimal CUSOBR set of G.*

Thus, each recurrent class of the regret matching dynamics with one-period memory corresponds to a distinct minimal CUSOBR set of G and vice versa. However, as shown in the following example, for longer memory lengths, there can exist a recurrent class $\hat{H}$ such that $A(\hat{H})$ is *not* a minimal CUSOBR set.

Example 4.3 Suppose $m = 100$. Consider the game in Fig. 3. Let the row and column players be denoted by i and j, respectively.

There are three CUSOBR sets in this game: (1) the set of all action profiles A, (2) $A \setminus \{(D, Z)\}$, and (3) $A \setminus \{(U, Y), (D, Z)\}$. The last set $A \setminus \{(U, Y), (D, Z)\}$ is therefore the unique minimal CUSOBR set of the game. Notice that $A \setminus \{(D, Z)\}$ is an ω-set that is neither a minimal CUSOBR set nor a minimal PCUSOBR set.

Since (D, X) belongs to all CUSOBR sets, Proposition 4.1 implies that the regret matching process with memory of 100 periods has a unique recurrent class $\hat{H}$ such that $(D, X) \in A(\hat{H})$. Due to inertia, the monomorphic state $((D, X), \ldots, (D, X)) \in \hat{H}$. Suppose that the dynamics is in this monomorphic state in some period $t \geq 100$. We argue that in finite time, the dynamics will reach a state in which (U, Y) is played.

Period t: State is $(\underbrace{(D, X), \ldots, (D, X)}_{100})$.

In this state, $R_j(Y) > 0$ since the average payoff of player j over the last 100 periods is $\Pi_j = 0$ but her average payoff had she played Y every time she played X in the last 100 periods is $\Pi_j(Y) = 10$. Thus, with a positive probability, player j switches to Y, and player i continues to play D in period $t + 1$.

Period $t + 1$: State is $(\underbrace{(D, X), \ldots, (D, X)}_{99}, (D, Y))$.

In this state, $R_i(M) > 0$ since $\Pi_i = 1.98$ and $\Pi_i(M) = 2$. Thus, with a positive probability, player i switches to M, and player j continues to play Y in period $t + 2$.

Period $t + 2$: State is $(\underbrace{(D, X), \ldots, (D, X)}_{98}, (D, Y), (M, Y))$.

In this state, $R_j(Z) > 0$ since $\Pi_j = 0.1$ and $\Pi_j(Z) = 1$. Thus, with a positive probability, player j switches to Z, and player i continues to play M in period $t + 3$.

Period $t + 3$: State is $(\underbrace{(D, X), \ldots, (D, X)}_{97}, (D, Y), (M, Y), (M, Z))$.

In this state, $R_i(U) > 0$ since $\Pi_i = 3.94$ and $\Pi_i(U) = 5.94$. Thus, with a positive probability, player i switches to U, and player j continues to play Z in period $t+4$.

Period $t+4$: State is $(\underbrace{(D,X),\ldots,(D,X)}_{96},(D,Y),(M,Y),(M,Z),(U,Z))$.

In this state, $R_j(X) > 0$ since $\Pi_j = 1.1$ and $\Pi_j(X) = 1.6$. Thus, with a positive probability, player j switches to X, and player i continues to play U in period $t+5$.

Period $t+5$: State is $(\underbrace{(D,X),\ldots,(D,X)}_{95},(D,Y),(M,Y),(M,Z),(U,Z),(U,X))$.

In this state, $R_j(Y) > 0$ since $\Pi_j = 1.6$ and $\Pi_j(Y) = 10.6$. Thus, with a positive probability, player j switches to Y, and player i continues to play U in period $t+6$.

Period $t+6$: State is $(\underbrace{(D,X),\ldots,(D,X)}_{94},(D,Y),(M,Y),(M,Z),(U,Z),(U,X),(U,Y))$.

Therefore, $(U,Y) \in A(\hat{H})$, but, as already shown, (U,Y) does not belong to the unique minimal CUSOBR set of the game.

4.1 Weakly Acyclic Games

A stronger result can be established if G is weakly acyclic under better replies. 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 $(a'_1,\ldots,a'_N)$ if and only if $(a_1,\ldots,a_N) \neq (a'_1,\ldots,a'_N)$ and $(a'_1,\ldots,a'_N) \in B_G(a_1,\ldots,a_N)$. A *sink* is a vertex with no outgoing edges. A *better-reply path* is a sequence of vertices $(a^1_1,\ldots,a^1_N),\ldots,(a^L_1,\ldots,a^L_N)$ such that there exists a directed edge from each $(a^l_1,\ldots,a^l_N)$ to $(a^{l+1}_1,\ldots,a^{l+1}_N)$. 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 easily obtain the following corollary from Proposition 4.1:

Corollary 4.4 *Suppose that 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.*

Given any directed edge from $(a_1,\ldots,a_N)$ to $(a'_1,\ldots,a'_N)$ in the better-reply graph, we have $a'_i \in B_i(a_1,\ldots,a_N)$ for all $i \in N$. Therefore, more than one players can switch their actions to better replies along any directed edge in the better-reply graph. A *single-better-reply graph* is a spanning subgraph of the better-reply graph (i.e., has the same set of vertices) with only those directed edges such that exactly one player switches her action to a better reply along that edge. A game is *weakly acyclic under single-better replies* if in the single-better-reply graph, there exists a path from any action profile to a pure Nash equilibrium.[5] Since a

[5] Friedman and Mezzetti [5] call this the weak finite improvement property. A stronger condition called the finite improvement property [16] is that there are no directed cycles in the single-better-reply graph.

single-better-reply graph is a subgraph of the better-reply graph, the class of games that are weakly acyclic under single-better replies is a subset of the class of games that are weakly acyclic under better replies. As mentioned in the introduction, Friedman and Mezzetti [5], Young [26], and Marden et al. [15] have studied better-reply dynamics that converge almost surely to pure Nash equilibria in games that are weakly acyclic under single-better replies. Corollary 4.4 implies that per-period play under regret matching with bounded recall converges almost surely to pure Nash equilibria in games that are weakly acyclic under better replies. Thus, we show convergence to pure Nash equilibria in a larger class of games.

Although it is possible to construct examples of several games that are weakly acyclic under better replies but not weakly acyclic under single-better replies (see [20] for one such game), it is nevertheless unclear how much larger the former class is. Some well-known games that are weakly acyclic under single-better replies, and hence also weakly acyclic under better replies and covered by our result, include ordinal potential games [16], supermodular games [5], second-price and first-price auctions, and Bertrand duopolies [20].

4.2 Random Sampling

We are 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 $\Pi_i^s = \frac{1}{s}\sum_{k=1}^{s} \pi_i(a^k)$ be the average payoff of player i 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, $\Pi_i^s(a_i') = \frac{1}{s}\sum_{k=1}^{s} \upsilon_i^k(a_i')$, 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 $1 - \sum_{a_i' \neq a_i^s} q(R_i^s(a_i')) > 0$.[6] This adaptive behavior is regret matching with bounded recall and random sampling.

As before, a state 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 4.5 *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.*

Thus, if the sample size is small enough, then per-period play under regret matching with bounded recall and random sampling will almost surely in finite time enter a minimal PCUSOBR set and then stay inside this set forever with every action profile in the set being played infinitely often.

[6] Again, several specifications of $q(\cdot)$ are possible. See footnote 3.

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

Hart and Mas-Colell [11] 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 $(a^1,\dots,a^t)$ 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, $\Pi_i^t(a_i') = \frac{1}{t}\sum_{k=1}^t \upsilon_i^k(a_i')$, where $\upsilon_i^k(a_i')$ 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

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

whereas she does not switch with probability

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

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 $(a_1,\dots,a_N)$,

$$\mu^t(a_1,\dots,a_N) = \frac{1}{t}\bigl|\bigl\{1\le k\le t | a^k=(a_1,\dots,a_N)\bigr\}\bigr|.$$

The main theorem in Hart and Mas-Colell [11] 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 [11, 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 converge to a pure Nash equilibrium (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 $\{\mu^t\}_{t\ge1}$ 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.[7]

To facilitate the comparison with regret matching with bounded recall (but no sampling), let us 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 [11], we have precise results about per-period play in the long run under regret matching with bounded recall. If G is weakly acyclic under better replies, Corollary 4.4 tells us that per-period play a^t will almost surely in finite time be a

[7] Hart and Mas-Colell [11] did not consider specific games, and it might well be that more precise convergence results are obtained for some classes of games.

Fig. 4 Empirical distribution of play need not converge to correlated equilibria when players have bounded recall

	X	Y	Z
U	$0, 20$	$50, 15$	$60, 20$
D	$10, 30$	$40, 35$	$60+\epsilon, 25$

pure Nash equilibrium—the particular equilibrium depends on the initial history. In general, for any game, pointwise convergence of per-period play can only happen to pure Nash equilibria. But in addition, 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 ω-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. Thus, μ^t will converge almost surely as $t \to \infty$ to a distribution with support over some ω-set. 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.

Example 5.1 Suppose that there are two players who repeatedly play the game in Fig. 4, where $\epsilon \geq 0$.

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 $a_j^k \neq Z$ for all $k = 1, \ldots, m$, then $i_{hh'}(m, \epsilon) = i_{hh'}(m, 0)$ for all ϵ. So suppose that h is such that player i played a_i^m in $\hat{m} \leq m$ periods and in those $\hat{m}$ periods, player j played X and Z, respectively, x and z times.

First, let $a_i^m = U$. Then, conditional on state h, $R_i(D) = \frac{10}{m}(2x + z - \hat{m}) + \epsilon \frac{z}{m}$. 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

$$\begin{cases} \frac{10}{cm}(2x + z - \hat{m}) + \epsilon \frac{z}{cm} < 1 & \text{if } \frac{10}{m}(2x + z - \hat{m}) \geq 0, \\ 0 & \text{if } \frac{10}{m}(2x + z - \hat{m}) < 0. \end{cases}$$

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

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

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

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

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

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

is

$$\begin{array}{ll} \frac{10}{cm}(\hat{m}-2x-z)-\epsilon\frac{z}{cm}\in(0,1) & \text{if } \frac{10}{m}(\hat{m}-2x-z)>0, \\ 0 & \text{if } \frac{10}{m}(\hat{m}-2x-z)\leq 0. \end{array}$$

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

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

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

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

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

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 ω-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

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

That is, $\mu(m)$ is an invariant distribution of the Markov process defined by $M(m,0)$. But if $\epsilon = 0$, then $\{(U,Z)\}$ is the game's unique ω-set; the other CUSOBR sets are A and $A\setminus\{(D,Z)\}$. Therefore, the Markov process defined by $M(m,0)$ has a unique invariant distribution, with support on the monomorphic state in which (U,Z) is played. Thus, 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,Z) is played.

For any $\epsilon > 0$, the game has a unique correlated equilibrium, in which each of the action profiles (U,X), (U,Y), (D,X), and (D,Y) has probability equal to 0.25. Thus, it follows from the above arguments that fixing a finite m as large as one wishes, there exists a small enough ϵ such that the limiting empirical distribution of play in the corresponding game is concentrated on the outcome (U,Z) 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 [11] 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 that $\hat{A}$ 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 that 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 that $\hat{H}$ 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 $(a'_1, \ldots, a'_N) \in B_G(a_1, \ldots, a_N)$. 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 that $A(\hat{H})$ is a CUSOBR set that contains a smaller PCUSOBR set $\hat{A}$. Then there exists a recurrent class H' of $\bar{\mathcal{M}}_G(q)$ 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 4.5 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 $\bar{\mathcal{M}}_G(q)$ such that $A(\hat{H}) \subseteq \hat{A}$.

We argue that if $\hat{H}$ is a recurrent class of $\bar{\mathcal{M}}_G(q)$, 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 $\hat{B}_G$, and to simplify notation, we instead write $\hat{B}$. Consider the iteration

$$\hat{B}\big(\{(a_1, \ldots, a_N)\}\big) \subseteq \hat{B}^2\big(\{(a_1, \ldots, a_N)\}\big) \subseteq \cdots \subseteq \hat{B}^l\big(\{(a_1, \ldots, a_N)\}\big) \cdots.$$

Since the set of action profiles is finite, there exists a finite l' such that for all $l \geq l'$, $\hat{B}^l(\{(a_1, \ldots, a_N)\}) = \hat{B}^{l+1}(\{(a_1, \ldots, a_N)\}) = \tilde{A}$. 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 an $(a^1, \ldots, a^m) \in \hat{H}$ such that $a^k = (a_1, \ldots, a_N)$ $\forall k = 1, \ldots, m$. Let $(a'_1, \ldots, a'_N) \in \hat{B}(\{(a_1, \ldots, a_N)\}) \setminus \{(a_1, \ldots, a_N)\}$. 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 an s-period sample in which only $(a_1, \ldots, a_N)$ is played. Let $(a''_1, \ldots, a''_N) \in \hat{B}(\{(a_1, \ldots, a_N)\}) \setminus \{(a_1, \ldots, a_N), (a'_1, \ldots, a'_N)\}$. 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 an 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 $\hat{B}(\{(a_1,\ldots,a_N)\}$ is played for at least s periods. Let $(\tilde{a}_1,\ldots,\tilde{a}_N) \in \hat{B}^2(\{(a_1,\ldots,a_N)\}) \setminus \hat{B}(\{(a_1,\ldots,a_N)\})$. Hence, for all i, there exists an $(\tilde{a}_i', \tilde{a}_{-i}') \in \hat{B}(\{(a_1,\ldots,a_N)\})$ such that $\tilde{a}_i \in B_i(\tilde{a}_i', \tilde{a}_{-i}')$. In each of the s periods following history h, there is a positive probability that player i will draw an 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 $A(\tilde{H})$, where $\tilde{H}$ is a recurrent class. But $\tilde{H} = \hat{H}$, and hence, $A(\hat{H}) = \tilde{A}$. Thus, the proposition is established. □

Acknowledgements We thank William Sandholm and an anonymous referee for very helpful comments. We also thank Antonio Cabrales, Sergiu Hart, David Levine, Andreu Mas-Colell, Karl Schlag, Andriy Zapechelnyuk, and audience at Fall 2010 Midwest Economic Theory Meeting (Madison) for their comments and encouragement.

References

1. Basu K, Weibull JW (1991) Strategy subsets closed under rational behavior. Econ Lett 36:141–146
2. Bernheim BD (1984) Rationalizable strategic behavior. Econometrica 52:1007–1028
3. Foster DP, Vohra RV (1997) Calibrated learning and correlated equilibrium. Games Econ Behav 21:40–55
4. Foster DP, Vohra RV (1998) Asymptotic calibration. Biometrika 85:379–390
5. Friedman JW, Mezzetti C (2001) Learning in games by random sampling. J Econ Theory 98:55–84
6. Fudenberg D, Levine DK (1995) Universal consistency and cautious fictitious play. J Econ Dyn Control 19:1065–1089
7. Fudenberg D, Levine DK (1998) The theory of learning in games. MIT Press, Cambridge
8. Fudenberg D, Levine DK (1999) Conditional universal consistency. Games Econ Behav 29:104–130
9. Hannan J (1957) Approximation to Bayes risk in repeated play. In: Dresher M et al (eds) Contributions to the theory of games III. Princeton University Press, Princeton, pp 97–139
10. Hart S (2005) Adaptive heuristics. Econometrica 73:1401–1430
11. Hart S, Mas-Colell A (2000) A simple adaptive procedure leading to correlated equilibrium. Econometrica 68:1127–1150
12. Hurkens S (1995) Learning by forgetful players. Games Econ Behav 11:304–329
13. Josephson J, Matros A (2004) Stochastic imitation in finite games. Games Econ Behav 49:244–259
14. Lehrer E, Solan E (2009) Approachability with bounded memory. Games Econ Behav 66:995–1004
15. Marden JR, Arslan G, Shamma JS (2007) Regret based dynamics: convergence in weakly acyclic games. In: AAMAS '07: proceedings of the 6th international joint conference on autonomous agents and multiagent systems. ACM, New York, pp 194–201
16. Monderer D, Shapley LS (1996) Potential games. Games Econ Behav 14:124–143
17. Pearce DG (1984) Rationalizable strategic behavior and the problem of perfection. Econometrica 52:1029–1050
18. Ritzberger K, Weibull JW (1995) Evolutionary selection in normal-form games. Econometrica 63:1371–1399

19. Sandholm WH (2009) Evolutionary game theory. In: Meyers R (ed) Encyclopedia of complexity and systems science. Springer, New York, pp 3176–3205
20. Saran R, Serrano R (2010) Ex-post regret learning in games with fixed and random matching: the case of private values. Working Paper, Brown University. URL: http://www.econ.brown.edu/faculty/serrano/pdfs/wp2010-11.pdf
21. Shapley LS (1964) Some topics in two-person games. In: Dresher M et al (eds) Advances in game theory. Annals of mathematical studies, vol 52. Princeton University Press, Princeton, pp 1–28
22. Viossat Y (2007) The replicator dynamics does not lead to correlated equilibria. Games Econ Behav 59:397–407
23. Viossat Y (2008) Evolutionary dynamics may eliminate all strategies used in correlated equilibrium. Math Soc Sci 56:27–43
24. Young HP (1993) The evolution of conventions. Econometrica 61:57–84
25. Young HP (1998) Individual strategy and social structure. Princeton University Press, Princeton
26. Young HP (2004) Strategic learning and its limits. Oxford University Press, Oxford
27. Zapechelnyuk A (2008) Better-reply dynamics with bounded recall. Math Oper Res 33:869–879