30 October 1997

A Simple Adaptive Procedure Leading to Correlated Equilibrium[1]

by

Sergiu Hart[2] and Andreu Mas-Colell[3]

Abstract

We propose a new and simple adaptive procedure for playing a game: "regret-matching." In this procedure, players 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. To compute these regret measures, a player needs to know his payoff function and the history of play. We also offer a variation where every player knows only his own realized payoff history (but not his payoff function).

---

[1] Previous versions: 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. We want to acknowledge the useful comments and suggestions of Robert Aumann, Antonio Cabrales, Dean Foster, Drew Fudenberg, David Levine, Al Roth, Reinhard Selten, Sylvain Sorin, the anonymous referees, and the participants at various seminars where this work was presented.
[2] Department of Economics; Department of Mathematics; and Center for Rationality and Interactive Decision Theory, The Hebrew University of Jerusalem, Givat Ram, 91904 Jerusalem, Israel.
e-mail: hart@sunset.ma.huji.ac.il http://www.ma.huji.ac.il/~hart
[3] Departament de Economia i Empresa and CREI, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain.
e-mail: mcolell@upf.es

1. Introduction

The leading non-cooperative 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. If the signals are (stochastically) independent across the players, this is just 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 practical point of view, it could be argued that correlated equilibrium may be the most relevant non-cooperative solution concept.

Indeed, with the possible exception of well controlled environments, it may be hard to eliminate a priori the possibility that correlating signals be 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?

A procedure with the desired convergence property was proposed by Foster and Vohra [1996]. Fudenberg and Levine [1996] offered a second one. We introduce here a procedure which we view as particularly simple and intuitive (see Section 5 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 would have played k every time in the past he chose j (and everything else remains unchanged). Then only those strategies k with V(k) larger than U are chosen, 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 as in the last period is chosen.

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, 1996] or calibrated learning [Foster and Vohra, 1996]; see Section 5 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 reach the smaller set of Nash equilibria? The answer is definitely no. On the one hand, in ours, as in most other procedures, there is a natural coordination device: the common history, observed by all players. It is thus very 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.[4] This is a new instrument in this field, which may well turn out to be widely applicable.

In Section 4 we consider the case where players do *not* know the game — neither their own payoff matrix nor the other players' payoff matrices. Each player observes (and remembers) only his own realized past actions and payoffs. The play (and the payoffs) of the other players are *not* observed. We present a modification[5] in this setup of the adaptive procedure of Section 2. Again, the dynamics leads to the set of correlated equilibria.

Section 5 contains a discussion of the literature, together with a number of relevant issues. The proof of the Main Theorem is relegated to Appendix 1, and the proof of the results for the unknown game case, to Appendix 2.

2. The Model and Main Result

Let $\Gamma = (N, (S^i)_{i\in N}, (u^i)_{i\in N})$ be an N-person finite 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 member of S is $s = (s^i)_{i\in N}$, and s^{-i} denotes the

---

[4] See Mertens, Sorin and Zamir [1995, Theorem 4.3] for a concise proof.
[5] The modification is based on a suggestion of Dean Foster; see Foster and Vohra [1993] and Auer *et al.* [1995].

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:[6]

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

If in the above inequality we replace 0the 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 non-empty, 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.

Suppose now that the game Γ is played repeatedly through time: t = 1, 2, ... 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[7] $p_{t+1}^i \in \Delta(S^i)$ which is defined as follows: 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 $\tau \leq t$, would then become

---

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

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

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

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

$$\begin{aligned} D^i_t(j,k) &:= \frac{1}{t}\sum_{\tau=1}^{t} W^i_\tau(j,k) - \frac{1}{t}\sum_{\tau=1}^{t} u^i(s_\tau), \qquad (2.1b) \\ &= \frac{1}{t} \sum_{\tau\le t: s^i_\tau = j} \left[u^i(k,s^{-i}_\tau) - u^i(s_\tau)\right]. \end{aligned}$$

Finally, denote

$$R^i_t(j,k) := [D^i_t(j,k)]^+ \equiv \max\ \{D^i_t(j,k),\ 0\}. \tag{2.1c}$$

The expression $R^i_t(j,k)$ 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.[8] Let $j \in S^i$ be the strategy last chosen by player i, i.e., $j = s^i_t$. Then define the probability distribution $p^i_{t+1} \in \Delta(S^i)$ used by i at time t+1 as

$$\begin{cases} p^i_{t+1}(k) := \frac{1}{\mu} R^i_t(j,k), & \text{for all } k \neq j, \\ p^i_{t+1}(j) := 1 - \sum\limits_{k\in S^i: k\neq j} p^i_{t+1}(k), & \end{cases} \tag{2.2}$$

---

[8] The parameter μ is fixed throughout the procedure (independent of time and history). It suffices to take μ so that $\mu > 2\, M^i\, (m^i - 1)$ 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).

Note that the choice of μ guarantees that $p^i_{t+1}(j) > 0$; that is, there is always a positive probability of playing the same strategy as in the previous period. The play $p^i_1 \in \Delta(S^i)$ at the initial period is chosen arbitrarily.[9]

The probabilities of choosing the different strategies are thus proportional to the regrets relative to the last period strategy. The factor of proportionality is constant. Note in particular that, if the regrets are small, then the probability of switching from previous 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[10] $s \in S$,

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

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

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

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

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 $(z_1, z_2, ..., z_t, ...)$ 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 equilibrium is non-empty 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 5):

(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; actually, the empirical distribution z_t of $(s_1, s_2, ..., s_t)$ suffices. Computationwise, player i needs to keep record of the $m^i(m^i - 1)$ numbers $D_t^i(j,k)$ for all $j \neq k$ in S^i (and update them every period).

(2) At every period the adaptive procedure that we propose randomizes only over the strategies which 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 is thus played with probability at least δ/m^i). The result still holds with the following two modifications: First, the empirical distribution z_t converges to the set of correlated ε-equilibria (of course, ε depends on δ, and it goes to zero as δ goes to zero); and second, the

convergence of the distance is only in probability (rather than almost surely). Thus, 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,[11] μ. This may be viewed as an "inertia" parameter (see also (g) and (h) in Section 5): A higher μ yields smaller probabilities of switching. The convergence to the set of correlated equilibria is always guaranteed (for large enough μ; see Footnote 8), 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[12] 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 which 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 [1996], of Fudenberg and Levine [1996], and of Theorem A in Section 3 below; see the discussion in Section 5. The delimitation of general classes of procedures converging to correlated equilibria seems, therefore, an interesting research problem.

---

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

[12] See the Proposition in Section 3.

3. No Regret and Blackwell Approachability

In this section (which can be viewed as a motivational preliminary) we will 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 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 of this section is interesting in its own right (see for instance the Remark following the statement of Theorem A).

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 $j \in S^i$. 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.,

$p^i_{t+1}(j) := q^i_t(j)$. 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^i_t *that satisfies* (3.1). *Then player* i*'s regrets* $R^i_t(j,k)$ *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 [1996, 1997], the adaptive procedure of this section is "conditionally universally consistent." For an extended discussion of this issue, see (d) in Section 5.

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:

Proposition: Let[13] $\varepsilon \geq 0$ and let $(s_t)_{t=1,2,...}$ be a sequence of plays (i.e., $s_t \in S$ for all t) such that $\limsup_{t\to\infty} R^i_t(j,k) \leq \varepsilon$ for every $i \in N$ and

---

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

every $j, k \in S^i$ with $j \neq k$. Then the sequence of empirical distributions z_t (defined by (2.3)) converges to the set of correlated ε-equilibria.

Proof: Let $\psi \in \Delta(S)$ be an accumulation point of the sequence z_t of empirical distributions of play along this history. Consider an arbitrary player i and arbitrary strategies $j \neq k$ in S^i. Then

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

$$= \sum_{s \in S : s^i = j} z_t(s) \left[u^i(k, s^{-i}) - u^i(j, s^{-i}) \right].$$

This becomes $\le \varepsilon$ in the limit (it is bounded above by $R_t^i(j,k)$). Taking the limit on the subsequence where the empirical distributions z_t converge to ψ yields:

$$\sum_{s \in S : s^i = j} \psi(s) \left[u^i(k, s^{-i}) - u^i(j, s^{-i}) \right] \le \varepsilon.$$

So $\psi \in \Delta(S)$ constitutes a correlated ε-equilibrium. ∎

Theorem A and the Proposition (for $\varepsilon = 0$) immediately imply

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 make a brief presentation of Blackwell's Approachability Theorem (or, rather, that part of it that is relevant here).

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 set S^{-i}. 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. The question is whether the decision-maker i can guarantee that the time average of the vector payoffs, namely $(1/t) \sum_{\tau \le t} v(s_\tau) \equiv (1/t) \sum_{\tau \le t} v(s_\tau^i, s_\tau^{-i})$, approach a pre-determined set (in $\mathbb{R}^L$).

Figure 1 about here

Let $\mathcal{C}$ be a non-empty, closed and convex subset of $\mathbb{R}^L$. For every point $D \in \mathbb{R}^L$, $D \notin \mathcal{C}$, denote by $F(D) \in \mathcal{C}$ the "foot" of D in $\mathcal{C}$, that is, the (unique) vector in $\mathcal{C}$ closest to D in the Euclidean distance. Suppose it to be the case that for every $D \notin \mathcal{C}$ the decision-maker i could resort to a mixed strategy (of the one-shot decision problem) $q(D) \in \Delta(S^i)$, having the following property: For all s^{-i} in S^{-i}, the expected one-shot payoff will be, relative to the hyperplane through F(D) with normal D – F(D), on the same side as $\mathcal{C}$ (i.e., in the half-space $\mathcal{H}$ — see Figure 1). Formally, q(D) satisfies[14]

$$\Big[\sum_{s^i \in S^i} [q(D)](s^i) \; v(s^i,s^{-i}) - F(D) \Big] \bullet \Big[D - F(D) \Big] \le 0, \tag{3.2}$$

---

[14] For two vectors x and y, their scalar product is denoted $x \bullet y$. We write $[q(D)](s^i)$ for the s^i-coordinate of q(D), i.e., the probability given by the mixed strategy q(D) to s^i.

for every $s^{-i} \in S^{-i}$. This is called the *Blackwell condition*. Blackwell's Approachability Theorem then tells us that if at every t+1 the decision maker i randomizes among his strategies according to[15] $q(D_t)$, for $D_t := (1/t) \sum_{\tau \le t} v(s_\tau)$, it follows that the average payoff vector D_t will approach the set $\mathcal{C}$ with probability one. Figure 2 suggests why this is so: If condition (3.2) is satisfied and t is large then[16]

$$\mathbf{E}[D_{t+1} | h_t] = \frac{t}{t+1} D_t + \frac{1}{t+1} \mathbf{E}[v(s^i_{t+1}, s^{-i}_{t+1})]$$

will be *inside* the circle of center $F(D_t)$ and radius $\|D_t - F(D_t)\|$. Hence ("dist" denotes Euclidean distance)

$$\begin{aligned} \text{dist}(\mathbf{E}[D_{t+1} | h_t], \mathcal{C}) &\le \|\mathbf{E}[D_{t+1} | h_t] - F(D_t)\| \\ &< \|D_t - F(D_t)\| = \text{dist}(D_t, \mathcal{C}) \end{aligned}$$

(the first inequality follows from the fact that $F(D_t)$ is in $\mathcal{C}$). A precise computation shows that the distance not only decreases, but actually goes to zero.[17]

Figure 2 about here

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

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

---

[15] If $D_t \in \mathcal{C}$, any choice at time t+1 will do.

[16] We write $\mathbf{E}[\cdot | h_t]$ for the conditional expectation given the history h_t up to time t.

[17] Note that one looks here at *expected* average payoffs; the Law of Large Numbers (for dependent random variables — ee.g., see Loève [1978, 32.1.E]) implies that the *actual* average payoffs also converge to the set $\mathcal{C}$.

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 $\mathcal{C}$ be the non-positive orthant $\mathbb{R}^L_- := \{x \in \mathbb{R}^L : x \le 0\}$ (this is a non-empty, closed, convex subset of $\mathbb{R}^L$). For any vector $D \notin \mathcal{C}$, let $R := [D]^+$ be its positive part (that is, $R(j,k) = \max\ \{D(j,k), 0\}$ for all $j \ne k$) and suppose that $q(D) \in \Delta(S^i)$ is an invariant vector for the non-negative matrix R; again, put formally $R(j,j) = 0$ (such a $q(D)$ always exists). That is, $q(D)$ satisfies

$$\sum_{k \in S^i} [q(D)](k)\ R(k,j) = [q(D)](j) \sum_{k \in S^i} R(j,k), \tag{3.3}$$

for every $j \in S^i$.

Figure 3 about here

We proceed to show that the probability vectors $q(D)$ satisfy Blackwell's condition (3.2) for the set $\mathcal{C} = \mathbb{R}^L_-$. Note first that, for any $D \notin \mathcal{C}$, we have $D - F(D) = [D]^+ = R$; see Figure 3. Then, for any $s^{-i} \in S^{-i}$, the expression in condition (3.2) takes the form

$$\sum_{(j,k) \in L} \ \sum_{s^i \in S^i} [q(D)](s^i)\ [v(s^i,s^{-i})](j,k)\ R(j,k) \le 0.$$

That is,

$$\sum_{(j,k) \in L} [q(D)](j) \left[u^i(k,s^{-i}) - u^i(j,s^{-i})\right] R(j,k) \le 0. \tag{3.4}$$

After collecting terms, the left-hand side of (3.4) can be written as

$$\sum_{j\in S^i} \alpha(j)\ u^i(j,s^{-i}), \tag{3.5a}$$

where

$$\alpha(j) := \sum_{k\in S^i} [q(D)](k)\ R(k,j) - [q(D)](j) \sum_{k\in S^i} R(j,k). \tag{3.5b}$$

By (3.3) it follows that $\alpha(j) = 0$ for all $j \in S^i$ and so inequality (3.4) holds true (as an equality[18]) for all $s^{-i} \in S^{-i}$. Thus Blackwell's condition (3.2) is satisfied. Because, for all t+1, the mixed strategy p^i_{t+1} played by player i at t+1 is precisely $q(D^i_t)$, it follows that $(1/t) \sum_{\tau=1}^{t} v(s_\tau)$ approaches $\mathcal{C} = \mathbb{R}^L_-$ a.s.; that is, $R^i_t(j,k) = \max \{(1/t) \sum_{\tau=1}^{t} [v(s_\tau)](j,k),\ 0\}$ goes to zero a.s., for every $j \neq k$. ∎

Remark: Blackwell's Approachability Theorem also provides bounds on the speed of convergence. In our case, we get the following: The expectation $\mathbf{E}[R^i_t(j,k)]$ 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 [1997, Section 2.3]). Clearly, a better speed of convergence[19] 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.

---

[18] Note that this is precisely Formula (2) in the Proof of Theorem 1 in Hart and Schmeidler [1989]; see (i) in Section 5.
[19] Up to a constant factor.

4. The Case of the Unknown Game

In this section we present a variation of our basic adaptive procedure of Section 2 that preserves the convergence property to correlated equilibria and where players do not know the game nor do they observe the choices of the other players. Thus, in choosing play probabilities at time t+1, each player uses information *only* on his own actual past play and payoffs (and not on possible payoffs that could have been obtained if past play had been different from what it was).

The setup is as in Section 2. However, the information of player i at time t+1 consists just of the t numbers $u^i(s_1)$, $u^i(s_2)$, ..., $u^i(s_t)$ (in addition to his t past choices s^i_1, s^i_2, ..., s^i_t). There will be two kinds of modifications on the definition of the adaptive procedure of Section 2. The first and most significant modification concerns the definition of the regret measures (2.1). Indeed, for each $j \neq k$ in S^i and each t, we use, instead of $D^i_t(j,k) = (1/t)\ \Sigma_{\tau \le t: s^i_\tau = j}\ [u^i(k,s^{-i}_\tau) - \Sigma_{\tau \le t: s^i_\tau = j}\ u^i(s_\tau)]$, the following:

$$C^i_t(j,k) := \frac{1}{t}\left[\sum_{\tau \le t: s^i_\tau = k} \frac{p^i_\tau(j)}{p^i_\tau(k)}\, u^i(s_\tau) - \sum_{\tau \le t: s^i_\tau = j} u^i(s_\tau)\right]. \qquad (4.1a)$$

The probabilities p^i_τ are defined in (4.2) below. We note that they are determined by the information available to player i at time τ, i.e., by $u^i(s_1)$, ..., $u^i(s_{\tau-1})$, and thus C^i_t depends on $u^i(s_1)$, ..., $u^i(s_t)$ only. We define the "modified regrets" by

$$Q^i_t(j,k) := [C^i_t(j,k)]^+ = \max\ \{C^i_t(j,k),\ 0\}. \qquad (4.1b)$$

That is, the modified regret for not having used k instead of j measures now the difference (specifically, the positive part) of the average payoff over the periods when k was used and the periods when j was used. In addition, the payoffs of these periods are normalized in a manner that, intuitively speaking, makes the length of the respective periods comparable.[20] We emphasize again that, in order to compute the modified regrets Q_t^i, player i needs to know only the actual payoffs $u^i(s_\tau)$ that he received in the past. In contrast, for the regrets R_t^i of Section 2, he needs in addition to know what the other players chose (i.e., s_τ^{-i}), as well as what his payoff would have been had he played differently than he actually did (i.e., $u^i(k,s_\tau^{-i})$) — which requires knowing the payoff matrix u^i.

The second kind of modification concerns the definition of the play probabilities. We make two changes. First, we need to guarantee that the sum of the proposed probabilities does not exceed one; multiplying the (modified) regrets by a factor $1/\mu$, which we still do, does not suffice because now regrets are not bounded.[21] Second, we need every strategy to be played with some minimal frequency. Thus we shall require that the transitions "tremble" over every strategy.

Let $\mu > 0$ be a sufficiently large number; as in Section 2, it suffices to take μ so that $\mu > 2\, M^i\, (m^i-1)$ for all i, where $m^i = |S^i|$ is the

---

[20] For instance, if k is chosen with a probability that is always twice the probability that j is chosen, then, in the long run, there will be roughly twice as many periods when k was played than when j was played. So, in order to compare the sum of the payoffs over the periods when j was played with the parallel sum when k was played, one needs to multiply the k-sum by 1/2. When the probabilities change over time, the formula in (4.1a) makes the correct normalization (see Steps B10(i) and B16 in the Proof of Theorem B in Appendix 2).

[21] Notice that here there is no a priori lower bound on the probabilities appearing in the denominators.

cardinality of the set S^i, and M^i is an upper bound on $|u^i(s)|$ for all $s \in S$. Let $0 < \delta < 1$, and define the play probabilities of player i at time t+1 by

$$\begin{cases} p^i_{t+1}(k) := (1-\delta) \min\{\frac{1}{\mu} Q^i_t(j,k), \frac{1}{m^i-1}\} + \delta \frac{1}{m^i}, & \text{for } k \neq j; \\ p^i_{t+1}(j) := 1 - \sum_{k \in S^i : k \neq j} p^i_{t+1}(k), \end{cases} \tag{4.2}$$

where $j := s^i_t$ is the choice of i in (the previous) period t. Note that $p^i_{t+1}(j)$ may be rewritten as $(1-\delta)\,[1 - \Sigma_{k \neq j} \min\{Q^i_t(j,k)/\mu,\ 1/(m^i-1)\}] + \delta/m^i$. As for the first period, the play probabilities at $t = 1$ are given by an arbitrary strictly positive vector $p^i_1 \in \Delta(S^i)$; for simplicity, assume that $p^i_1(j) \geq \delta/m^i$ for all j in S^i.

Formula (4.2) says that p^i_{t+1} is a weighted average of two probability vectors. The first, with weight $1-\delta$, is given by the modified regrets in a manner similar to (2.2) (the minimum with $1/(m^i-1)$ guarantees that the resulting sum does not exceed 1). The second term, with weight δ, is just the uniform distribution over S^i: Each $k \in S^i$ has equal probability $1/m^i$. This "uniform tremble" guarantees that all probabilities at time t+1 are at least $\delta/m^i > 0$.

The empirical distributions of play up to t are denoted, as in Section 2, by $z_t \in \Delta(S)$ (see (2.3)). Our result is

Theorem B: *For every* $\varepsilon > 0$ *there is a* $\delta_0 \equiv \delta_0(\varepsilon) > 0$ *such that for all* $\delta < \delta_0$, *if every player* i *plays according to the adaptive procedure* (4.2), *then the empirical distributions of play* z_t *converge in probability to the set of correlated* ε*-equilibria of the game* Γ.

Thus the probability that z_t is a correlated 2ε-equilibrium converges to 1 as t goes to infinity. As in the Main Theorem, both μ and δ may be taken to differ for different players.

To obtain in the limit correlated equilibria — rather than correlated *approximate* equilibria — we need δ to decrease to 0 as t increases. A simple way to do this is to replace δ in (4.2) by $\delta\, t^{-\gamma}$, where γ is a number strictly between 0 and 1/3. Thus, we now define

$$\begin{cases} p^i_{t+1}(k) := (1 - \delta\, t^{-\gamma}) \min\{\frac{1}{\mu} Q^i_t(j,k), \frac{1}{m^i-1}\} + \delta\, t^{-\gamma} \frac{1}{m^i}, \quad \text{for } k \neq j; \\ p^i_{t+1}(j) := 1 - \sum_{k\in S^i : k\neq j} p^i_{t+1}(k), \end{cases} \tag{4.3}$$

where, again, $j := s^i_t$ is the choice of i in period t. At t = 1, let p^i_1 be an arbitrary vector in $\Delta(S^i)$ with $p^i_1(j) \geq \delta/m^i$ for all j in S^i. We have

Theorem B': *If every player* i *plays according to the adaptive procedure* (4.3), *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.

We have chosen the particular type of sequence $\delta_t := \delta\, t^{-\gamma}$ for simplicity and convenience only. What matters is, on the one hand, that δ_t converge to 0. On the other hand, it should do so sufficiently slowly (otherwise the modified regrets C^i_t may become too large; recall that the probabilities p^i_t, which are bounded from below by δ_{t-1}/m^i, appear in the denominators). This explains the need for an upper bound on γ (it turns out that $\gamma < 1/3$ suffices;

the reader should consult the proof in Appendix 2). Moreover, we note that one may well take different sequences δ_t^i for different players i (see the Remark at the end of Appendix 2).

The results of this section can be summarized as follows: Consider games where each player knows only his set of strategies (but knows nothing about other players nor about his own payoff function). Then there exist simple adaptive procedures, based for each player solely on his own past realized payoffs, that lead to correlated equilibria.

5. Discussion

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

(a) The seminal paper in this field of research is Foster and Vohra [1996]. The motivation and the formulation of their work is in terms of best-replying to forecasting rules enjoying good properties — namely, "calibration" — and is quite different from ours. Nonetheless, their results are close to ours (specifically, to our Theorem A), since their calibrated forecasts are also based on regret measures.[22]

(b) The next important paper is Fudenberg and Levine [1996] (see also their forthcoming book [1997]). 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

---

[22] 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 matrix is of fixed size, $m^i \times m^i$.

convergence to the set of correlated ε-equilibria (however, 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 conditional consistency."

(c) 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

$$\limsup_{t \to \infty} \left[\max_{k \in S^i} \frac{1}{t} \sum_{\tau=1}^{t} u^i(k, s_\tau^{-i}) - \frac{1}{t} \sum_{\tau=1}^{t} u^i(s_\tau) \right] \leq 0 \quad \text{a.s.} \tag{5.1}$$

In other words, i's average payoff is no worse than if he were to play any *constant* strategy $k \in S^i$ for all $\tau \leq t$. 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, (5.1) 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

$$D^i_t(k) := \frac{1}{t} \sum_{\tau=1}^{t} \left[u^i(k, s^{-i}_\tau) - u^i(s_\tau) \right], \tag{5.2a}$$

$$p^i_{t+1}(k) := \frac{[D^i_t(k)]^+}{\sum_{k' \in S^i} [D^i_t(k')]^+}, \tag{5.2b}$$

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

Theorem C. *The adaptive procedure* (5.2) *is universally consistent for player* i.

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

Fudenberg and Levine [1995] propose a class of procedures which turn out to be universally ε-consistent:[23] "smooth fictitious play." Player i follows a smooth fictitious play behavior rule if at time t he plays a mixed strategy

---

[23] That is, the right-hand side of (5.1) is ε instead of 0.

$\sigma^i \in \Delta(S^i)$ that maximizes the sum of his expected payoff (with the actions of the remaining 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 large probability and the remaining strategies with small but positive probability). The procedure is, therefore, clearly distinct from (5.2): In (5.2) all the better, even if not best, replies are played with significant probability; also, in (5.2) 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 (5.2) has no trembles.

(d) The idea of "conditional universal consistency," also introduced by Fudenberg and Levine [1996], 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 which are conditional on some relevant characteristic of the history up to t. To be specific, for the case that interests us, the regret is conditioned on the strategy played by i himself. Recall the Proposition of Section 3: If such conditionally universally ε-consistent strategies are played by all players, then all regrets become in the limit $\leq \varepsilon$, and thus the convergence to correlated ε-equilibria is guaranteed.

The procedure of Theorem A is conditionally universally consistent. So is (up to ε) the "calibrated smooth fictitious play" of Fudenberg and Levine [1995]. The two procedures stand to each other as, in the unconditional version, Theorem C stands to "smooth fictitious play."

The procedure of our Main Theorem is not conditionally universally consistent. In particular, if player i follows the procedure we cannot conclude that all his regrets go to zero; sophisticated adversaries, knowing the procedure used by player i, could keep his regrets positive.[24] Such highly complex strategies of the other players are however outside the framework of our study — which deals with simple adaptive behavior. Furthermore, it turns out that the procedure of our Main Theorem is guaranteed to be conditionally consistent not just against opponents using the same procedure, but also against any behavior that is not too sophisticated.[25]

(e) 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) The procedure of our Main Theorem differs from all the other procedures (including that of our Theorem A) in this important aspect: It

---

[24] 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 [1997, Section 8.10].

[25] An examination of Steps M4 - M6 of the Proof of the Main Theorem in Appendix 1 shows that what is needed is that the other players do not change their behavior too quickly; namely, that their transition probabilities at times t and t+w are close enough.

does not require the player to compute, at every step, an invariant (eigen-) vector for an appropriate positive matrix. We view the simplicity of (2.2) as an essential property when discussing non-sophisticated behavior[26] — this is the reason we have sought this result as our Main Theorem.

(g) A specific and most distinctive feature by which the procedure in our Main Theorem differs from those in Theorem A and in 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).[27] Further, our procedure exhibits "friction": there is always a positive probability of staying with the period t strategy.[28]

(h) To understand the role played by friction,[29] 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

---

[26] 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.
[27] 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).
[28] See Sanchirico [1996] and Section 4.6 in Fudenberg and Levine [1997] for a related point in a best-reply context.
[29] See Step M7 in the Proof of the Main Theorem in Appendix 1.

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 will alternate 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 described above, is obtained not by ergodic arguments but by the probability of staying put. What matters is that the diagonal of 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 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 non-empty. Hart and Schmeidler [1989] (see also Nau and McCardle [1990]) provide an elementary (non-fixed point) proof of the non-emptiness 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 of 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).[30] 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.[31]

(j) Theorems B and B' are versions of the Main Theorem when the game is unknown. As pointed out by Fudenberg and Levine [1997, Section 4.8.4], results of this type can be interpreted in terms of "stimulus-response" decision behavior models.[32] In contrast, models where the players know the game — and make use of this knowledge — as in our Main Theorem (or in Theorem A), are models where the players do not simply respond to stimuli (i.e., the obtained payoffs) but undertake some thinking on the game (and its strategic aspects). Although parallel to the Main Theorem, and with a parallel proof, we point out that the result of Theorem B is not simply a transposition of the Main Theorem. Note, for example, that it requires

---

[30] 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 $\Delta(S^i)$).

[31] For another interesting use of the auxiliary two-person zero-sum game, see Myerson [1995].

[32] For related discussions see Roth and Erev [1995], Erev and Roth [1997], Camerer and Ho [1996], and Marimon [1996, Section 3.4].

stronger hypotheses than the latter. In particular, it is now essential that there be some tremble of actions (since there has to be some variation for the sampling to be effective).

To conclude, let us return to the "universal consistency" setting of (c) above. With precedents in Baños [1968] and Megiddo [1980], Auer *et al.* [1995] have recently proved that a result of the type (5.1) can also be guaranteed in the unknown game setup.[33] It is worth pointing out that our universally consistent adaptive procedure (5.2) can be modified to fit the unknown game setup, in the same way that (4.1)–(4.2) were obtained from (2.1)–(2.2). Fix $0 < \delta < 1$ and $0 < \gamma < 1/2$. For each $t = 1, 2, \ldots$, let $\delta_t := \delta\, t^{-\gamma}$, and define for each $k \in S^i$

$$C^i_t(k) := \frac{1}{t}\left[\sum_{\tau \le t:\, s^i_\tau = k} \frac{1}{p^i_\tau(k)}\, u^i(s_\tau) - \sum_{\tau=1}^{t} u^i(s_\tau)\right], \tag{5.3a}$$

$$q^i_{t+1}(k) := \frac{[C^i_t(k)]^+}{\sum_{k' \in S^i} [C^i_t(k')]^+}, \tag{5.3b}$$

$$p^i_{t+1}(k) := (1-\delta_t)\, q^i_{t+1}(k) + \delta_t\, \frac{1}{m^i}. \tag{5.3c}$$

Again, if the denominator of $q^i_{t+1}(k)$ vanishes, take $q^i_{t+1} \in \Delta(S^i)$ to be arbitrary. The strategy of player i is to choose k in S^i at time t+1 with probability $p^i_{t+1}(k)$; at time 1, $p^i_1 \in \Delta(S^i)$ is arbitrary with $p^i_1(k) \ge \delta/m^i$ for all k. We then have:

---

[33] Specifically: Baños and Megiddo establish the slightly weaker property that the minimax payoff is guaranteed. As for the procedure of Auer *et al.*, it depends on the length of the game t, and it yields (5.1) in probability only; compare with (5.3) and Theorem C* below.

Theorem C*. *The adaptive procedure* (5.3) *is universally consistent for player* i.

The proof of Theorem C* is similar to (and simpler than) the proof of Theorem B and is omitted. We conclude by noting that in the same way that Theorem A is the conditional parallel to Theorem C, we could, rather straightforwardly, formulate an unknown game conditional parallel to Theorem C*, a Theorem A*. Note, however, that this Theorem A* would be a result different from Theorem B (the unknown game version of the Main Theorem); in particular, it will use invariant (eigen-) vectors.

Appendix 1. 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 rather complex[34] 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.

Fix player i in N. To simplify notation we drop reference to the index i whenever this cannot cause confusion (thus we write D_t and R_t instead of D^i_t and R^i_t, and so on). Let $m := |S^i|$ be the number of strategies of player i, and let M be an upper bound on i-th possible payoffs: $M \geq |u^i(s)|$ for all s in S. Denote $L := \{(j,k) \in S^i \times S^i : j \neq k\}$, then $\mathbb{R}^L$ is the

[34] We only said that the adaptive *procedure* is simple.

m(m–1) - dimensional Euclidean space with coordinates indexed by L. For each $t = 1, 2, \ldots$ and each (j,k) in L, put[35]

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

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

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

We will write A_t for the vector $(A_t(j,k))_{j\neq k} \in \mathbb{R}^L$; the same goes for D_t, D_t^+, R_t, and so on. Let $\Pi_t(\cdot,\cdot)$ denote the transition probabilities from t to t+1 (these are computed after period t, based on h_t):

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

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

$$\rho_t := [\mathrm{dist}(D_t, \mathbb{R}_-^L)]^2$$

be the squared distance (in $\mathbb{R}^L$) of the vector D_t to the non-positive orthant $\mathbb{R}_-^L$. Since the closest point to D_t in $\mathbb{R}_-^L$ is[36] D_t^-, we have $\rho_t = \|D_t - D_t^-\|^2 = \|D_t^+\|^2 = \Sigma_{j\neq k} [D_t^+(j,k)]^2$.

---

[35] We write $\mathbf{1}_G$ for the indicator of the event G.

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

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 $|f(x)| \leq K\, g(x)$ for all x in[37] 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. Step M11, together with the Proposition of Section 3, completes the proof. The reader will find a short "intuitive guide" to the proof following the statements of the steps.

- *Step M1:* $\mathbf{E}[(t+v)^2\, \rho_{t+v} \,|\, h_t] \leq t^2\, \rho_t + 2t \sum_{w=1}^{v} R_t \bullet \mathbf{E}[A_{t+w} \,|\, h_t] + O(v^2)$.

Define

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

- *Step M2:* $R_t \bullet \mathbf{E}[A_{t+w} \,|\, h_t] = \mu \sum_{s^{-i} \in S^{-i}} \sum_{j \in S^i} \alpha_{t,w}(j,s^{-i})\, u^i(j,s^{-i})$.

- *Step M3:* $R_{t+v}(j,k) - R_t(j,k) = O(\frac{v}{t})$.

---

[37] 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 $|f(t,v)| \leq K\, v$ for all v and all t. The constants K will always depend only on the game (through N, m, M, and so on) and on the parameter μ.

For each $t > 0$ and each history h_t, define an auxiliary stochastic process $(\hat{s}_{t+w})_{w=0,1,2,...}$ with values in S as follows: The initial value is $\hat{s}_t = s_t$, and the transition probabilities are[38]

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

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 M4:* $\mathbf{P}[s_{t+w} = s \,|\, h_t] - \mathbf{P}[\hat{s}_{t+w} = s \,|\, h_t] = O(\frac{w^2}{t})$.

Define

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

- *Step M5:* $\alpha_{t,w}(j,s^{-i}) - \hat{\alpha}_{t,w}(j,s^{-i}) = O(\frac{w^2}{t})$.

- *Step M6:* $\hat{\alpha}_{t,w}(j,s^{-i}) = \mathbf{P}[\hat{s}_{t+w}^{-i} = s^{-i} \,|\, h_t] \ [\Pi_t^{w+1} - \Pi_t^{w}](s_t^i, j)$,

where $\Pi_t^w \equiv (\Pi_t)^w$ is the w-th 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 M7:* $\hat{\alpha}_{t,w}(j,s^{-i}) = O(w^{-1/2})$.

- *Step M8:* $\mathbf{E}[(t+v)^2 \ \rho_{t+v} \,|\, h_t] \leq t^2 \ \rho_t + O(v^3 + t \ v^{1/2})$.

---

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

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

- *Step M9:* $\mathbf{E}[t_{n+1}^2\, \rho_{t_{n+1}} | h_{t_n}] \le t_n^2\, \rho_{t_n} + O(n^2)$.

- *Step M10:* $\mathbf{E}[\rho_{t_n}] = O(n^{-1/3})$ and $\lim_{n \to \infty} \rho_{t_n} = 0$ a.s.

- *Step M11:* $\lim_{t \to \infty} R_t(j,k) = 0$ a.s.

We now provide a short intuitive guide to the proof. Step M1 is the basic recursion equation, similar to that used for the standard proof of Blackwell's Approachability Theorem. However, we need to consider v periods together, rather than just one. The reader may want to keep in mind that both t and v are very large, and that v is relatively small compared to t. This implies that, in the right-hand expression of M1, only the middle term has to be bounded (from above). Blackwell's condition (3.2) precisely says that the middle term in the corresponding formula is ≤ 0. Here this is not so (we are *not* in the stylized version of Section 3), so we need to estimate this term. Step M2 shows that we get an expression similar to (3.5), but the coefficients α depend also on the moves of the other players. Given h_t, the choices s^i_{t+w} and s^{-i}_{t+w} are *not* independent when $w > 1$ (since the transition probabilities change with time). Therefore we replace the process $(s_{t+w})_{0 \le w \le v}$ by another process $(\hat{s}_{t+w})_{0 \le w \le v}$, 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), yielding the approximation of Step M5. In 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 $\hat{\alpha}$ and obtain a formula comparable to (3.5). However, instead of the invariant distribution, we have the transition probabilities after w periods and after w+1 periods. The difference between them is shown in Step M7 to be small, since w is large and the transition matrix has all its diagonal elements strictly positive.[39] This yields the final recursive formula of Step M8. The proof is now completed in Steps M9 - M11, by considering a carefully chosen subsequence of periods $(t_n)_{n=1,2,...}$.

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

• *Proof of Step M1:* Because $D_t^- \in \mathbb{R}_-^L$ we have:

$$\rho_{t+v} \leq \|D_{t+v} - D_t^-\|^2 = \left\|\frac{t}{t+v} D_t + \frac{1}{t+v} \sum_{w=1}^{v} A_{t+w} - D_t^-\right\|^2$$

$$= \frac{t^2}{(t+v)^2} \|D_t - D_t^-\|^2 + \frac{2t}{(t+v)^2} \sum_{w=1}^{v} (A_{t+w} - D_t^-) \bullet (D_t - D_t^-)$$

$$+ \frac{v^2}{(t+v)^2} \left\|\frac{1}{v} \sum_{w=1}^{v} A_{t+w} - D_t^-\right\|^2$$

$$\leq \frac{t^2}{(t+v)^2} \rho_t + \frac{2t}{(t+v)^2} \sum_{w=1}^{v} A_{t+w} \bullet R_t + \frac{v^2}{(t+v)^2} m\,(m-1)\, 4\, M^2.$$

Indeed: $|u^i(s)| \leq M$, so $|A_{t+w}(j,k)| \leq 2M$ and $|D_t(j,k)| \leq 2M$, yielding the upper bound on the third term. As for the second term, note that $R_t = D_t^+ = D_t - D_t^-$ and $D_t^- \bullet D_t^+ = 0$. Taking conditional expectation given the history h_t (so ρ_t and R_t are known) completes the proof. ▮

• *Proof of Step M2:* We have

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

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

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

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

$$= \sum_{s^{-i}} \sum_j u^i(j,s^{-i}) \Big[\sum_{k \neq j} R_t(k,j) \; \phi(k,s^{-i}) - \sum_{k \neq j} R_t(j,k) \; \phi(j,s^{-i}) \Big]$$

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

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

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

• *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 $|A_{t+w}(j,k)| \leq 2M$ and $|D_t(j,k)| \leq 2M$. ▮

• *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

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

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

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

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

Proof of Lemma: We write $\mathbf{P}_X$ and $\mathbf{P}_Y$ for the probabilities of the two processes $(X_n)_n$ and $(Y_n)_n$, respectively (thus $\mathbf{P}_X[b_{n+w} | b_0, \ldots, b_{n-1}]$ stands for $\mathbf{P}[X_{n+w} = b_{n+w} | X_0 = b_0, \ldots, X_{n-1} = b_{n-1}]$, and so on). We use induction on w.

$$\begin{aligned}
\mathbf{P}_X[b_{n+w} | b_0, \ldots, b_{n-1}] &= \sum_{b_n} \mathbf{P}_X[b_{n+w} | b_0, \ldots, b_n] \; \mathbf{P}_X[b_n | b_0, \ldots, b_{n-1}] \\
&\leq \sum_{b_n} \mathbf{P}_Y[b_{n+w} | b_0, \ldots, b_n] \; \mathbf{P}_X[b_n | b_0, \ldots, b_{n-1}] + |B| \sum_{r=1}^{w} \beta_{n+r} \\
&\leq \sum_{b_n} \mathbf{P}_Y[b_{n+w} | b_0, \ldots, b_n] \; (\mathbf{P}_Y[b_n | b_0, \ldots, b_{n-1}] + \beta_n) + |B| \sum_{r=1}^{w} \beta_{n+r} \\
&\leq \mathbf{P}_Y[b_{n+w} | b_0, \ldots, b_{n-1}] + |B| \; \beta_n + |B| \sum_{r=1}^{m} \beta_{n+r}
\end{aligned}$$

(the first inequality is by the induction hypothesis). Exchanging the roles of X and Y completes the proof. ▮

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 of the s-process by at most O(w/t), which, when composed by the Lemma, yields $|N|\ w\ |S|\ O(w/t) = O(w^2/t)$. ▮

• *Proof of Step M5:* Immediate by Step M4. ▮

• *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:

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

implying that

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

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

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

completing the proof. ▮

• *Proof of Step M7:* It follows from M6 using the following Lemma[40] (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 $j, k = 1, ..., m$.

Proof of the Lemma: 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 $\gamma_r \leq 0$ if $r \leq q$; together with $0 \leq \Lambda^r(j,k) \leq 1$, we get

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

---

[40] If Π were a strictly positive matrix, this would be a standard result, because then Π^w would converge to the invariant matrix. However we only know 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).

Consider the left-most sum. It equals

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

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):

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

where

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

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. ∎

• *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

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

Adding over w = 1, 2, ..., v, and substituting into Step M1 gives the result (recall that $\sum_{w=1}^{v} w^{\gamma} = O(v^{\gamma+1})$ for $\gamma \neq -1$). ∎

• *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 $t\, v^{1/2} = O(n^{5/3+1/3}) = O(n^2)$, and the result follows. ▮

• *Proof of Step M10:* Taking expectation in the expression of Step M9 yields

$$\mathbf{E}[t_{n+1}^2\, \rho_{t_{n+1}}] \le \mathbf{E}[t_n^2\, \rho_{t_n}] + O(n^2).$$

So, by iterating, $\mathbf{E}[t_n^2\, \rho_{t_n}] = O(n^3)$, or $\mathbf{E}[\rho_{t_n}] = O(n^3 / n^{10/3}) = O(n^{-1/3})$.

To obtain the almost sure convergence, divide the formula of Step M9 by t_{n+1}^2, to get (note that $t_n^2 / t_{n+1}^2 \le 1$)

$$\mathbf{E}[\rho_{t_{n+1}} | h_{t_n}] \le \rho_{t_n} + K\, n^{-4/3}$$

for an appropriate constant $K < \infty$. Define

$$X_n := \rho_{t_n} + K \sum_{\nu=n}^{\infty} \nu^{-4/3}.$$

Then $(X_n)_{n=1,2,...}$ is a non-negative super-martingale, adapted to the sequence of fields generated by the histories h_{t_n}, i.e., $\mathbf{E}[X_{n+1} | h_{t_n}] \le X_n$. It therefore converges a.s. as $n \to \infty$ (see Loève [1978, 32.3.A(i)]). The limit X_∞ satisfies by Fatou's lemma $\mathbf{E}[X_\infty] \le \lim_n \mathbf{E}[X_n] = \lim_n \mathbf{E}[\rho_{t_n}] = 0$ (the series $\sum_\nu \nu^{-4/3}$ converges, so its tail goes to 0). Thus $X_\infty = 0$ a.s., or $X_n \to 0$, so $\rho_{t_n} \to 0$. ▮

• *Proof of Step M11:* Since $\rho_{t_n} = \sum_{j \ne k} [R_{t_n}(j,k)]^2$, the previous Step M10 implies that $R_{t_n}(j,k) \to 0$ a.s. as $n \to \infty$, for all $j \ne k$. 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. as $t \to \infty$. ▮

Appendix 2. Proof of the Results for the Unknown Game Case

In this Appendix we will prove the two Theorems B and B' of Section 4 together. Let

$$\delta_t := \frac{\delta}{t^\gamma},$$

where $\delta > 0$ and $0 \le \gamma < 1/3$. For Theorem B take $\gamma = 0$, and for Theorem B', $\gamma > 0$. As in the proof of the Main Theorem in Appendix 1, we fix player i in N, and drop reference to the index i whenever this cannot cause confusion. In addition to the notation introduced there, we define for each t = 1, 2, ... and each (j,k) in L

$$Z_t(j,k) := \frac{p_t^i(j)}{p_t^i(k)} \mathbf{1}_{\{s_t^i=k\}} - \mathbf{1}_{\{s_t^i=j\}} ;$$

$$B_t(j,k) := Z_t(j,k)\, u^i(s_t)$$

(thus $C_t(j,k) = (1/t) \sum_{\tau \le t} B_\tau(j,k)$); and

$$\Pi_t(j,k) := \begin{cases} (1-\delta_t) \min\{\frac{1}{\mu} C_t^+(j,k), \frac{1}{m-1}\} + \delta_t \frac{1}{m}, & \text{if } k \neq j, \\ (1-\delta_t)\left[1 - \sum_{k' \neq j} \min\{\frac{1}{\mu} C_t^+(j,k'), \frac{1}{m-1}\}\right] + \delta_t \frac{1}{m}, & \text{if } k = j. \end{cases}$$

Note that $\Pi_t(j,\cdot) \in \Delta(S^i)$ for all $j \in S^i$, thus Π_t is a probability transition matrix on S^i. Both procedures (4.2) and (4.3) satisfy $p_{t+1}^i(k) = \Pi_t(s_t^i,k)$ for all k (where, again, $\gamma = 0$ corresponds to (4.2) and $\gamma > 0$ to (4.3)). Let

$$\rho_t := [\mathrm{dist}(C_t, \mathbb{R}_-^L)]^2$$

be the squared distance (in $\mathbb{R}^L$) of the vector C_t from the non-positive orthant $\mathbb{R}^L_-$. Thus $\rho_t = \Sigma_{j\neq k} [C_t^+(j,k)]^2$.

The proof of Theorems B and B' will be divided into 16 steps, B1 - B16. Step B16, together with the Proposition of Section 3, completes the proof. We note that Steps B1 - B10 hold for *any* non-increasing sequence $\delta_t > 0$, whereas Steps B11 - B16 make use of the rate of the special form $\delta_t = \delta\, t^{-\gamma}$. A guide to the proof follows the statement of the steps.

- *Step B1:* $\mathbf{E}[(t+v)^2 \rho_{t+v} | h_t] \leq t^2 \rho_t + 2t \sum_{w=1}^{v} C_t^+ \bullet \mathbf{E}[B_{t+w} | h_t] + O(\frac{v^2}{\delta^2_{t+v}})$.

Define[41]

$$\beta_{t+w}(j) := \sum_{k\neq j} \frac{1}{\mu} C_t^+(k,j)\, p^i_{t+w}(k) - \sum_{k\neq j} \frac{1}{\mu} C_t^+(j,k)\, p^i_{t+w}(j).$$

- *Step B2:* $C_t^+ \bullet \mathbf{E}[B_{t+w} | h_t] = \mu\, \mathbf{E}\Big[\sum_{j\in S^i} u^i(j,s^{-i}_{t+w})\, \beta_{t+w}(j) \Big| h_t\Big]$,

- *Step B3:* $\Pi_{t+v}(j,k) - \Pi_t(j,k) = O(\frac{v}{t\,\delta_{t+v}} + [\delta_t - \delta_{t+v}])$.

Define

$$\vartheta := \frac{1}{m-1} - \frac{2M}{\mu} > 0;\quad \bar{Z}_t(j,k) := \frac{1}{t} \sum_{\tau=1}^{t} Z_t(j,k);\quad \text{and}$$

$$Y_t := \sum_{(j,k)\in L} [\frac{M}{\mu} \bar{Z}_t(j,k) - \vartheta]^+.$$

---

[41] Note that $\beta_{t+w}(j)$ is measurable with respect to h_{t+w} (actually, it depends only on h_{t+w-1} and s^{-i}_{t+w}, but not on s^i_{t+w}).

• *Step B4:* $\Pi_t(j,k) - \frac{1}{\mu} C_t^{+}(j,k) = O(Y_t + \delta_t)$ for all $j \neq k$.

• *Step B5:* $$C_t^{+} \bullet \mathbf{E}[B_{t+w} | h_t] = \mu \, \mathbf{E}\Big[\sum_{j \in S^i} u^i(j, s_{t+w}^{-i}) \, [\Pi_t^{\,2} - \Pi_t](s_{t+w-1}^{i}, j) \Big| h_t \Big] + O(Y_t + \delta_t + \frac{w}{t\,\delta_{t+w}}).$$

Recall the auxiliary process $(\hat{s}_{t+w})_{w=0,1,2,...}$ as defined in Step M4 in Appendix 1.

• *Step B6:* $\mathbf{P}[s_{t+w} = s | h_t] - \mathbf{P}[\hat{s}_{t+w} = s | h_t] = O(\frac{w^2}{t\,\delta_{t+w}} + w\,[\delta_t - \delta_{t+w}])$.

• *Step B7:* $$C_t^{+} \bullet \mathbf{E}[B_{t+w} | h_t] = \mu \, \mathbf{E}\Big[\sum_{j \in S^i} u^i(j, \hat{s}_{t+w}^{-i}) \, [\Pi_t^{\,2} - \Pi_t](\hat{s}_{t+w-1}^{i}, j) \Big| h_t \Big] + O(Y_t + \delta_t + \frac{w^2}{t\,\delta_{t+w}} + w\,[\delta_t - \delta_{t+w}]).$$

• *Step B8:*[42] $$\mathbf{E}\Big[\sum_{j \in S^i} u^i(j, \hat{s}_{t+w}^{-i}) \, [\Pi_t^{\,2} - \Pi_t](\hat{s}_{t+w-1}^{i}, j) \Big| h_t \Big] = \sum_{s^{-i}} \mathbf{P}[\hat{s}_{t+w}^{-i} = s^{-i} | h_t] \, [\Pi_t^{\,w+1} - \Pi_t^{\,w}](s_t^i, j) = O((1-\delta_t)^w).$$

• *Step B9:* $$\mathbf{E}[(t+v)^2 \rho_{t+v} | h_t] \le t^2 \rho_t + O(t\,v\,Y_t + t\,v\,\delta_t + \frac{v^3}{\delta_{t+v}} + t\,v^2\,[\delta_t - \delta_{t+v}] + \frac{t}{\delta_t} + \frac{v^2}{\delta_{t+v}^2}).$$

• *Step B10:* (i) $\mathbf{E}[Z_t(j,k) | h_{t-1}] = 0$;

(ii) $\mathrm{Var}[Z_t(j,k)] = O(\delta_t^{-1})$; and

---

42 One could use here another estimate, $O(w^{-1/2})$, as in Steps M6 and M7 in the proof of the Main Theorem; it would have made no difference.

(iii) $\mathbf{E}[Y_t] = O(t^{-1/2}\ \delta_t^{-1/2})$.

- *Step B11:* $\bar{Z}_t(j,k) \to 0$ a.s. as $t \to \infty$.

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

- *Step B12:* $\mathbf{E}[t_{n+1}^2\ \rho_{t_{n+1}} | h_{t_n}] \le t_n^2\ \rho_{t_n} + O(\delta\ n^{2-3\gamma/2} + Y_{t_n}\ n^2 + n^{3/2+3\gamma/2})$.

- *Step B13:* $\mathbf{E}[\rho_{t_n}] = O(\delta\ n^{-3\gamma/2} + n^{3\gamma/2-1/2})$.

- *Step B14:* If $\delta_t = \delta$ for all t, then $\mathbf{E}[\rho_t] = O(\delta)$.

- *Step B15:* If $\delta_t \to 0$ (i.e., $\gamma > 0$), then $\mathbf{E}[\rho_{t_n}] = O(n^{-\lambda})$, where $\lambda := \min\ \{3\gamma/2,\ (1-3\gamma)/2\} > 0$, and $\rho_{t_n} \to 0$ a.s. as $n \to \infty$.

- *Step B16:* (i) If $\delta_t = \delta$ for all t, then $\limsup_{t\to\infty} \mathbf{E}[D_t^+(j,k)] = O(\sqrt{\delta})$.

  (ii) If $\delta_t \to 0$ (i.e., if $\gamma > 0$), then $\lim_{t\to\infty} D_t^+(j,k) = 0$.

We now provide a short intuitive overview of the steps of the proof. The proof is based on the proof of the Main Theorem in Appendix 1, with a number of additional steps needed to take care of the modifications. Most of the proof is devoted to showing that the modified regrets $Q_t^i \equiv [C_t^i]^+$ are small (see Steps B14 and B15). From this one readily gets in Step B16 that the actual regrets $R_t^i \equiv [D_t^i]^+$ are also small, since the difference $C_t^i - D_t^i$ is a martingale converging almost surely to 0 (see Steps B10 and B11). The main steps in the proof are as follows: We start with the basic recursion equation in Step B1 (similar to M1). Next, we estimate the "middle term" on the right-hand of B1 by approximating the s-process with the $\hat{s}$-process, which is independent across players (Steps B2 - B7; parallel to M2 - M6). This leads to a formula comparable to (3.5), except that the invariant distribution in (3.5b) is replaced here by the transitions after w and w+1 periods, which are close to one another (Step B8; compare with M7). Finally, we obtain the recursion formula in Step B9. When comparing this formula with the parallel one in M8, we see that in B9 there are additional terms; one of them (that comes about because the modified regrets are not bounded) contains a random variable Y_t. Step B10(iii), Step B11 and the proof of Step B15 show that this term also goes to zero. The next steps, B12 - B15, complete the proof for the modified regrets, in a manner that is similar (though more complicated, partly because of the Y_t terms) to M9 - M11. As we have indicated above, Step B16 yields the result for the actual regrets.

We now proceed to the proofs of B1 - B16.

• *Proof of Step B1:* Similar to that of Step M1, with B and C, respectively, replacing A and D there. Note that $Z_{t+w}(j,k)$ is bounded by m/δ_{t+w} which, for $w \le v$, is at most m/δ_{t+v} (the sequence δ_t is non-increasing), therefore $\|B_{t+w}\|$ and $\|C_t\|$ are $O(\delta_{t+v}^{-2})$ — which yields the bound on the third term. ▮

• *Proof of Step B2:* Note that $B_{t+w}(j,k)$ vanishes except when $s^i_{t+w} = j$ or k. We condition on h_{t+w-1} and s^{-i}_{t+w} (i.e., on the whole history h_{t+w} *except* player i's choice at time t+w):

$$\mathbf{E}[B_{t+w}(j,k) \,|\, h_{t+w-1},\, s^{-i}_{t+w} = s^{-i}]$$

$$= p^i_{t+w}(k)\, \frac{p^i_{t+w}(j)}{p^i_{t+w}(k)}\, u^i(k,s^{-i}) - p^i_{t+w}(j)\, u^i(j,s^{-i})$$

$$= p^i_{t+w}(j)\, [u^i(k,s^{-i}) - u^i(j,s^{-i})].$$

Hence

$$C_t^+ \bullet \mathbf{E}[B_{t+w} | h_{t+w-1},\ s_{t+w}^{-i} = s^{-i}] = \sum_j \sum_{k\neq j} C_t^+(j,k)\ p_{t+w}^i(j)\ [u^i(k,s^{-i}) - u^i(j,s^{-i})]$$

$$= \sum_j u^i(j,s^{-i}) \left[\sum_{k\neq j} C_t^+(k,j)\ p_{t+w}^i(k) - \sum_{k\neq j} C_t^+(j,k)\ p_{t+w}^i(j)\right]$$

(we have collected together all terms containing $u^i(j,s^{-i})$). Conditioning now on h_t yields the result. ▌

• *Proof of Step B3:* We have

$$(t+v)\ |C_{t+v}(j,k) - C_t(j,k)| \le \sum_{w=1}^{v} |B_{t+w}(j,k) - C_t(j,k)|.$$

Since both $B_{t+w}(j,k)$ and $C_t(j,k)$ are $O(\delta_{t+v}^{-1})$, we get $C_{t+v}(j,k) - C_t(j,k) = O(vt^{-1}\delta_{t+v}^{-1})$. For $j \neq k$, the difference between $\Pi_{t+v}(j,k)$ and $\Pi_t(j,k)$ is therefore at most $(1-\delta_t)\ O(vt^{-1}\delta_{t+v}^{-1}) + (\delta_t - \delta_{t+v})\ (m-1)^{-1}$. For $j = k$, it is at most $m-1$ times this amount. ▌

• *Proof of Step B4:* If $(1/\mu)\ C_t^+(j,k) \le 1/(m-1)$, then

$$\Pi_t(j,k) - \frac{1}{\mu} C_t^+(j,k) = \delta_t\ [\frac{1}{m} - \frac{1}{\mu} C_t^+(j,k)],$$

and this lies between 0 and δ_t/m. If $(1/\mu)\ C_t^+(j,k) \ge 1/(m-1)$, then $\Pi_t(j,k) = (1-\delta_t)/(m-1) + \delta_t/m \le 1/(m-1)$, so

$$\frac{1}{\mu} C_t^+(j,k) \le \frac{1}{\mu}\ |C_t(j,k)| \le \frac{1}{\mu}\frac{1}{t} \sum_{\tau=1}^{t} |Z_\tau(j,k)|\ |u^i(s_\tau)|$$

$$\le \frac{2M}{\mu} + \frac{M}{\mu}\bar{Z}_t(j,k) = \frac{1}{m-1} - \vartheta + \frac{M}{\mu}\bar{Z}_t(j,k)$$

$$\le \Pi_t(j,k) + O(\delta_t) + [\frac{M}{\mu}\bar{Z}_t(j,k) - \vartheta]^+,$$

where we have used the inequality $|Z_\tau(j,k)| \leq 2 + Z_\tau(j,k)$ (since the only possible negative value of $Z_\tau(j,k)$ is -1). Together with

$$\Pi_t(j,k) \leq \frac{1}{m-1} \leq \frac{1}{\mu}\, C_t^+(j,k),$$

the proof is complete. ∎

- *Proof of Step B5:* Denote s^i_{t+w-1} by r, then

$$\beta_{t+w}(j) = \sum_{k\neq j} \frac{1}{\mu}\, C_t^+(k,j)\; \Pi_{t+w-1}(r,k) - \sum_{k\neq j} \frac{1}{\mu}\, C_t^+(j,k)\; \Pi_{t+w-1}(r,j).$$

Also,

$$\begin{aligned}
[\Pi_t^2 - \Pi_t](r,j) &= [\Pi_t^2](r,j) - \Pi_t(r,j) \\
&= \sum_{k\in S^i} \Pi_t(k,j)\; \Pi_t(r,k) - \Pi_t(r,j) \sum_{k\in S^i} \Pi_t(j,k) \\
&= \sum_{k\neq j} \Pi_t(k,j)\; \Pi_t(r,k) - \sum_{k\neq j} \Pi_t(j,k)\; \Pi_t(r,j)
\end{aligned}$$

(we have subtracted the j-th term from both sums). Comparing the last expression with $\beta_{t+w}(j)$, we see that $[\Pi_t^2 - \Pi_t](r,j)$ is obtained by replacing each C_t^+/μ and each Π_{t+w-1} in $\beta_{t+w}(j)$ by Π_t. Thus

$$\begin{aligned}
\beta_{t+w}(j) - [\Pi_t^2 - \Pi_t](r,j) = &\sum_{k\neq j} [\tfrac{1}{\mu}\, C_t^+(k,j) - \Pi_t(k,j)]\; \Pi_{t+w-1}(r,k) \\
&+ \sum_{k\neq j} \Pi_t(k,j)\; [\Pi_{t+w-1}(r,k) - \Pi_t(r,k)] \\
&- \sum_{k\neq j} [\tfrac{1}{\mu}\, C_t^+(j,k) - \Pi_t(j,k)]\; \Pi_{t+w-1}(r,j)
\end{aligned}$$

$$- \sum_{k \neq j} \Pi_t(j,k) \; [\Pi_{t+w-1}(r,j) - \Pi_t(r,j)]$$

$$= O(Y_t + \delta_t + \frac{w}{t\,\delta_{t+w}}),$$

by the estimates of Steps B3 and B4. It only remains to substitute this into the formula of Step B2. ∎

• *Proof of Step B6:* Identical to the proof of Step M4, using the estimate of Step B3 instead of that of Step M3. ∎

• *Proof of Step B7:* Replacing $(s_{t+w})_w$ by $(\hat{s}_{t+w})_w$ in the formula of Step B5 gives an additional error that is estimated in Step B6 (note that the two processes $(s_{t+w})_w$ and $(\hat{s}_{t+w})_w$ start from the same history h_t). ∎

• *Proof of Step B8:* As in Step M6, $\hat{s}^{-i}_{t+w}$ and $\hat{s}^{i}_{t+w-1}$ are independent given h_t. Hence:

$$\mathbf{E}\Big[\sum_{j \in S^i} u^i(j,\hat{s}^{-i}_{t+w}) \; [\Pi_t^{\,2} - \Pi_t](\hat{s}^i_{t+w-1},j) \Big| h_t \Big]$$

$$= \sum_{s^{-i} \in S^{-i}} \mathbf{P}[\hat{s}^{-i}_{t+w} = s^{-i} | h_t] \sum_{r \in S^i} \mathbf{P}[\hat{s}^i_{t+w-1} = r | h_t] \; [\Pi_t^{\,2} - \Pi_t](r,j)$$

$$= \sum_{s^{-i}} \mathbf{P}[\hat{s}^{-i}_{t+w} = s^{-i} | h_t] \sum_{r} \Pi_t^{\,w-1}(s^i_t,r) \; [\Pi_t^{\,2} - \Pi_t](r,j)$$

$$= \sum_{s^{-i}} \mathbf{P}[\hat{s}^{-i}_{t+w} = s^{-i} | h_t] \; [\Pi_t^{\,w+1} - \Pi_t^{\,w}](s^i_t,j).$$

Now Π_t is an $m \times m$ stochastic matrix with all entries at least $\delta_t/m > 0$, implying, as it is well-known, that $|[\Pi_t^{w+1}](k,j) - [\Pi_t^{w}](k,j)| \leq (1 - m\ \delta_t/m)^w$. ▮

• *Proof of Step B9:* Putting together the estimates of Steps B8 and B9 (note that $\sum_{w=1}^{v} (1-\delta_t)^w \leq 1/\delta_t$) yields

$$2t \sum_{w=1}^{v} C_t^+ \bullet \mathbf{E}[B_{t+w}|h_t] = O(t\ v\ Y_t + t\ v\ \delta_t + \frac{t\ v^3}{\delta_{t+v}} + t\ v^2\ [\delta_t - \delta_{t+v}] + \frac{t}{\delta_t}).$$

Recalling the formula of Step B1 completes the proof. ▮

• *Proof of Step B10:* By definition $Z_t(j,k)$,

$$\mathbf{E}[Z_t(j,k)|h_{t-1}] = \frac{p_t^i(j)}{p_t^i(k)}\ p_t^i(k) - p_t^i(j) = 0.$$

Therefore $(Z_t(j,k))_t$ are martingale differences, and $\mathbf{E}[Z_t(j,k)] = 0$. Also

$$\mathrm{Var}[Z_t(j,k)] = \mathbf{E}[Z_t^2(j,k)] = \mathbf{E}\Big[\mathbf{E}[Z_t^2(j,k)|h_{t-1}]\Big]$$

$$= \mathbf{E}\left[\frac{(p_t^i(j))^2}{(p_t^i(k))^2}\ p_t^i(k) + (-1)^2\ p_t^i(j)\right] \leq \mathbf{E}\Big[\frac{1}{p_t^i(k)}\Big] \leq \frac{m}{\delta_t}.$$

From this we get

$$\mathrm{Var}[\bar{Z}_t(j,k)] = \frac{1}{t^2} \sum_{\tau=1}^{t} \mathrm{Var}[Z_\tau(j,k)] \leq \frac{m}{t\ \delta_t}$$

(the variance of the sum of martingale differences is the sum of their variances). Finally, since $\vartheta > 0$ we have $[\zeta - \vartheta]^+ \leq |\zeta|$, so

$$\mathbf{E}[\frac{M}{\mu}\,\bar{Z}_t(j,k) - \vartheta]^+ \le \mathbf{E}|\frac{M}{\mu}\,\bar{Z}_t(j,k)| \le \left[\mathrm{Var}[\frac{M}{\mu}\,\bar{Z}_t(j,k)]\right]^{1/2} \le \left[\frac{m}{t\,\delta_t}\right]^{1/2}.$$

Summing over all $j \ne k$ gives the estimate for $\mathbf{E}[Y_t]$. ∎

• *Proof of Step B11:* By Step B10,

$$\sum_{t=1}^{\infty} \frac{1}{t^2}\,\mathrm{Var}[Z_t(j,k)] \le \sum_{t=1}^{\infty} \frac{m}{t^2\,\delta_t}\,.$$

This series converges, since $\gamma < 1/3$. Therefore $\bar{Z}_t(j,k) \to 0$ a.s. (see Loève [1978, 32.1.E]). ∎

• *Proof of Step B12:* Apply the inequality of Step B10 with $t = t_n = n^{3/2}$ and $v = t_{n+1} - t_n$. Then: $v = O(n^{1/2})$; $\delta_t = \delta\,(n^{3/2})^{-\gamma} = \delta\,n^{-3\gamma/2}$; $\delta_{t+v} = \delta\,(n+1)^{-3\gamma/2} = O(n^{-3\gamma/2})$; $\delta_t - \delta_{t+v} = O(n^{-3\gamma/2-1})$, so

$$\begin{aligned}\mathbf{E}[t_{n+1}^2\,\rho_{t_{n+1}}|h_{t_n}] &\le t_n^2\,\rho_{t_n}\\ &\quad + O(Y_{t_n}\,n^2 + \delta\,n^{2-3\gamma/2} + n^{3/2+3\gamma/2} + n^{3/2-3\gamma/2} + n^{3/2+3\gamma/2} + n^{1+3\gamma}).\end{aligned}$$

Now $3/2 - 3\gamma/2$ and $1 + 3\gamma$ are both $< 3/2 + 3\gamma/2$ (recall that $\gamma < 1/3$). ∎

• *Proof of Step B13:* By Step B10, $\mathbf{E}[Y_{t_n}] = O(t_n^{-1/2}\,\delta_{t_n}^{-1/2}) = O(n^{-3/4+3\gamma/4}) = o(n^{-1/2})$ (since[43] $\gamma < 1/3$); thus $n^2\,\mathbf{E}[Y_{t_n}] = o(n^{3/2})$. Taking expectation in the formula of Step B12, we thus get

$$\mathbf{E}[t_{n+1}^2\,\rho_{t_{n+1}}] \le \mathbf{E}[t_n^2\,\rho_{t_n}] + O(\delta\,n^{2-3\gamma/2} + n^{3/2+3\gamma/2}).$$

So, by iterating, $\mathbf{E}[t_n^2\,\rho_{t_n}] = O(\delta\,n^{3-3\gamma/2} + n^{5/2+3\gamma/2})$, or $\mathbf{E}[\rho_{t_n}] = O(\delta\,n^{-3\gamma/2} + n^{-1/2+3\gamma/2})$. ∎

---

[43] We write "f(n) = o(g(n))" (note: small "o") to mean that $f(n)/g(n) \to 0$ as $n \to \infty$.

• *Proof of Step B14:* By Step B13, $\mathbf{E}[\rho_{t_n}] = O(\delta)$. When $t_n \le t \le t_{n+1}$, we have $\mathbf{E}[C_t^+(j,k)] - \mathbf{E}[C_{t_n}^+(j,k)] = O(\delta)$ for all $j \ne k$ (combine the estimates of Steps B3, B4 and B10), so $\mathbf{E}[\rho_t] - \mathbf{E}[\rho_{t_n}] = O(\delta^2)$. ▮

• *Proof of Step B15:* By Step B13, $\mathbf{E}[\rho_{t_n}] = O(n^{-\lambda}) \to 0$ as $n \to \infty$. Step B12 gives

$$\mathbf{E}[t_{n+1}^2 \, \rho_{t_{n+1}} | h_{t_n}] \le t_n^2 \, \rho_{t_n} + O(n^{2-\lambda} + Y_{t_n} \, n^2).$$

Divide by t_{n+1}^2, to get (note that $t_n^2 \,/\, t_{n+1}^2 \le 1$)

$$\mathbf{E}[\rho_{t_{n+1}} | h_{t_n}] \le \rho_{t_n} + \eta_n,$$

where $\eta_n := K \,(n^{-1-\lambda} + n^{-1} \; Y_{t_n})$ for an appropriate constant $K < \infty$.

By Step B10, the following statements hold with probability 1: $\bar{Z}_t(j,k)$ converges to 0 as t goes to ∞; therefore, $[(M/\mu)\; \bar{Z}_t(j,k) - \vartheta]^+ = 0$ for all except finitely many t's (recall that $\vartheta > 0$); so the series $\Sigma_t \; \eta_t$ converges. Moreover, the estimate for $\mathbf{E}[Y_t]$ in Step B10 implies that $\Sigma_t \; \mathbf{E}[\eta_t]$ also converges.

Let

$$X_n := \rho_{t_n} - \sum_{\nu=1}^{n-1} \eta_\nu.$$

Then $\mathbf{E}[X_{n+1} | h_{t_n}] \le X_n$, so $(X_n)_{n=1,2,...}$ is a super-martingale (adapted to the sequence of fields generated by the histories $(h_{t_n})_n$). Moreover, $\sup_n \mathbf{E}[X_n^-] \le \sup_n \mathbf{E}[\Sigma_\nu^{n-1} \; \eta_\nu] \le \Sigma_{\nu=1}^\infty \; \mathbf{E}[\eta_\nu] < \infty$, so $(X_n)_n$ converges a.s. as as $n \to \infty$ (see Loève [1978, 32.3.A(i)]). Call the limit X_∞, and define

$$\rho_\infty := X_\infty + \sum_{\nu=1}^{\infty} \eta_\nu.$$

Then

$$\rho_{t_n} + \sum_{\nu=n}^{\infty} \eta_\nu = X_n + \sum_{\nu=1}^{\infty} \eta_\nu \to X_\infty + \sum_{\nu=1}^{\infty} \eta_\nu = \rho_\infty.$$

But $\Sigma_{\nu=n}^{\infty} \eta_\nu \to 0$ as $n \to \infty$, so $\rho_{t_n} \to \rho_\infty$. Fatou's lemma implies $\mathbf{E}[\rho_\infty] \le \lim_n \mathbf{E}[\rho_{t_n}] = 0$, so $\rho_\infty = 0$ a.s. ▮

- *Proof of Step B16:* We have $\mathbf{1}_{\{s_t^i=k\}}\, u^i(s_t) = \mathbf{1}_{\{s_t^i=k\}}\, u^i(k,s_t^{-i})$, so

$$B_t(j,k) - A_t(j,k)$$
$$= \left[\frac{p_t^i(j)}{p_t^i(k)} \mathbf{1}_{\{s_t^i=k\}} - \mathbf{1}_{\{s_t^i=j\}} \right] u^i(s_t) - \mathbf{1}_{\{s_t^i=j\}} [u^i(k,s_t^{-i}) - u^i(s_t)]$$
$$= \left[\frac{p_t^i(j)}{p_t^i(k)} \mathbf{1}_{\{s_t^i=k\}} - \mathbf{1}_{\{s_t^i=j\}} \right] u^i(k,s_t^{-i}) = Z_t(j,k)\, u^i(k,s_t^{-i}).$$

But s_t^i and s_t^{-i} are independent given h_{t-1}, therefore $\mathbf{E}[B_t(j,k) - A_t(j,k)\,|\,h_{t-1}] = \mathbf{E}[Z_t(j,k)\,|\,h_{t-1}]\ \mathbf{E}[u^i(k,s_t^{-i})\,|\,h_{t-1}] = 0$, since the first term is 0 by Step B10. So $(B_t(j,k) - A_t(j,k))_t$ are martingale differences, and

$$\mathrm{Var}[B_t(j,k) - A_t(j,k)] = \mathbf{E}[Z_t^2(j,k)\ (u^i(k,s_t^{-i}))^2]$$
$$\le M^2\ \mathbf{E}[Z_t^2(j,k)] = O(\delta_t^{-1}).$$

Thus the series $\Sigma_t \mathrm{Var}[B_t(j,k) - A_t(j,k)] / t^2$ converges, implying that $C_t(j,k) - D_t(j,k) = (1/t)\ \Sigma_{\tau\le t}\ [B_\tau(j,k) - A_\tau(j,k)] \to 0$ a.s. as $t \to \infty$ (again, see Loève [1978, 32.1.E]).

In the case $\gamma > 0$ we have $C_t^+(j,k) \le \sqrt{\rho_t} \to 0$ by Step B15, so $D_t^+(j,k) \to 0$ a.s. In the case $\delta_t = \delta$ for all t we get $\mathbf{E}^2[C_t^+(j,k)] \le \mathbf{E}[[C_t^+(j,k)]^2] \le \mathbf{E}[\rho_t] = O(\delta)$ by Step B14, completing the proof since $\mathbf{E}[D_t(j,k)] = \mathbf{E}[C_t(j,k)]$. ▮

Remark: For simplicity, we have assumed that all players use the same sequence $(\delta_t)_t$. In case there are different sequences $(\delta_t(i'))_t$ for the different players i', where $\delta_t(i') = \delta(i')\, t^{-\gamma(i')}$ for some $\delta(i') > 0$ and $0 \le \gamma(i') < 1/3$, then it is straightforward to check that the estimate of B6 becomes $\sum_{i' \in N} O(w^2\, t^{-1}\, (\delta_{t+w}(i'))^{-1} + w\, [\delta_t(i') - \delta_{t+w}(i')])$. This implies that the right-hand side in Step B12 is $t_n^2\, \rho_{t_n} + O(\delta(i)\, n^{2-3\gamma(i)/2} + Y_{t_n}\, n^2 + n^{3/2+3\bar{\gamma}/2})$, where $\bar{\gamma} := \max_{i' \in N} \gamma(i')$, and the final result in B16 becomes:

(i) If $\delta_t(i) = \delta(i)$ for all t, then $\limsup_t \mathbf{E}[[D_t^i]^+(j,k)] = O(\sqrt{\delta(i)})$.

(ii) If $\delta_t(i) \to 0$ (i.e., if $\gamma(i) > 0$), then $\lim_t\, [D_t^i]^+(j,k) = 0$.

REFERENCES

Auer, P., N. Cesa-Bianchi, Y. Freund and R.E. Schapire [1995], Gambling in a Rigged Casino: The Adversarial Multi-Armed Bandit Problem, *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.

Aumann, R.J. [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.

Blackwell, D. [1956b], Controlled Random Walks, *Proceedings of the International Congress of Mathematicians 1954, Vol. III*, North-Holland, 335-338.

Brown, G.W. [1951], Iterative Solutions of Games by Fictitious Play, *Activity Analysis of Production and Allocation*, T.C. Koopmans (ed.), Cowles Commission Monograph 13, Wiley, 374-376.

Camerer, C. and T.-H. Ho [1996], Experience-Weighted Attraction Learning in Games: A Unified Approach (mimeo).

Erev, I. and A.E. Roth [1997], Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria, *American Economic Review* (forthcoming).

Feller, W. [1965], *An Introduction to Probability Theory and its Applications*, *Vol. II*, 2nd edition, Wiley.

Foster, D. and R.V. Vohra [1991], Asymptotic Calibration, *Biometrika* (forthcoming).

Foster, D. and R.V. Vohra [1993], A Randomization Rule for Selecting Forecasts, *Operations Research* 41, 704-709.

Foster, D. and R.V. Vohra [1996], Calibrated Learning and Correlated Equilibrium, (1st version: 1993); *Games and Economic Behavior* (forthcoming).

Foster, D. and R.V. Vohra [1997], Regret in the On-line Decision Problem (mimeo).

Fudenberg, D. and D.K. Levine [1995], Universal Consistency and Cautious Fictitious Play, *Journal of Economic Dynamics and Control* 19, 1065-1090.

Fudenberg, D. and D.K. Levine [1996], Conditional Universal Consistency (mimeo).

Fudenberg, D. and D.K. Levine [1997], *Theory of Learning in Games* (mimeo).

Hannan, J. [1957], Approximation to Bayes Risk in Repeated Play, *Contributions to the Theory of Games, Vol. III* (*Annals of Mathematics Studies* 39),

M. Dresher, A.W. Tucker and P. Wolfe (eds.), Princeton University Press, 97-139.

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, Springer-Verlag.

Luce, R.D. and H. Raiffa [1957], *Games and Decisions*, Wiley.

Marimon, R. [1996], Learning from Learning in Economics, *Advances in Economic Theory*, D. Kreps (ed.), 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. [1995], Dual Reduction and Elementary Games, DP-1113, Northwestern University (mimeo).

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.