האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

BETTER-REPLY STRATEGIES WITH BOUNDED RECALL

by

ANDRIY ZAPECHELNYUK

Discussion Paper # 449 March 2007

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Better-reply strategies with bounded recall

Andriy Zapechelnyuk [*, †]

Hebrew University and Kyiv School of Economics

March 11, 2007

Abstract

A decision maker (an *agent*) is engaged in a repeated interaction with Nature. The objective of the agent is to guarantee to himself the long-run average payoff as large as the best-reply payoff to Nature's empirical distribution of play, no matter what Nature does. An agent with perfect recall can achieve this objective by a simple better-reply strategy. In this paper we demonstrate that the relationship between perfect recall and bounded recall is not straightforward: An agent with bounded recall may fail to achieve this objective, no matter how long recall he has and no matter what better-reply strategy he employs.

JEL classification: C73; D81; D83

Keywords: Better-reply strategies; regret; bounded recall; fictitious play; approachability

[*] I thank Dean Foster, Sergiu Hart, Eilon Solan, Tymofiy Mylovanov, Peyton Young, and participants of the seminars at the Hebrew University and Tel Aviv University for helpful comments and suggestions. I gratefully acknowledge the financial support from Lady Davis and Golda Meir Fellowship Funds, the Hebrew University.

[†] Center for Rationality, the Hebrew University, Givat Ram, Jerusalem 91904, Israel. *E-mail:* andriy@vms.huji.ac.il

1 Introduction

In every (discrete) period of time a decision maker (for short, an *agent*) makes a decision and, simultaneously, Nature selects a state of the world. The agent receives a payoff which depends on both his action and the state. Nature's behavior is ex-ante unknown to the agent, it may be as simple as an i.i.d. environment or as sophisticated as a strategic play of a rational player. The agent's objective is to select a sequence of decisions which guarantees to him the long-run average payoff as large as the best-reply payoff against Nature's empirical distribution of play, *no matter what Nature does*. A behavior rule of the agent which fulfills this objective is called *universally consistent* [1]: the rule is "consistent" if it is optimized against the empirical play of Nature; the word "universally" refers to its applicability to *any* behavior of Nature.

A range of problems can be described within this framework. One example, known as the *on-line decision problem*, deals with predicting a sequence of states of Nature, where at every period t the agent makes a prediction based on information known before t. The classical problem of predicting the sequence of 0's and 1's with "few" mistakes has been a subject of study in statistics, computer science and game theory for more than 40 years. In a more general problem, an agent's goal is to predict a sequence of states of Nature at least as well as the best expert from a given pool of experts [2] (see Littlestone and Warmuth, 1994; Freund and Schapire, 1996; Cesa-Bianchi et al., 1997; Vovk, 1998). Another example is *no-regret learning* in game-theory. A regret [3] of an agent for action a is his average gain had he played constant action a instead of his actual past play; the agent's goal is to play a sequence of actions so that he has "no regrets" (e.g., Hannan, 1957; Fudenberg and Levine, 1995; Foster and Vohra, 1999; Hart and Mas-Colell, 2000, 2001; Cesa-Bianchi and Lugosi, 2003).

---

[1] The term "universal consistency" is due to Fudenberg and Levine (1995).

[2] By an "expert" we understand a given deterministic on-line prediction algorithm. Thus, "to do as well as the best expert" means to make predictions, on average, as close to the true sequence of states as the best of the given prediction algoritms.

[3] This paper deals with the simplest notion of regret known as *external* (or *unconditional*) regret (see, e.g., Foster and Vohra, 1999).

Action a is called a *better reply* to Nature's empirical play if the agent could have improved upon his average past play had he played action a instead of what he actually played in the past. In this paper, we assume that in every period the agent plays a *better reply* to Nature's past play. The better-reply play is a natural adaptive behavior of an unsophisticated, myopic, non-Bayesian decision maker. The class of better-reply strategies encompasses a big variety of behavior rules, such as fictitious play and smooth fictitious play[4]; Hart and Mas-Colell (2000)'s "no-regret" strategy of playing an action with probability proportional to the regret for that action; the logistic (or exponential-weighted) algorithms used in both game theory and computer science (see Littlestone and Warmuth, 1994; Freund and Schapire, 1996; Cesa-Bianchi et al., 1997; Vovk, 1998); the polynomial (l_p-norm) "no-regret" strategies and potential-based strategies of Hart and Mas-Colell (2001) (see also Cesa-Bianchi and Lugosi, 2003).

The agent is said to have m-recall if he is capable of remembering the play of m last periods, and the empirical frequency of Nature's play to which the agent "better-replies" is the simple average across the time interval not exceeding the last m periods. A special case of agent with perfect recall ($m = \infty$) is well studied in the literature, and universally consistent better-reply strategies of an agent with perfect recall are well known (see Hannan, 1957; Foster and Vohra, 1999; Hart and Mas-Colell, 2000, 2001; Cesa-Bianchi and Lugosi, 2003).

The question that we pose in this paper is whether there are better-reply strategies for an agent with *bounded recall* ($m < \infty$) which are (nearly) universally consistent if the agent has sufficiently large length of recall. We show that an agent with long enough recall can approach the best reply to any i.i.d. environment. However, by a simple example we demonstrate that an agent cannot optimize his average play against general (non-i.i.d.) environment, no matter how long (yet, bounded) recall he has and no matter what better-reply strategy he employs. Formally, we say that a family of better-reply strategies with bounded recall is *asymptotically universally consistent* if for every $\varepsilon > 0$ and every sufficiently large $m = m(\varepsilon)$ an agent with recall length m has an

[4] In the original (Fudenberg and Levine, 1995)'s definition, the smooth fictitious play is not a better-reply strategy; however, certain versions of it (e.g., l_p-norm strategy with large p) are better-reply strategies (see Section 3).

ε-universally consistent strategy in this family. We prove the following statement.

There is no family of bounded-recall better-reply strategies which is asymptotically universally consistent.

The statement is proven by a counterexample. We construct a game where if Nature plays a certain form of the fictitious play, then, regardless of what better-reply strategy the agent uses, for every agent's recall length m the limit play forms a cycle. The average payoff of the agent along the cycle is bounded away from the best-reply payoff by a uniform bound for all m. Intuitively, the reason for a cyclical behavior is that in every period t the agent's learns a new observation, a pair (a_t, ω_t), and forgets another observation, (a_{t-m}, ω_{t-m}). An addition of the new observation shifts, in expectation, the agent's average payoff (across the last m periods) in a "better" direction, however, the loss of (a_{t-m}, ω_{t-m}) shifts it in an arbitrary direction. Since the magnitude of the two effects is the same, $1/m$, it may lead to a cyclical behavior of the play. Note that with unbounded recall, $m = \infty$, the second effect does not exist, i.e., the agent does not forget anything, and, consequently, a cyclical behavior is not possible.

A closely related work of Lehrer and Solan (2003) assumes bounded recall of agents and studies a certain form of a better-reply behavior. Lehrer and Solan describe an ε-universally consistent strategy, where the agent periodically "wipes out" his memory. Comparison of this work with our results brought into our paper an interesting insight that "better memory multiplies regrets": an agent can achieve a better average payoff by not using, or deliberately forgetting some information about the past (see Section 6 for further discussion).

2 Preliminaries

In every discrete period of time $t = 1, 2, \ldots$ a decision maker (or an *agent*) chooses an action, a_t, from a finite set A of actions, and Nature chooses a state, ω_t, from a finite set Ω of states. Let $u : A \times \Omega \to \mathbb{R}$ be the agent's payoff function; $u(a_t, \omega_t)$ is the agent's payoff at period t. Denote by $h_t :=$

$((a_1, \omega_1), \ldots, (a_t, \omega_t))$ the history of play up to t. Let $H_t = (A \times \Omega)^t$ be the set of histories of length t and let $H = \bigcup_{t=1}^{\infty} H_t$ be the set of all histories.

Let $p : H \to \Delta(A)$ and $q : H \to \Delta(\Omega)$ be behavior rules of the agent and Nature, respectively. For every period t, we will denote by $p_{t+1} := p(h_t)$ the next-period mixed action of the agent and by $q_{t+1} := q(h_t)$ the next-period distribution of states of Nature. A pair (p, q) and an initial history h_{t_0} induce a probability measure over H_t for all $t > t_0$.

We assume that the agent does not know q, that is, he plays against an unknown environment. We consider better-reply behavior rules, according to which the agent plays actions which are "better" than his actual past play against the observed empirical behavior of Nature. Formally, for every $a \in A$ and every period t define $R_t^m(a) \in \mathbb{R}_+$ as the average gain of the agent had he played a over the last m periods instead of his actual past play. Namely, let [5]

$$R_t^m(a) = \left[\frac{1}{m} \sum\nolimits_{k=t-m+1}^{t} \left(u(a, \omega_k) - u(a_k, \omega_k)\right)\right]^+ \quad \text{for all } t \geq m$$

and

$$R_t^m(a) = \left[\frac{1}{t} \sum\nolimits_{k=1}^{t} \left(u(a, \omega_k) - u(a_k, \omega_k)\right)\right]^+ \quad \text{for all } t < m.$$

We will refer to $R_t^m(a)$ as the agent's *regret for action* a.

The parameter $m \in \{1, 2, \ldots\} \cup \{\infty\}$ is the agent's length of recall. An agent with a specified m is said to have *m-recall*. We shall distinguish the cases of *perfect recall* ($m = \infty$) and *bounded recall* ($m < \infty$).

Consider an agent with m-recall. Action a is called a better reply to Nature's empirical play if the agent could have improved upon his average past play had he played action a instead of what he actually played in the last m periods.

Definition 1. Action $a \in A$ is a *better-reply action* if $R_t^m(a) > 0$.

A behavior rule is called a better-reply rule if the agent plays only better-reply actions, as long as there are such.

Definition 2. Behavior rule p is a *better-reply rule* if for every period t, whenever $\max_{a \in A} R_t^m(a) > 0$,

$$R_t^m(a) = 0 \quad \Rightarrow \quad p_{t+1}(a) = 0, \quad a \in A.$$

[5] We write $[x]^+$ for the positive part of a scalar x, i.e., $[x]^+ = \max\{0, x\}$.

The focus of our study is how well better-reply rules perform against an unknown, possibly, hostile environment. To assess performance of a behavior rule, we use Fudenberg and Levine (1995)'s criterion of ε-universal consistency defined below.

An agent's behavior rule p is said to be *consistent with* q if the agent's long-run average payoff is at least as large as the best-reply payoff to the average empirical play of Nature which plays q.

Definition 3. Let $\varepsilon > 0$. A behavior rule p of the agent with m-recall is *ε-consistent with* q if for every initial history h_{t_0} there exists T such that for every [6] $t \geq T$

$$\Pr_{(p,q,h_{t_0})}\left[\max_{a\in A} R_t^{\infty}(a) < \varepsilon\right] > 1-\varepsilon.$$

A behavior rule p is *consistent with* q if it is *ε-consistent with* q for every $\varepsilon > 0$.

Let $\mathcal{Q}$ be the class of all behavior rules. An agent's behavior rule p is said to be *universally consistent* if it is consistent with *any* behavior of Nature.

Definition 4. A behavior rule p of the agent with m-recall is (ε-) *universally consistent* if it is (ε-) *consistent with* q for every $q \in \mathcal{Q}$.

3 Perfect recall and prior results

Suppose that the agent has perfect recall ($m = \infty$). This case has been extensively studied in the literature, starting from Hannan (1957), who proved the following theorem. [7]

Theorem 1 (Hannan, 1957). *There exists a better-reply rule which is universally consistent.*

[6] $\Pr_{(p,q,h)}[E]$ denotes the probability of event E induced by strategies p and q, and initial history h.

[7] The statements of theorems of Hannan (1957) and Hart and Mas-Colell (2001) presented in this section are sufficient for this paper, though the authors obtained stronger results.

Hart and Mas-Colell (2000) showed that the following rule is universally consistent:

$$p_{t+1}(a) := \begin{cases} \frac{R_t^\infty(a)}{\sum_{a' \in A} R_t^\infty(a')}, \text{ if } \sum_{a' \in A} R_t^\infty(a') > 0, \\ \text{arbitrary}, \quad \text{otherwise.} \end{cases} \tag{1}$$

According to this rule, the agent assigns probability on action a proportional to his regret for a; if there are no regrets, his play is arbitrary. This result is based on Blackwell (1956)'s Approachability Theorem. We shall refer to p in (1) as the *Blackwell strategy.*

The above result has been extended by Hart and Mas-Colell (2001) as follows. A behavior rule p is called a (*stationary*) *regret-based rule* if for every period t the agent's next-period behavior depends only on the current regret vector. That is, for every history h_t, the next-period mixed action of the agent is a function of $R_t^\infty = (R_t^\infty(a))_{a \in A}$ only: $p_{t+1} = \sigma(R_t^\infty)$. Hart and Mas-Colell proved that among better-reply rules, all "well-behaved" stationary regret-based rules are universally consistent.

Theorem 2 (Hart and Mas-Colell, 2001). *Suppose that a better-reply rule p satisfies the following:*

(i) *p is a stationary regret-based rule given for every t by $p_{t+1} = \sigma(R_t^\infty)$; and*
(ii) *There exists a continuously differential potential $P : \mathbb{R}_+^{|A|} \to \mathbb{R}_+$ such that $\sigma(x)$ is positively proportional to $\nabla P(x)$ for every $x \in \mathbb{R}_+^{|A|}$, $x \neq 0$.*

Then p is universally consistent.

The class of universally consistent behavior rules (or "no regret" strategies) which satisfy conditions of Theorem 2 includes the logistic (or exponential adjustment) strategy used by Littlestone and Warmuth (1994), Freund and Schapire (1996), Cesa-Bianchi et al. (1997), Vovk (1998) and others, its better-reply form is given for every t and every $a \in A$ by

$$p_{t+1}(a) = \frac{\exp(\eta R_t^m(a)) - 1}{\sum_{b \in A} (\exp(\eta R_t^m(b)) - 1)}, \tag{2}$$

$\eta > 0$; the polynomial (l_p-norm) strategies and other strategies based on a separable potential (Hart and Mas-Colell, 2001; Cesa-Bianchi and Lugosi, 2003); the smooth fictitious play [8].

[8] For instance, the l_p-norm strategy with large p, and the strategy (2) with large η

4 Bounded recall and i.i.d. environment

The previous section shows that the universal consistency can be achieved for agents with perfect recall. Considering the perfect recall as the limit of m-recall as $m \to \infty$, one may wonder whether the universal consistency can be approached by bounded-recall agents with sufficiently large m.

We start with a result that establishes existence of better-reply rules which are consistent with any i.i.d. environment. Nature's behavior rule q is called an i.i.d. rule if $q_t = q_{t'}$ for all t, t', independently of the history. Let $\mathcal{Q}_{i.i.d.} \subset \mathcal{Q}$ be the set of all i.i.d. behavior rules. An agent's behavior rule p is said to be *i.i.d. consistent* if it is consistent with any i.i.d. behavior of Nature.

Definition 5. A behavior rule p of the agent with m-recall is (ε-) *i.i.d. consistent* if it is (ε-) *consistent with* q for every $q \in \mathcal{Q}_{i.i.d.}$.

Denote by $\mathcal{P}^m$ the class of all better-reply rules for an agent with m-recall, $m \in \mathbb{N}$. Consider an indexed family of better-reply rules $\mathbf{p} = (p^1, p^2, \ldots)$, where $p^m \in \mathcal{P}^m$, $m \in \mathbb{N}$.

Definition 6. A family $\mathbf{p}$ is *asymptotically i.i.d consistent* if for every $\varepsilon > 0$ there exists m such that for every $m' \geq m$ rule $p^{m'}$ is ε-i.i.d. consistent.

Theorem 3. *There exists a family* $\mathbf{p}$ *of better-reply rules which is asymptotically i.i.d. consistent.*

Proof Let $q^* \in \Delta(\Omega)$ and suppose that $q_t = q^*$ for all t. Denote by $\bar{q}_t^m$ the empirical distribution of Nature's play over the last m periods,

$$\bar{q}_t^m(\omega) = \frac{1}{m} \left| k \in \{t - m + 1, \ldots, t\} : \omega_k = \omega \right|, \quad \omega \in \Omega.$$

Suppose that the agent plays the fictitious play with m-recall. Namely, the agent's next-period play, p_{t+1}^m, assigns probability 1 on an action in $\operatorname{argmax}_{a \in A} u(a, \bar{q}_t^m)$, ties are resolved arbitrarily. Thus, the agent plays in every period a best reply to the average realization of m i.i.d. random variables with mean q^*. Since $\max_{a \in A} u(a, x)$ is continuous in x for $x \in \Delta(\Omega)$, the Law of Large Numbers implies that in every period the agent obtains an expected payoff which is ε_m-close to the best reply payoff to q^* with probability at least $1 - \varepsilon_m$, with

approximate the fictitious play.

$\varepsilon_m \to 0$ as $m \to \infty$. $\square$

5 A negative result

In this section we demonstrate that an agent with bounded recall cannot guarantee his play to be ε-optimized against the empirical play of Nature, no matter how large recall length he has and no matter what better-reply rule he uses.

Definition 7. Family $\mathbf{p} = (p^1, p^2, \ldots)$ of better-reply rules is *asymptotically universally consistent* if for every $\varepsilon > 0$ there exists m such that for every $m' \geq m$ rule $p^{m'}$ is ε-universally consistent.

Theorem 4. *There is no family of better-reply rules which is asymptotically universally consistent.*

The theorem is proven by a counterexample.

	L	M	R
U	1,0	0,1	$0,\frac{3}{4}$
D	0,1	1,0	$0,\frac{3}{4}$

Fig. 1.

Consider a repeated game Γ with the stage game given by Fig. 1, where the row player is the agent and the column player is Nature. For every m denote by p^m and q^m be the behavior rules of the agent and Nature, respectively. We shall show that for every $m_0 \in N$ there exists $m \geq m_0$ such that the following holds.

Suppose that the agent with recall length m and Nature play game Γ. Then for every agent's better-reply rule p^m there exist behavior rule q^m of Nature, initial history h_{t_0} and period T such that for all $t \geq T$

$$\Pr_{(p^m, q^m, h_{t_0})} \left[\max_{a \in \{\mathrm{U},\mathrm{D}\}} R_t^\infty(a) \geq \frac{1}{32} \right] \geq \frac{1}{32}.$$

Let $M = \{4j + 2 | j = 2, 3, \ldots\}$. For every $m \in M$, let p^m be an arbitrary better-reply rule, and let q^m be the fictitious play with m-recall. Namely,

denote by u_N the payoff function of Nature as given by Fig. 1, and denote by $\bar{p}_t$ the empirical distribution of the agent's play over the last m periods,

$$\bar{p}_t(a) = \frac{1}{m} \left| k \in \{t-m+1, \ldots, t\} : a_k = a \right|, \quad a \in A.$$

Then q^m_{t+1} assigns probability 1 to a state in $\underset{\omega \in \{\mathrm{L,M,R}\}}{\operatorname{argmax}} u_N(\bar{p}_t, \omega)$ (ties are resolved arbitrarily). Let P^m be the Markov chain with state space $H^m := (A \times \Omega)^m$ induced by p^m and q^m and an initial state h_{t_0}. A history of the last m periods, $h^m_t \in H^m$ will be called, for short, *history at* t. Denote by $H^m_C \subset H^m$ the set of states generated along the following cycle (Fig. 2). The

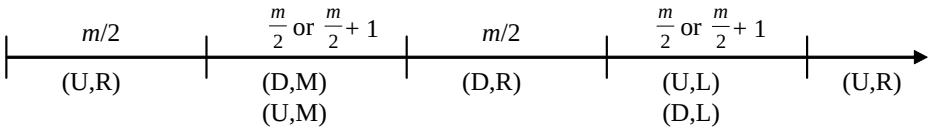

Fig. 2. Closed cycle of Markov chain P^m

cycle has four phases. In two phases labeled (U,R) and (D,R), the play is deterministic, and the duration of each phase is exactly $m/2$ periods. In the two other phases, the play may randomize between two profiles (one written above the other), and the duration of each phase is $m/2$ or $m/2+1$ periods. First, we show that this cycle is closed in P^m, i.e., $h^m_t \in H^m_C$ implies $h^m_{t'} \in H^m_C$ for every $t' > t$.

Lemma 1. *For every $m \in M$, the set H^m_C is closed in P^m.*

The proof is in the Appendix.

Next, we show that the average regrets generated by this cycle are bounded away from zero by a uniform bound for all m.

Lemma 2. *For every $m \in M$, if $h_{t_0} \in H^m_C$, then there exists period T such that for all $t \geq T$*

$$\Pr_{(p^m, q^m, h_{t_0})} \left[\max_{a \in \{U,D\}} R^\infty_t(a) \geq \frac{1}{32} \right] \geq \frac{1}{32}.$$

The proof is in the Appendix. Lemmata 1 and 2 entail the statement of Theorem 4.

Remark 1 In the proof of Theorem 4, Nature plays the fictitious play with m-recall, which is a better-reply strategy for every m. Consequently, an agent

with bounded recall cannot guarantee a nearly optimized behavior even if Nature's behavior is constrained to be in the class of better-reply strategies.

Remark 2 The result can be strengthened as follows. Suppose that whenever an agent has no regrets, then he plays a fully mixed action, i.e.,

$$\max_{a' \in A} R_t^m(a') = 0 \;\Rightarrow\; p_{t+1}^m(a) > 0 \;\text{ for all } a \in A. \tag{3}$$

The next lemma shows that if in game Γ the agent plays a better-reply strategy p^m which satisfies (3) and Nature plays the fictitious play with m-recall, then the Markov chain P^m converges to the cycle H_C^m *regardless of an initial history.* Thus the above negative result is not an isolated phenomenon, it is not peculiar to a small set of initial histories.

Lemma 3. *For every $m \in M$, if p^m satisfies (3), then for every initial history h_{t_0} the process P^m converges to H_C^m with probability* 1.

The proof is in the Appendix.

To see that the statement of Lemma 3 does not hold if p^m fails to satisfy (3), consider again game Γ with the agent playing a better-reply strategy p^m and Nature playing the fictitious play with m-recall, q^m. In addition, suppose that whenever $\max_{a' \in A} R_t^m(a') = 0$, $p_{t+1}^m(\mathrm{U}) = 1$ if t is odd and 0 if t is even. Let t be even and let h_t consist of alternating (UR) and (DR). Clearly, $R_t^m(\mathrm{U}) = R_t^m(\mathrm{D}) = 0$, and Nature's best reply is R, thus, $q_{t+1}(\mathrm{R}) = 1$. The following play is deterministic, alternating between (UR) and (DR) forever.

6 Concluding remarks

We conclude the paper with a few remarks.

1. Why does the better-reply play of an agent with bounded recall fail to exhibit a (nearly) optimized behavior (against Nature's empirical play)?

For every $a \in A$ denote by $v_t(a)$ the one-period regret for action a,

$$v_t(a) = u(a, \omega_t) - u(a_t, \omega_t),$$

and let $v_t = (v_t(a))_{a \in A}$. Since $R_{t-1}^m = \frac{1}{m} \sum_{k=t-m}^{t-1} v_k$, we can consider how the

regret vector changes from period $t-1$ to period t:

$$R_t^m = R_{t-1}^m + \frac{1}{m} v_t - \frac{1}{m} v_{t-m}.$$

Since the play at period t is a better reply to the empirical play over time interval $t-m, \ldots, t-1$, the term $\frac{1}{m} v_t(a)$ shifts the regret vector, in expectation, towards zero, however, the term $-\frac{1}{m} v_{t-m}$ shifts the regret vector in an arbitrary direction. A carefully constructed example, as in Section 5, causes the regret vector to display a cyclical behavior.

2. The following model was introduced by Lehrer and Solan (2003). Suppose that the agent has bounded recall m. Divide the time into blocks of size m: the first block contains periods $1, \ldots, m$, the second block contains periods $m+1, \ldots, 2m$, etc. Let $n(t)$ be the first period of the current block,[9] $n(t) = m \lceil t/m \rceil + 1$. The agent's regret for action $a \in A$ is defined by

$$\hat{R}_t^m(a) = \frac{1}{t - n(t) + 1} \sum\nolimits_{\tau=n(t)}^{t} \left(u(a, \omega_\tau) - u(a_\tau, \omega_\tau) \right). \tag{4}$$

That is, $\hat{R}_t^m(a)$ is the agent's average increase in payoff had he played a constantly instead of his actual past play *within* in the current block. Let p^* be the Blackwell strategy (1) with better replies computed relative to (4). Clearly, this strategy can be implemented by the agent with m-recall. However, the agent behaves as if he remembers only the history of the current block, and at the beginning of a new block he "wipes out" the content of his memory. Notice that the induced probability distribution over histories within every block is identical between blocks and equal to the probability distribution over histories within first m periods in the model with a perfect-recall agent. The Blackwell (1956)'s Approachability Theorem (which is behind the result of Hart and Mas-Colell (2000) on the universal consistency of p^*) gives the rate of convergence of $1/\sqrt{t}$, hence, within each block the agent can approach $1/\sqrt{m}$-best reply to the empirical distribution of Nature's play.

This result is a surprising contrast to the counterexample in Section 5. It shows that *an agent can achieve better average payoff by not using, or deliberately forgetting some information about the past.* Indeed, according to the example presented in Section 5, if the agent uses full information that he remembers,

[9] $\lceil x \rceil$ denotes a number x rounded up to the nearest integer.

the play may eventually enter the cycle with far-from-optimal behavior, no matter with what initial history he starts.

3. Hart and Mas-Colell (2001) used a slightly different notion of better reply. Consider an agent with perfect recall and define for every period t and every $a \in A$

$$D_t^m(a) = \frac{1}{t} \sum_{k=1}^{t} \left(u(a, \omega_k) - u(a_k, \omega_k) \right).$$

Note that $R_t^m(a) = [D_t^m(a)]^+$. Action a is a *strict* better reply (to the empirical distribution of Nature's play) if $D_t^m(a) > 0$ and it is a *weak* better reply if $D_t^m(a) \geq 0$. According to Hart and Mas-Colell, behavior rule p is a better-reply rule if whenever there exist actions which are weak better replies, only such actions are played; formally, whenever $\max_{a \in A} D_t^m(a) \geq 0$,

$$D_t^m(a) < 0 \quad \Rightarrow \quad p_{t+1}(a) = 0, \quad a \in A.$$

The definition of a better-reply rule used in this paper is the same as Hart and Mas-Colell's, except that the word "weak" is replaced by "strict"; formally, whenever $\max_{a \in A} D_t^m(a) > 0$,

$$D_t^m(a) \leq 0 \quad \Rightarrow \quad p_{t+1}(a) = 0, \quad a \in A.$$

These notions are very close, and one does not imply the other. To the best of our knowledge, all specific better-reply rules mentioned in the literature satisfy both notions of better reply. It can be verified that our results remain intact with either notion.

Appendix

A-1 Proof of Lemma 1.

Let $k = \frac{m-2}{4}$. Denote by z_t the empirical distribution of play, that is, for every $(a, \omega) \in A \times \Omega$, $z_t(a, \omega)$ is the frequency of (a, ω) in the history at t,

$$z_t(a, \omega) := \frac{1}{m} \left| \{ \tau \in \{t - m + 1, \ldots, t\} : (a_\tau, \omega_\tau) = (a, \omega) \right|.$$

Let ζ_t be is the frequency of play of U in the last m periods, $\zeta_t = z_t(\text{U,L}) + z_t(\text{U,M}) + z_t(\text{U,R})$.

Fact 1. For every period t,

$$\omega_{t+1} = \begin{cases} \text{L, if } \zeta_t < \frac{1}{4}, \\ \text{M, if } \zeta_t > \frac{3}{4}, \\ \text{R, if } \frac{1}{4} < \zeta_t < \frac{3}{4}. \end{cases}$$

Proof. Note that

$$\begin{aligned} u_N(\bar{p}_t, \text{L}) &= z_t(\text{D,L}) + z_t(\text{D,M}) + z_t(\text{D,R}) = 1 - \zeta_t, \\ u_N(\bar{p}_t, \text{M}) &= z_t(\text{U,L}) + z_t(\text{U,M}) + z_t(\text{U,R}) = \zeta_t, \\ u_N(\bar{p}_t, \text{R}) &= \frac{3}{4}. \end{aligned}$$

Since Nature plays fictitious play, at $t+1$ it selects $\omega_{t+1} \in \underset{\omega \in \{\text{L,M,R}\}}{\text{argmax}}\ u_N(\bar{p}_t, \omega)$. Note that ties never occur, since $m \in M$ and ζ_t is a multiple of $\frac{1}{m}$, thus $\zeta_t \neq \frac{1}{4}$ or $\frac{3}{4}$. □

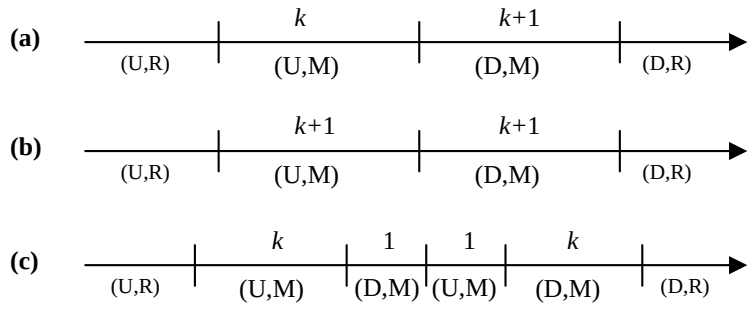

Fig. 3. Three forms of the (U,M)/(D,M) phase

Fact 2. Suppose that $h_t^m \in H_C^m$ such that t is the last period of the (D,R) phase, and suppose that the (U,M)/(D,M) phase preceding the (D,R) phase has form (a), (b) or (c), as shown in Fig. 3. Then the play for the next $m/2$, $m/2+1$, or $m/2+2$ periods constitute the full cycle as shown in Fig. 2, where phases (D,L)/(U,L) and (U,M)/(D,M) have forms [10] (a), (b) or (c).

Proof. Suppose that h_t^m contains $m/2$ (D,R)'s, preceded by the (U,M)/(D,M) phase in form (a), (b), or (c). We shall show that the play in the next $m/2$ or

[10] The forms of the (D,L)/(U,L) phase are symmetric to those of (U,M)/(D,M), obtained by replacement of (U,M) by (D,L) and (D,M) by (U,L).

$m/2+1$ periods constitute phase (D,L)/(U,L) in form (a), (b) or (c), followed by $m/2$ (U,R)'s. Once this is established, by considering the last period of phase (U,R) and repeating the arguments, we obtain Fact 2.

Case 1. Phase (U,M)/(D,M) preceding phase (D,R) has form (a) or (b). Note that whether the (U,M)/(D,M) phase has form (a) or (b), h_t^m is the same, since it contains only $2k+1 \equiv m/2$ last periods of the (U,M)/(D,M) phase. Let t be the last period of the (D,R) phase. We have $\zeta_t = \frac{k}{m} < \frac{1}{4}$, thus by Fact 1, ω_{t+1} =L. Also,

$$\begin{aligned} R_t(\mathrm{U}) &= z_t(\mathrm{D,L}) - z_t(\mathrm{D,M}) = -z_t(\mathrm{D,M}) = -\frac{k+1}{m}, \\ R_t(\mathrm{D}) &= z_t(\mathrm{U,M}) - z_t(\mathrm{U,L}) = z_t(\mathrm{U,M}) = \frac{k}{m}, \end{aligned}$$

hence a_{t+1} =D. Further, in every period $t+j$, $j = 1, \ldots, k$, $(a_{t+j}, \omega_{t+j}) = (\mathrm{D,L})$ is played and $(a_{t+j-m}, \omega_{t+j-m}) = (\mathrm{U,M})$ disappears from the history. At period $t+k$ we have

$$\begin{aligned} R_{t+k}(\mathrm{U}) &= z_{t+k}(\mathrm{D,L}) - z_{t+k}(\mathrm{D,M}) = \frac{k}{m} - \frac{k+1}{m} = -\frac{1}{m}, \\ R_{t+k}(\mathrm{D}) &= z_{t+k}(\mathrm{U,M}) - z_{t+k}(\mathrm{U,L}) = 0 - 0 = 0. \end{aligned}$$

There are no regrets, and therefore both (U,L) and (D,L) may occur at $t+k+1$. Suppose that (D,L) occurs. Since $(a_{t+k-m}, \omega_{t+k-m}) = (\mathrm{D,M})$, it will disappear from the history at $t+k+1$, so, we have

$$\begin{aligned} R_{t+k+1}(\mathrm{U}) &= \frac{k+1}{m} - \frac{k}{m} = \frac{1}{m}, \\ R_{t+k+1}(\mathrm{D}) &= 0 - 0 = 0, \end{aligned}$$

and (U,L) occurs in periods $k+2, \ldots, 2k+2$, until we reach $\zeta_{t+2k+2} = \frac{k+1}{m} > 1/4$. Thus, the phase (D,L)/(U,L) has $k+1$ (D,L)'s, then $k+1$ (U,L)'s, i.e., it takes form (b). If instead at $t+k+1$ action profile (U,L) occurs, then

$$\begin{aligned} R_{t+k+1}(\mathrm{U}) &= \frac{k}{m} - \frac{k}{m} = 0, \\ R_{t+k+1}(\mathrm{D}) &= 0 - \frac{1}{m} = -\frac{1}{m}, \end{aligned}$$

and, again, there are no regrets and both (U,L) and (D,L) may occur at $t+1$. If (U,L) occurs, then

$$R_{t+k+2}(\mathrm{U}) = \frac{k}{m} - \frac{k-1}{m} = \frac{1}{m},$$
$$R_{t+k+1}(\mathrm{D}) = 0 - \frac{2}{m} = -\frac{2}{m},$$

and (U,L) occurs in periods $k+3, \ldots, 2k+1$, until we reach $\zeta_{t+2k+1} = \frac{k+1}{m} > 1/4$. Thus, the phase (D,L)/(U,L) has k (D,L)'s, then $k+1$ (U,L)'s, i.e., it takes form (a). Finally, if at $t+k+2$ (D,L) occurs, then

$$R_{t+k+1}(\mathrm{U}) = \frac{k+1}{m} - \frac{k-1}{m} = \frac{2}{m},$$
$$R_{t+k+1}(\mathrm{D}) = 0 - \frac{1}{m} = -\frac{1}{m},$$

and (U,L) occurs in periods $k+3, \ldots, 2k+2$, until we reach $\zeta_{t+2k+2} = \frac{k+1}{m} > 1/4$. Thus, the phase (D,L)/(U,L) has k (D,L)'s, then single (U,L), then single (D,L), and then k (U,L)'s, i.e., it takes form (c).

Case 2. Phase (U,M)/(D,M) preceding phase (D,R) has form (c). Then, similarly to Case 1, we have $\zeta_t = \frac{k}{m} < \frac{1}{4}$, and (D,L) is deterministically played $k+1$ times, until

$$R_{t+k+1}(\mathrm{U}) = z_{t+k+1}(\mathrm{D,L}) - z_{t+k+1}(\mathrm{D,M}) = \frac{k+1}{m} - \frac{k}{m} = \frac{1}{m},$$
$$R_{t+k+1}(\mathrm{D}) = z_{t+k+1}(\mathrm{U,M}) - z_{t+k+1}(\mathrm{U,L}) = 0 - 0 = 0.$$

After that, (U,L) is played in periods $k+2, \ldots, 2k+2$, until we reach $\zeta_{t+2k+2} = \frac{k+1}{m} > 1/4$. Thus, the phase (D,L)/(U,L) has $k+1$ (D,L)'s and then $k+1$ (U,L)'s, i.e., it takes form (b).

Let $t_1 = t+2k+1$ if the phase (D,L)/(U,L) had form (a) and $t_1 = t+2k+2$ if (b) or (c). Notice that at the end of the phase (D,L)/(U,L) we have $z_{t_1}(\mathrm{U,M}) = z_{t_1}(\mathrm{D,M}) = 0$, hence

$$R_{t_1}(\mathrm{U}) = z_{t_1}(\mathrm{D,L}) - z_{t_1}(\mathrm{D,M}) > 0,$$
$$R_{t_1}(\mathrm{D}) = z_{t_1}(\mathrm{U,M}) - z_{t_1}(\mathrm{U,L}) < 0,$$

Thus, (U,R) is played for the next $m/2 = 2k+1$ periods, until we reach $\zeta_{t_1+m/2} = \frac{3k+2}{m} > 3/4$, and phase (U,M)/(D,M) begins. □

A-2 Proof of Lemma 2.

By Lemma 1, $h_{t_0} \in H_C^m$ implies $h_t^m \in H_C^m$ for all $t > t_0$. Let $h_t^m \in H_C^m$ such that t is the period at the end of the (D,R) phase. Since the history at t contains only (U,M)/(D,M) and (D,R) phases, we have $z_t(\text{D,L}) = z_t(\text{U,L}) = 0$. Also, since at the end of the (D,R) phase the number of U in the history is $\frac{m+2}{4}$, it implies that $z_t(\text{U,M}) = \frac{1}{4} + \frac{1}{2m}$. Therefore,

$$R_t(\text{D}) = z_t(\text{U,M}) - z_t(\text{U,L}) = z_t(\text{U,M}) = \frac{1}{4} + \frac{1}{2m} \equiv C$$

For every period τ, $|R_\tau(\text{D}) - R_{\tau+1}(\text{D})| \leq \frac{2}{m}$, therefore, in periods $t-j$ and $t+j$ the regret for D must be at least $R_t(\text{D}) - 2j/m$. Since the duration of every cycle is at most $2m+2$, the average regret for D during the cycle is at least

$$\frac{1}{2m+2}\left(C + 2\left[\left(C - \frac{2}{m}\right) + \left(C - \frac{4}{m}\right) + \ldots + \left(C - \frac{2(m/4-2)}{m}\right)\right]\right) \geq$$
$$\geq \frac{1}{2m}\left(\frac{m}{2}C - \frac{2}{m}\frac{m^2-4}{32}\right) \geq \frac{1}{32}. \tag{5}$$

Let γ^m be the limit frequency of periods where at least one of the regrets exceeds ε,

$$\gamma^m = \lim_{t\to\infty} \frac{1}{t}\left|\tau \in \{1,\ldots,t\} : \max_{a\in\{\text{U,D}\}} R_\tau^m(a) \geq \varepsilon\right|.$$

Clearly, $\gamma^m > \varepsilon$ implies that for all large enough t

$$\Pr_{(p^m,q^m,h_{t_0})}\left[\max_{a\in\{\text{U,D}\}} R_t^\infty(a) \geq \varepsilon\right] \geq \varepsilon.$$

Combining (5) with the fact that γ^m is at least as large as the average regret for D during the cycle, we obtain $\gamma^m \geq 1/32$. □

A-3 Proof of Lemma 3.

We shall prove that, regardless of the initial history, some event $H_E^m \subset H^m$ occurs infinitely often, and whenever it occurs, the process reaches the cycle, H_C^m, within at most $2m$ periods with strictly positive probability. It follows that the process reaches the cycle with probability 1 from any initial history.

Fact 3. Regardless of an initial state, L and M occur infinitely often.

Proof. Suppose that M never occurs from some time on. Then at any t

$$R_t(\mathrm{U}) = z_t(\mathrm{D,L}) - z_t(\mathrm{D,M}) = z_t(\mathrm{D,L}) \geq 0,$$
$$R_t(\mathrm{D}) = z_t(\mathrm{U,M}) - z_t(\mathrm{U,L}) = -z_t(\mathrm{U,L}) \leq 0.$$

Case 1. $z_t(\mathrm{D,L}) > 0$. Suppose that L occurred last time at $t-j$, $0 \leq j \leq m-1$. After that U must be played with probability 1 in every period $j' = t-j+1, \ldots,$ until frequency of U increases above $\frac{3}{4}$ and, by Fact 1, Nature begins playing M. Contradiction.

Case 2. $z_t(\mathrm{D,L}) = 0$, That is, the agent has no regrets, his play is defined arbitrarily. By assumption (3), $p^m_{t+1}(\mathrm{U}) > 0$, and thus there is a positive probability that U occurs sufficiently many times that the frequency of U increases above $\frac{3}{4}$ and M is played. Contradiction.

The proof that L occurs infinitely often is analogous. □

Fact 4. If ω_t =L and ω_{t+j} =M, then $j > \frac{m}{2}$. Symmetrically, if ω_t =M and ω_{t+j} =L, then $j > \frac{m}{2}$.

Proof. Suppose that ω_t =L, then by Fact 1, $\zeta_{t-1} < \frac{1}{4}$. Clearly, it requires $j > \frac{m}{2}$ periods to reach ζ_{t+j-1} greater than $\frac{3}{4}$, which is required to have ω_{t+j} =M. The second part of the fact is proved analogously. □

Fact 5. Regardless of an initial state, the event $\{\omega_t$ =L and there are no more L in $h^m_t\}$ occurs infinitely often.

Proof. By Fact 3, both L and M occur infinitely often. By Fact 4, the minimal interval of occurrence of L and M is $\frac{m}{2}$, hence if L occurs first time after M, previous occurrence of L is at least $m+1$ periods ago. □

Fact 6. Suppose that ω_t =L and there are no more L in the history. Then after $j < m$ periods we obtain $\frac{1}{4} < \zeta_{t+j} < \frac{1}{4} + \frac{1}{m}$, and with strictly positive probability $R_{t+j}(\mathrm{U}) > 0$ and $R_{t+j}(\mathrm{D}) \leq 0$.

Proof. We have

$$R_t(\mathrm{U}) = z_t(\mathrm{D,L}) - z_t(\mathrm{D,M}),$$
$$R_t(\mathrm{D}) = z_t(\mathrm{U,M}) - z_t(\mathrm{U,L}).$$

By Fact 1, ω_t =L implies $\zeta_{t-1} < \frac{1}{4}$, that is, U occurs at most k times in the history at $t-1$, thus $z_t(\text{U,M}) \leq z_{t-1}(\text{U,M}) \leq \frac{k}{m}$.

Case 1. $R_t(\text{D}) > 0$ and $R_t(\text{U}) > 0$ Then both (D,L) and (U,L) may be played. Since history at $t-1$ does not contain L, regardless of what disappears from the history, we have $R_t(\text{U})$ nondecreasing and $R_t(\text{D})$ nonincreasing. Thus, with positive probability, both (D,L) and (U,L) are played for j periods, until we obtain $\frac{1}{4} < \zeta_{t+j} < \frac{1}{4} + \frac{1}{m}$, $R_{t+j}(\text{U}) > 0$ and $R_{t+j}(\text{D}) \leq 0$. Note that $j < \frac{3}{4}m + 1$, since by Fact 4 the interval between the last occurrence of M and the first occurrence of L is at least $m/2$, thus after period $t + m/2$ there are no M in the history, $R_{t+m/2}(\text{U}) > 0$, $R_{t+m/2}(\text{D}) < 0$, and (U,L) is played at most $k + 1 = \frac{m+2}{4}$ times until the frequency of U becomes above 1/4.

Case 2. $R_t(\text{D}) > 0$, $R_t(\text{U}) \leq 0$. Then (D,L) is played for the next $j' = (z_t(\text{D,L}) - z_t(\text{D,M})) \cdot m + 1$ periods. At period $t + j'$ we have $R_{t+j'}(\text{D}) > 0$ and $R_{t+j'}(\text{U}) > 0$, and proceed similarly to Case 1.

Case 3. $R_t(\text{D}) \leq 0$, $R_t(\text{U}) \leq 0$. That is, the agent has no regrets, his play is defined arbitrarily. By assumption, $p_{t+1}(\text{D}) > 0$, hence there is a positive probability that (D,L) occurs for $j' = z_t(\text{D,M}) \cdot m$ periods which will yield $R_{t+j'}(\text{U}) > 0$, Case 2.

Case 4. $R_t(\text{D}) \leq 0$, $R_t(\text{U}) > 0$. Then (U,L) is played for $j = 1$ or 2 periods (depending whether $(a_t, \omega_t) = (\text{D,L})$ or (U,L)), and we have $\frac{1}{4} < \zeta_{t+j} < \frac{1}{4} + \frac{1}{m}$, $R_{t+j}(\text{U}) = R_t(\text{U}) > 0$ and $R_{t+j}(\text{D}) < R_t(\text{D}) \leq 0$. □

Using Fact 6, we can now analyze the dynamics of the process. Suppose that $\frac{1}{4} < \zeta_t < \frac{1}{4} + \frac{1}{m}$, $R_t(\text{U}) > 0$, $R_t(\text{D}) \leq 0$. Then

I. (U,R) is played in the next $j_{UR} \geq \frac{m}{2}$ periods, and we obtain $\frac{3}{4} < \zeta_{t+j_{UR}} < \frac{3}{4} + \frac{1}{m}$. Since by now M has disappeared from the history, the regrets are

$$R_{t+j_{UR}}(\text{U}) \geq z_t(\text{D,L}) > 0,$$
$$R_{t+j_{UR}}(\text{D}) \leq -z_t(\text{U,L}) \leq 0.$$

II. (U,M) is played for the next $j_{UM} = k + 1$ periods. Since $j_{UR} + j_{UM} \geq \frac{m}{2} + k + 1 = 3k + 1$, it implies that $z_{t+j_{UR}+j_{UM}}(\text{U,L}) \leq k$, and

$$
\begin{aligned}
R_{t+j_{UR}+j_{UM}}(\mathrm{D}) &= z_{t+j_{UR}+j_{UM}}(\mathrm{U,M}) - z_{t+j_{UR}+j_{UM}}(\mathrm{U,L}) \\
&\geq \frac{k+1}{m} - \frac{k}{m} = \frac{1}{m} > 0.
\end{aligned}
$$

III. With positive probability, (D,M) is played for the next $j_{DM} = k+1$ periods, and, since by now L is not in the history, we have

$$
\begin{aligned}
\zeta_{t+j_{UR}+j_{UM}+j_{DM}} &= 1 - \frac{j_{DM}}{m} = \frac{3k+1}{m} < \frac{3}{4}, \\
R_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{U}) &= -z_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{D,M}) < 0, \\
R_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{D}) &= z_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{U,M}) > 0.
\end{aligned}
$$

Notice that at period $t + j_{UR} + j_{UM} + j_{DM}$ the last m periods correspond to phases (U,R) and (U,M)/(D,M) of the cycle (the latter is in form (b)). □

References

Blackwell, D. (1956). An analog of the minmax theorem for vector payoffs. *Pacific Journal of Mathematics 6*, 1–8.

Cesa-Bianchi, N., Y. Freund, D. Helmbold, D. Haussler, R. Shapire, and M. Warmuth (1997). How to use expert advice. *Journal of the ACM 44*, 427–485.

Cesa-Bianchi, N. and G. Lugosi (2003). Potential-based algorithms in on-line prediction and game theory. *Machine Learning 51*, 239–261.

Foster, D. and R. Vohra (1999). Regret in the online decision problem. *Games and Economic Behavior 29*, 7–35.

Freund, Y. and R. Schapire (1996). Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pp. 325–332.

Fudenberg, D. and D. Levine (1995). Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control 19*, 1065–1089.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (Eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, pp. 97–139. Princeton University Press.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*, 1127–1150.

Hart, S. and A. Mas-Colell (2001). A general class of adaptive procedures. *Journal of Economic Theory 98*, 26–54.

Lehrer, E. and E. Solan (2003). No regret with bounded computational capacity. The Center for Mathematical Studies in Economics and Management Science, Northwestern University. Discussion Paper 1373.

Littlestone, N. and M. Warmuth (1994). The weighted majority algorithm. *Information and Computation 108*, 212–261.

Vovk, V. (1998). A game of prediction with expert advice. *Journal of Computer and System Sciences 56*, 153–173.