Econometrica, Vol. 68, No. 5 (September, 2000), 1127–1150

A SIMPLE ADAPTIVE PROCEDURE LEADING TO CORRELATED EQUILIBRIUM[1]

BY SERGIU HART AND ANDREU MAS-COLELL[2]

We propose a new and simple adaptive procedure for playing a game: "regret-matching." In this procedure, players may depart from their current play with probabilities that are proportional to measures of regret for not having used other strategies in the past. It is shown that our adaptive procedure guarantees that, with probability one, the empirical distributions of play converge to the set of correlated equilibria of the game.

KEYWORDS: Adaptive procedure, correlated equilibrium, no regret, regret-matching, simple strategies.

1. INTRODUCTION

THE LEADING NONCOOPERATIVE EQUILIBRIUM NOTIONS for N-person games in strategic (normal) form are Nash equilibrium (and its refinements) and correlated equilibrium. In this paper we focus on the concept of correlated equilibrium.

A *correlated equilibrium*—a notion introduced by Aumann (1974)—can be described as follows: Assume that, before the game is played, each player receives a private signal (which does not affect the payoffs). The player may then choose his action in the game depending on this signal. A correlated equilibrium of the original game is just a Nash equilibrium of the game with the signals. Considering all possible signal structures generates all correlated equilibria. If the signals are (stochastically) independent across the players, it is a Nash equilibrium (in mixed or pure strategies) of the original game. But the signals could well be correlated, in which case new equilibria may obtain.

Equivalently, a correlated equilibrium is a probability distribution on N-tuples of actions, which can be interpreted as the distribution of play instructions given to the players by some "device" or "referee." Each player is given—privately—instructions for his own play only; the joint distribution is known to all of them. Also, for every possible instruction that a player receives, the player realizes that the instruction provides a best response to the random estimated play of the other players—assuming they all follow their instructions.

There is much to be said for correlated equilibrium. See Aumann (1974, 1987) for an analysis and foundational arguments in terms of rationality. Also, from a

[1] October 1998 (minor corrections: June 1999). Previous versions: February 1998; November 1997; December 1996; March 1996 (handout). Research partially supported by grants of the U.S.-Israel Binational Science Foundation, the Israel Academy of Sciences and Humanities, the Spanish Ministry of Education, and the Generalitat de Catalunya.

[2] We want to acknowledge the useful comments and suggestions of Robert Aumann, Antonio Cabrales, Dean Foster, David Levine, Alvin Roth, Reinhard Selten, Sylvain Sorin, an editor, the anonymous referees, and the participants at various seminars where this work was presented.

practical point of view, it could be argued that correlated equilibrium may be the most relevant noncooperative solution concept. Indeed, with the possible exception of well-controlled environments, it is hard to exclude a priori the possibility that correlating signals are amply available to the players, and thus find their way into the equilibrium.

This paper is concerned with dynamic considerations. We pose the following question: *Are there simple adaptive procedures always leading to correlated equilibrium*?

Foster and Vohra (1997) have obtained a procedure converging to the set of correlated equilibria. The work of Fudenberg and Levine (1999) led to a second one. We introduce here a procedure that we view as particularly simple and intuitive (see Section 4 for a comparative discussion of all these procedures). It does not entail any sophisticated updating, prediction, or fully rational behavior. Our procedure takes place in discrete time and it specifies that players adjust strategies probabilistically. This adjustment is guided by ''regret measures'' based on observation of past periods. Players know the past history of play of all players, as well as their own payoff matrix (but not necessarily the payoff matrices of the other players). Our Main Theorem is: The adaptive procedure generates trajectories of play that almost surely converge to the set of correlated equilibria.

The procedure is as follows: At each period, a player may either continue playing the same strategy as in the previous period, or switch to other strategies, with probabilities that are proportional to how much higher his accumulated payoff would have been had he always made that change in the past. More precisely, let U be his total payoff up to now; for each strategy k different from his last period strategy j, let $V(k)$ be the total payoff he would have received if he had played k every time in the past that he chose j (and everything else remained unchanged). Then only those strategies k with $V(k)$ larger than U may be switched to, with probabilities that are proportional to the differences $V(k) - U$, which we call the ''regret'' for having played j rather than k. These probabilities are normalized by a fixed factor, so that they add up to strictly less than 1; with the remaining probability, the same strategy j is chosen as in the last period.

It is worthwhile to point out three properties of our procedure. First, its simplicity; indeed, it is very easy to explain and to implement. It is not more involved than fictitious play (Brown (1951) and Robinson (1951); note that in the two-person zero-sum case, our procedure also yields the minimax value). Second, the procedure is *not* of the ''best-reply'' variety (such as fictitious play, smooth fictitious play (Fudenberg and Levine (1995, 1999)) or calibrated learning (Foster and Vohra (1997)); see Section 4 for further details). Players do not choose only their ''best'' actions, nor do they give probability close to 1 to these choices. Instead, all ''better'' actions may be chosen, with probabilities that are proportional to the apparent gains, as measured by the regrets; the procedure could thus be called ''*regret-matching*.'' And third, there is ''inertia.'' The strategy played in the last period matters: There is always a positive probability of

continuing to play this strategy and, moreover, changes from it occur only if there is reason to do so.

At this point a question may arise: Can one actually guarantee that the smaller set of Nash equilibria is always reached? The answer is definitely "no." On the one hand, in our procedure, as in most others, there is a natural coordination device: the common history, observed by all players. It is thus reasonable to expect that, at the end, independence among the players will not obtain. On the other hand, the set of Nash equilibria is a mathematically complex set (a set of fixed-points; by comparison, the set of correlated equilibria is a convex polytope), and simple adaptive procedures cannot be expected to guarantee the global convergence to such a set.

After this introductory section, in Section 2 we present the model, describe the adaptive procedure, and state our result (the Main Theorem). Section 3 is devoted to a "stylized variation" of the procedure of Section 2. It is a variation that lends itself to a very direct proof, based on Blackwell's (1956a) Approachability Theorem. This is a new instrument in this field, which may well turn out to be widely applicable.

Section 4 contains a discussion of the literature, together with a number of relevant issues. The proof of the Main Theorem is relegated to the Appendix.

2. THE MODEL AND MAIN RESULT

Let $\Gamma = (N, (S^i)_{i \in N}, (u^i)_{i \in N})$ be a finite N-person game in strategic (normal) form: N is the set of players, S^i is the set of strategies of player i, and $u^i : \prod_{i \in N} S^i \to \mathbb{R}$ is player i's payoff function. All sets N and S^i are assumed to be finite. Denote by $S := \prod_{i \in N} S^i$ the set of N-tuples of strategies; the generic element of S is $s = (s^i)_{i \in N}$, and s^{-i} denotes the strategy combination of all players except i, i.e., $s^{-i} = (s^{i'})_{i' \neq i}$. We focus attention on the following solution concept:

DEFINITION: A probability distribution ψ on S is a *correlated equilibrium* of Γ if, for every $i \in N$, every $j \in S^i$ and every $k \in S^i$ we have[3]

$$\sum_{s \in S : s^i = j} \psi(s)[u^i(k, s^{-i}) - u^i(s)] \leq 0.$$

If in the above inequality we replace the right-hand side by an $\varepsilon > 0$, then we obtain the concept of a *correlated ε-equilibrium*.

Note that every Nash equilibrium is a correlated equilibrium. Indeed, Nash equilibria correspond to the special case where ψ is a product measure, that is, the play of the different players is independent. Also, the set of correlated equilibria is nonempty, closed and convex, and even in simple games (e.g., "chicken") it may include distributions that are not in the convex hull of the Nash equilibrium distributions.

[3] We write $\sum_{s \in S : s^i = j}$ for the sum over all N-tuples s in S whose ith coordinate s^i equals j.

Suppose now that the game Γ is played repeatedly through time: $t = 1, 2, \ldots$. At time $t+1$, given a history of play $h_t = (s_\tau)_{\tau=1}^t \in \prod_{\tau=1}^t S$, we postulate that each player $i \in N$ chooses $s_{t+1}^i \in S^i$ according to a probability distribution[4] $p_{t+1}^i \in \Delta(S^i)$ which is defined in the following way:

For every two different strategies $j, k \in S^i$ of player i, suppose i were to replace strategy j, every time that it was played in the past, by strategy k; his payoff at time τ, for $\tau \le t$, would become

$$W_\tau^i(j,k) := \begin{cases} u^i(k, s_\tau^{-i}), & \text{if } s_\tau^i = j, \\ u^i(s_\tau), & \text{otherwise.} \end{cases} \tag{2.1a}$$

The resulting difference in i's average payoff up to time t is then

$$\begin{aligned} D_t^i(j,k) &:= \frac{1}{t}\sum_{\tau=1}^t W_\tau^i(j,k) - \frac{1}{t}\sum_{\tau=1}^t u^i(s_\tau) \\ &= \frac{1}{t}\sum_{\tau \le t:\, s_\tau^i = j} \left[u^i(k, s_\tau^{-i}) - u^i(s_\tau)\right]. \end{aligned} \tag{2.1b}$$

Finally, denote

$$R_t^i(j,k) := \left[D_t^i(j,k)\right]^+ = \max\{D_t^i(j,k), 0\}. \tag{2.1c}$$

The expression $R_t^i(j,k)$ has a clear interpretation as a measure of the (average) "regret" at period t for not having played, every time that j was played in the past, the different strategy k.

Fix $\mu > 0$ to be a large enough number.[5] Let $j \in S^i$ be the strategy last chosen by player i, i.e., $j = s_t^i$. Then the probability distribution $p_{t+1}^i \in \Delta(S^i)$ used by i at time $t+1$ is defined as

$$\begin{cases} p_{t+1}^i(k) := \dfrac{1}{\mu} R_t^i(j,k), & \text{for all } k \ne j, \\ p_{t+1}^i(j) := 1 - \displaystyle\sum_{k \in S^i:\, k \ne j} p_{t+1}^i(k). \end{cases} \tag{2.2}$$

Note that the choice of μ guarantees that $p_{t+1}^i(j) > 0$; that is, there is always a positive probability of playing the same strategy as in the previous period. The play $p_1^i \in \Delta(S^i)$ at the initial period is chosen arbitrarily.[6]

[4] We write $\Delta(Q)$ for the set of probability distributions over a finite set Q.

[5] The parameter μ is fixed throughout the procedure (independent of time and history). It suffices to take μ so that $\mu > 2M^i(m^i - 1)$ for all $i \in N$, where M^i is an upper bound for $|u^i(\cdot)|$ and m^i is the number of strategies of player i. Even better, we could let μ satisfy $\mu > (m^i - 1)|u^i(k, s^{-i}) - u^i(j, s^{-i})|$ for all $j, k \in S^i$, all $s^{-i} \in S^{-i}$, and all $i \in N$ (and moreover we could use a different μ^i for each player i).

[6] Actually, the procedure could start with any finite number of periods where the play is arbitrary.

Informally, (2.2) may be described as follows. Player i starts from a "reference point": his current actual play. His choice next period is governed by propensities to depart from it. It is natural therefore to postulate that, if a change occurs, it should be to actions that are perceived as being better, relative to the current choice. In addition, and in the spirit of adaptive behavior, we assume that all such better choices get positive probabilities; also, the better an alternative action seems, the higher the probability of choosing it next time. Further, there is also inertia: the probability of staying put (and playing the same action as in the last period) is always positive.

More precisely, the probabilities of switching to different strategies are proportional to their regrets relative to the current strategy. The factor of proportionality is constant. In particular, if the regrets are small, then the probability of switching from current play is also small.

For every t, let $z_t \in \Delta(S)$ be the empirical distribution of the N-tuples of strategies played up to time t. That is, for every[7] $s \in S$,

$$z_t(s) := \frac{1}{t}|\{\tau \leq t : s_\tau = s\}| \tag{2.3}$$

is the relative frequency that the N-tuple s has been played in the first t periods. We can now state our main result.

MAIN THEOREM: *If every player plays according to the adaptive procedure* (2.2), *then the empirical distributions of play z_t converge almost surely as $t \to \infty$ to the set of correlated equilibrium distributions of the game Γ.*

Note that convergence to the *set* of correlated equilibria does not imply that the sequence z_t converges to *a point*. The Main Theorem asserts that the following statement holds with probability one: For any $\varepsilon > 0$ there is $T_0 = T_0(\varepsilon)$ such that for all $t > T_0$ we can find a correlated equilibrium distribution ψ_t at a distance less than ε from z_t. (Note that this T_0 depends on the history; it is an "a.s. finite stopping time.") That is, the Main Theorem says that, with probability one, for any $\varepsilon > 0$, the (random) trajectory $(z_1, z_2, \ldots, z_t, \ldots)$ enters and then stays forever in the ε-neighborhood in $\Delta(S)$ of the set of correlated equilibria. Put differently: Given any $\varepsilon > 0$, there exists a constant (i.e., independent of history) $t_0 = t_0(\varepsilon)$ such that, with probability at least $1 - \varepsilon$, the empirical distributions z_t for *all* $t > t_0$ are in the ε-neighborhood of the set of correlated equilibria. Finally, let us note that because the set of correlated equilibria is nonempty and compact, the statement "the trajectory (z_t) converges to the set of correlated equilibria" is equivalent to the statement "the trajectory (z_t) is such that for any $\varepsilon > 0$ there is $T_1 = T_1(\varepsilon)$ with the property that z_t is a correlated ε-equilibrium for all $t > T_1$."

We conclude this section with a few comments (see also Section 4):

(1) Our adaptive procedure (2.2) requires player i to know his own payoff matrix (but not those of the other players) and, at time $t+1$, the history h_t;

[7] We write $|Q|$ for the number of elements of a finite set Q.

actually, the empirical distribution z_t of $(s_1, s_2, \ldots, s_t)$ suffices. In terms of computation, player i needs to keep record of the time t together with the $m^i(m^i - 1)$ numbers $D_t^i(j, k)$ for all $j \neq k$ in S^i (and update these numbers every period).

(2) At every period the adaptive procedure that we propose randomizes only over the strategies that exhibit positive regret relative to the most recently played strategy. Some strategies may, therefore, receive zero probability. Suppose that we were to allow for trembles. Specifically, suppose that at every period we put a $\delta > 0$ probability on the uniform tremble (each strategy thus being played with probability at least δ/m^i). It can be shown that in this case the empirical distributions z_t converge to the set of correlated ε-equilibria (of course, ε depends on δ, and it goes to zero as δ goes to zero). In conclusion, unlike most adaptive procedures, ours does not rely on trembles (which are usually needed, technically, to get the "ergodicity" properties); moreover, our result is robust with respect to trembles.

(3) Our adaptive procedure depends only on one parameter,[8] μ. This may be viewed as an "inertia" parameter (see Subsections 4(g) and 4(h)): A higher μ yields lower probabilities of switching. The convergence to the set of correlated equilibria is always guaranteed (for any large enough μ; see footnote 5), but the speed of convergence changes with μ.

(4) We know little about additional convergence properties for z_t. It is easy to see that the empirical distributions z_t either converge to a Nash equilibrium in pure strategies, or must be infinitely often outside the set of correlated equilibria (because, if z_t is a correlated equilibrium from some time on, then[9] all regrets are 0, and the play does not change). This implies, in particular, that interior (relative to $\Delta(S)$) points of the set of correlated equilibria that are not pure Nash equilibria are unreachable as the limit of some z_t (but it is possible that they are reachable as limits of a *subsequence* of z_t).

(5) There are other procedures enjoying convergence properties similar to ours: the procedures of Foster and Vohra (1997), of Fudenberg and Levine (1999), and of Theorem A in Section 3 below; see the discussion in Section 4. The delimitation of general classes of procedures converging to correlated equilibria seems, therefore, an interesting research problem.[10]

3. NO REGRET AND BLACKWELL APPROACHABILITY

In this section (which can be viewed as a motivational preliminary) we shall replace the adaptive procedure of Section 2 by another procedure that, while related to it, is more stylized. Then we shall analyze it by means of Blackwell's (1956a) Approachability Theorem, and prove that it yields convergence to the

[8] Using a parameter μ (rather than a fixed normalization of the payoffs) was suggested to us by Reinhard Selten.

[9] See the Proposition in Section 3.

[10] See Hart and Mas-Colell (1999) and Cahn (2000) for such results.

set of correlated equilibria. In fact, the Main Theorem stated in Section 2, and its proof in Appendix 1, were inspired by consideration and careful study of the result of this section. Furthermore, the procedure here is interesting in its own right (see, for instance, the Remark following the statement of Theorem A, and (d) in Section 4).

Fix a player i and recall the procedure of Section 2: At time $t+1$ the transition probabilities, from the strategy played by player i in period t to the strategies to be played at $t+1$, are determined by the stochastic matrix defined by the system (2.2). Consider now an invariant probability vector $q_t^i = (q_t^i(j))_{j \in S^i} \in \Delta(S^i)$ for this matrix (such a vector always exists). That is, q_t^i satisfies

$$q_t^i(j) = \sum_{k \neq j} q_t^i(k) \frac{1}{\mu} R_t^i(k, j) + q_t^i(j) \left[1 - \sum_{k \neq j} \frac{1}{\mu} R_t^i(j, k) \right],$$

for every $j \in S^i$. By collecting terms, multiplying by μ, and formally letting $R_t^i(j, j) := 0$, the above expression can be rewritten as

$$\sum_{k \in S^i} q_t^i(k) R_t^i(k, j) = q_t^i(j) \sum_{k \in S^i} R_t^i(j, k), \tag{3.1}$$

for every $j \in S^i$.

In this section we shall assume that play at time $t+1$ by player i is determined by a solution q_t^i to the system of equations (3.1); i.e., $p_{t+1}^i(j) := q_t^i(j)$. In a sense, we assume that player i at time $t+1$ goes instantly to the invariant distribution of the stochastic transition matrix determined by (2.2). We now state the key result.

THEOREM A: *Suppose that at every period $t+1$ player i chooses strategies according to a probability vector q_t^i that satisfies* (3.1). *Then player i's regrets $R_t^i(j, k)$ converge to zero almost surely for every j, k in S^i with $j \neq k$.*

REMARK: Note that—in contrast to the Main Theorem, where every player uses (2.2)—no assumption is made in Theorem A on how players different from i choose their strategies (except for the fact that for every t, given the history up to t, play is independent among players). In the terminology of Fudenberg and Levine (1999, 1998), the adaptive procedure of this section is "(universally) calibrated." For an extended discussion of this issue, see Subsection 4(d).

What is the connection between regrets and correlated equilibria? It turns out that a necessary and sufficient condition for the empirical distributions to converge to the set of correlated equilibria is precisely that all regrets converge to zero. More generally, we have the following proposition.

PROPOSITION: *Let $(s_t)_{t=1,2,\ldots}$ be a sequence of plays (i.e., $s_t \in S$ for all t) and let*[11] $\varepsilon \geq 0$. *Then*: $\limsup_{t \to \infty} R_t^i(j, k) \leq \varepsilon$ *for every $i \in N$ and every $j, k \in S^i$ with*

[11] Note that both $\varepsilon > 0$ and $\varepsilon = 0$ are included.

$j \neq k$, if and only if the sequence of empirical distributions z_t (defined by (2.3)) *converges to the set of correlated ε-equilibria.*

PROOF: For each player i and every $j \neq k$ in S^i we have

$$D_t^i(j,k) = \frac{1}{t} \sum_{\tau \leq t : s_\tau^i = j} \left[u^i(k, s_\tau^{-i}) - u^i(j, s_\tau^{-i}) \right]$$
$$= \sum_{s \in S : s^i = j} z_t(s)[u^i(k, s^{-i}) - u^i(j, s^{-i})].$$

On any subsequence where z_t converges, say $z_{t'} \to \psi \in \Delta(S)$, we get

$$D_{t'}^i(j,k) \to \sum_{s \in S : s^i = j} \psi(s)[u^i(k, s^{-i}) - u^i(j, s^{-i})].$$

The result is immediate from the definition of a correlated ε-equilibrium and (2.1c). *Q.E.D.*

Theorem A and the Proposition immediately imply the following corollary.

COROLLARY: *Suppose that at each period $t+1$ every player i chooses strategies according to a probability vector q_t^i that satisfies* (3.1). *Then the empirical distributions of play z_t converge almost surely as $t \to \infty$ to the set of correlated equilibria of the game Γ.*

Before addressing the formal proof of Theorem A, we shall present and discuss Blackwell's Approachability Theorem.

The basic setup contemplates a decision-maker i with a (finite) action set S^i. For a finite indexing set L, the decision-maker receives an $|L|$-dimensional vector payoff $v(s^i, s^{-i}) \in \mathbb{R}^L$ that depends on his action $s^i \in S^i$ and on some external action s^{-i} belonging to a (finite) set S^{-i} (we will refer to $-i$ as the "opponent"). The decision problem is repeated through time. Let $s_t = (s_t^i, s_t^{-i}) \in S^i \times S^{-i}$ denote the choices at time t; of course, both i and $-i$ may use randomizations. The question is whether the decision-maker i can guarantee that the time average of the vector payoffs, $D_t := (1/t)\sum_{\tau \leq t} v(s_\tau) \equiv (1/t)\sum_{\tau \leq t} v(s_\tau^i, s_\tau^{-i})$, approaches a predetermined set (in $\mathbb{R}^L$).

Let C be a convex and closed subset of $\mathbb{R}^L$. The set C is *approachable* by the decision-maker i if there is a procedure[12] for i that guarantees that the average vector payoff D_t approaches the set C (i.e.,[13] $\operatorname{dist}(D_t, C) \to 0$ almost surely as $t \to \infty$), regardless of the choices of the opponent $-i$. To state Blackwell's result,

[12] In the repeated setup, we refer to a (behavior) strategy as a "procedure."

[13] $\operatorname{dist}(x, A) := \min\{\|x - a\| : a \in A\}$, where $\|\cdot\|$ is the Euclidean norm. Strictly speaking, Blackwell's definition of approachability requires also that the convergence of the distance to 0 be uniform over the procedures of the opponent; i.e., there is a procedure of i such that for every $\varepsilon > 0$ there is $t_0 \equiv t_0(\varepsilon)$ such that for any procedure of $-i$ we have $P[\operatorname{dist}(D_t, C) < \varepsilon$ for all $t > t_0] > 1 - \varepsilon$. The Blackwell procedure (defined in the next Theorem) guarantees this as well.

let w_C denote the *support function* of the convex set C, i.e., $w_C(\lambda) := \sup\{\lambda \cdot c : c \in C\}$ for all λ in $\mathbb{R}^L$. Given a point $x \in \mathbb{R}^L$ which is not in C, let $F(x)$ be the (unique) point in C that is closest to x in the Euclidean distance, and put $\lambda(x) := x - F(x)$; note that $\lambda(x)$ is an outward normal to the set C at the point $F(x)$.

BLACKWELL'S APPROACHABILITY THEOREM: *Let $C \subset \mathbb{R}^L$ be a convex and closed set, with support function w_C. Then C is approachable by i if and only if for every $\lambda \in \mathbb{R}^L$ there exists a mixed strategy $q_\lambda \in \Delta(S^i)$ such that*[14]

$$\lambda \cdot v(q_\lambda, s^{-i}) \leq w_C(\lambda), \qquad \textit{for all } s^{-i} \in S^{-i}. \tag{3.2}$$

Moreover, the following procedure of i guarantees that $\operatorname{dist}(D_t, C)$ *converges almost surely to* 0 *as* $t \to \infty$: *At time* $t+1$, *play* $q_{\lambda(D_t)}$ *if* $D_t \notin C$, *and play arbitrarily if* $D_t \in C$.

We will refer to the condition for approachability given in the Theorem as the *Blackwell condition*, and to the procedure there as the *Blackwell procedure*. To get some intuition for the result, assume that D_t is not in C, and let $H(D_t)$ be the half-space of $\mathbb{R}^L$ that contains C (and not D_t) and is bounded by the supporting hyperplane to C at $F(D_t)$ with normal $\lambda(D_t)$; see Figure 1. When i uses the Blackwell procedure, it guarantees that $v(q_{\lambda(D_t)}, s^{-i})$ lies in $H(D_t)$ for all s^{-i} in S^{-i} (by (3.2)). Therefore, given D_t, the expectation of the next period

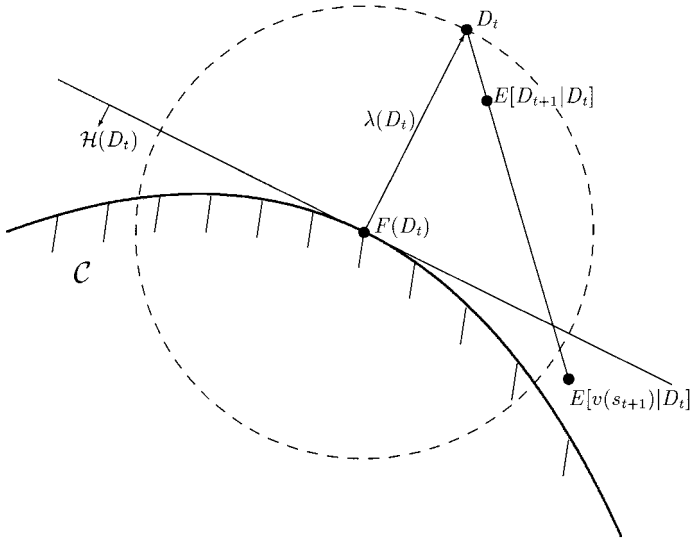

FIGURE 1.—Approaching the set C.

[14] $v(q, s^{-i})$ denotes the expected payoff, i.e., $\sum_{s^i \in S^i} q(s^i) v(s^i, s^{-i})$. Of course, only $\lambda \neq 0$ with $w_C(\lambda) < \infty$ need to be considered in (3.2).

payoff $E[v(s_{t+1})|D_t]$ will lie in the half-space $H(D_t)$ for any pure choice s_{t+1}^{-i} of $-i$ at time $t+1$, and thus also for any randomized choice of $-i$. The expected average vector payoff at period $t+1$ (conditional on D_t) is

$$E[D_{t+1}|D_t] = \frac{t}{t+1} D_t + \frac{1}{t+1} E[v(s_{t+1})|D_t].$$

When t is large, $E[D_{t+1}|D_t]$ will thus be *inside* the circle of center $F(D_t)$ and radius $\|\lambda(D_t)\|$. Hence

$$\begin{aligned}\operatorname{dist}(E[D_{t+1}|D_t], C) &\leq \| E[D_{t+1}|D_t] - F(D_t)\| < \|\lambda(D_t)\| \\ &= \operatorname{dist}(D_t, C)\end{aligned}$$

(the first inequality follows from the fact that $F(D_t)$ is in C). A precise computation shows that the distance not only decreases, but actually goes to zero.[15] For proofs of Blackwell's Approachability Theory, see[16] Blackwell (1956a), or Mertens, Sorin, and Zamir (1995, Theorem 4.3).

We now prove Theorem A.

PROOF OF THEOREM A: As mentioned, the proof of this Theorem consists of an application of Blackwell's Approachability Theorem. Let

$$L := \{(j,k) \in S^i \times S^i : j \neq k\},$$

and define the vector payoff $v(s^i, s^{-i}) \in \mathbb{R}^L$ by letting its $(j,k) \in L$ coordinate be

$$[v(s^i, s^{-i})](j,k) := \begin{cases} u^i(k, s^{-i}) - u(j, s^{-i}), & \text{if } s^i = j, \\ 0, & \text{otherwise}. \end{cases}$$

Let C be the nonpositive orthant $\mathbb{R}_-^L := \{x \in \mathbb{R}^L : x \leq 0\}$. We claim that C is approachable by i. Indeed, the support function of C is given by $w_C(\lambda) = 0$ for all $\lambda \in \mathbb{R}_+^L$ and $w_C(\lambda) = \infty$ otherwise; so only $\lambda \in \mathbb{R}_+^L$ need to be considered. Condition (3.2) is

$$\sum_{(j,k) \in L} \lambda(j,k) \sum_{s^i \in S^i} q_\lambda(s^i)[v(s^i, s^{-i})](j,k) \leq 0,$$

or

$$\sum_{(j,k) \in L} \lambda(j,k) q_\lambda(j)[u^i(k, s^{-i}) - u^i(j, s^{-i})] \leq 0 \tag{3.3}$$

[15] Note that one looks here at *expected* average payoffs; the Strong Law of Large Numbers for Dependent Random Variables—see the Proof of Step M10 in the Appendix—implies that the *actual* average payoffs also converge to the set C.

[16] The Blackwell condition is usually stated as follows: For every $x \notin C$ there exists $q(x) \in \Delta(S^i)$ such that $[x - F(x)] \cdot [v(q(x), s^{-i}) - F(x)] \leq 0$, for all $s^{-i} \in S^{-i}$. It is easy to verify that this is equivalent to our formulation. We further note a simple way of stating the Blackwell result: A convex set C is approachable if and only if any half-space containing C is approachable.

for all $s^{-i} \in S^{-i}$. After collecting terms, the left-hand side of (3.3) can be written as

$$\sum_{j \in S^i} \alpha(j) u^i(j, s^{-i}), \tag{3.4a}$$

where

$$\alpha(j) := \sum_{k \in S^i} q_\lambda(k) \lambda(k, j) - q_\lambda(j) \sum_{k \in S^i} \lambda(j, k). \tag{3.4b}$$

Let $q_\lambda \in \Delta(S^i)$ be an invariant vector for the nonnegative $S^i \times S^i$ matrix with entries $\lambda(j, k)$ for $j \neq k$ and 0 for $j = k$ (such a q_λ always exists). That is, q_λ satisfies

$$\sum_{k \in S^i} q_\lambda(k) \lambda(k, j) = q_\lambda(j) \sum_{k \in S^i} \lambda(j, k), \tag{3.5}$$

for every $j \in S^i$. Therefore $\alpha(j) = 0$ for all $j \in S^i$, and so inequality (3.3) holds true (as an equality[17]) for all $s^{-i} \in S^{-i}$. The Blackwell condition is thus satisfied by the set $C = \mathbb{R}^L_-$.

Consider D_t, the average payoff vector at time t. Its (j, k)-coordinate is $(1/t)\Sigma_{\tau \leq t}[v(s_\tau)](j, k) = D^i_t(j, k)$. If $D_t \notin \mathbb{R}^L_-$, then the closest point to D_t in $\mathbb{R}^L_-$ is $F(D_t) = [D_t]^-$ (see Figure 2), hence $\lambda(D_t) = D_t - [D_t]^- = [D_t]^+ = (R^i_t(j, k))_{(j, k) \in L}$, which is the vector of regrets at time t. Now the given strategy

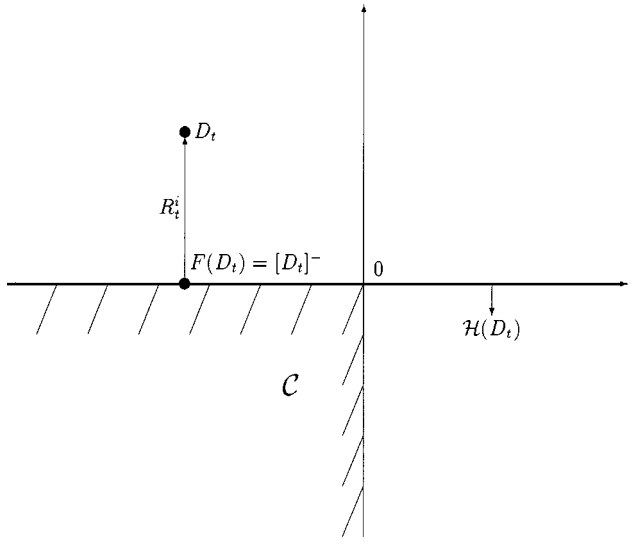

FIGURE 2.—Approaching $C = \mathbb{R}^L_-$.

[17] Note that this is precisely Formula (2) in the Proof of Theorem 1 in Hart and Schmeidler (1989); see Subsection 4(i).

of i at time $t+1$ satisfies (3.1), which is exactly condition (3.5) for $\lambda = \lambda(D_t)$. Hence player i uses the Blackwell procedure for $\mathbb{R}^L_-$, which guarantees that the average vector payoff D_t approaches $\mathbb{R}^L_-$, or $R^i_t(j,k) \to 0$ a.s. for every $j \neq k$. *Q.E.D.*

REMARK: The proof of Blackwell's Approachability Theorem also provides bounds on the speed of convergence. In our case, one gets the following: The expectation $E[R^i_t(j,k)]$ of the regrets is of the order of $1/\sqrt{t}$, and the probability that z_t is a correlated ε-equilibrium for all $t > T$ is at least $1 - ce^{-cT}$ (for an appropriate constant $c > 0$ depending on ε; see Foster and Vohra (1999, Section 4.1)). Clearly, a better speed of convergence[18] for the expected regrets cannot be guaranteed, since, for instance, if the other players play stationary mixed strategies, then the errors are of the order $1/\sqrt{t}$ by the Central Limit Theorem.

4. DISCUSSION

This section discusses a number of important issues, including links and comparisons to the relevant literature.

(a) *Foster and Vohra*. The seminal paper in this field of research is Foster and Vohra (1997). They consider, first, "forecasting rules"—on the play of others—that enjoy good properties, namely, "calibration." Second, they assume that each player best-replies to such calibrated forecasts. The resulting procedure leads to correlated equilibria. The motivation and the formulation are quite different from ours; nonetheless, their results are close to our results (specifically, to our Theorem A), since their calibrated forecasts are also based on regret measures.[19]

(b) *Fudenberg and Levine*. The next important paper is Fudenberg and Levine (1999) (see also their book (1998)). In that paper they offer a class of adaptive procedures, called "calibrated smooth fictitious play," with the property that for every $\varepsilon > 0$ there are procedures in the class that guarantee almost sure convergence to the set of correlated ε-equilibria (but the conclusion does not hold for $\varepsilon = 0$). The formal structure of these procedures is also similar to that of our Theorem A, in the sense that the mixed choice of a given player at time t is determined as an invariant probability vector of a transition matrix. However, the transition matrix (and therefore the stochastic dynamics) is different from the regret-based transition matrix of our Theorem A. To understand further the similarities and differences between the Fudenberg and Levine procedures and our own, the next two Subsections, (c) and (d), contain a detour on the concepts of "universal consistency" and "universal calibration."

[18] Up to a constant factor.

[19] These regrets are defined on an ε-grid on $\Delta(S^{-i})$, with ε going to zero as t goes to infinity. Therefore, at each step in their procedure one needs to compute the invariant vector for a matrix of an increasingly large size; by comparison, in our Theorem A the size of the matrix is fixed, $m^i \times m^i$.

(c) *Universal Consistency.* The term "universal consistency" is due to Fudenberg and Levine (1995). The concept goes back to Hannan (1957), who proved the following result: There is a procedure (in the setup of Section 2) for player i that guarantees, no matter what the other players do, that

$$\limsup_{t\to\infty}\left[\max_{k\in S^i}\frac{1}{t}\sum_{\tau=1}^{t}u^i(k,s_\tau^{-i})-\frac{1}{t}\sum_{\tau=1}^{t}u^i(s_\tau)\right]\le 0 \quad \text{a.s.} \tag{4.1}$$

In other words, i's average payoff is, in the limit, no worse than if he were to play any *constant* strategy $k\in S^i$ for all $\tau\le t$. This property of the Hannan procedure for player i is called *universal consistency* by Fudenberg and Levine (1995) (it is "universal" since it holds no matter how the other players play). Another universally consistent procedure was shown by Blackwell (1956b) to result from his Approachability Theorem (see also Luce and Raiffa (1957, pp. 482–483)).

The adaptive procedure of our Theorem A is also universally consistent. Indeed, for each j in S^i, (4.1) is guaranteed even when restricted to those periods when player i chose that particular j; this being true for all j in S^i, the result follows. However, the application of Blackwell's Approachability Theorem in Section 3 suggests the following particularly simple procedure.

At time t, for each strategy k in S^i, let

$$D_t^i(k):=\frac{1}{t}\sum_{\tau=1}^{t}\left[u^i(k,s_\tau^{-i})-u^i(s_\tau)\right], \tag{4.2a}$$

$$p_{t+1}^i(k):=\frac{\left[D_t^i(k)\right]^+}{\sum_{k'\in S^i}\left[D_t^i(k')\right]^+}, \tag{4.2b}$$

if the denominator is positive, and let $p_{t+1}^i\in\Delta(S^i)$ be arbitrary otherwise. The strategy of player i is then, at time $t+1$, to choose k in S^i with probability $p_{t+1}^i(k)$. These probabilities are thus proportional to the "unconditional regrets" $[D_t^i(k)]^+$ (by comparison to the "conditional on j" regrets of Section 2). We then have the following theorem.

THEOREM B: *The adaptive procedure* (4.2) *is universally consistent for player* i.

The proof of Theorem B is similar to the proof of Theorem A in Section 3 and is omitted.

Fudenberg and Levine (1995) propose a class of procedures that turn out to be universally ε-consistent:[20] "smooth fictitious play." Player i follows a smooth fictitious play behavior rule if at time t he plays a mixed strategy $\sigma^i\in\Delta(S^i)$ that maximizes the sum of his expected payoff (with the actions of the remaining

[20] That is, the right-hand side of (4.1) is $\varepsilon>0$ instead of 0.

players distributed as in the empirical distribution up to t) and $\lambda v^i(\sigma^i)$, where $\lambda > 0$ and v^i is a strictly concave smooth function defined on i's strategy simplex, $\Delta(S^i)$, with infinite length gradient at the boundary of $\Delta(S^i)$. The result of Fudenberg and Levine is then that, given any $\varepsilon > 0$, there is a sufficiently small λ such that universal ε-consistency obtains for player i. Observe that, for small λ, smooth fictitious play is very close to fictitious play (it amounts to playing the best response with high probability and the remaining strategies with low but positive probability). The procedure is, therefore, clearly distinct from (4.2): In (4.2) all the better, even if not best, replies are played with significant probability; also, in (4.2) the inferior replies get zero probability. Finally, it is worth emphasizing that the tremble from best response is required for the Fudenberg and Levine result, since fictitious play is not guaranteed to be consistent. In contrast, the procedure of (4.2) has no trembles.

The reader is referred to Hart and Mas-Colell (1999), where a wide class of universally consistent procedures is exhibited and characterized (including as special cases (4.2) as well as smooth fictitious play).

(d) *Universal Calibration.* The idea of "universal calibration," also introduced[21] by Fudenberg and Levine (1998, 1999), is that, again, regret measures go to zero irrespective of the other players' play. The difference is that, now, the set of regret measures is richer: It consists of regrets that are conditional on the strategy currently played by i himself. Recall the Proposition of Section 3: If such universally calibrated strategies are played by all players, then all regrets become nonpositive in the limit, and thus the convergence to the correlated equilibrium set is guaranteed.

The procedure of Theorem A is universally calibrated; so (up to ε) is the "calibrated smooth fictitious play" of Fudenberg and Levine (1999). The two procedures stand to each other as, in the unconditional version, Theorem B stands to "smooth fictitious play."

The procedure (2.2) of our Main Theorem is not universally calibrated. If only player i follows the procedure, we cannot conclude that all his regrets go to zero; adversaries who know the procedure used by player i could keep his regrets positive.[22] Such sophisticated strategies of the other players, however, are outside the framework of our study—which deals with simple adaptive behavior. In fact, it turns out that the procedure of our Main Theorem is guaranteed to be calibrated not just against opponents using the same procedure, but also against a wide class of behaviors.[23]

We regard the simplicity of (2.2) as a salient point. Of course, if one needs to guarantee calibration even against sophisticated adversaries, one may have to give up on simplicity and resort to the procedure of Theorem A instead.

[21] They actually call it "calibration"; we prefer the term "universal calibration," since it refers to *any* behavior of the opponents (as in their "[conditional] universal consistency").

[22] At each time $t+1$, let them play an $(N-1)$-tuple of strategies that minimizes the expected (relative to p^i_{t+1}) payoff of player i; for an example, see Fudenberg and Levine (1998, Section 8.10).

[23] Namely, such that the dependence of any one choice of $-i$ on any one past choice of i is small, relative to the number of periods; see Cahn (2000).

(e) *Better-reply vs. Best-reply*. Note that all the procedures in the literature reviewed above are best-reply-based: A player uses (almost) exclusively actions that are (almost) best-replies to a certain belief about his opponents. In contrast, our procedure gives significant probabilities to any actions that are just better (rather than best). This has the additional effect of making the behavior continuous, without need for approximations.

(f) *Eigenvector Procedures*. The procedure of our Main Theorem differs from all the other procedures leading to correlated equilibria (including that of our Theorem A) in an important aspect: It does not require the player to compute, at every step, an invariant (eigen-) vector for an appropriate positive matrix. Again, the simplicity[24] of (2.2) is an essential property when discussing nonsophisticated behavior; this is the reason we have sought this result as our Main Theorem.

(g) *Inertia*. A specific and most distinctive feature by which the procedure of our Main Theorem differs from those of Theorem A and the other works mentioned above is that in the former the individual decisions privilege the most recent action taken: The probabilities used at period $t+1$ are best thought of as propensities to depart from the play at t.

Viewed in this light, our procedure has significant inertial characteristics. In particular, there is a positive probability of moving from the strategy played at t only if there is another that appears better (in which case the probabilities of playing the better strategies are proportional to the regrets relative to the period t strategy).[25]

(h) *Friction*. The procedure (2.2) exhibits "friction": There is always a positive probability of continuing with the period t strategy.[26] To understand the role played by friction,[27] suppose that we were to modify the procedure (2.2) by requiring that the switching probabilities be rescaled in such a way that a switch occurs if and only if there is at least one better strategy (i.e., one with positive regret). Then the result of the Main Theorem may not hold. For example, in the familiar two-person 2×2 coordination game, if we start with an uncoordinated strategy pair, then the play alternates between the two uncoordinated pairs. However, no distribution concentrated on these two pairs is a correlated equilibrium.

It is worth emphasizing that in our result the breaking away from a bad cycle, like the one just described, is obtained not by ergodic arguments but by the probability of staying put (i.e., by friction). What matters is that the diagonal of

[24] For a good test of the simplicity of a procedure, try to explain it verbally; in particular, consider the procedure of our Main Theorem vs. those requiring the computation of eigenvectors.

[25] It is worth pointing out that if a player's last choice was j, then the relative probabilities of switching to k or to k' do not depend only on the average utilities that would have been obtained if j had been changed to k or to k' in the past, but also on the average utility that was obtained in those periods by playing j itself (it is the magnitude of the increases in moving from j to k or to k' that matters).

[26] See Sanchirico (1996) and Section 4.6 in Fudenberg and Levine (1998) for a related point in a best-reply context.

[27] See Step M7 in the Proof of the Main Theorem in the Appendix.

the transition matrix be positive, rather than that all the entries be positive (which, indeed, will not hold in our case).

(i) *The set of correlated equilibria*. The set of correlated equilibria of a game is, in contrast to the set of Nash equilibria, geometrically simple: It is a convex set (actually, a convex polytope) of distributions. Since it includes the Nash equilibria we know it is nonempty. Hart and Schmeidler (1989) (see also Nau and McCardle (1990)) provide an elementary (nonfixed point) proof of the nonemptiness of the set of correlated equilibria. This is done by using the Minimax Theorem. Specifically, Hart and Schmeidler proceed by associating to the given N-person game an auxiliary two-person zero-sum game. As it turns out, the correlated equilibria of the original game correspond to the maximin strategies of player I in the auxiliary game. More precisely, in the Hart–Schmeidler auxiliary game, player I chooses a distribution over N-tuples of actions, and player II chooses a pair of strategies for one of the N original players (interpreted as a play and a suggested deviation from it). The payoff to auxiliary player II is the expected gain of the designated original player if he were to follow the change suggested by auxiliary player II. In other words, it is the "regret" of that original player for not deviating. The starting point for our research was the observation that fictitious play applied to the Hart–Schmeidler auxiliary game must converge, by the result of Robinson (1951), and thus yield optimal strategies in the auxiliary game, in particular for player I—hence, correlated equilibria in the original game. A direct application of this idea does not, however, produce anything that is simple and separable across the N players (i.e., such that the choice of each player at time t is made independently of the other players' choices at t—an indispensable requirement).[28] Yet, our adaptive procedure is based on "no-regret" ideas motivated by this analysis and it is the direct descendant—several modifications later—of this line of research.[29]

(j) *The case of the unknown game*. The adaptive procedure of Section 2 can be modified[30] to yield convergence to correlated equilibria also in the case where players neither know the game, nor observe the choices of the other players.[31] Specifically, in choosing play probabilities at time $t+1$, a player uses information *only* on his own actual past play and payoffs (and *not* on the payoffs that would have been obtained if his past play had been different). The construction

[28] This needed "decoupling" across the N original players explains why applying linear programming-type methods to reach the convex polytope of correlated equilibria is not a fruitful approach. The resulting procedures operate in the space of N-tuples of strategies S (more precisely, in $\Delta(S)$), whereas adaptive procedures should be defined for each player i separately (i.e., on $\Delta(S^i)$).

[29] For another interesting use of the auxiliary two-person zero-sum game, see Myerson (1997).

[30] Following a suggestion of Dean Foster.

[31] For similar constructions, see: Baños (1968), Megiddo (1980), Foster and Vohra (1993), Auer et al. (1995), Roth and Erev (1995), Erev and Roth (1998), Camerer and Ho (1998), Marimon (1996, Section 3.4), and Fudenberg and Levine (1998, Section 4.8). One may view this type of result in terms of "stimulus-response" decision behavior models.

is based on replacing $D_t^i(j,k)$ (see (2.1b)) by

$$C_t^i(j,k) := \frac{1}{t}\left[\sum_{\tau \le t : s_\tau^i = k} \frac{p_\tau^i(j)}{p_\tau^i(k)} u^i(s_\tau) - \sum_{\tau \le t : s_\tau^i = j} u^i(s_\tau)\right].$$

Thus, the payoff that player i would have received had he played k rather than j is estimated by the actual payoffs he obtained when he did play k in the past.

For precise formulations, results and proofs, as well as further discussions, the reader is referred to Hart and Mas-Colell (2000).

Center for Rationality and Interactive Decision Theory, Dept. of Economics, and Dept. of Mathematics, The Hebrew University of Jerusalem, Feldman Bldg., Givat-Ram, 91904 Jerusalem, Israel; hart@math.huji.ac.il; http://www.ma.huji.ac.il/~hart

and

Dept. de Economia i Empresa, and CREI, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain; mcolell@upf.es; http://www.econ.upf.es/crei/mcolell.htm

Manuscript received November, 1997; final revision received July, 1999.

APPENDIX: PROOF OF THE MAIN THEOREM

This appendix is devoted to the proof of the Main Theorem, stated in Section 2. The proof is inspired by the result of Section 3 (Theorem A). It is however more complex on account of our transition probabilities not being the invariant measures that, as we saw in Section 3, fitted so well with Blackwell's Approachability Theorem.

As in the standard proof of Blackwell's Approachability Theorem, the proof of our Main Theorem is based on a recursive formula for the distance of the vector of regrets to the negative orthant. However, our procedure (2.2) does not satisfy the Blackwell condition; it is rather a sort of iterative approximation to it. Thus, a simple one-period recursion (from t to $t+1$) does not suffice, and we have to consider instead a multi-period recursion where a large "block" of periods, from t to $t+v$, is combined together. Both t and v are carefully chosen; in particular, t and v go to infinity, but v is relatively small compared to t.

We start by introducing some notation. Fix player i in N. For simplicity, we drop reference to the index i whenever this cannot cause confusion (thus we write D_t and R_t instead of D_t^i and R_t^i, and so on). Let $m := |S^i|$ be the number of strategies of player i, and let M be an upper bound on i's possible payoffs: $M \ge |u^i(s)|$ for all s in S. Denote $L := \{(j,k) \in S^i \times S^i : j \ne k\}$; then $\mathbb{R}^L$ is the $m(m-1)$-dimensional Euclidean space with coordinates indexed by L. For each $t = 1, 2, \ldots$ and each (j,k) in L, put[32]

$$A_t(j,k) = 1_{\{s_t^i = j\}}[u^i(k, s_t^{-i}) - u^i(s_t)],$$

$$D_t(j,k) = \frac{1}{t}\sum_{\tau=1} A_\tau(j,k),$$

$$R_t(j,k) = D_t^+(j,k) \equiv [D_t(j,k)]^+.$$

We shall write A_t for the vector $(A_t(j,k))_{j \ne k} \in \mathbb{R}^L$; the same goes for D_t, D_t^+, R_t, and so on. Let

[32] We write 1_G for the indicator of the event G.

$\Pi_t(\cdot,\cdot)$ denote the transition probabilities from t to $t+1$ (these are computed after period t, based on h_t):

$$\Pi_t(j,k) := \begin{cases} \dfrac{1}{\mu} R_t(j,k), & \text{if } k \neq j, \\ 1 - \displaystyle\sum_{k' \neq j} \dfrac{1}{\mu} R_t(j,k'), & \text{if } k = j. \end{cases}$$

Thus, at time $t+1$ the strategy used by player i is to choose each $k \in S^i$ with probability $p^i_{t+1}(k) = \Pi_t(s^i_t, k)$. Note that the choice of μ guarantees that $\Pi_t(j,j) > 0$ for all $j \in S^i$ and all t. Finally, let

$$\rho_t := [\operatorname{dist}(D_t, \mathbb{R}^L_-)]^2$$

be the squared distance (in $\mathbb{R}^L$) of the vector D_t to the nonpositive orthant $\mathbb{R}^L_-$. Since the closest point to D_t in $\mathbb{R}^L_-$ is[33] D_t^-, we have $\rho_t = \|D_t - D_t^-\|^2 = \|D_t^+\|^2 = \sum_{j \neq k} [D_t^+(j,k)]^2$.

It will be convenient to use the standard "O" notation: For two real-valued functions $f(\cdot)$ and $g(\cdot)$ defined on a domain X, "$f(x) = O(g(x))$" means that there exists a constant $K < \infty$ such that $|f(x)| \leq Kg(x)$ for all x in[34] X. We write P for Probability, and E for Expectation. From now on, t, v, and w will denote positive integers; $h_t = (s_\tau)_{\tau \leq t}$ will be histories of length t; j, k, and s^i will be elements of S^i; s and s^{-i} will be elements of S and S^{-i}, respectively. Unless stated otherwise, all statements should be understood to hold "for all t, v, h_t, j, k, etc."; where histories h_t are concerned, only those that occur with positive probability are considered.

We divide the proof of the Main Theorem into 11 steps, M1–M11, which we now state formally; an intuitive guide follows.

- *Step M*1:

(i) $$E[(t+v)^2 \rho_{t+v} | h_t] \leq t^2 \rho_t + 2t \sum_{w=1}^{v} R_t \cdot E[A_{t+w} | h_t] + O(v^2); \quad \text{and}$$

(ii) $$(t+v)^2 \rho_{t+v} - t^2 \rho_t = O(tv + v^2).$$

Define

$$\alpha_{t,w}(j, s^{-i}) := \sum_{k \in S^i} \Pi_t(k,j) P[s_{t+w} = (k, s^{-i}) | h_t] - P[s_{t+w} = (j, s^{-i}) | h_t].$$

- *Step M*2:

$$R_t \cdot E[A_{t+w} | h_t] = \mu \sum_{s^{-i} \in S^{-i}} \sum_{j \in S^i} \alpha_{t,w}(j, s^{-i}) u^i(j, s^{-i}).$$

- *Step M*3:

$$R_{t+v}(j,k) - R_t(j,k) = O\left(\frac{v}{t}\right).$$

For each $t > 0$ and each history h_t, define an auxiliary stochastic process $(\hat{s}_{t+w})_{w=0,1,2,\ldots}$ with values in S as follows: The initial value is $\hat{s}_t = s_t$, and the transition probabilities are[35]

$$P[\hat{s}_{t+w} = s | \hat{s}_t, \ldots, \hat{s}_{t+w-1}] := \prod_{i' \in N} \Pi_t^{i'}(\hat{s}^{i'}_{t+w-1}, s^{i'}).$$

[33] We write $[x]^-$ for $\min\{x, 0\}$, and D_t^- for the vector $([D_t(j,k)]^-)_{(j,k) \in L}$.

[34] The domain X will usually be the set of positive integers, or the set of vectors whose coordinates are positive integers. Thus when we write, say, $f(t,v) = O(v)$, it means $|f(t,v)| \leq Kv$ for all v and t. The constants K will always depend only on the game (through N, m, M, and so on) and on the parameter μ.

[35] We write $\Pi_t^{i'}$ for the transition probability matrix of player i' (thus Π_t is Π_t^i).

(The $\hat{s}$-process is thus stationary: It uses the transition probabilities of period t at each period $t+w$, for all $w \geq 0$.)

- *Step M*4:

$$P[s_{t+w} = s \mid h_t] - P[\hat{s}_{t+w} = s \mid h_t] = O\left(\frac{w^2}{t}\right).$$

Define

$$\hat{\alpha}_{t,w}(j, s^{-i}) := \sum_{k \in S^i} \Pi_t(k, j) P[\hat{s}_{t+w} = (k, s^{-i}) \mid h_t] - P[\hat{s}_{t+w} = (j, s^{-i}) \mid h_t].$$

- *Step M*5:

$$\alpha_{t,w}(j, s^{-i}) - \hat{\alpha}_{t,w}(j, s^{-i}) = O\left(\frac{w^2}{t}\right).$$

- *Step M*6:

$$\hat{\alpha}_{t,w}(j, s^{-i}) = P[\hat{s}_{t+w}^{-i} = s^{-i} \mid h_t][\Pi_t^{w+1} - \Pi_t^{w}](s_t^i, j),$$

where $\Pi_t^w \equiv (\Pi_t)^w$ is the wth power of the matrix Π_t, and $[\Pi_t^{w+1} - \Pi_t^{w}](s_t^i, j)$ denotes the (s_t^i, j) element of the matrix $\Pi_t^{w+1} - \Pi_t^{w}$.

- *Step M*7:

$$\hat{\alpha}_{t,w}(j, s^{-1}) = O(w^{-1/2}).$$

- *Step M*8:

$$E[(t+v)^2 \rho_{t+v} \mid h_t] \leq t^2 \rho_t + O(v^3 + tw^{1/2}).$$

For each $n = 1, 2, \ldots$, let $t_n := \lfloor n^{5/3} \rfloor$ be the largest integer not exceeding $n^{5/3}$.

- *Step M*9:

$$E[t_{n+1}^2 \rho_{t_{n+1}} \mid h_{t_n}] \leq t_n^2 \rho_{t_n} + O(n^2).$$

- *Step M*10:

$$\lim_{n \to \infty} \rho_{t_n} = 0 \quad \text{a.s.}$$

- *Step M*11:

$$\lim_{t \to \infty} R_t(j, k) = 0 \quad \text{a.s.}$$

We now provide an intuitive guide to the proof. The first step (M1(i)) is our basic recursion equation. In Blackwell's Theorem, the middle term on the right-hand side vanishes (it is ≤ 0 by (3.2)). This is not so in our case; Steps M2–M8 are thus devoted to estimating this term. Step M2 yields an expression similar to (3.4), but here the coefficients α depend also on the moves of the other players. Indeed, given h_t, the choices s_{t+w}^i and s_{t+w}^{-i} are *not* independent when $w > 1$ (since the transition probabilities change with time). Therefore we replace the process $(s_{t+w})_{0 \leq w \leq v}$ by another process $(\hat{s}_{t+w})_{0 \leq w \leq v}$, with a *stationary* transition matrix (that of period t). For w small relative to t, the change in probabilities is small (see Steps M3 and M4), and we estimate the total difference (Step M5). Next (Step M6), we factor out the moves of the other players (which, in the $\hat{s}$-process, are independent of the moves of player i) from the coefficients $\hat{\alpha}$. At this point we get the

difference between the transition probabilities after w periods and after $w+1$ periods (for comparison, in formula (3.4) we would replace both by the invariant distribution, so the difference vanishes). This difference is shown (Step M7) to be small, since w is large and the transition matrix has all its diagonal elements strictly positive.[36] Substituting in M1(i) yields the final recursive formula (Step M8). The proof is now completed (Steps M9–M11) by considering a carefully chosen subsequence of periods $(t_n)_{n=1,2,\ldots}$.

The rest of this Appendix contains the proofs of the Steps M1–M11.

- PROOF OF STEP M1: Because $D_t^- \in \mathbb{R}_-^L$ we have

$$\begin{aligned}\rho_{t+v} \le \|D_{t+v} - D_t^-\|^2 &= \left\| \frac{t}{t+v} D_t + \frac{1}{t+v} \sum_{w=1}^{v} A_{t+w} - D_t^- \right\|^2 \\ &= \frac{t^2}{(t+v)^2} \|D_t - D_t^-\|^2 + \frac{2t}{(t+v)^2} \sum_{w=1}^{v} (A_{t+w} - D_t^-) \cdot (D_t - D_t^-) \\ &\quad + \frac{v^2}{(t+v)^2} \left\| \frac{1}{v} \sum_{w=1}^{v} A_{t+w} - D_t^- \right\|^2 \\ &\le \frac{t^2}{(t+v)^2} \rho_t + \frac{2t}{(t+v)^2} \sum_{w=1}^{v} A_{t+w} \cdot R_t + \frac{v^2}{(t+v)^2} m(m-1) 16 M^2.\end{aligned}$$

Indeed: $|u^i(s)| \le M$, so $|A_{t+w}(j,k)| \le 2M$ and $|D_t(j,k)| \le 2M$, yielding the upper bound on the third term. As for the second term, note that $R_t = D_t^+ = D_t - D_t^-$ and $D_t^- \cdot D_t^+ = 0$. This gives the bound of (ii). To get (i), take conditional expectation given the history h_t (so ρ_t and R_t are known). *Q.E.D.*

- PROOF OF STEP M2: We have

$$E[A_{t+w}(j,k) | h_t] = \sum_{s^{-i}} \phi(j, s^{-i}) [u^i(k, s^{-i}) - u^i(j, s^{-i})],$$

where $\phi(j, s^{-i}) := P[s_{t+w} = (j, s^{-i}) | h_t]$. So

$$\begin{aligned}R_t \cdot E[A_{t+w} | h_t] &= \sum_j \sum_{k \ne j} R_t(j,k) \sum_{s^{-i}} \phi(j, s^{-i}) [u^i(k, s^{-i}) - u^i(j, s^{-i})] \\ &= \sum_{s^{-i}} \sum_j u^i(j, s^{-i}) \left[\sum_{k \ne j} R_t(k,j) \phi(k, s^{-i}) - \sum_{k \ne j} R_t(j,k) \phi(j, s^{-i}) \right]\end{aligned}$$

(we have collected together all terms containing $u^i(j, s^{-i})$). Now, $R_t(k,j) = \mu \Pi_t(k,j)$ for $k \ne j$, and $\sum_{k \ne j} R_t(j,k) = \mu(1 - \Pi_t(j,j))$ by definition, so

$$R_t \cdot E[A_{t+w} | h_t] = \mu \sum_{s^{-i}} \sum_j u^i(j, s^{-i}) \left[\sum_k \Pi_t(k,j) \phi(k, s^{-i}) - \phi(j, s^{-i}) \right]$$

(note that the last sum is now over *all* k in S^i). *Q.E.D.*

- PROOF OF STEP M3: This follows immediately from

$$(t+v)[D_{t+v}(j,k) - D_t(j,k)] = \sum_{w=1}^{v} A_{t+w}(j,k) - v D_t(j,k),$$

together with $|A_{t+w}(j,k)| \le 2M$ and $|D_t(j,k)| \le 2M$. *Q.E.D.*

[36] For further discussion on this point, see the Proof of Step M7.

• PROOF OF STEP M4: We need the following Lemma, which gives bounds for the changes in the w-step transition probabilities as a function of changes in the 1-step transitions.

LEMMA: *Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be two stochastic processes with values in a finite set B. Assume $X_0 = Y_0$ and*

$$|P[X_n = b_n | X_0 = b_0, \ldots, X_{n-1} = b_{n-1}] - P[Y_n = b_n | Y_0 = b_0, \ldots, Y_{n-1} = b_{n-1}]| \leq \beta_n$$

for all $n \geq 1$ and all $b_0, \ldots, b_{n-1}, b_n \in B$. Then

$$|P[X_{n+w} = b_{n+w} | X_0 = b_0, \ldots, X_{n-1} = b_{n-1}]$$
$$- P[Y_{n+w} = b_{n+w} | Y_0 = b_0, \ldots, Y_{n-1} = b_{n-1}]| \leq |B| \sum_{r=0}^{w} \beta_{n+r}$$

for all $n \geq 1$, $w \geq 0$, and all $b_0, \ldots, b_{n-1}, b_{n+w} \in B$.

PROOF: We write P_X and P_Y for the probabilities of the two processes $(X_n)_n$ and $(Y_n)_n$, respectively (thus $P_X[b_{n+w}|b_0, \ldots, b_{n-1}]$ stands for $P[X_{n+w} = b_{n+w} | X_0 = b_0, \ldots, X_{n-1} = b_{n-1}]$, and so on). The proof is by induction on w.

$$\begin{aligned}
&P_X[b_{n+w}|b_0, \ldots, b_{n-1}] \\
&\quad = \sum_{b_n} P_X[b_{n+w}|b_0, \ldots, b_n] P_X[b_n|b_0, \ldots, b_{n-1}] \\
&\quad \leq \sum_{b_n} P_Y[b_{n+w}|b_0, \ldots, b_n] P_X[b_n|b_0, \ldots, b_{n-1}] + |B| \sum_{r=1}^{w} \beta_{n+r} \\
&\quad \leq \sum_{b_n} P_Y[b_{n+w}|b_0, \ldots, b_n](P_Y[b_n|b_0, \ldots, b_{n-1}] + \beta_n) + |B| \sum_{r=1}^{w} \beta_{n+r} \\
&\quad \leq P_Y[b_{n+w}|b_0, \ldots, b_{n-1}] + |B| \beta_n + |B| \sum_{r=1}^{w} \beta_{n+r}
\end{aligned}$$

(the first inequality is by the induction hypothesis). Exchanging the roles of X and Y completes the proof. *Q.E.D.*

We proceed now with the proof of Step M4. From t to $t+w$ there are $|N|w$ transitions (at each period, think of the players moving one after the other, in some arbitrary order). Step M3 implies that each transition probability for the $\hat{s}$-process differs from the corresponding one for the s-process by at most $O(w/t)$, which yields, by the Lemma, a total difference of $|N|w|S|O(w/t) = O(w^2/t)$. *Q.E.D.*

• PROOF OF STEP M5: Immediate by Step M4. *Q.E.D.*

• PROOF OF STEP M6: Given h_t, the random variables $(\hat{s}^{i'}_{t+w})_w$ are independent over the different players i' in N; indeed, the transition probabilities are all determined at time t, and the players randomize independently. Hence:

$$P[\hat{s}_{t+w} = (j, s^{-i}) | h_t] = P[\hat{s}^{-i}_{t+w} = s^{-i} | h_t] P[\hat{s}^{i}_{t+w} = j | h_t],$$

implying that

$$\hat{\alpha}_{t,w}(j, s^{-i}) = P[\hat{s}^{-i}_{t+w} = s^{-i} | h_t] \left[\sum_{k \in S^i} \Pi_t(k, j) P[\hat{s}^{i}_{t+w} = k | h_t] - P[\hat{s}^{i}_{t+w} = j | h_t] \right].$$

Now $P[\hat{s}^i_{t+w} = j \mid h_t]$ is the probability of reaching j in w steps starting from s^i_t, using the transition probability matrix Π_t. Therefore $P[\hat{s}^i_{t+w} = j \mid h_t]$ is the (s^i_t, j)-element of the wth power $\Pi_t{}^w \equiv (\Pi_t)^w$ of Π_t, i.e., $[\Pi_t{}^w](s^i_t, j)$. Hence

$$\begin{aligned}\hat{\alpha}_{t,w}(j, s^{-i}) &= P[\hat{s}^{-i}_{t+w} = s^{-i} \mid h_t]\left[\sum_{k \in S^i} \Pi_t(k, j)[\Pi_t{}^w](s^i_t, k) - [\Pi_t{}^w](s^i_t, j)\right] \\ &= P[\hat{s}^{-i}_{t+w} = s^{-i} \mid h_t][[\Pi_t{}^{w+1}](s^i_t, j) - [\Pi_t{}^w](s^i_t, j)],\end{aligned}$$

completing the proof. *Q.E.D.*

• PROOF OF STEP M7: It follows from M6 using the following Lemma (recall that $\Pi_t(j, j) > 0$ for all $j \in S^i$).

LEMMA: *Let Π be an $m \times m$ stochastic matrix with all of its diagonal entries positive. Then $[\Pi^{w+1} - \Pi^w](j, k) = O(w^{-1/2})$ for all $j, k = 1, \ldots, m$.*

PROOF:[37] Let $\beta > 0$ be a lower bound on all the diagonal entries of Π, i.e., $\beta := \min_j \Pi(j, j)$. We can then write $\Pi = \beta I + (1 - \beta)\Lambda$, where Λ is also a stochastic matrix. Now

$$\Pi^w = \sum_{r=0}^{w} \binom{w}{r} \beta^{w-r}(1-\beta)^r \Lambda^r,$$

and similarly for Π^{w+1}. Subtracting yields

$$\Pi^{w+1} - \Pi^w = \sum_{r=0}^{w+1} \gamma_r \binom{w}{r} \beta^{w-r}(1-\beta)^r \Lambda^r,$$

where $\gamma_r := \beta(w+1)/(w+1-r) - 1$. Now $\gamma_r > 0$ if $r > q := (w+1)(1-\beta)$, and $\gamma_r \le 0$ if $r \le q$; together with $0 \le \Lambda^r(j, k) \le 1$, we get

$$\sum_{r \le q} \gamma_r \binom{w}{r} \beta^{w-r}(1-\beta)^r \le [\Pi^{w+1} - \Pi^w](j, k) \le \sum_{r > q} \gamma_r \binom{w}{r} \beta^{w-r}(1-\beta)^r.$$

Consider the left-most sum. It equals

$$\sum_{r \le q} \binom{w+1}{r} \beta^{w+1-r}(1-\beta)^r - \sum_{r \le q} \binom{w}{r} \beta^{w-r}(1-\beta)^r = G_{w+1}(q) - G_w(q),$$

where $G_n(\cdot)$ denotes the cumulative distribution function of a sum of n independent Bernoulli random variables, each one having the value 0 with probability β and the value 1 with probability $1 - \beta$. Using the normal approximation yields (Φ denotes the standard normal cumulative distribution function):

$$G_{w+1}(q) - G_w(q) = \Phi(x) - \Phi(y) + O\left(\frac{1}{\sqrt{(w+1)}}\right) + O\left(\frac{1}{\sqrt{w}}\right),$$

where

$$x := \frac{q - (w+1)(1-\beta)}{\sqrt{(w+1)\beta(1-\beta)}} \qquad \text{and} \qquad y := \frac{q - w(1-\beta)}{\sqrt{w\beta(1-\beta)}};$$

[37] If Π were a strictly positive matrix, then $\Pi^{w+1} - \Pi^w \to 0$ would be a standard result, because then Π^w would converge to the invariant matrix. However, we know only that the diagonal elements are positive. This implies that, if w is large, then with high probability there will be a positive fraction of periods when the process does not move. But this number is random, so the probabilities of going from j to k in w steps or in $w + 1$ steps should be almost the same (since it is like having r "stay put" transitions versus $r + 1$).

the two error terms $O((w+1)^{-1/2})$ and $O(w^{-1/2})$ are given by the Berry-Esséen Theorem (see Feller (1965, Theorem XVI.5.1)). By definition of q we have $x=0$ and $y=O(w^{-1/2})$. The derivative of Φ is bounded, so $\Phi(x)-\Phi(y)=O(x-y)=O(w^{-1/2})$. Altogether, the left-most sum is $O(w^{-1/2})$. A similar computation applies to the right-most sum. *Q.E.D.*

• PROOF OF STEP M8: Steps M5 and M7 imply $\alpha_{t,w}(j,s^{-i})=O(w^2/t+w^{-1/2})$. The formula of Step M2 then yields

$$R_t \cdot E[A_{t+w}|h_t]=O\left(\frac{w^2}{t}+w^{-1/2}\right).$$

Adding over $w=1,2,\ldots,v$ (note that $\sum_{w=1}^{v} w^{\lambda}=O(v^{\lambda+1})$ for $\lambda\neq -1$) and substituting into Step M1(i) gives the result. *Q.E.D.*

• PROOF OF STEP M9: We use the inequality of Step M8 for $t=t_n$ and $v=t_{n+1}-t_n$. Because $v=\lfloor (n+1)^{5/3}\rfloor-\lfloor n^{5/3}\rfloor=O(n^{2/3})$, we have $v^3=O(n^2)$ and $tv^{1/2}=O(n^{5/3+1/3})=O(n^2)$, and the result follows. *Q.E.D.*

• PROOF OF STEP M10: We use the following result (see Loève (1978, Theorem 32.1.E)):

THEOREM (Strong Law of Large Numbers for Dependent Random Variables): *Let X_n be a sequence of random variables and b_n a sequence of real numbers increasing to ∞, such that the series $\sum_{n=1}^{\infty} \operatorname{var}(X_n)/b_n^2$ converges. Then*

$$\frac{1}{b_n}\sum_{\nu=1}^{n}[X_\nu-E[X_\nu|X_1,\ldots,X_{\nu-1}]]\xrightarrow[n\to\infty]{}0 \quad a.s.$$

We take $b_n:=t_n^2$, and $X_n:=b_n\rho_{t_n}-b_{n-1}\rho_{t_{n-1}}=t_n^2\rho_{t_n}-t_{n-1}^2\rho_{t_{n-1}}$. By Step M1(ii) we have $|X_n|\le O(t_n v_n+v_n^2)=O(n^{7/3})$, thus $\sum_n \operatorname{var}(x_n)/b_n^2=\sum_n O(n^{14/3})/n^{20/3}=\sum_n O(1/n^2)<\infty$. Next, Step M9 implies

$$(1/b_n)\sum_{\nu\le n}E[X_\nu|X_1,\ldots,X_{\nu-1}]\le O\left(n^{-10/3}\sum_{\nu\le n}\nu^2\right)=O(n^{-10/3}n^3)=O(n^{-1/3})\to 0.$$

Applying the Theorem above thus yields that ρ_{t_n}, which is nonnegative and equals $(1/b_n)\sum_{\nu\le n}X_\nu$, must converge to 0 a.s. *Q.E.D.*

• PROOF OF STEP M11: Since $\rho_{t_n}=\sum_{j\neq k}[R_{t_n}(j,k)]^2$, the previous Step M10 implies that $R_{t_n}(j,k)\to 0$ a.s. $n\to\infty$, for all $j\neq k$. When $t_n\le t\le t_{n+1}$, we have $R_t(j,k)-R_{t_n}(j,k)=O(n^{-1})$ by the inequality of Step M3, so $R_t(j,k)\to 0$ a.s. $t\to\infty$. *Q.E.D.*

REFERENCES

AUER, P., N. CESA-BIANCHI, Y. FREUND, AND R. E. SCHAPIRE (1995): "Gambling in a Rigged Casino: The Adversarial Multi-Armed Bandit Problem," in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 322–331.

AUMANN, R. J. (1974): "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1, 67–96.

——— (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18.

BAÑOS, A. (1968): "On Pseudo-Games," *The Annals of Mathematical Statistics*, 39, 1932–1945.

BLACKWELL, D. (1956a): "An Analog of the Minmax Theorem for Vector Payoffs," *Pacific Journal of Mathematics*, 6, 1–8.

——— (1956b): "Controlled Random Walks," in *Proceedings of the International Congress of Mathematicians 1954, Vol. III*, ed. by E. P. Noordhoff. Amsterdam: North-Holland, pp. 335–338.

BROWN, G. W. (1951): "Iterative Solutions of Games by Fictitious Play," in *Activity Analysis of Production and Allocation*, Cowles Commission Monograph 13, ed. by T. C. Koopmans. New York: Wiley, pp. 374–376.

CAHN, A. (2000): "General Procedures Leading to Correlated Equilibria," The Hebrew University of Jerusalem, Center for Rationality DP-216.

CAMERER, C., AND T.-H. HO (1998): "Experience-Weighted Attraction Learning in Coordination Games: Probability Rules, Heterogeneity, and Time-Variation," *Journal of Mathematical Psychology*, 42, 305–326.

EREV, I., AND A. E. ROTH (1998): "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review*, 88, 848–881.

FELLER, W. (1965): *An Introduction to Probability Theory and its Applications, Vol. II*, 2nd edition. New York: Wiley.

FOSTER, D., AND R. V. VOHRA (1993): "A Randomization Rule for Selecting Forecasts," *Operations Research*, 41, 704–709.

——— (1997): "Calibrated Learning and Correlated Equilibrium," *Games and Economic Behavior*, 21, 40–55.

——— (1998): "Asymptotic Calibration," *Biometrika*, 85, 379–390.

——— (1999): "Regret in the On-line Decision Problem," *Games and Economic Behavior*, 29, 7–35.

FUDENBERG, D., AND D. K. LEVINE (1995): "Universal Consistency and Cautious Fictitious Play," *Journal of Economic Dynamics and Control*, 19, 1065–1089.

——— (1998): *Theory of Learning in Games*. Cambridge, MA: The MIT Press.

——— (1999): "Conditional Universal Consistency," *Games and Economic Behavior*, 29, 104–130.

HANNAN, J. (1957): "Approximation to Bayes Risk in Repeated Play," in *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, ed. by M. Dresher, A. W. Tucker, and P. Wolfe. Princeton: Princeton University Press, pp. 97–139.

HART, S., AND A. MAS-COLELL (1999): "A General Class of Adaptive Strategies," The Hebrew University of Jerusalem, Center for Rationality DP-192, forthcoming in *Journal of Economic Theory*.

——— (2000): "A Stimulus-Response Procedure Leading to Correlated Equilibrium," The Hebrew University of Jerusalem, Center for Rationality (mimeo).

HART, S., AND D. SCHMEIDLER (1989): "Existence of Correlated Equilibria," *Mathematics of Operations Research*, 14, 18–25.

LOÈVE, M. (1978): *Probability Theory, Vol. II*, 4th edition. Berlin: Springer-Verlag.

LUCE, R. D., AND H. RAIFFA (1957): *Games and Decisions*. New York: Wiley.

MARIMON, R. (1996): "Learning from Learning in Economics," in *Advances in Economic Theory*, ed. by D. Kreps. Cambridge: Cambridge University Press.

MEGIDDO, N. (1980): "On Repeated Games with Incomplete Information Played by Non-Bayesian Players," *International Journal of Game Theory*, 9, 157–167.

MERTENS, J.-F., S. SORIN, AND S. ZAMIR (1995): "Repeated Games, Part A," CORE DP-9420 (mimeo).

MYERSON, R. B. (1997): "Dual Reduction and Elementary Games," *Games and Economic Behavior*, 21, 183–202.

NAU, R. F., AND K. F. MCCARDLE (1990): "Coherent Behavior in Noncooperative Games," *Journal of Economic Theory*, 50, 424–444.

ROBINSON, J. (1951): "An Iterative Method of Solving a Game," *Annals of Mathematics*, 54, 296–301.

ROTH, A. E., AND I. EREV (1995): "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior*, 8, 164–212.

SANCHIRICO, C. W. (1996): "A Probabilistic Model of Learning in Games," *Econometrica*, 64, 1375–1393.