A Reinforcement Procedure Leading to Correlated Equilibrium*

Sergiu Hart[1] and Andreu Mas-Colell[2]

[1] Center for Rationality and Interactive Decision Theory; Department of Mathematics; and Department of Economics, The Hebrew University of Jerusalem, Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
[2] Department of Economics and Business and CREI, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain

Abstract. We consider repeated games where at any period each player knows only his set of actions and the stream of payoffs that he has received in the past. He knows neither his own payoff function, nor the characteristics of the other players (how many there are, their strategies and payoffs). In this context, we present an adaptive procedure for play — called "modified-regret-matching" — which is interpretable as a stimulus-response or reinforcement procedure, and which has the property that any limit point of the empirical distribution of play is a correlated equilibrium of the stage game.

1 Introduction

Werner Hildenbrand has repeatedly emphasized the usefulness, conceptual and technical, of carrying out equilibrium analysis by means of distributional notions. For social situations modelled as N-person games, the concept of correlated equilibrium, an extension of Nash equilibrium introduced by Aumann, can be viewed in a most natural way as imposing equilibrium conditions on the distribution of action combinations of the different players. Correlated equilibrium will be the subject of the current paper, and the style of our analysis should not surprise anybody familiar with Hildenbrand's research.

This paper continues the study of Hart and Mas-Colell [2000], where the notion of correlated equilibrium and that of approachability by means of simple "rules of thumb" or "adaptive procedures" of play were linked. We showed there that if a game is repeatedly played and the players determine their stage probabilities of play according to a procedure called "regret-matching," then the empirical distribution of play converges with probability one to the set

* Dedicated with great admiration to Werner Hildenbrand on his 65th birthday. Previous versions of these results were included in the Center for Rationality Discussion Papers #126 (December 1996) and #166 (March 1998). We thank Dean Foster for suggesting the use of "modified regrets." The research is partially supported by grants of the Israel Academy of Sciences and Humanities; the Spanish Ministry of Education; the Generalitat de Catalunya; CREI; and the EU-TMR Research Network.

of correlated equilibria of the given game. *Regret-matching* means that if at period t player i has taken action j, then the probability of playing another action k at time $t+1$ will be proportional to the "regret for not having played k instead of j," which is defined simply as the increase, if any, in the average payoff that would result if all past plays of action j were replaced by action k (and everything else remained unaltered).

The implementation of regret-matching by a player requires that player to observe the actions taken by *all* players in the past. But that is not all. The player should also be able to carry out the thought experiment of computing what his payoffs would have been, had his actions in the past been different from what they really were. For this he needs to know what game he is playing (or, at the very least, his own payoff function).

In this paper we will show that this level of sophistication and knowledge is not really necessary to obtain correlated equilibria. We shall modify the regret-matching procedure in such a way that the play probabilities are determined from the *actual realizations only.* Specifically, each player only needs to know the payoffs he received in past periods. He need not know the game he is playing — neither his own payoff function nor the others players' payoffs; in fact, he may well be oblivious of the fact that there is a game at all. The procedure that we shall examine is a "stimulus-response" or "reinforcement" procedure, in the sense that a relatively high payoff at period t will tend to increase the probability of playing at period $t+1$ the same action that was played at t.

In Section 2 we present the basic model and state our results. Section 3 contains some further discussions, and the proofs are given in Section 4.

To summarize: We consider repeated games where each player knows only his set of available choices (but nothing about the other players), and observes only his own actually realized payoffs. We exhibit simple strategies of play whereby the long-run frequencies of the various action combinations being played are tantamount to a correlated equilibrium. Of course, this sophisticated "macro"-picture that emerges cannot be seen at the individual level. This is a conclusion, we trust, with which Werner Hildenbrand can truly sympathize.

2 Model and results

We consider finite N-person games in strategic (or "normal") form. The set of players is a finite set N, the action[1] set of each player $i \in N$ is a finite set S^i, and the payoff function of i is $u^i : S \to \mathbb{R}$, where[2] $S := \prod_{\ell \in N} S^\ell$. We will denote this game $\left\langle N, (S^i)_{i \in N}, (u^i)_{i \in N} \right\rangle$ by Γ.

[1] To avoid confusion, we will use the term "action" for the one-shot game, and "strategy" for the multi-stage game.

[2] $\mathbb{R}$ denotes the real line.

A correlated equilibrium — a concept introduced by Aumann [1974] — is nothing other than a Nash equilibrium where each player may receive a private signal before the game is played (the signals do not affect the payoffs; and the players may base their choices on the signals received). This may be equivalently described as follows: Assume that the signal of each player i is in fact a "play recommendation" $s^i \in S^i$, where the N-tuple $s = \left(s^i\right)_{i \in N}$ is selected (by a "device" or "mediator") according to a commonly known probability distribution. A correlated equilibrium results if each player realizes that the best he can do is to follow the recommendation, provided that all the other players do likewise.

Equivalently, a probability distribution ψ on the set S of action N-tuples (i.e., $\psi(s) \geq 0$ for all $s \in S$ and $\sum_{s \in S} \psi(s) = 1$) is a *correlated equilibrium* of the game $\mathit{\Gamma}$ if, for every player $i \in N$ and every two actions $j, k \in S^i$ of i, we have

$$\sum_{s \in S : s^i = j} \psi(s) \left(u^i\left(k, s^{-i}\right) - u^i(s)\right) \leq 0, \tag{1}$$

where $s^{-i} \in S^{-i} := \prod_{\ell \neq i} S^\ell$ denotes the action combination of all players except i (thus $s = \left(s^i, s^{-i}\right)$). The inequality (1) (after dividing the expression there by $\sum_{s \in S : s^i = j} \psi(s)$) means that when the recommendation to player i is to choose action j, then choosing k instead of j cannot yield a higher expected payoff to i. Finally, a *correlated ε-equilibrium* obtains when 0 is replaced by ε on the right-hand side of (1).

Now assume that the game $\mathit{\Gamma}$ is played repeatedly in discrete time $t = 1, 2, \ldots$. Denote by $s_t^i \in S^i$ the (realized) choice of player i at time t, and put $s_t = (s_t^i)_{i \in N} \in S$. The payoff of player i in period t is denoted $U_t^i := u^i(s_t)$.

The basic setup of this paper assumes that the information of each player i consists just of his set of available choices S^i. He does not know the game $\mathit{\Gamma}$; in fact, he does not know how many players there are besides himself,[3] what their choices are and what the payoff functions (his own as well as the others') are. The only thing player i observes is his own realized payoffs and actions. That is, in determining his play probabilities at time $t+1$, player i's information consists of his own realized payoffs $U_1^i, U_2^i, \ldots, U_t^i$ in the previous t periods, as well as his actual actions $s_1^i, s_2^i, \ldots, s_t^i$ there.

In Hart and Mas-Colell [2000] — henceforth [HM]— we introduced *regret-matching procedures*. These are simple adaptive strategies where the play probabilities are proportional to the "regrets" for not having played other actions. Specifically, for any two distinct actions $j \neq k$ in S^i and every time t, the *regret* of i at t from j to k is

$$R_t^i(j, k) := \left[D_t^i(j, k)\right]^+ \equiv \max\left\{D_t^i(j, k), 0\right\}, \tag{2}$$

[3] Perhaps none: it could be a game against Nature.

where

$$D_t^i(j,k) := \frac{1}{t} \sum_{\tau \le t : s_\tau^i = j} \left(u^i \left(k, s_\tau^{-i} \right) - u^i \left(s_\tau \right) \right). \tag{3}$$

This is the change in the average payoff of i that would have resulted if he had played action k every time in the past that he actually chose j. Rewriting this as

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

shows that player i can compute the second term: it is just $(1/t) \sum_{\tau \le t : s_\tau^i = j} U_\tau^i$. But i cannot compute the first term, since he knows neither his own payoff function u^i nor the choices s_τ^{-i} of the other players.

Therefore we replace the first term by an "estimate" that can be computed on the basis of the available information; namely, we define

$$C_t^i(j,k) := \frac{1}{t} \sum_{\tau \le t : s_\tau^i = k} \frac{p_\tau^i(j)}{p_\tau^i(k)} U_\tau^i - \frac{1}{t} \sum_{\tau \le t : s_\tau^i = j} U_\tau^i, \tag{4}$$

where p_τ^i denotes the play probabilities at time t (thus $p_\tau^i(j)$ is the probability that i chose j at period τ); these probabilities p_τ^i are defined below (they will indeed depend only on $U_1^i, ..., U_t^i$). As in (2), the *modified regrets* are then

$$Q_t^i(j,k) := \left[C_t^i(j,k) \right]^+ \equiv \max \left\{ C_t^i(j,k), 0 \right\}. \tag{5}$$

In words, the modified regret for not having used k instead of j measures the difference (strictly speaking, its 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.[4]

Next we define the play probabilities, based on these modified regrets. Recall the regret-matching procedure of [HM]: If $j := s_t^i$ is the action chosen by i at time t, then the probability of switching at time $t+1$ to another action $k \ne j$ is proportional (with a fixed factor $1/\mu$) to the regret from j to k; with the remaining probability, the same action j is chosen again. Here we shall make two changes relative to [HM]. 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

[4] For instance, if the probability of k being chosen is always twice that of j, then in the long run k will be played twice as often as j. So, in order to compare the sum of the payoffs over those periods when j is played with the parallel sum when k is played, one needs to multiply the k-sum by $1/2$. When the probabilities change over time, Formula (4) uses the correct normalization: the difference between the regrets and the modified regrets has conditional expectation equal to 0 (see Step 10(i) and the Proof of Step 11(iii) in Section 4).

because the modified regrets are not bounded.[5] Second, we need every action to be played with some minimal frequency.[6] Thus we shall require that the transitions "tremble" over every action.

Let therefore $\mu > 0$ be a sufficiently large number; as in [HM], it suffices to take μ so that $\mu > 2M^i \left(m^i - 1\right)$ for all i, where $m^i := \left|S^i\right|$ is[7] the number of actions of i, and M^i is an upper bound on $\left|u^i\left(s\right)\right|$ for all $s \in S$. Let $0 < \delta < 1$, and define the play probabilities of player i at time $t+1$ by

$$\begin{aligned} p^i_{t+1}\left(k\right) &:= \left(1-\delta\right)\min\left\{\frac{1}{\mu}Q^i_t\left(j,k\right), \frac{1}{m^i-1}\right\} + \delta\frac{1}{m^i}, \text{ for } k \neq j; \text{ and} \\ p^i_{t+1}\left(j\right) &:= 1 - \sum_{k\in S^i : k\neq j} p^i_{t+1}\left(k\right), \end{aligned} \tag{6}$$

where $j := s^i_t$ is the choice of i in (the previous) period t. As for the first period, the play probabilities at $t = 1$ are given by an arbitrary strictly positive vector $p^i_1 \in \Delta\left(S^i\right)$; for simplicity, assume that $p^i_1\left(j\right) \geq \delta/m^i$ for all j in S^i.

Formula (6) says that p^i_{t+1} is a weighted average of two probability vectors (note that $p^i_{t+1}\left(j\right) = \left(1-\delta\right)\left(1 - \sum_{k\neq j}\min\left\{Q^i_t\left(j,k\right)/\mu, 1/\left(m^i-1\right)\right\}\right) + \delta/m^i$). The first, with weight $1-\delta$, is given by the modified regrets in a manner similar to [HM, Formula (2.2)] (note that taking the minimum with $1/\left(m^i-1\right)$ 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 $\delta/m^i > 0$.

Putting (4), (5) and (6) together implies that p^i_{t+1} depends only on the previously realized payoffs $U^i_1, ..., U^i_t$ and play probabilities $p^i_1, ..., p^i_t$. Therefore, by recursion, p^i_{t+1} does indeed depend only on $U^i_1, ..., U^i_t$.

Let[8] $z_t \in \Delta\left(S\right)$ be the empirical distribution of play up to t; that is, for every $s \in S$,

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

is the relative frequency that the N-tuple of actions s has been played in the first t stages.

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

[6] Roughly speaking, one needs this exogenous statistical "noise" to be able to estimate the contingent payoffs using only the realized payoffs (recall the first term in the modified regret). In fact, this is quite delicate, since the actions of the other players are never observed.

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

[8] We write $\Delta\left(A\right) := \left\{p \in \mathbb{R}^A_+ : \sum_{a\in A} p(a) = 1\right\}$ for the set of probability distributions over the finite set A.

Our first result is:

Theorem 1. *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 adaptive procedure (6), then the empirical distributions of play z_t converge almost surely as $t \to \infty$ to the set of correlated ε-equilibria of the game Γ.*

Thus: for almost every trajectory, z_t is close to a correlated ε-equilibrium for all large enough t; equivalently, z_t is a correlated 2ε-equilibrium from some time on (this time may depend on the trajectory; the theorem states that, with probability one, it is finite). As in the Main Theorem of [HM], note that the convergence is not to a *point*, but to a *set*; that is, the distance to the set goes to 0. Finally, we remark that both μ and δ can be taken to be different 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 (6) by δ/t^{γ}, where γ is a number strictly between 0 and[9] 1/4. Thus, we now define

$$\begin{aligned} p_{t+1}^{i}(k) &:= \left(1-\frac{\delta}{t^{\gamma}}\right)\min\left\{\frac{1}{\mu}Q_{t}^{i}(j,k),\frac{1}{m^{i}-1}\right\}+\frac{\delta}{t^{\gamma}}\frac{1}{m^{i}}, \text{ for } k \neq j; \text{ and} \\ p_{t+1}^{i}(j) &:= 1-\sum_{k\in S^{i}:k\neq j}p_{t+1}^{i}(k), \end{aligned} \tag{7}$$

where, again, $j := s_t^i$ is the choice of i in period t. At $t = 1$, let p_1^i be an arbitrary vector in $\Delta\left(S^i\right)$ with $p_1^i(j) \geq \delta/m^i$ for all j in S^i. The second result is:

Theorem 2. *If every player i plays according to adaptive procedure (7), 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, first, that δ_t converge to 0, and second, that it do so sufficiently slowly (otherwise the modified regrets C_t^i may become too large; recall that the probabilities p_t^i, 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 $\gamma < 1/4$ suffices; see the proof in Section 4). Moreover, we note that one may well take different sequences δ_t^i for different players i (cf. the Remark at the end of Section 4).

[9] The reason for this restriction will be explained below.

3 Discussion

(a) Reinforcement and stimulus-response

The play procedures of this paper can be interpreted as "reinforcement" or "stimulus-response" procedures (see, for example, Bush and Mosteller [1955], Roth and Erev [1995], Borgers and Sarin [1995], Erev and Roth [1998], Fudenberg and Levine [1998, Section 4.8]). Indeed, the behavior of our players is very far from the ideal of a sophisticated decision-maker that makes optimal decisions given his (more or less well-formed) beliefs about the environment. The behavior we postulate is, in fact, much closer to the model of a reflex-oriented individual that, from a very limited conception of the world in which he lives, simply reacts to stimuli by reinforcing those behaviors with "pleasurable" consequences.

In order to be specific and to illustrate the above point further, let us assume that the payoffs are positive and focus on the limit case of our procedure where probabilities are chosen in exact proportion to the modified regrets (5). Suppose that player i has played action j at period t. We then have for every $k \neq j$:

- if $Q_t^i(j,k) > 0$ then $Q_{t+1}^i(j,k) < Q_t^i(j,k)$; and
- if $Q_t^i(j,k) = 0$ then $Q_{t+1}^i(j,k) = 0$.

Thus all the modified regrets decrease (or stay null) from t to $t+1$. Hence the probability of choosing j at $t+1$ gets "reinforced" (i.e., increases relative to the same probability at t) by the occurrence of j at t, while the probabilities of the other actions $k \neq j$ decrease. Moreover, as can easily be seen from Definitions (4) and (5), the higher the payoff obtained at time t (when j was played), the greater this reinforcement. Finally, all these effects decrease with time — since we average over t in (4).

(b) Related work

There is by now a substantive body of work on not fully optimal behavior in repeated games (see for instance the book of Fudenberg and Levine [1998]). In particular, strategies that lead to the set of correlated equilibria have been proposed by Foster and Vohra [1997] and by Fudenberg and Levine [1999]. There is also an older tradition, beginning with Hannan [1957], that focuses on another form of regrets (of an "unconditional" kind — see (c) below) and on strategies that asymptotically take them down to zero. Clearly, our work (here and in Hart and Mas-Colell [2000, 2001]) belongs to these lines of research. Since the main difference between [HM] and the current paper is that here the players do not know the game, we want to point out that this "unknown game case" has already been studied. Specifically, in the context of Hannan regrets, see Foster and Vohra [1993, 1998], Auer *et al.* [1995] and Fudenberg and Levine [1998, Section 4.8] (and also Baños [1968] and Megiddo [1980] for related work).

(c) Hannan-consistency

We say that a strategy of a player is *Hannan-consistent* if it guarantees that his long-run average payoff is as large as the highest payoff that can be obtained by playing a constant action; that is, it is no less than the one-shot best-reply payoff against the empirical distribution of play of the other players. Formally, the *Hannan regret* of player i at time t for action $k \in S^i$ is defined as

$$DH_t^i(k) := \frac{1}{t}\sum_{\tau=1}^{t} u^i(k, s_\tau^{-i}) - \frac{1}{t}\sum_{\tau=1}^{t} U_\tau^i = u^i(k, z_t^{-i}) - \frac{1}{t}\sum_{\tau=1}^{t} U_\tau^i,$$

where $z_t^{-i} \in \Delta(S^{-i})$ is the empirical distribution of the actions chosen by the other players in the past.[10,11] A strategy of player i is then called *Hannan-consistent* if, as t increases, all Hannan-regrets are guaranteed to become almost surely non-positive in the limit, no matter what the other players do; that is, with probability one, $\limsup_{t\to\infty} DH_t^i(k) \leq 0$ for all $k \in S^i$. The reader is referred to Hart and Mas-Colell [2001] for detailed discussions and results.

In the setup of the current paper, the "modified-regret-matching" approach leads to a simple reinforcement strategy that is Hannan-consistent (recall [HM, Theorem B]): For every $k \in S^i$ define

$$CH_t^i(k) := \frac{1}{t}\sum_{\tau \leq t : s_\tau^i = k} \frac{1}{p_\tau^i(k)} U_\tau^i - \frac{1}{t}\sum_{\tau=1}^{t} U_\tau^i,$$

and

$$p_{t+1}^i(k) := (1-\delta_t)\frac{\left[CH_t^i(k)\right]_+}{\sum_{j\in S^i}\left[CH_t^i(j)\right]_+} + \delta_t \frac{1}{m^i}. \tag{8}$$

Here $\delta_t = \delta/t^\gamma$ for some $\delta > 0$ and $0 < \gamma < 1/2$; we take $p_{t+1} \in \Delta(S^i)$ to be arbitrary for $t = 0$ and whenever the denominator vanishes. We have

Theorem 3. *The strategy (8) is Hannan-consistent.*

The proof of this theorem is parallel to, and simpler than, the proof of Theorem 2 below, and therefore omitted.

[10] I.e., $z_t^{-i}(s^{-i}) := \left|\{\tau \leq t : s_\tau^{-i} = s^{-i}\}\right|/t$ for each $s^{-i} \in S^{-i}$.

[11] Note that $DH_t^i(k) = \sum_{j\neq k} D_t^i(j,k)$; we can thus refer to $DH_t^i(k)$ as the "unconditional regret for k," and to $D_t^i(j,k)$ as the "regret for k, conditional on j."

4 Proof

In this section we will prove Theorems 1 and 2 of Section 2 together. Let

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

where $\delta > 0$ and $0 \leq \gamma < 1/4$. For Theorem 1 take $\gamma = 0$, and for Theorem 2, $\gamma > 0$.

We introduce some notations, in addition to those of the previous sections. Fix player i in N; for simplicity, we drop reference to the index i whenever this does not create confusion (thus we write C_t instead of C_t^i, and so on). Recall that $m := \left|S^i\right|$ is the number of strategies of i, and M is an upper bound on the payoffs: $M \geq \left|u^i(s)\right|$ for all $s \in S$. Denote $L := \left\{(j,k) \in S^i \times S^i : j \neq k\right\}$; then $\mathbb{R}^L$ is the $m(m-1)$ Euclidean space with coordinates indexed by L.

For each $t = 1, 2, ...$ and each (j,k) in L, denote[12]

$$\begin{aligned}
Z_t(j,k) &:= \frac{p_t^i(j)}{p_t^i(k)} \mathbf{1}_{\left\{s_t^i = k\right\}} - \mathbf{1}_{\left\{s_t^i = j\right\}}; \\
B_t(j,k) &:= Z_t(j,k)\, u^i(s_t); \\
A_t(j,k) &:= \mathbf{1}_{\left\{s_t^i = j\right\}} \left(u^i\left(k, s_t^{-i}\right) - u^i(s_t)\right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
C_t(j,k) &= \frac{1}{t} \sum_{\tau \leq t} B_\tau(j,k) \text{ and} \\
D_t(j,k) &= \frac{1}{t} \sum_{\tau \leq t} A_\tau(j,k).
\end{aligned}$$

We shall write B_t for the vector $\left(B_t(j,k)\right)_{(j,k)\in L} \in \mathbb{R}^L$; and similarly for the other vectors A_t, C_t, and so on. Next, define

$$\Pi_t(j,k) := \begin{cases} (1-\delta_t) \min\left\{\frac{1}{\mu} C_t^+(j,k), \frac{1}{m-1}\right\} + \delta_t \frac{1}{m}, & \text{if } k \neq j, \\[2ex] (1-\delta_t)\left(1 - \sum\limits_{k' \neq j} \min\left\{\frac{1}{\mu} C_t^+(j,k'), \frac{1}{m-1}\right\}\right) + \delta_t \frac{1}{m}, & \text{if } k = j. \end{cases}$$

Note that $\Pi_t(j,\cdot) \in \Delta\left(S^i\right)$ for all $j \in S^i$; thus Π_t is a transition probability matrix on S^i. Both procedures (6) and (7) satisfy $p_{t+1}^i(k) = \Pi_t\left(s_t^i, k\right)$ for all k (where, again, $\gamma = 0$ corresponds to (6) and $\gamma > 0$ to (7)). Let

$$\rho_t := \left(\operatorname{dist}\left(C_t, \mathbb{R}_-^L\right)\right)^2 \equiv \sum_{j \neq k} \left(C_t^+(j,k)\right)^2$$

[12] We write $\mathbf{1}_E$ for the *indicator* of the event E (i.e., $\mathbf{1}_E = 1$ if E occurs and $= 0$ otherwise).

be the squared Euclidean distance (in $\mathbb{R}^L$) of the vector C_t from the non-positive orthant $\mathbb{R}^L_-$.

We will use the standard "**O**" notation: For two real-valued functions $f(\cdot)$ and $g(\cdot)$ defined on a domain X, we take "$f(x) = \mathbf{O}(g(x))$" to mean that there exists a constant $K < \infty$ such that $|f(x)| \leq Kg(x)$ for all x in X. From now on, t, v, 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 strategies of i (i.e., 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. Finally, **P** stands for Probability, **E** for Expectation and **Var** for Variance.

The proof of Theorems 1 and 2 will be divided into 14 steps. Step 14 shows that the regrets are "small" in the limit; the Proposition of Section 3 of [HM] then implies that the empirical distributions are correlated approximate equilibria. We note that Steps 1–11 hold for *any* non-increasing sequence $\delta_t > 0$, whereas Steps 12–14 make use of the special form $\delta_t = \delta/t^\gamma$. A guide to the proof follows the statement of the steps.

- **Step 1:**

(i) $$\mathbf{E}\left[(t+v)^2 \rho_{t+v} \mid h_t\right] \leq t^2 \rho_t + 2t \sum_{w=1}^{v} C_t^+ \cdot \mathbf{E}\left[B_{t+w} \mid h_t\right] + \mathbf{O}\left(\frac{v^2}{\delta_{t+v}^2}\right).$$

(ii) $$(t+v)^2 \rho_{t+v} - t^2 \rho_t = \mathbf{O}\left(\frac{tv + v^2}{\delta_{t+v}^2}\right).$$

Define[13]

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

- **Step 2:**

$$C_t^+ \cdot \mathbf{E}\left[B_{t+w} \mid h_t\right] = \mu \mathbf{E}\left[\sum_{j \in S^i} u^i\left(j, s_{t+w}^{-i}\right) \beta_{t+w}(j) \mid h_t\right].$$

- **Step 3:**

(i) $$C_{t+v}(j, k) - C_t(j, k) = \mathbf{O}\left(\frac{v}{t\delta_{t+v}}\right).$$

(ii) $$\Pi_{t+v}(j, k) - \Pi_t(j, k) = \mathbf{O}\left(\frac{v}{t\delta_{t+v}} + (\delta_t - \delta_{t+v})\right).$$

[13] 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_{t+w}^{-i}, but not on s_{t+w}^i).

Define

$$\begin{aligned}\overline{Z}_t\left(j,k\right) &:= \frac{1}{t}\sum_{\tau=1}^{t} Z_\tau\left(j,k\right);\ \text{and}\\ Y_t &:= \sum_{(j,k)\in L}\left|\overline{Z}_t\left(j,k\right)\right|.\end{aligned}$$

- **Step 4:**

$$\Pi_t\left(j,k\right)-\frac{1}{\mu}C_t^+\left(j,k\right)=\mathbf{O}\left(\delta_t+Y_t\right)\ \text{ for all } j\neq k.$$

- **Step 5:**[14]

$$\begin{aligned}C_t^+\cdot\mathbf{E}\left[B_{t+w}\mid h_t\right] &= \mu\mathbf{E}\left[\sum_{j\in S^i}u^i\left(j,s_{t+w}^{-i}\right)\left(\left(\Pi_t\right)^2-\Pi_t\right)\left(s_{t+w-1}^i,j\right)\mid h_t\right]\\ &\quad+\mathbf{O}\left(\delta_t+Y_t+\frac{w}{t\delta_{t+w}}\right).\end{aligned}$$

For each $t>0$ and each history h_t, we define an auxiliary stochastic process $\left(\hat{s}_{t+w}\right)_{w=0,1,2,...}$ with values in S as follows: The initial state is $\hat{s}_t=s_t$, and the transition probabilities are[15]

$$\mathbf{P}\left[\hat{s}_{t+w}=s\mid\hat{s}_t,...,\hat{s}_{t+w-1}\right]:=\prod_{\ell\in N}\Pi_t^\ell\left(\hat{s}_{t+w-1}^\ell,s^\ell\right).$$

That is, the $\hat{s}$-process is stationary: It uses the transition probabilities of period t at each period $t+1,t+2,...,t+w,....$

- **Step 6:**

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

- **Step 7:**

$$\begin{aligned}C_t^+\cdot\mathbf{E}\left[B_{t+w}\mid h_t\right] &= \mu\mathbf{E}\left[\sum_{j\in S^i}u^i\left(j,\hat{s}_{t+w}^{-i}\right)\left(\left(\Pi_t\right)^2-\Pi_t\right)\left(\hat{s}_{t+w-1}^i,j\right)\mid h_t\right]\\ &\quad+\mathbf{O}\left(\delta_t+Y_t+\frac{w^2}{t\delta_{t+w}}+w\left(\delta_t-\delta_{t+w}\right)\right).\end{aligned}$$

[14] $\left(\Pi_t\right)^2$ is the second power of the matrix Π_t (i.e., $\Pi_t\Pi_t$), and $\left(\left(\Pi_t\right)^2-\Pi_t\right)(k,j)$ is the (k,j)-element of the matrix $\left(\Pi_t\right)^2-\Pi_t$.

[15] We write Π_t^ℓ for the transition probability matrix of player ℓ (thus Π_t is Π_t^i).

- **Step 8:**

$$\mathbf{E}\left[\sum_{j\in S^i} u^i\left(j,\hat{s}^{-i}_{t+w}\right)\left(\left(\mathit{\Pi}_t\right)^2-\mathit{\Pi}_t\right)\left(\hat{s}^i_{t+w-1},j\right)\mid h_t\right]$$
$$=\sum_{s^{-i}\in S^{-i}}\mathbf{P}\left[\hat{s}^{-i}_{t+w}=s^{-i}\mid h_t\right]\left(\left(\mathit{\Pi}_t\right)^{w+1}-\left(\mathit{\Pi}_t\right)^w\right)\left(s^i_t,j\right)=\mathbf{O}\left(\frac{1}{\sqrt{w}}\right)$$

- **Step 9:**

$$\mathbf{E}\left[\left(t+v\right)^2\rho_{t+v}\mid h_t\right]\leq$$

$$t^2\rho_t+\mathbf{O}\left(tv\delta_t+tvY_t+\frac{v^3}{\delta_{t+v}}+tv^2\left(\delta_t-\delta_{t+v}\right)+t\sqrt{v}+\frac{v^2}{\delta^2_{t+v}}\right).$$

- **Step 10:**

$$\textbf{(i)}\ \ \mathbf{E}\left[Z_t\left(j,k\right)\mid h_{t-1}\right]=0.$$
$$\textbf{(ii)}\ \ \mathbf{Var}\left[Z_t\left(j,k\right)\right]=\mathbf{O}\left(\frac{1}{\delta_t}\right).$$

- **Step 11:**

$$\textbf{(i)}\ \ \lim_{t\to\infty}\overline{Z}_t\left(j,k\right)=0\text{ a.s.}$$
$$\textbf{(ii)}\ \ \lim_{t\to\infty}Y_t=0\text{ a.s.}$$
$$\textbf{(iii)}\ \ \lim_{t\to\infty}\left(C_t\left(j,k\right)-D_t\left(j,k\right)\right)=0\text{ a.s.}$$

Let ξ satisfy[16]

$$1<\xi<\min\left\{\frac{2}{1+\gamma},\frac{1}{4\gamma}\right\};\tag{9}$$

such a ξ exists since $0\leq\gamma<1/4$. For each $n=1,2,...$, let $t_n:=\left\lfloor n^\xi\right\rfloor$ be the largest integer not exceeding n^ξ.

- **Step 12:** There exists $\eta<2\xi-1$ such that

$$\mathbf{E}\left[t^2_{n+1}\rho_{t_{n+1}}\mid h_{t_n}\right]\leq t^2_n\rho_{t_n}+\mathbf{O}\left(\delta n^{2\xi-\xi\gamma-1}+Y_{t_n}n^{2\xi-1}+n^\eta\right).$$

- **Step 13:**

$$\textbf{(i)}\ \ \text{If }\gamma=0\text{ then }\limsup_{n\to\infty}\rho_{t_n}=\mathbf{O}\left(\delta\right)\text{ a.s.}$$
$$\textbf{(ii)}\ \ \text{If }\gamma>0\text{ then }\lim_{n\to\infty}\rho_{t_n}=0\text{ a.s.}$$

[16] When $\gamma=0$, (9) is $1<\xi<2$.

- **Step 14:**

$$\textbf{(i)}\ \text{If } \gamma = 0 \text{ then } \limsup_{t\to\infty} R_t(j,k) = \mathbf{O}\left(\sqrt{\delta}\right) \text{ a.s.}$$

$$\textbf{(ii)}\ \text{If } \gamma > 0 \text{ then } \lim_{t\to\infty} R_t(j,k) = 0 \text{ a.s.}$$

We now provide a short intuitive overview of the steps of the proof. The proof is based on the Proof of the Main Theorem of [HM] (see Steps M1–M11 in the Appendix there) — which in turn is inspired by the Approachability Theorem of Blackwell [1956] — 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 $Q_t \equiv C_t^+$ are small. From this one readily gets in Step 14 that the actual regrets $R_t \equiv D_t^+$ are also small, since the difference $C_t - D_t$ is a martingale converging almost surely to 0 (see Step 11(iii)). The main steps in the proof are as follows: We start with the basic recursion equation in Step 1(i) (similar to [HM, Step M1(i)]). Next, we estimate the "middle term" on the right-hand side of 1(i) by approximating the s-process with the $\hat{s}$-process, which is independent across players (Steps 2–7; parallel to [HM, Steps M2–M6]). This leads to a formula similar to [HM, Formula (3.4)], except that the invariant distribution q_λ in [HM, (3.4b)] is replaced here by the transitions after w and $w+1$ periods, which are close to one another (Step 8; compare with [HM, Step M7]). Finally, we obtain the recursion formula in Step 9. On comparing this formula with the parallel one in [HM, Step M8], we see that in Step 9 there are additional terms; one of them, which comes about because the modified regrets are not bounded, contains a random variable Y_t. Step 10, Step 11 and the Proof of Step 13 show that this term also goes to zero. Steps 12 and 13 complete the proof for the modified regrets, in a manner that is similar to [HM, Steps M9–M11] (though more complicated, partly because of the Y_t terms). As we have indicated above, Step 14 yields the needed result for the actual regrets.

We now proceed to the proofs of Steps 1–14.

- Proof of Step 1: Because $C_t^- \in \mathbb{R}_-^L$ we have

$$\begin{aligned}
\rho_{t+v} &\le \left\|C_{t+w} - C_t^-\right\|^2 = \left\|\frac{t}{t+v}C_t + \frac{1}{t+v}\sum_{w=1}^{v} B_{t+w} - C_t^-\right\|^2 \\
&= \frac{t^2}{(t+v)^2}\left\|C_t - C_t^-\right\|^2 + \frac{2t}{(t+v)^2}\sum_{w=1}^{v}\left(B_{t+w} - C_t^-\right)\cdot\left(C_t - C_t^-\right) \\
&\quad + \frac{v^2}{(t+v)^2}\left\|\frac{1}{v}\sum_{w=1}^{v} B_{t+w} - C_t^-\right\|^2 \\
&\le \frac{t^2}{(t+v)^2}\rho_t + \frac{2t}{(t+v)^2}\sum_{w=1}^{v} B_{t+w}\cdot C_t^+ + \frac{v^2}{(t+v)^2}m(m-1)\frac{4M^2m^2}{\delta_{t+v}^2}.
\end{aligned}$$

Indeed: For the second term, note that $C_t^- \cdot \left(C_t - C_t^-\right) = C_t^- \cdot C_t^+ = 0$. As for the third term, we have $\left|u^i\left(s\right)\right| \leq M$ and $\left|Z_{t+w}\left(j,k\right)\right| \leq m/\delta_{t+w} \leq m/\delta_{t+v}$ for $w \leq v$ (since the sequence δ_t is non-increasing); therefore $B_{t+w}\left(j,k\right)$ and $C_t\left(j,k\right)$ are each bounded by $2Mm/\delta_{t+v}$. This yields (ii). To get (i), take the conditional expectation given h_t (thus ρ_t and C_t are fixed). □

• Proof of Step 2: Note that $B_{t+w}\left(j,k\right)$ vanishes except when $s_{t+w}^i = j,k$. We condition on h_{t+w-1} and s_{t+w}^{-i} (i.e., on the whole history h_{t+w} *except* player i's choice at time $t+w$):

$$\begin{aligned}
&\mathbf{E}\left[B_{t+w}\left(j,k\right) \mid h_{t+w-1}, s_{t+w}^{-i} = s^{-i}\right] \\
&\quad = p_{t+w}^i\left(k\right)\frac{p_{t+w}^i\left(j\right)}{p_{t+w}^i\left(k\right)}u^i\left(k,s^{-i}\right) - p_{t+w}^i\left(j\right)u^i\left(j,s^{-i}\right) \\
&\quad = p_{t+w}^i\left(j\right)\left(u^i\left(k,s^{-i}\right) - u^i\left(j,s^{-i}\right)\right)
\end{aligned}$$

Hence

$$\begin{aligned}
&C_t^+ \cdot \mathbf{E}\left[B_{t+w} \mid h_{t+w-1}, s_{t+w}^{-i} = s^{-i}\right] \\
&\quad = \sum_j \sum_{k \neq j} C_t^+\left(j,k\right) p_{t+w}^i\left(j\right)\left(u^i\left(k,s^{-i}\right) - u^i\left(j,s^{-i}\right)\right) \\
&\quad = \sum_j u^i\left(j,s^{-i}\right)\left(\sum_{k \neq j} C_t^+\left(k,j\right) p_{t+w}^i\left(k\right) - \sum_{k \neq j} C_t^+\left(j,k\right) p_{t+w}^i\left(j\right)\right)
\end{aligned}$$

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

• Proof of Step 3: We have

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

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

• Proof of Step 4: We distinguish two cases. First, when $\left(1/\mu\right)C_t^+\left(j,k\right) \leq 1/\left(m-1\right)$, we have

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

and this is $\mathbf{O}\left(\delta_t\right)$ (it lies between 0 and δ_t/m).

Second, when $(1/\mu)\, C_t^+(j,k) \geq 1/(m-1)$, we have

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

For the opposite inequality, note that $|Z_\tau(j,k)| \leq 2 + Z_\tau(j,k)$ (since the only possible negative value of $Z_\tau(j,k)$ is -1), thus

$$\begin{aligned} \frac{1}{\mu} C_t^+(j,k) &\leq \frac{1}{\mu} |C_t(j,k)| \leq \frac{1}{\mu t} \sum_{\tau=1}^{t} |Z_\tau(j,k)| \left|u^i(s_\tau)\right| \\ &\leq \frac{2M}{\mu} + \frac{M}{\mu} \overline{Z}_t(j,k) < \frac{1}{m-1} + \frac{M}{\mu} \overline{Z}_t(j,k) \\ &= \Pi_t(j,k) + \frac{\delta_t}{m(m-1)} + \frac{M}{\mu} \overline{Z}_t(j,k) \end{aligned}$$

(recall that $\mu > 2M(m-1)$). □

- Proof of Step 5: Denote s_{t+w-1}^i 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,

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

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

$$\begin{aligned} \beta_{t+w}(j) - \left((\Pi_t)^2 - \Pi_t \right)(r,j) &= \sum_{k \neq j} \left(\frac{1}{\mu} C_t^+(k,j) - \Pi_t(k,j) \right) \Pi_{t+w-1}(r,k) \\ &\quad + \sum_{k \neq j} \Pi_t(k,j) \left(\Pi_{t+w-1}(r,k) - \Pi_t(r,k) \right) \\ &\quad - \sum_{k \neq j} \left(\frac{1}{\mu} C_t^+(j,k) - \Pi_t(j,k) \right) \times \\ &\quad \times \Pi_{t+w-1}(r,j) \\ &\quad - \sum_{k \neq j} \Pi_t(j,k) \left(\Pi_{t+w-1}(r,j) - \Pi_t(r,j) \right) \\ &= \mathbf{O}\left(\delta_t + Y_t + \frac{w}{t \delta_{t+w}} \right), \end{aligned}$$

by the estimates of Steps 3(ii) and 4. It only remains to substitute this into the formula of Step 2. □

• Proof of Step 6: We use the Lemma of [HM, Step M4], which implies that if the 1-step transition probabilities of two Markov processes on S differ by at most β, then the w-step transition probabilities differ by at most $|S|\, w\beta$. Applying this to the $\hat{s}$- and s-processes, with β given by Step 3(ii), yields the result. □

• Proof of Step 7: Replacing $(s_{t+w})_w$ by $(\hat{s}_{t+w})_w$ in the formula of Step 5 gives an additional error that is estimated in Step 6 (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 8: Given h_t, the random variables $\hat{s}^{-i}_{t+w}$ and $\hat{s}^{i}_{t+w-1}$ are independent, since the transition probabilities of the $\hat{s}$-process are all determined at time t, and the players randomize independently. Hence:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{j\in S^i} u^i\left(j, \hat{s}^{-i}_{t+w}\right)\left(\left(\Pi_t\right)^2 - \Pi_t\right)\left(\hat{s}^{i}_{t+w-1}, j\right) \mid h_t\right] \\
&= \sum_{s^{-i}\in S^{-i}} \mathbf{P}\left[\hat{s}^{-i}_{t+w} = s^{-i} \mid h_t\right] \sum_{r\in S^i} \mathbf{P}\left[\hat{s}^{i}_{t+w-1} = r \mid h_t\right]\left(\left(\Pi_t\right)^2 - \Pi_t\right)(r, j) \\
&= \sum_{s^{-i}} \mathbf{P}\left[\hat{s}^{-i}_{t+w} = s^{-i} \mid h_t\right] \sum_{r} \left(\Pi_t\right)^{w-1}\left(s^i_t, r\right)\left(\left(\Pi_t\right)^2 - \Pi_t\right)(r, j) \\
&= \sum_{s^{-i}} \mathbf{P}\left[\hat{s}^{-i}_{t+w} = s^{-i} \mid h_t\right]\left(\left(\Pi_t\right)^{w+1} - \left(\Pi_t\right)^{w}\right)\left(s^i_t, j\right).
\end{aligned}$$

The estimate of $\mathbf{O}\left(1/\sqrt{w}\right)$ is obtained by the Lemma of [HM, Step M7].[17] □

• Proof of Step 9: Putting together the estimates of Steps 7 and 8 and recalling that $\sum_{w=1}^{v} w^{\lambda} = \mathbf{O}\left(v^{\lambda+1}\right)$ for $\lambda \neq -1$ yields

$$2t\sum_{w=1}^{v} C_t^{+} \cdot \mathbf{E}\left[B_{t+w} \mid h_t\right] = \mathbf{O}\left(tv\delta_t + tvY_t + \frac{tv^3}{\delta_{t+v}} + tv^2\left(\delta_t - \delta_{t+v}\right) + t\sqrt{v}\right).$$

Recalling the formula of Step 1(i) completes the proof. □

[17] Which is based on a Central Limit Theorem estimate. Note that here (unlike the Main Theorem of [HM]) Π_t is a strictly positive stochastic matrix: all its entries are $\geq \delta_t/m > 0$. It can then be shown that $\left|\left(\Pi_t\right)^{w+1}(k, j) - \left(\Pi_t\right)^{w}(k, j)\right| \leq \left(1 - \delta_t/m\right)^{w}$. This alternative estimate can be used instead of $O\left(w^{-1/2}\right)$ (but we then need $\gamma < 1/5$ rather than $\gamma < 1/4$).

• Proof of Step 10: Part (i) follows immediately from the definition of $Z_t(j,k)$:

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

Therefore $\mathbf{E}\left[Z_t(j,k)\right] = 0$, and

$$\begin{aligned}\mathbf{Var}\left[Z_t(j,k)\right] &= \mathbf{E}\left[Z_t^2(j,k)\right] = \mathbf{E}\left[\mathbf{E}\left[Z_t^2(j,k) \mid h_{t-1}\right]\right] \\ &= \mathbf{E}\left[\frac{\left(p_t^i(j)\right)^2}{\left(p_t^i(k)\right)^2} p_t^i(k) + (-1)^2 p_t^i(j)\right] \leq \mathbf{E}\left[\frac{1}{p_t^i(k)}\right] \leq \frac{m}{\delta_t},\end{aligned}$$

which gives (ii). □

• Proof of Step 11: We will use the following Strong Law of Large Numbers for Dependent Random Variables; see Loève [1978, Theorem 32.1.E]:

Theorem 4. *Let $(X_n)_{n=1,2,...}$ be a sequence of random variables and $(b_n)_{n=1,2,...}$ a sequence of real numbers increasing to ∞, such that the series $\sum_{n=1}^{\infty} \mathbf{Var}(X_n)/b_n^2$ converges. Then*

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{\nu=1}^{n} \left(X_\nu - \mathbf{E}\left[X_\nu \mid X_1, ..., X_{\nu-1}\right]\right) = 0 \ \textit{a.s.}$$

In our case, we have by Step 10(ii)

$$\sum_{t=1}^{\infty} \frac{1}{t^2} \mathbf{Var}\left[Z_t(j,k)\right] \leq \sum_{t=1}^{\infty} \frac{m}{t^2 \delta_t} = \sum_{t=1}^{\infty} \frac{m}{\delta t^{2-\gamma}}.$$

This series converges, since $\gamma < 1/4 < 1$. Therefore

$$\frac{1}{t} \sum_{\tau \leq t} \left(Z_\tau(j,k) - \mathbf{E}\left[Z_\tau(j,k) \mid Z_1(j,k), ..., Z_{\tau-1}(j,k)\right]\right) \to 0 \text{ a.s.}$$

and thus, by Step 10(i), $\overline{Z}_t(j,k) \to 0$ a.s. This yields (i) and (ii).

To get (iii), note that $\mathbf{1}_{\{s_t^i = k\}} u^i(s_t) = \mathbf{1}_{\{s_t^i = k\}} u^i\left(k, s_t^{-i}\right)$, so

$$\begin{aligned}&B_t(j,k) - A_t(j,k) \\ &\quad = \left(\frac{p_t^i(j)}{p_t^i(k)} \mathbf{1}_{\{s_t^i = k\}} - \mathbf{1}_{\{s_t^i = j\}}\right) u^i(s_t) - \mathbf{1}_{\{s_t^i = j\}} \left(u^i\left(k, s_t^{-i}\right) - u^i(s_t)\right) \\ &\quad = \left(\frac{p_t^i(j)}{p_t^i(k)} \mathbf{1}_{\{s_t^i = k\}} - \mathbf{1}_{\{s_t^i = j\}}\right) u^i\left(k, s_t^{-i}\right) = Z_t(j,k)\, u^i\left(k, s_t^{-i}\right).\end{aligned}$$

But s_t^i and s_t^{-i} are independent given h_{t-1}, therefore

$$\mathbf{E}\left[B_t(j,k) - A_t(j,k) \mid h_{t-1}\right] = \mathbf{E}\left[Z_t(j,k) \mid h_{t-1}\right] \mathbf{E}\left[u^i\left(k, s_t^{-i}\right) \mid h_{t-1}\right] = 0,$$

since the first term is 0 by Step 10(i). Moreover,

$$\begin{aligned}\mathbf{Var}\left[B_t(j,k) - A_t(j,k)\right] &= \mathbf{E}\left[Z_t^2(j,k)\left(u^i\left(k, s_t^{-i}\right)\right)^2\right] \\ &\le M^2 \mathbf{E}\left[Z_t^2(j,k)\right] = \mathbf{O}\left(\frac{1}{\delta_t}\right).\end{aligned}$$

It follows that the series $\sum_t \mathbf{Var}\left[B_t(j,k) - A_t(j,k)\right]/t^2$ converges, implying that[18] $C_t(j,k) - D_t(j,k) = (1/t)\sum_{\tau \le t}\left(B_\tau(j,k) - A_\tau(j,k)\right) \to 0$ a.s. $t \to \infty$ (argument as in the proof of (i) above). □

• Proof of Step 12: Apply the inequality of Step 9 with $t = t_n = \lfloor n^\xi \rfloor$ and $v = t_{n+1} - t_n$. Then: $v = \mathbf{O}\left(n^{\xi-1}\right)$; $\delta_t \approx \delta n^{-\xi\gamma}$; $\delta_{t+v} \approx \delta(n+1)^{-\xi\gamma} = \mathbf{O}\left(n^{-\xi\gamma}\right)$; and[19] $\delta_t - \delta_{t+v} = \mathbf{O}\left(n^{-\xi\gamma-1}\right)$. Therefore

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

To complete the proof, note that the definition (9) of ξ implies that: $3\xi - \xi\gamma - 3 \le 3\xi + \xi\gamma - 3 < 2\xi - 1$ since $\xi < 2/(1+\gamma)$; $(3\xi - 1)/2 < 2\xi - 1$ since $\xi > 1$; and $2\xi + 2\xi\gamma - 2 < 2\xi - 1$ since $\xi < 1/(4\gamma) \le 1/(2\gamma)$. Therefore we take $\eta := \max\{3\xi + \xi\gamma - 3, (3\xi-1)/2, 2\xi + 2\xi\gamma - 2\} < 2\xi - 1$. □

• Proof of Step 13: Let $b_n := t_n^2 \approx n^{2\xi}$ and $X_n := b_n \rho_{t_n} - b_{n-1}\rho_{t_{n-1}} = t_n^2 \rho_{t_n} - t_{n-1}^2 \rho_{t_{n-1}}$. By Step 1(ii) we have $X_n = \mathbf{O}\left(\left(t_n v_n + v_n^2\right)/\delta_{t_{n+1}}^2\right) = \mathbf{O}\left(n^{2\xi+2\xi\gamma-1}\right)$; thus $\sum_n \mathbf{Var}(X_n)/b_n^2 = \sum_n \mathbf{O}\left(n^{4\xi+4\xi\gamma-2}/n^{4\xi}\right) = \sum_n \mathbf{O}\left(n^{4\xi\gamma-2}\right)$ converges (since $4\xi\gamma < 1$ by the choice of ξ).

Next, consider $(1/b_n)\sum_{\nu \le n} \mathbf{E}\left[X_\nu \mid X_1, ..., X_{\nu-1}\right]$. The inequality of Step 12 yields three terms: The first is $\mathbf{O}\left(n^{-2\xi}\delta \sum_{\nu \le n} \nu^{2\xi-\xi\gamma-1}\right) = \mathbf{O}\left(\delta n^{-\xi\gamma}\right)$; the second one converges to 0 a.s. as $t \to \infty$, since $Y_{t_n} \to 0$ by Step 10(ii) and Lemma 1 below (with $y_n = Y_{t_n}$ and $a_n = n^{2\xi-1}$); and the third one is $\mathbf{O}\left(n^{\eta-(2\xi-1)}\right) \to 0$ since $\eta < 2\xi - 1$. Altogether, we get $\mathbf{O}(\delta)$ when $\gamma = 0$, and 0 when $\gamma > 0$.

[18] It is interesting to note that, while the regrets are invariant to a change of origin for the utility function (i.e., adding a constant to all payoffs), this is not so for the modified regrets. Nonetheless, Step 13 shows that the resulting difference is just a martingale converging a.s. to 0.

[19] When $\gamma = 0$ (and thus $\delta_t - \delta_{t+v} = 0$), this yields an (over)estimate of $O\left(n^{-1}\right)$, which will turn out however not to matter.

The proof is completed by applying again the Strong Law of Large Numbers for Dependent Random Variables (Theorem 4) and noting that $0 \leq \rho_{t_n} = (1/b_n) \sum_{\nu \leq n} X_\nu$. □

Lemma 1. *Assume: (i) $y_n \to 0$ as $n \to \infty$; (ii) $a_n > 0$ for all n; and (iii) $\sum_{n=1}^{\infty} a_n = \infty$. Then $c_n := \sum_{\nu=1}^{n} a_\nu y_\nu / \sum_{\nu=1}^{n} a_\nu \to 0$ as $n \to \infty$.*

Proof. Given $\varepsilon > 0$, let n_0 be such that $|y_n| < \varepsilon$ for all $n > n_0$. Then $c_n = \sum_{\nu \leq n_0} a_\nu y_\nu / \sum_{\nu \leq n} a_\nu + \sum_{n_0 < \nu \leq n} a_\nu y_\nu / \sum_{\nu \leq n} a_\nu$. The first term converges to 0 (since the numerator is fixed), and the second is bounded by ε. □

• Proof of Step 14: When $t_n \leq t \leq t_{n+1}$, we have by Step 3(i): $C_t(j,k) - C_{t_n}(j,k) = \mathbf{O}(v_n/(t_n \delta_{n+1})) = \mathbf{O}\left(n^{\xi\gamma - 1}\right) \to 0$, for all $j \neq k$. Thus $\limsup_{t \to \infty} \rho_t$
$= \limsup_{n \to \infty} \rho_{t_n}$. Recalling Step 11(iii) completes the proof. □

Remark. For simplicity, we have assumed that all players use the same sequence $(\delta_t)_t$. In the case of different sequences $\left(\delta_t^\ell\right)_t$ for the different players $\ell \in N$, with $\delta_t^\ell = \delta^\ell / t^{\gamma^\ell}$ for some $\delta^\ell > 0$ and $0 \leq \gamma^\ell < 1/4$, it is straightforward to check that the estimate of Step 6 becomes now $\sum_{\ell \in N} \mathbf{O}\left(w^2 / \left(t\delta_{t+w}^\ell\right) + w\left(\delta_t^\ell - \delta_{t+w}^\ell\right)\right)$. Choosing ξ to satisfy $1 < \xi < \min\{2/(1+\overline{\gamma}), 1/(4\overline{\gamma})\}$, where $\overline{\gamma} := \max_{\ell \in N} \gamma^\ell$, yields, on the right-hand side of Step 12, $t_n^2 \rho_{t_n} + \mathbf{O}\left(\delta^i n^{2\xi - \xi\gamma^i - 1} + Y_{t_n} n^{2\xi - 1}\right) + \mathbf{O}\left(n^{2\xi - 1}\right)$, and Steps 13 and 14 go through with δ^i and γ^i instead of δ and γ, respectively. The final result in Step 14 becomes:

(i) If $\delta_t^i = \delta^i$ for all t, then $\limsup_t R_t^i(j,k) = \mathbf{O}\left(\sqrt{\delta^i}\right)$ a.s.

(ii) If $\delta_t^i \to 0$ (i.e. if $\gamma(i) > 0$, then $\lim_t R_t^i(j,k) = 0$ a.s.

References

1. 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.
2. Aumann, R. J. [1974], Subjectivity and Correlation in Randomized Strategies, *Journal of Mathematical Economics* 1, 67–96.
3. Baños, A. [1968], On Pseudo-Games, *The Annals of Mathematical Statistics* 39, 1932–1945.
4. Blackwell, D. [1956], An Analog of the Minmax Theorem for Vector Payoffs, *Pacific Journal of Mathematics* 6, 1–8.
5. Borgers, T. and R. Sarin [1995], Naive Reinforcement Learning with Endogenous Aspirations, University College London (mimeo).
6. Bush, R. and F. Mosteller [1955], *Stochastic Models for Learning*, Wiley.

7. Erev, I. and A. E. Roth [1998], Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategies, *American Economic Review* 88, 848–881.
8. Foster, D. and R. V. Vohra [1993], A Randomized Rule for Selecting Forecasts, *Operations Research* 41, 704–709.
9. Foster, D. and R. V. Vohra [1997], Calibrated Learning and Correlated Equilibrium, *Games and Economic Behavior* 21, 40–55.
10. Foster, D. and R. V. Vohra [1998], Asymptotic Calibration, *Biometrika* 85, 379–390.
11. Fudenberg, D. and D. K. Levine [1998], *Theory of Learning in Games*, MIT Press.
12. Fudenberg, D. and D. K. Levine [1999], Conditional Universal Consistency, *Games and Economic Behavior* 29, 104–130.
13. Hannan, J. [1957], Approximation to Bayes Risk in Repeated Play, in *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.
14. Hart, S. and A. Mas-Colell [2000], A Simple Adaptive Procedure Leading to Correlated Equilibrium, *Econometrica*.
15. Hart, S. and A. Mas-Colell [2001], A General Class of Adaptive Strategies, *Journal of Economic Theory*.
16. Loève, M. [1978], *Probability Theory*, Vol. II, 4th Edition, Springer-Verlag.
17. Megiddo, N. [1980], On Repeated Games with Incomplete Information Played by Non-Bayesian Players, *International Journal of Game Theory* 9, 157–167.
18. 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.