MATHEMATICS OF OPERATIONS RESEARCH
Vol. 33, No. 4, November 2008, pp. 869–879
ISSN 0364-765X | EISSN 1526-5471 | 08 | 3304 | 0869

DOI 10.1287/moor.1080.0323

Better-Reply Dynamics with Bounded Recall

Andriy Zapechelnyuk
Kyiv School of Economics, 03113 Kyiv, Ukraine,
andriy@vms.huji.ac.il, http://www.gtcenter.org/people/andriy

A decision maker is engaged in a repeated interaction with Nature. The objective of the decision maker is to guarantee to himself the average payoff as large as the best-reply payoff to Nature's empirical distribution of play, no matter what Nature does. The decision maker 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: The decision maker with bounded recall may fail to achieve this objective, no matter how long his recall and no matter what better-reply strategy he uses.

Key words: better-reply dynamics; regret; bounded recall; fictitious play; approachability
MSC2000 subject classification: Primary: 91A35, 91A20; secondary: 91A50
OR/MS subject classification: Primary: decision analysis-sequential; games/group decisions-noncooperative
History: Received October 28, 2007; revised March 20, 2008. Published online in *Articles in Advance* October 10, 2008.

1. Introduction. In every (discrete) period of time a decision maker (for short, *Agent*) makes a decision and, simultaneously, Nature selects a state of the world. Agent receives a payoff that depends on both his action and the state. Nature's behavior is ex ante unknown to Agent; it may be as simple as an i.i.d. environment or as sophisticated as a strategic play of a rational player. Agent's objective is to select a sequence of decisions that guarantees the 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 Agent that 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 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, 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 [16], Freund and Schapire [8], Cesa-Bianchi et al. [6], Vovk [20]). Another example is *no-regret learning* in game-theory. A regret[3] of Agent for action a is his average gain had he played constant action a instead of his actual past play; Agent's goal is to play a sequence of actions so that he has "no regrets" (e.g., Hannan [10], Fudenberg and Levine [9], Foster and Vohra [7], Hart and Mas-Colell [11, 12], Cesa-Bianchi and Lugosi [5]).

Action a is called a *better reply* to Nature's empirical play if Agent could have improved on 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 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 [11]'s "no-regret" strategy of playing an action with probability proportional to the regret for that action; some forms of the logistic (or exponential-weighted) algorithms used in both game theory and computer science (see Littlestone and Warmuth [16], Freund and Schapire [8], Cesa-Bianchi et al. [6], Vovk [20]); the polynomial (l_p-norm) "no-regret" strategies, and potential-based strategies of Hart and Mas-Colell [12] (see also Cesa-Bianchi and Lugosi [5]).

Agent is said to have m-recall if he is capable of remembering the play of m last periods; the empirical frequency of Nature's play to which 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

[1] The term "universal consistency" is from Fudenberg and Levine [9].

[2] Through 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 algorithms.

[3] This paper deals with the simplest notion of regret known as *external* (or *unconditional*) regret (see, e.g., Foster and Vohra [7]).

[4] In the original (Fudenberg and Levine [9])'s definition, the smooth fictitious play is *not* a better-reply strategy; however, certain versions of it, such as the l_p-norm strategy with large p (Hart and Mas-Colell [12], Cesa-Bianchi and Lugosi [5]) are better-reply strategies.

literature, and universally consistent better-reply strategies of Agent with perfect recall are well known (see Hannan [10], Foster and Vohra [7], Hart and Mas-Colell [11, 12], Cesa-Bianchi and Lugosi [5]). The case of bounded-recall strategies is considered by Lehrer and Solan [15] whose work is very close to our paper and will be discussed later on. There is also an extensive literature on bounded-recall strategies (e.g., Lehrer [13], Aumann and Sorin [2], Lehrer [14], Watson [21]) and, more generally, strategies implemented by finite automata (e.g., Aumann [1], Rubinstein [19], Ben-Porath [3], Neyman [17], Neyman and Okada [18]), which studies what equilibria can be achieved (or what payoffs can be guaranteed) in repeated games, extending the Folk Theorem to the case when players have "bounded capacity." In this literature, players are not constrained to such simplistic strategies as playing a better reply to the opponents' average behavior.

The question that we pose in this paper is whether there are better-reply strategies for Agent with *bounded recall* ($m < \infty$) which are (nearly) universally consistent if Agent has sufficiently large length of recall. We show that Agent with long enough recall can approach the best reply to any i.i.d. environment. However, by a simple example we demonstrate that Agent cannot optimize his average play against general (non-i.i.d.) environment, no matter how long (yet, bounded) recall, and no matter what better-reply strategy he uses. 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 $m = m(\varepsilon)$ Agent with recall length m has an 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 Agent with m-recall is allowed to play any better-reply strategy; Nature is assumed to play the *fictitious play with m-recall*, i.e., in every period it plays the best reply to Agent's average play over the last m periods. Thus, given an initial history and strategies of Agent and Nature, the joint play constitutes a finite Markov chain whose state space is the set of all histories of length m. We show that there exists a closed set of states of the Markov chain (which forms a cyclical play over the action profiles in the game), where in every state the average payoff of Agent (over the last m periods) is bounded away from the best-reply payoff by a uniform bound for every finite m. Intuitively, the reason for a cyclical behavior is that in every period t Agent 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, 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: Agent does not forget anything, and, consequently, a cyclical behavior is not possible.

A setting very similar to ours is considered by Lehrer and Solan [15], who also assume bounded recall of a player,[5] however, they do not constrain the player to play a better reply to the opponents' average play over the *full* history within the recall limit. Lehrer and Solan construct an ε-universally consistent strategy where the player periodically "wipes out" his memory. The idea of their strategy is that the player divides time into blocks of size equal to her recall length m, and plays in every period a better-reply to the opponents' average play *within the current block*, behaving as if she recalls only the history of the current block. By contrast, in this paper we prove that *any* better-reply strategy to the average play over the full history *within the recall limit* need not be ε-universally consistent.

The comparison of our result with Lehrer and Solan's [15] leads to the following conclusion: *Sometimes Agent can be better off by not using, or deliberately forgetting, some information about the past*. The analysis of the situation[6] shows that in periods $t = 1, \ldots, m$, when Agent only accumulates information without forgetting anything, he can approach the best reply to the opponent's average play with rate $1/\sqrt{t}$. However, from period $t = m + 1$ on, Agent's memory is full, and in every period he forgets the oldest observation, which can drive his average payoff away from the best reply and lock him in a nonoptimal cyclical play. In this situation, periodic restarts from scratch help Agent to get out of this vicious cycle.

2. Preliminaries. In every discrete period of time $t = 1, 2, \ldots$ 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 Agent's payoff function; $u(a_t, \omega_t)$ is 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.

[5] In fact, Lehrer and Solan [15] deal with a more general problem of the set approachability by bounded-recall strategies or by finite automata in vector-payoff games.

[6] See §6 for more details.

Let $p\colon H \to \Delta(A)$ and $q\colon H \to \Delta(\Omega)$ be behavior rules of Agent and Nature, respectively. For every period t, we will denote by $p_{t+1} := p(h_t)$ the next-period mixed action of 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 Agent does not know q, that is, he plays against an unknown environment. We consider better-reply behavior rules, according to which 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 Agent had he played a over the last m periods instead of his actual past play. Namely, let[7]

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

and

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

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

The parameter $m \in \{1, 2, \dots\} \cup \{\infty\}$ is Agent's length of recall. 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 Agent with m-recall. Action a is called a better reply to Nature's empirical play if Agent could have improved on his average past play had he played action a instead of what he actually played in the last m periods.

Definition 2.1. 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 Agent plays only better-reply actions, as long as there are such.

Definition 2.2. Behavior rule p is a *better-reply rule* if for every period t, whenever $\max_{a\in A} R_t^m(a) > 0$,

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

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 [9]'s criterion of ε-universal consistency defined below.

Agent's behavior rule p is said to be *consistent with* q if Agent's average payoff (over the past that he remembers) tends to be at least as large as the best-reply payoff to the average empirical play of Nature which plays q.

Definition 2.3. Let $\varepsilon > 0$. A behavior rule p of Agent with m-recall is *ε-consistent with q* if for every initial history h_{t_0} there exists T such that for every[8] $t \geq T$

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

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. Agent's behavior rule p is said to be *universally consistent* if it is consistent with *any* behavior of Nature.

Definition 2.4. A behavior rule p of 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 Agent has perfect recall ($m = \infty$). This case has been extensively studied in the literature, starting with Hannan [10], who proved the following theorem.[9]

Theorem 3.1 (Hannan [10]). *There exists a better-reply rule which is universally consistent.*

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

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

[9] The statements of theorems of Hannan [10] and Hart and Mas-Colell [12] presented in this section are sufficient for this paper, though the author obtained stronger results.

Hart and Mas-Colell [11] showed that the following rule is universally consistent:

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

According to this rule, 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 [4]'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 [12] as follows. A behavior rule p is called a (*stationary*) *regret-based rule* if for every period t 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 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 [12] proved that among better-reply rules, all "well-behaved" stationary regret-based rules are universally consistent.

Theorem 3.2 (Hart and Mas-Colell [12]). *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 3.2 includes the logistic (or exponential adjustment) strategy given for every t and every $a \in A$ by

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

$\eta > 0$, used by Littlestone and Warmuth [16], Freund and Schapire [8], Cesa-Bianchi et al. [6], Vovk [20], and others; the smooth fictitious play;[10] the polynomial (l_p-norm) strategies, and other strategies based on a separable potential (Hart and Mas-Colell [12], Cesa-Bianchi and Lugosi [5]).

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 the existence of better-reply rules that 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 $\mathcal{Q}_{\text{i.i.d.}} \subset \mathcal{Q}$ be the set of all i.i.d. behavior rules. 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 4.1. A behavior rule p of Agent with m-recall is (ε-) *i.i.d. consistent* if it is (ε-) *consistent with* q for every $q \in \mathcal{Q}_{\text{i.i.d.}}$.

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 $\mathbf{p} = (p^1, p^2, \dots)$, where $p^m \in \mathcal{P}^m$, $m \in \mathbb{N}$.

Definition 4.2. 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 4.1. *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,

$$\bar{q}_t^m(\omega) = \frac{1}{m} \big| k \in \{t - m + 1, \dots, t\} \colon \omega_k = \omega \big|, \quad \omega \in \Omega.$$

Suppose that Agent plays the fictitious play with m-recall. Namely, Agent's next-period play, p_{t+1}^m, assigns probability 1 on an action in $\arg\max_{a \in A} u(a, \bar{q}_t^m)$, ties are resolved arbitrarily. Thus, 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 uniformly continuous in x for $x \in \Delta(\Omega)$, the Law of Large Numbers implies that in every period 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 $\varepsilon_m \to 0$ as $m \to \infty$. □

[10] See Footnote 4.

TABLE 1. A counterexample.

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

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

DEFINITION 5.1. Family $\mathbf{p} = (p^1, p^2, \dots)$ 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 5.1. *There is no family of better-reply rules that is asymptotically universally consistent.*

The theorem is proven by a counterexample.

Consider a repeated game Γ with the stage game given by Table 1, where the row player is Agent and the column player is Nature. For every m denote by p^m and q^m be the behavior rules of 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 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* $t \geq T$

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

Let $M = \{4j + 2 \mid j = 2, 3, \dots\}$. 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 Table 1, and denote by $\bar{p}_t$ the empirical distribution of Agent's play over the last m periods,

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

Then q_{t+1}^m assigns probability 1 to a state in $\arg\max_{\omega \in \{\mathrm{L}, \mathrm{M}, \mathrm{R}\}} u_N(\bar{p}_t, \omega)$ (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, $h_t^m \in H^m$ will be called, for short, *history at* t. Denote by $H_C^m \subset H^m$ the set of states generated along the following cycle (Figure 1).

The 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., $h_t^m \in H_C^m$ implies $h_{t'}^m \in H_C^m$ for every $t' > t$.

LEMMA 5.1. *For every* $m \in M$, *the set* H_C^m *is closed in* P^m.

The proof is in the appendix.

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

LEMMA 5.2. *For every* $m \in M$, *if* $h_{t_0} \in H_C^m$, *then there exists period* T *such that for all* $t \geq T$

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

The proof is in the appendix. Lemmas 5.1 and 5.2 entail the statement of Theorem 5.1.

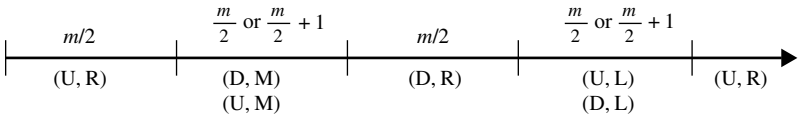

FIGURE 1. Closed cycle of Markov chain P^m.

Remark 1. In the proof of Theorem 5.1, Nature plays the fictitious play with m-recall, which is a better-reply strategy for every m. Consequently, 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 Agent has no regrets, he plays a fully mixed action, i.e.,

$$\max_{a' \in A} R_t^m(a') = 0 \;\Rightarrow\; p_{t+1}^m(a) > 0 \quad \text{for all } a \in A. \tag{2}$$

The next lemma shows that if in game Γ Agent plays a better-reply strategy p^m which satisfies (2) 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 5.3. *For every $m \in M$, if p^m satisfies (2), 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 5.3 does not hold if p^m fails to satisfy (2), consider again game Γ with 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 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 $R_{t-1}^m = (1/m) \sum_{k=t-m}^{t-1} v_k$, 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 $(1/m)v_t(a)$ shifts the regret vector, in expectation, towards zero, however, the term $-(1/m)v_{t-m}$ shifts the regret vector in an arbitrary direction. A carefully constructed example, as in §5, causes the regret vector to display a cyclical behavior.

2. The following behavior rule was introduced by Lehrer and Solan [15]. Suppose that 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,[11] $n(t) = m\lceil t/m \rceil + 1$. Agent's regret for action $a \in A$ is defined by

$$\hat{R}_t^m(a) = \frac{1}{t - n(t) + 1} \sum_{\tau = n(t)}^{t} (u(a, \omega_\tau) - u(a_\tau, \omega_\tau)).$$

That is, $\hat{R}_t^m(a)$ is Agent's average increase in payoff had he played a constantly instead of his actual past play *within* in the current block. Let $\hat{R}_t^m(a) = (\hat{R}_t^m(a))_{a \in A}$, and let p^m be a behavior rule, where in every period t, p_{t+1}^m is the function of $\hat{R}_t^m$ only,[12] $p_{t+1}^m = \sigma(\hat{R}_t^m)$. Clearly, this rule can be implemented by Agent with m-recall. However, 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. Lehrer and Solan [15] show that for every $\varepsilon > 0$ and large enough m there exists an m-recall ε-universally consistent rule p^m. Indeed, let p^m be the Blackwell [4] strategy (1) with R_t^∞ replaced by $\hat{R}_t^m$. Note that the induced probability distribution over histories within every block is identical to the probability distribution over histories within first m periods in the model

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

[12] Note that the described rule is nonstatonary, as p_{t+1}^m actually depends on the starting period of the current block. Lehrer and Solan [15] also construct a *stationary* rule of the same kind, where the beginning of the block is "marked" by a specific sequence of actions that is unlikely to occur in the course of a regular better-reply play.

with a perfect-recall agent. The Blackwell [4]'s Approachability Theorem (which is behind the result of Hart and Mas-Colell [11] on the universal consistency of p^m) gives the rate of convergence of $1/\sqrt{t}$, hence, within each block 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 §5. It shows that *Agent can achieve a better average payoff by not using, or deliberately forgetting, some information about the past*. Indeed, according to the example presented in §5, if Agent uses full information that he remembers, the play may eventually enter the cycle with far-from-optimal behavior. A deliberate forgetting of past information may help Agent to get out of this cyclical behavior.

3. Hart and Mas-Colell [12] used a slightly different notion of better reply. Consider Agent with perfect recall and define for every period t and every $a \in A$

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

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 a *weak* better reply if $D_t^m(a) \geq 0$. According to Hart and Mas-Colell [12], behavior rule p is a better-reply rule if, whenever there exist actions that are weak better replies, only such actions are played; formally, whenever $\max_{a\in A} D_t^m(a) \geq 0$,

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

The definition of a better-reply rule used in this paper is the same as Hart and Mas-Colell's [12], except that the word "weak" is replaced by "strict"; formally, whenever $\max_{a\in A} D_t^m(a) > 0$,

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

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.

A.1. Proof of Lemma 5.1. Let $k = (m-2)/4$. 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,

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

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

Fact 1. For every period t,

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

Proof. Note that

$$u_N(\bar{p}_t,\mathrm{L}) = z_t(\mathrm{D},\mathrm{L}) + z_t(\mathrm{D},\mathrm{M}) + z_t(\mathrm{D},\mathrm{R}) = 1 - \zeta_t,$$
$$u_N(\bar{p}_t,\mathrm{M}) = z_t(\mathrm{U},\mathrm{L}) + z_t(\mathrm{U},\mathrm{M}) + z_t(\mathrm{U},\mathrm{R}) = \zeta_t,$$
$$u_N(\bar{p}_t,\mathrm{R}) = \tfrac{3}{4}.$$

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

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 $(\mathrm{U},\mathrm{M})/(\mathrm{D},\mathrm{M})$ phase preceding the (D,R) phase has form (a), (b), or (c), as shown in Figure A.1. Then the play for the next $2m$, $2m+1$, or $2m+2$ periods, constitutes the full cycle as shown in Figure 1, where phases $(\mathrm{D},\mathrm{L})/(\mathrm{U},\mathrm{L})$ and $(\mathrm{U},\mathrm{M})/(\mathrm{D},\mathrm{M})$ have forms[13] (a), (b), or (c).

[13] The forms of the $(\mathrm{D},\mathrm{L})/(\mathrm{U},\mathrm{L})$ phase are symmetric to those of $(\mathrm{U},\mathrm{M})/(\mathrm{D},\mathrm{M})$ obtained by replacement of (U,M) by (D,L) and (D,M) by (U,L).

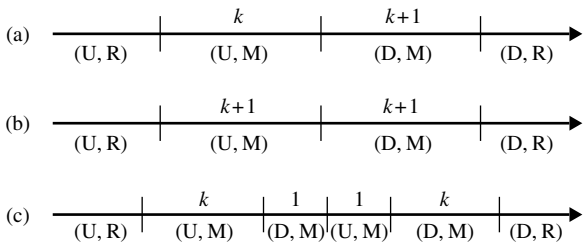

FIGURE A.1. Three forms of the (U, M)/(D, M) phase.

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 $m/2+1$ periods constitutes 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 $\zeta_t = k/m < \frac{1}{4}$, thus by Fact 1, $\omega_{t+1} = \mathrm{L}$. Also,

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

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

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

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

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

and (U, L) occurs in periods $k+2, \ldots, 2k+2$, until we reach $\zeta_{t+2k+2} = (k+1)/m > 1/4$. 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

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

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

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

and (U, L) occurs in periods $k+3, \ldots, 2k+1$, until we reach $\zeta_{t+2k+1} = (k+1)/m > 1/4$. 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 $k+3, \dots, 2k+2$, until we reach $\zeta_{t+2k+2} = (k+1)/m > 1/4$. 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, similar to Case 1, we have $\zeta_t = k/m < \frac{1}{4}$, and (D, L) is deterministically played $k+1$ times, until

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

After that, (U, L) is played in periods $k+2, \dots, 2k+2$, until we reach $\zeta_{t+2k+2} = (k+1)/m > 1/4$. 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). Note that at the end of the phase (D, L)/(U, L) we have $z_{t_1}(\mathrm{U},\mathrm{M}) = z_{t_1}(\mathrm{D},\mathrm{M}) = 0$, hence

$$R^m_{t_1}(\mathrm{U}) = z_{t_1}(\mathrm{D},\mathrm{L}) - z_{t_1}(\mathrm{D},\mathrm{M}) > 0,$$
$$R^m_{t_1}(\mathrm{D}) = z_{t_1}(\mathrm{U},\mathrm{M}) - z_{t_1}(\mathrm{U},\mathrm{L}) < 0.$$

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

A.2. Proof of Lemma 5.2. By Lemma 5.1, $h_{t_0} \in H^m_C$ implies $h^m_t \in H^m_C$ for all $t > t_0$. Let $h^m_t \in H^m_C$ 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 $z_t(\mathrm{D},\mathrm{L}) = z_t(\mathrm{U},\mathrm{L}) = 0$. Also, since at the end of the (D, R) phase the number of U in the history is $(m+2)/4$, it implies that $z_t(\mathrm{U},\mathrm{M}) = \frac{1}{4} + 1/(2m)$. Therefore,

$$R^m_t(\mathrm{D}) = z_t(\mathrm{U},\mathrm{M}) - z_t(\mathrm{U},\mathrm{L}) = z_t(\mathrm{U},\mathrm{M}) = \frac{1}{4} + \frac{1}{2m} \equiv C.$$

For every period τ, $|R^m_\tau(\mathrm{D}) - R^m_{\tau+1}(\mathrm{D})| \le 2/m$, therefore, in periods $t-j$ and $t+j$ the regret for D must be at least $R^m_t(\mathrm{D}) - 2j/m$. Since the duration of every cycle is at most $2m+2$, the average regret for D during the cycle is at least

$$\frac{1}{2m+2}\left(C + 2\left[\left(C - \frac{2}{m}\right) + \left(C - \frac{4}{m}\right) + \cdots + \left(C - \frac{2(m/4-2)}{m}\right)\right]\right) \ge \frac{1}{2m}\left(\frac{m}{2}C - \frac{2}{m}\frac{m^2-4}{32}\right) \ge \frac{1}{32}. \quad \text{(A.1)}$$

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

$$\gamma^m = \lim_{t\to\infty} \frac{1}{t}\left|\tau \in \{1,\dots,t\}\colon \max_{a\in\{\mathrm{U},\mathrm{D}\}} R^m_\tau(a) \ge \varepsilon\right|.$$

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

$$\Pr_{(p^m, q^m, h_{t_0})}\left[\max_{a\in\{\mathrm{U},\mathrm{D}\}} R^m_t(a) \ge \varepsilon\right] \ge \varepsilon.$$

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

A.3. Proof of Lemma 5.3. We shall prove that, regardless of the initial history, some event $H^m_E \subset H^m$ occurs infinitely often, and whenever it occurs, the process reaches the cycle, H^m_C, 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

$$R^m_t(\mathrm{U}) = z_t(\mathrm{D},\mathrm{L}) - z_t(\mathrm{D},\mathrm{M}) = z_t(\mathrm{D},\mathrm{L}) \ge 0,$$
$$R^m_t(\mathrm{D}) = z_t(\mathrm{U},\mathrm{M}) - z_t(\mathrm{U},\mathrm{L}) = -z_t(\mathrm{U},\mathrm{L}) \le 0.$$

Case 1. $z_t(\mathrm{D},\mathrm{L}) > 0$. Suppose that L occurred last time at $t-j$, $0 \le j \le m-1$. 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 (see proof of Lemma 5.1), Nature begins playing M. Contradiction.

Case 2. $z_t(\mathrm{D},\mathrm{L}) = 0$, That is, Agent has no regrets; his play is defined arbitrarily. By Assumption (2), $p^m_{t+1}(\mathrm{U}) > 0$, 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 $\omega_t = \mathrm{L}$ and $\omega_{t+j} = \mathrm{M}$, then $j > m/2$. Symmetrically, if $\omega_t = \mathrm{M}$ and $\omega_{t+j} = \mathrm{L}$, then $j > m/2$.

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

Fact 5. Regardless of an initial state, the event $\{\omega_t = \mathrm{L}$ 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 $m/2$, hence if L occurs first time after M, previous occurrence of L is at least $m+1$ periods ago. □

Fact 6. Suppose that $\omega_t = \mathrm{L}$ 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 $R^m_{t+j}(\mathrm{U}) > 0$ and $R^m_{t+j}(\mathrm{D}) \le 0$.

Proof. We have

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

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

Case 1. $R^m_t(\mathrm{D}) > 0$ and $R^m_t(\mathrm{U}) > 0$ 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 $R^m_t(\mathrm{U})$ nondecreasing and $R^m_t(\mathrm{D})$ nonincreasing. Thus, with positive probability, both (D,L) and (U,L) are played for j periods, until we obtain $\frac{1}{4} < \zeta_{t+j} < \frac{1}{4} + 1/m$, $R^m_{t+j}(\mathrm{U}) > 0$ and $R^m_{t+j}(\mathrm{D}) \le 0$. 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, $R^m_{t+m/2}(\mathrm{U}) > 0$, $R^m_{t+m/2}(\mathrm{D}) < 0$, and (U,L) is played at most $k+1 = (m+2)/4$ times until the frequency of U becomes above $1/4$.

Case 2. $R^m_t(\mathrm{D}) > 0$, $R^m_t(\mathrm{U}) \le 0$. Then (D,L) is played for the next $j' = (z_t(\mathrm{D},\mathrm{L}) - z_t(\mathrm{D},\mathrm{M})) \cdot m + 1$ periods. At period $t+j'$ we have $R^m_{t+j'}(\mathrm{D}) > 0$ and $R^m_{t+j'}(\mathrm{U}) > 0$, and proceed similar to Case 1.

Case 3. $R^m_t(\mathrm{D}) \le 0$, $R^m_t(\mathrm{U}) \le 0$. That is, Agent has no regrets; his play is defined arbitrarily. By Assumption (2), $p_{t+1}(\mathrm{D}) > 0$, hence there is a positive probability that (D,L) occurs for $j' = z_t(\mathrm{D},\mathrm{M}) \cdot m$ periods which will yield $R^m_{t+j'}(\mathrm{U}) > 0$, Case 2.

Case 4. $R^m_t(\mathrm{D}) \le 0$, $R^m_t(\mathrm{U}) > 0$. Then (U,L) is played for $j = 1$ or 2 periods (depending whether $(a_t, \omega_t) = (\mathrm{D},\mathrm{L})$ or (U,L)), and we have $\frac{1}{4} < \zeta_{t+j} < \frac{1}{4} + 1/m$, $R^m_{t+j}(\mathrm{U}) = R^m_t(\mathrm{U}) > 0$, and $R^m_{t+j}(\mathrm{D}) < R^m_t(\mathrm{D}) \le 0$. □

Using Fact 6, we can now analyze the dynamics of the process. Suppose that $\frac{1}{4} < \zeta_t < \frac{1}{4} + 1/m$, $R^m_t(\mathrm{U}) > 0$, $R^m_t(\mathrm{D}) \le 0$. Then

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

$$R^m_{t+j_{UR}}(\mathrm{U}) \ge z_t(\mathrm{D},\mathrm{L}) > 0,$$
$$R^m_{t+j_{UR}}(\mathrm{D}) \le -z_t(\mathrm{U},\mathrm{L}) \le 0.$$

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

$$R^m_{t+j_{UR}+j_{UM}}(\mathrm{D}) = z_{t+j_{UR}+j_{UM}}(\mathrm{U},\mathrm{M}) - z_{t+j_{UR}+j_{UM}}(\mathrm{U},\mathrm{L})$$
$$\ge \frac{k+1}{m} - \frac{k}{m} = \frac{1}{m} > 0.$$

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

$$\zeta_{t+j_{UR}+j_{UM}+j_{DM}} = 1 - \frac{j_{DM}}{m} = \frac{3k+1}{m} < \frac{3}{4},$$
$$R^m_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{U}) = -z_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{D}, \mathrm{M}) < 0,$$
$$R^m_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{D}) = z_{t+j_{UR}+j_{UM}+j_{DM}}(\mathrm{U}, \mathrm{M}) > 0.$$

Note 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)). □

Acknowledgments. The author thanks Dean Foster, Sergiu Hart, Tymofiy Mylovanov, Eilon Solan, Peyton Young, and seminar participants at the Hebrew University and Tel Aviv University for helpful discussions and suggestions, as well as an anonymous referee and an associate editor for valuable comments. The author also thanks the Center for Rationality, the Hebrew University, for its hospitality while this research was being conducted, and the Lady Davis and Golda Meir Fellowship Funds, the Hebrew University, for their financial support.

References

[1] Aumann, R. J. 1981. Survey of repeated games. V. Bohm, ed. *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern*. Bibliographisches Institut, Mannheim, 11–42.
[2] Aumann, R. J., S. Sorin. 1989. Cooperation and bounded recall. *Games Econom. Behav.* **1**(1) 5–39.
[3] Ben-Porath, E. 1993. Repeated games with finite automata. *J. Econom. Theory* **59**(1) 17–32.
[4] Blackwell, D. 1956. An analog of the minmax theorem for vector payoffs. *Pacific J. Math.* **6**(1) 1–8.
[5] Cesa-Bianchi, N., G. Lugosi. 2003. Potential-based algorithms in on-line prediction and game theory. *Machine Learning* **51**(3) 239–261.
[6] Cesa-Bianchi, N., Y. Freund, D. Helmbold, D. Haussler, R. Shapire, M. Warmuth. 1997. How to use expert advice. *J. ACM* **44** 427–485.
[7] Foster, D., R. Vohra. 1999. Regret in the online decision problem. *Games Econom. Behav.* **29**(1-2) 7–35.
[8] Freund, Y., R. Schapire. 1996. Game theory, on-line prediction and boosting. *Proc. Ninth Annual Conf. Computational Learn. Theory, June 28–July 1*, 1996 Desenzano del Garda, Italy, 325–332.
[9] Fudenberg, D., D. Levine. 1995. Universal consistency and cautious fictitious play. *J. Econom. Dynam. Control* **19**(5-7) 1065–1089.
[10] Hannan, J. 1957. Approximation to Bayes risk in repeated play. M. Dresher, A. W. Tucker, P. Wolfe, eds. *Contributions to the Theory of Games*, Vol. 3. *Annals of Mathematics Studies* 39. Princeton University Press, Princeton, NJ, 97–139.
[11] Hart, S., A. Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* **68**(5) 1127–1150.
[12] Hart, S., A. Mas-Colell. 2001. A general class of adaptive procedures. *J. Econom. Theory* **98**(1) 26–54.
[13] Lehrer, E. 1988. Repeated games with stationary bounded recall strategies. *J. Econom. Theory* **46**(1) 130–144.
[14] Lehrer, E. 1994. Finitely many players with bounded recall in infinitely repeated games. *Games Econom. Behav.* **7**(3) 390–405.
[15] Lehrer, E., E. Solan. 2003. No regret with bounded computational capacity. Discussion paper 1373, The Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston, IL.
[16] Littlestone, N., M. Warmuth. 1994. The weighted majority algorithm. *Inform. Comput.* **108**(2) 212–261.
[17] Neyman, A. 1998. Finitely repeated games with finite automata. *Math. Oper. Res.* **23**(3) 513–552.
[18] Neyman, A., D. Okada. 2000. Repeated games with bounded entropy. *Games Econom. Behav.* **30**(2) 228–247.
[19] Rubinstein, A. 1986. Finite automata play the repeated prisoner's dilemma. *J. Econom. Theory* **39**(1) 83–96.
[20] Vovk, V. 1998. A game of prediction with expert advice. *J. Comput. System Sci.* **56**(2) 153–173.
[21] Watson, J. 1994. Cooperation in the infinitely repeated prisoner's dilemma with perturbations. *Games Econom. Behav.* **7**(2) 260–285.