Global Nash convergence of Foster and Young's regret testing

Fabrizio Germano
fabrizio.germano@upf.edu

Gábor Lugosi
lugosi@upf.es

Departament d'Economia i Empresa
Universitat Pompeu Fabra
Ramon Trias Fargas 25-27
08005 Barcelona, Spain

September 2005 (first version: October 2004)

Abstract

We construct an uncoupled randomized strategy of repeated play such that, if every player plays according to it, mixed action profiles converge almost surely to a Nash equilibrium of the stage game. The strategy requires very little in terms of information about the game, as players' actions are based only on their own past payoffs. Moreover, in a variant of the procedure, players need not know that there are other players in the game and that payoffs are determined through other players' actions. The procedure works for finite generic games and is based on appropriate modifications of a simple stochastic learning rule introduced by Foster and Young [12].

Keywords Regret testing; Regret-based learning; Random search; Stochastic dynamics; Uncoupled dynamics; Global convergence to Nash equilibria. *JEL Classification* C72, C73, D81, D83.

1 Introduction

We construct a stochastic learning rule such that, if all players play according to it, then mixed action profiles will converge, almost surely, to a Nash equilibrium of a generic game. An important feature is that it requires very little in terms of what players need to know about the underlying game. Moreover, in a variant of the basic rule, convergence obtains even if players do not know whether they are playing against other players or whether there are other players. What they do need to know are their own past realized payoffs which they need to observe over sufficiently long periods of time. The paper thus contributes to the theory of learning, by showing existence of globally converging learning rules, possibly providing intuition for why some large interactive systems might be at or close to Nash equilibrium behavior.

The procedure is a variant of the *regret testing* learning rule introduced by Foster and Young [12]. Essentially, time is divided into sufficiently long periods such that, at the beginning of each period, each player chooses a mixed action at random and plays according to the corresponding distribution for the duration of the period. If the player could not have performed much better by playing some other fixed action throughout the (just elapsed) period, then it repeats the previously played mixed action for the next period; otherwise the player randomly selects a new mixed action and plays it during the next period. The procedure thus implements a kind of exhaustive search with agents separately testing their own actions through summary statistics of past payoffs. (In this sense, it is related to reinforcement or aspiration models such as Erev and Roth [8] and Börgers and Sarin [4], see also Fudenberg and Levine [15].) The basic variation we study adds *experimentation* to Foster and Young's procedure so that, with small probability, players sample a new mixed action even if they could not have done much better with any fixed action over the previous period. (This is similar to some of the learning models with mutations or persistent randomness such as Kandori, Mailath and Rob [30] and Young [37], see also Fudenberg and Levine [15].) Among other things, this guarantees that the process of mixed action profiles taken at the beginning of each period is an irreducible Markov chain.

More specifically, the setup is the following. We consider repeated play of a finite N-player normal form game. At each time instant $t = 1, 2, \ldots,$ player $i \in N$ chooses a mixed action $\sigma_t^i \in \Sigma_i$ depending on the history and selects an action s_t^i randomly according to the distribution of σ_t^i, $i \in N$. In the basic setup, we assume that after taking an action at time t, player i observes the actions s_t^{-i} played by the rest of the players. (This assumption of *standard monitoring* is significantly weakened in Section 6.) However, we focus our attention on *uncoupled* procedures in the sense that each player i knows its own payoff function γ^i but ignores the payoff functions of the rest of the players, see Hart and Mas-Colell [23, 24]. We also allow for *randomized* procedures in the sense that, at each time instant t, player i has access to a random variable $\chi_{i,t}$ whose value it can use in determining σ_t^i where the $\chi_{i,t}$ are independent and (say) uniformly distributed over the interval $[0, 1]$.

Our main objective here is to see whether uncoupled randomized procedures can lead to Nash equilibrium. Or, more precisely, does there exist a randomized uncoupled strategy[1] such that, regardless of what the underlying game is, if all players follow such a strategy, mixed action profiles $\sigma_t = (\sigma_t^1, \ldots, \sigma_t^N)$ converge, almost surely, to a Nash equilibrium of the stage game? We answer this in the affirmative for generic games, thus providing a strong possibility or existence result. Previous work on uncoupled procedures either did not obtain global convergence for all (or almost all) N-player games, or obtained convergence to weaker notions of equilibrium and with weaker notions of convergence; see for example the discussions in Foster and Young [12] and Hart and Mas-Colell [23]. Moreover, as with Foster and Young's regret testing, our variant also extends to the case where players observe their own past realized payoffs but not the actions of the other players.

[1] Throughout the paper we use the term (mixed) action to denote the distribution σ_t^i used to play the stage game at any time instant t and use the term strategy for the repeated game strategy. In the terminology of Hart [18], our procedure belongs to the class of adaptive heuristics and is to be located between evolutionary dynamics and sophisticated learning dynamics in terms of the sophistication of the players; see also Fudenberg and Levine [15] on this.

We refer to this case as the *unknown game model*.

Perhaps the first such universal convergence result was shown by Foster and Vohra [9], who proved the existence of adaptive procedures such that the joint empirical frequencies of play, $\widehat{P}_{s,t} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{I}_{s_\tau = s}$, $s \in S$, converge to the set of correlated equilibria of the game, see also Foster and Vohra [10], Fudenberg and Levine [14, 16], Hart and Mas-Colell [19, 20, 22], Stoltz and Lugosi [35], and Cahn [5]. The original result of Foster and Vohra shows that if players base their actions on a calibrated forecast of the other players' actions then convergence to correlated equilibria takes place in the above mentioned sense. Kakade and Foster [28] take these ideas further and show that if all players play according to a best response to a certain common "almost deterministic" well-calibrated forecaster (the existence of which they also prove) then the joint empirical frequencies of play converge not only to the set of correlated equilibria but, in fact, to the convex hull of the set of Nash equilibria. Foster and Young [11, 12] introduce two procedures in which, asymptotically, the joint mixed strategy profiles are within distance ϵ of the set of Nash equilibria in a fraction of at least $1-\epsilon$ of time, though almost sure convergence is not achieved.

On the negative side, Hart and Mas-Colell [23] show that it is impossible to achieve convergence to Nash equilibrium for all games if one is restricted to deterministic uncoupled strategies. More recently, in [24] they extend the impossibility result to stationary uncoupled randomized strategies that have bounded recall. By "bounded recall" they mean that there is a finite integer T such that each player bases its play only on the last T rounds of play. At the same time, by relaxing the bounded recall assumption, for every $\epsilon > 0$, they show a randomized stationary uncoupled procedure for which mixed actions converge almost surely to an ϵ-Nash equilibrium. Their procedure relies heavily on the assumption that other players's actions are observable. In contrast, our procedure, while extending to the unknown game case, is not stationary (and neither satisfies bounded recall). Overall, their results reveal that there is a fine line between what is possible in terms of convergence to Nash equilibrium by uncoupled strategies and what is not. Our paper further contributes to the filling of this gap.

More specifically, in Theorem 1 we prove almost sure convergence of mixed action profiles to a Nash equilibrium for *generic* games, which include almost all games in the sense of the Lebesgue measure over the set of all finite N-player normal form games. Theorem 2 establishes the existence of an uncoupled randomized strategy that achieves almost sure convergence to an ϵ-Nash equilibrium without any restriction on the game. Finally, in Theorem 3 we drop the standard monitoring assumption and show convergence in the senses above in the unknown game model. Hart and Mas-Colell [21] show almost sure convergence of the empirical frequencies of play to the set of correlated equilibria in this case, and Foster and Young [12] show convergence in probability of the mixed action profiles to the set of ϵ-Nash equilibria of two player games. It is their ideas that we extend here.

The rest of the paper is organized as follows. Section 2 introduces the experimental regret testing procedure. Section 3 shows some basic properties, including that empirical frequencies converge to the convex hull of the set of ϵ–Nash equilibria. The main convergence results are in Sections 4–6. Section 6 deals with the case in which players observe their own realized payoffs but not other players' actions. Section 7 contains the proofs.

2 Preliminary definitions

We consider N-player normal form games, where N also denotes the set of players $\{1, .., N\}$. S_i denotes player i's space of pure actions with cardinality $K_i = \#S_i$, and $S = \times_{i \in N} S_i$ denotes the space of pure action profiles with cardinality $K = \sum_{i \in N} K_i$; Σ_i denotes the set of probability measures (or mixed actions) on S_i, $\Sigma = \times_{i \in N} \Sigma_i$ denotes the space of mixed action profiles. Set also $S_{-i} = \times_{j \neq i} S_j$ and $\Sigma_{-i} = \times_{j \neq i} \Sigma_j$, and for $J \subset N$, $S_J = \times_{i \in J} S_i$ and $\Sigma_J = \times_{i \in J} \Sigma_i$.

Given N and each K_i finite, we identify a game with a point in Euclidean space $\gamma \in \mathbb{R}^{\kappa N}$, where $\kappa = \prod_{i=1}^{N} K_i$. We also denote by $\gamma^i \in \mathbb{R}^{\kappa}$ the payoff array of player i and, by slight abuse of notation, also the payoff function of player i at game γ. Without loss of generality, we may assume that all payoffs take values in $[0,1]$ so that the space of games reduces to $[0,1]^{\kappa N}$. Let $B^i(\gamma) \subset \Sigma$ denote the graph of i's best reply correspondence at γ and $B^i_\epsilon(\gamma) \subset \Sigma$ the graph of i's ϵ–best reply correspondence; $\mathcal{N}(\gamma) = \cap_{i \in N} B^i(\gamma)$ denotes the set of Nash equilibria and $\mathcal{N}_\epsilon(\gamma) = \cap_{i \in N} B^i_\epsilon(\gamma)$ the set of ϵ–Nash equilibria of γ. Let $\mathcal{N}^c_\epsilon(\gamma) = \Sigma \setminus \mathcal{N}_\epsilon(\gamma)$ denote its complement in Σ; we will often suppress the argument γ. μ denotes uniform probability measure over either Σ or $[0,1]^{\kappa N}$, according to the context.

The following learning dynamics is based on the regret testing dynamics of Foster and Young [12] and coincides with it when $\lambda = 0$.

Definition 1 Experimental regret testing *with parameters* (T, ρ, λ), *where* $T \in \mathbb{N}$, $\rho \in \mathbb{R}_{++}$, *and* $\lambda \in (0,1)$, *is defined by the following algorithm.*
1. Initialization: Set $t = 0$. *Each player chooses* $\sigma^i_0 \in \Sigma_i$ *uniformly at random.*
2. Loop:

(a) Each player plays according to $\sigma^i_t \in \Sigma_i$ *for* $T \geq 1$ *periods, where in each of the* T *periods an action* $s^i_\tau \in S_i$ *is chosen according to the distribution* σ^i_t.

(b) Each player computes its vector of average regrets over the T *periods*

$$r^i_{t,k} = \frac{1}{T} \sum_{\tau=t+1}^{t+T} \left(\gamma^i(k, s^{-i}_\tau) - \gamma^i(s_\tau) \right) , \qquad k = 1, \ldots, K_i \tag{1}$$

where $s_\tau = (s^1_\tau, \ldots, s^N_\tau)$ *is the* N*-tuple of pure strategies played by the* N *players at round* τ *and* s^{-i}_τ *is the* $(N-1)$*-tuple obtained from* s_τ *by excluding* s^i_τ.

(c) Each player chooses $\sigma^i_{t+T} \in \Sigma_i$ *as follows: if* $r^i_{t,k} \geq \rho$ *for some* $k = 1, \ldots, K_i$*, then randomly select* $\sigma^i_{t+T} \in \Sigma_i$ *according to the uniform distribution over* Σ_i. *If* $r^i_{t,k} < \rho$ *for all* $k = 1, \ldots, K_i$*, then, with probability* $1 - \lambda$*, set* $\sigma^i_{t+T} = \sigma^i_t$ *and, with probability* λ*, randomly select* $\sigma^i_{t+T} \in \Sigma_i$ *according to the uniform distribution over* Σ_i.

(d) Set $t = t + T$ *and repeat the loop.*

In words, experimental regret testing with parameters (T, ρ, λ) is defined by an updating algorithm, where every T periods each player computes its vector of recent average regrets. If one of the components exceeds ρ, then a new action is drawn from the uniform distribution on the player's action simplex, and this action is played for the next T periods. If, on the other hand, none of the components exceeds ρ, then, with probability $1 - \lambda$, it continues to play according to the previous action for further T periods, and, with probability λ, a new action is drawn from the uniform distribution on the action simplex and is played for the next T periods.

Note that while the procedure of experimental regret testing is clearly *uncoupled* in the sense that the actions of each player only depend on the players' own past payoffs and not on the payoffs of the other players (see Hart and Mas-Colell [23, 24]), it also requires some amount of coordination, since it is assumed that all players use the same parameters (T, ρ, λ) and that the intervals of length T over which they keep their mixed actions fixed are synchronized.

The difference of this dynamics from the regret testing dynamics of Foster and Young is that in our case, with a small positive probability λ, players select a new action even if their current action does not lead to regrets above the threshold ρ. This ensures that there is some amount of experimentation by all the players throughout the learning process.

3 Properties of experimental regret testing

We state some key properties of experimental regret testing that will be used throughout the paper. The proofs are all in Section 7.

One of the key properties of experimental regret testing needed to prove such convergence is that the process of mixed action profiles $\sigma_0, \sigma_T, \sigma_{2T}, \ldots$ is a geometrically mixing Markov chain, as summarized in the following lemma. Denote by μ the uniform probability measure over the set Σ of mixed action profiles.

Lemma 1 *The stochastic process $\{\sigma_t\}$, $t = 0, T, 2T, \ldots$, defined by experimental regret learning with $0 < \lambda < 1$, is a recurrent and irreducible (L^1) Markov chain satisfying Doeblin's condition. In particular, for any measur-*

able set $A \subset \Sigma$,

$$P(\sigma \to A) \geq \lambda^N \mu(A)$$

for every $\sigma \in \Sigma$ where $P(\sigma \to A) = \mathbb{P}\{\sigma_{(m+1)T} \in A | \sigma_{mT} = \sigma\}$ denotes the transition probabilities of the Markov chain. (Here m is an arbitrary nonnegative integer.)

An immediate corollary is the following (see, e.g., Meyn and Tweedie [32, Theorem 16.2.4]).

Corollary 1 *For $m = 0, 1, 2, \ldots$ let P_m denote the distribution of σ_{mT}, that is, $P_m(A) = \mathbb{P}\{\sigma_{mT} \in A\}$. Then there exists a unique probability distribution π over Σ (the stationary distribution of the Markov process) such that*

$$\sup_A |P_m(A) - \pi(A)| \leq (1 - \lambda^N)^m$$

where the supremum is taken over all measurable sets $A \subset \Sigma$.

The main idea behind Foster and Young's heuristics is that, after a not very long search period, by pure chance, the mixed action profile σ_{mT} will be an ϵ-Nash equilibrium, and then, since all players have a small expected regret, the process gets stuck with this value for a much longer time than the search period. The main technical result needed to justify such a statement is summarized in Lemma 3 which will imply that the length of the search period is negligible compared to the length of time the process spends in an ϵ-Nash equilibrium. A similar result was used by Foster and Young [12] for the case of two players.

Throughout the paper we work with generic games in the following sense. Given a game $\gamma \in [0,1]^{\kappa N}$, we say a game $\gamma' \in [0,1]^{\kappa' N}$ is a *pure subgame* of γ if $S' \subset S$, $\kappa' = \prod_{i \in N} K'_i$, where $K'_i = \#S'_i \geq 1$, and where the payoffs are the ones induced by γ, that is, $\gamma' = \gamma_{|S'}$. For an arbitrary set $J \subset N$ and arbitrary mixed action profile $\sigma^J \in \Sigma_J$, let γ_{σ^J}, denote the subgame where players in J play the fixed mixed action σ^J. We call an N-player normal form game $\gamma \in [0,1]^{\kappa N}$ *generic* if every pure subgame has only regular Nash equilibria and for every pure subgame γ' of γ, we have for almost every mixed action profile $\sigma^J \in \Sigma_J$, $J \subset N$, that the subgame γ'_{σ^J} of γ' also only has regular Nash equilibria. The notion of regular Nash equilibrium we use is as in Ritzberger [33] or van Damme [36]; essentially we require that the system of equations defining a given equilibrium be invertible.

Lemma 2 *Almost every game $\gamma \in [0,1]^{\kappa N}$ is generic.*

Let $\mathcal{N}_\epsilon^c(\gamma) = \Sigma \setminus \mathcal{N}_\epsilon(\gamma)$ denote the complement of the set of ϵ-Nash equilibria. The next lemma is essential for the convergence results.

Lemma 3 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic N-player normal form game. Then there exist positive constants c_1, c_2 such that, for all sufficiently small $\rho > 0$, the N–step transition probabilities of experimental regret testing satisfy*

$$P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho) \geq c_1 \rho^{c_2} \ .$$

(where we use the notation $P^{(N)}(A \to B) = \mathbb{P}\{\sigma_{(m+N)T} \in B | \sigma_{mT} \in A\}$ for the N-step transition probabilities).

One more technical result is needed before we state the main properties of experimental regret testing.

The next basic proposition shows that after sufficiently many rounds of play the distribution of the joint mixed actions σ concentrates in the neighborhood of the set of Nash equilibria. It extends the main result of Foster and Young [12] to generic games of an arbitrary number of players.

Proposition 1 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic N-player normal form game. There exists a positive number ϵ_0 such that for all $\epsilon < \epsilon_0$ the following holds: there exist positive constants $c_1, \ldots, c_4$ such that if the experimental regret testing procedure is used with parameters*

$$\rho \in (\epsilon, \epsilon + \epsilon^{c_1}) \ , \quad \lambda \leq c_2 \epsilon^{c_3} \ , \quad \textit{and} \quad T \geq -\frac{1}{2(\rho - \epsilon)^2} \log \left(c_4 \epsilon^{c_3}\right) \ ,$$

then for all $M \geq \log(\epsilon/2)/\log(1 - \lambda^N)$,

$$P_M(\mathcal{N}_\epsilon^c) = \mathbb{P}\{\sigma_{MT} \notin \mathcal{N}_\epsilon\} \leq \epsilon \ .$$

An implication of this theorem concerns the long-term joint empirical frequencies of play. If all players play according to the experimental regret testing procedure, then the joint empirical frequencies of play converge almost surely to a mixed action profile $\overline{P}$ that is in the convex hull of ϵ-Nash equilibria taken in $\Delta(S)$.

Recall that, for each $i \in N$, $\tau = 1, 2, \ldots$, $s_\tau^i \in S_i$ is the pure action played by the ith player, and where s_τ^i is drawn randomly according to the mixed action σ_{mT}^i whenever $\tau \in \{mT+1, \ldots, (m+1)T\}$. Consider the joint empirical distribution of plays $\widehat{P}_t$ defined by

$$\widehat{P}_{s,t} = \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{I}_{s_\tau = s} \ , \quad s \in S \ .$$

Denote the convex hull taken in $\Delta(S)$ by $\mathrm{co}(\cdot)$. We can state the following.

Corollary 2 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic N-player normal form game. For every $\epsilon > 0$ there exists a choice of the parameters (T, ρ, λ) such that there is a $\overline{P} \in \mathrm{co}(\mathcal{N}_\epsilon) \subset \Delta(S)$ such that the joint empirical frequencies of play of experimental regret testing satisfy*

$$\lim_{t \to \infty} \widehat{P}_t \to \overline{P} \quad \textit{almost surely.}$$

Remark. (INITIALIZATION). In the definition of experimental regret testing we assumed that each player chooses its initial mixed action σ_0^i uniformly at random. The reason for the choice of the uniform distribution is merely simplicity, and it is easy to see that Proposition 1 remains true under the weaker assumption that the distribution of σ_0 is absolutely continuous with respect to the uniform measure on Σ. This observation will be relevant in Section 4.

Remark. (UNCOUPLEDNESS). Corollary 2 guarantees, for any fixed ϵ, the existence of parameters (T, ρ, λ) such that the empirical frequencies of play converge to $\mathcal{N}_\epsilon$. Moreover, it is clear from the proof that these parameters depend not only on ϵ but also on properties of the overall game, and therefore, the procedure using these parameters is not uncoupled. In a fully uncoupled procedure, the players should be able to determine the parameters based solely on the value of ϵ. In the following sections we introduce fully uncoupled versions of this strategy. Proposition 1 should be treated as a main technical tool for further analysis.

Remark. (RATES OF CONVERGENCE). The bounds established in Proposition 1 also allow us to estimate the length of play MT, as a function of ϵ, to achieve that the mixed action profile is an ϵ-Nash equilibrium with a probability at least $1-\epsilon$. The bounds reveal that experimental regret testing with appropriately chosen parameters achieves this after $O\left((1/\epsilon)^C\right)$ rounds of play where the constant C depends, in a complicated way, on the properties of the game. However, a closer look at the proof reveals that C is at least proportional with $K = \sum_{i=1}^{N} K_i$ (the sum of the number of actions of all players) and therefore the speed of convergence is at least exponentially slow as a function of the number of players and the number of actions of each player. This slow rate of convergence is in sharp contrast with the rates of convergence achievable to approximate correlated equilibria. In fact, it follows from results of Cesa-Bianchi and Lugosi [6] that there exists an uncoupled way of play such that, after $O(\epsilon^{-2}\log(K/\epsilon))$ rounds of play the joint empirical frequencies of play form, with probability at least $1-\epsilon$, an ϵ-correlated equilibrium.

4 Convergence in generic games

The purpose of this section is to derive a regret-based method that guarantees that the mixed action profiles σ_t, $t = 1, 2, \ldots$ converge almost surely to the set $\mathcal{N}$ of Nash equilibria of a generic game. Thus, we not only claim convergence of the empirical frequencies of plays but also of the actual mixed action profiles σ_t. Also, we show convergence to $\mathcal{N}$ and not only to the convex hull $\mathrm{co}(\mathcal{N}_\epsilon)$ of ϵ-Nash equilibria for a fixed ϵ. Actually, our proposed method guarantees convergence of $\{\sigma_t\}$ to just one Nash equilibrium, though in case of multiple Nash equilibria the limiting equilibrium may depend on the actual (random) realization of the sequence of plays.

The basic idea is to "anneal" experimental regret testing such that first it is used with some parameters (T_1, ρ_1, λ_1) for a number M_1 of periods of length T_1, then change the parameters to (T_2, ρ_2, λ_2) (by increasing T and decreasing ρ and λ properly), use experimental regret testing for a number $M_2 \gg M_1$ of periods (of length T_2), etc. However, this is not sufficient to guarantee almost sure convergence as at each change of parameters the process is reinitialized and therefore there is an infinite set of indices t such that σ_t is far away from any Nash equilibrium. The solution we propose is a careful modification of experimental regret testing that guarantees that for any ϵ, $\sigma_t \notin \mathcal{N}_\epsilon$ only occurs a finite number of times, almost surely. This is achieved by "localizing" the search after each change of parameters such that each player limits its choice to a small neighborhood of the mixed action played right before the change of parameters (unless a player experiences a large regret in which case the search is extended again to the whole simplex).

Another challenge we must face is that the values of the parameters of the procedure (i.e., $T_\ell, \rho_\ell, \lambda_\ell$, and M_ℓ, $\ell = 1, 2, \ldots$) cannot depend on the parameters of the game, since by requiring uncoupledness we must assume that the players only know their payoff function but not those of the other players.

Next we define the annealed localized experimental regret testing process. To this end, let $\epsilon_1 > \epsilon_2 > \cdots$ be a decreasing sequence of positive numbers such that $\sum_{\ell=1}^{\infty} \epsilon_\ell < \infty$. For the sake of concreteness, for each $\ell = 1, 2, \ldots$, take $\epsilon_\ell = 2^{-\ell}$, and define

$$\rho_\ell = \epsilon_\ell + \epsilon_\ell^\ell \ , \quad \lambda_\ell = \epsilon_\ell^\ell \ , \quad \text{and} \quad T_\ell = \left\lceil -\frac{1}{2\epsilon_\ell^{2\ell}} \log\left(\epsilon_\ell^\ell\right) \right\rceil \ .$$

Introduce also

$$M_\ell = 2 \left\lceil \frac{\log \frac{2}{\epsilon_\ell}}{\log \frac{1}{1-\lambda_\ell}} \right\rceil \ ,$$

and denote by $\sigma_{[\ell]}^i$ the mixed action played by player i at the end of the $(\ell-1)$-st regime, by $D_\infty^i(\sigma^i, \epsilon)$ the L_∞–ball of radius ϵ around $\sigma^i \subset \Sigma_i$ and by $D_\infty(\sigma, \epsilon) = \max_{i \in N} D_\infty^i(\sigma^i, \epsilon)$ the L_∞–ball of radius ϵ around $\sigma \subset \Sigma$. For simplicity, let also $r_t^i = \max_{k=1,\ldots,K_i} r_{t,k}^i$.

Definition 2 ANNEALED LOCALIZED EXPERIMENTAL REGRET TESTING.
1. Initialization: Each player chooses $\sigma_0^i \in \Sigma_i$ uniformly at random.
2. Loop: There are different regimes indexed by $\ell = 1, 2, \ldots$. In the ℓ-th regime, each player plays according to the loop of experimental regret testing with parameters $(T_\ell, \rho_\ell, \lambda_\ell)$ during M_ℓ periods of length T_ℓ with step (c) of experimental regret testing replaced by the following,

(c) Each player chooses $\sigma_{t+T_\ell}^i \in \Sigma_i$ as follows:

(c1) if $r_t^i \geq \epsilon_\ell^{2/3}$, then select $\sigma_{t+T_\ell}^i$ randomly according to the uniform distribution over Σ_i;

(c2) if $\rho_\ell \leq r_t^i < \epsilon_\ell^{2/3}$, then select $\sigma_{t+T_\ell}^i$ randomly according to the uniform distribution over Σ_i if, for some $t' < t$ of the current (ℓ-th) regime, $\sigma_{t'+T_\ell}^i$ has been selected randomly and uniformly from Σ_i, and otherwise select $\sigma_{t+T_\ell}^i$ randomly according to the uniform distribution over $D_\infty^i(\sigma_{[\ell]}^i, \sqrt{\epsilon_\ell})$;

(c3) if $r_t^i < \rho_\ell$, then with probability $1-\lambda_\ell$ set $\sigma_{t+T_\ell}^i = \sigma_t^i$ and with probability λ_ℓ select $\sigma_{t+T_\ell}^i \in D_\infty^i(\sigma_{[\ell]}^i, \sqrt{\epsilon_\ell})$ randomly according to the uniform distribution.

The main result of this section is the following theorem which establishes almost sure convergence of the procedure described above to Nash equilibria.

Theorem 1 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic N-player normal form game and let $\{\epsilon_\ell\}_{\ell=1}^{\infty}$ be defined by $\epsilon_\ell = 2^{-\ell}$. If each player plays according to annealed localized experimental regret testing, then the sequence of mixed action profiles converges almost surely, and*

$$\lim_{t\to\infty} \sigma_t \in \mathcal{N} \quad \textit{almost surely.}$$

In case of multiple Nash equilibria the value of the limit may depend on the randomization used in the procedure.

Remark. (ANNEALING AND LOCALIZATION). As mentioned above, annealing and localization are both necessary to get almost sure convergence to (exact) Nash equilibrium. Localization allows players who are experiencing small regrets over long periods of time to narrow their search (including experimentation) to decreasing neighborhoods of the low-regret actions. It is important to make sure these neighborhoods eventually always contain a Nash equilibrium of the game. The distinction of case (c2) ensures that localization does not have players searching too frequently within neighborhoods not containing a Nash equilibrium.

Remark. (UNCOUPLEDNESS REVISITED). Note that the procedure is fully uncoupled as the only parameter is the sequence $\{\epsilon_\ell\}_{\ell=1}^{\infty}$, which is independent of the properties of the game. This is to be contrasted with the corresponding remark after Corollary 2.

Remark. (PLAUSIBLE STRATEGIES). The specific parameters given in the definition of the procedure of annealed experimental regret testing make it unlikely that one finds agents under "natural" circumstances that follow such a strategy. While we recognize that the specific details of the procedure may be quite unnatural, we emphasize that the main message of this paper is that there exists an uncoupled strategy that leads to Nash equilibrium for "most" games even in the model of unknown games, and Theorem 1 should be regarded as an existence result, not more. Nevertheless, the main ingredients of the procedure, such as random search, experimentation, and localization are quite natural, and appear in many learning systems. As an interesting topic for future research, it remains to see whether there exist more attractive uncoupled procedures that lead to Nash equilibrium. In particular, it would be important to find strategies that do not require synchronization between the players.

5 Non-generic games

All results presented up to this point require the game to be generic in the sense specified above. However, since almost all games are generic (with respect to the Lebesque measure over the set $[0,1]^{\kappa N}$ of all games), it is easy to construct a randomized uncoupled procedure such that convergence to an ϵ-Nash equilibrium is achieved for *all* games.

Theorem 2 *Let $\gamma \in [0,1]^{\kappa N}$ be an arbitrary N-player normal form game and let $\epsilon > 0$. There exists an uncoupled randomized learning procedure such that the mixed action profiles converge almost surely to a profile $\sigma \in \Sigma$ that is an ϵ–Nash equilibrium of γ.*

Proof. The idea is that before starting to play, each player slightly perturbes the values of its payoff function and then plays as if its payoff were the perturbed values. For example, define, for each player $i \in N$ and pure action profile $s \in S$,

$$\tilde{\gamma}^i(s) = \gamma^i(s) + U_{i,s},$$

where the $U_{i,s}$ are i.i.d. random variables uniformly distributed in the interval $[-\epsilon, \epsilon]$. Clearly, the perturbed game $\tilde{\gamma}$ is generic almost surely. Therefore, if all players play according to annealed localized experimental regret testing described in Section 4 but based on the payoffs of $\tilde{\gamma}$, then by Theorem 1 the mixed action profiles σ_t converge, with probability one, to a Nash equilibrium of $\tilde{\gamma}$. However, since for all $i \in N$, $s \in S$, we have $|\tilde{\gamma}^i(s) - \gamma^i(s)| < \epsilon$, every Nash equilibrium of $\tilde{\gamma}$ is an ϵ-Nash equilibrium of γ. □

Remark. (nash convergence for all games). Even though we only prove convergence to ϵ-Nash equilibria in the case of non-generic games, it seems plausible that, by a refinement of the same idea as in Theorem 2, it is also possible to achieve almost sure convergence to exact Nash equilibria. The idea is that, in annealed localized experimental regret testing, each time the parameters $(T_\ell, \rho_\ell, \lambda_\ell)$ are updated, the payoffs of the game γ are perturbed by a new noise $U_{(i,s),\ell}$ whose magnitude decreases with ℓ in an appropriately calibrated way. However, such a calibration is far from being trivial, as it requires a fine control of the constants from Lemma 5 and we leave its study for future research.

6 Unknown games

Next we show that all the results shown up to this point extend easily to the significantly more general case where the actions of each player can depend only on own past realized payoffs, without seeing the actions taken by the rest of the players. This model is sometimes referred to as "unknown game" as the players need not be aware of any characteristics of the game, like, for example, the number of overall players or the number of actions other players can choose from. The setup is closely related to the multi-armed bandit problem where, at each time instance, a player chooses an action and receives a reward but cannot check what reward it would have obtained had it chosen some other action (see, e.g., Auer, Cesa-Bianchi, Freund, and Schapire [1]).

Formally, an action for player i is now a sequence of functions that, at time t, assigns a mixed action σ_t^i to the payoff function γ^i, the history of payoffs $(\gamma^i(s_1), \gamma^i(s_2), \ldots, \gamma^i(s_{t-1}))$, and the randomizing variable $\chi_{i,t}$. Just as before, at time t, player i chooses action s_t^i randomly according to the mixed action σ_t^i.

Foster and Young [12] show that their regret testing procedure adapts to the unknown game model. Their idea also extends to our modifications. In order to adjust the procedures of experimental regret testing and annealed localized experimental regret testing, note that the only place in which the players look at the past is when they calculate the regrets $r_{t,k}^i$ in (1). However, each player may also estimate its regret in a simple way: at each time instant, player i flips a biased coin and if the outcome is head (whose probability is very small), then instead of choosing an action according to the mixed action σ_t^i, it chooses one uniformly. At these time instants, the player collects sufficient information to estimate the regret with respect to each fixed action $k \in K_i$.

To formalize this, consider a period between times $(m-1)T+1$ and mT and denote $t = (m-1)T$. During this period, player i draws n_i samples for each $k = 1, \ldots, K_i$ actions. Define the random variables $U_{i,\tau} \in \{0, 1, \ldots, K_i\}$, where, for τ between $(m-1)T+1$ and mT, for each $k = 1, \ldots, K_i$, there are exactly n_i values of τ such that $U_{i,\tau} = k$, and all such configurations are equally probable; for the remaining τ, $U_{i,\tau} = 0$. (In other words, for each $k = 1, \ldots, K_i$, n_i values of τ are chosen randomly, without replacement, such that these values are disjoint for different k's.) Then, at time τ, player i draws an action s^i_τ as follows: conditionally on the past up to time $\tau - 1$,

$$s^i_\tau \quad \begin{cases} \text{is distributed as } \sigma^i_\tau & \text{if } U_{i,\tau} = 0 \\ \text{equals } k & \text{if } U_{i,\tau} = k \ . \end{cases}$$

The regret $r^i_{t,k}$ may be estimated by

$$\widehat{r}^i_{t,k} = \frac{1}{n_i} \sum_{\tau=t+1}^{t+T} \mathbb{I}_{U_{i,\tau}=k} \gamma^i(k, s^{-i}_\tau) - \frac{1}{T - K_i n_i} \sum_{\tau=t+1}^{t+T} \gamma^i(s_\tau) \mathbb{I}_{U_{i,\tau}=0} \ , \qquad (2)$$

$k = 1, \ldots, K_i$. Observe that $\widehat{r}^i_{t,k}$ only depends on the past payoffs experienced by player i and therefore these estimates are feasible in the unknown game model.

After checking that Proposition 1 goes through in the unknown game model, it is easy to see by inspecting the proofs that the rest of the arguments go through without modification, and therefore the results of Theorems 1 and 2 as well as of Corollary 2 are true in this more general case. In particular, we can state

Theorem 3 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic (arbitrary) N-player normal form game (and let ϵ¿0). Then there exists an uncoupled randomized learning procedure satisfying the unknown game model, such that the mixed action profiles converge almost surely to a profile $\sigma \in \Sigma$ that is a Nash equilibrium (ϵ–Nash equilibrium) of the game γ.*

Remark. (BAYESIAN GAMES). The unknown game model can be adapted to encompass the case of Bayesian games, i.e., where payoffs depend on action profiles chosen as well as players' types. The latter are assumed to be drawn by nature from a finite set and according to a fixed distribution. We only need to require that (i) agents observe their own types and can condition their actions on those types, and (ii) the game is repeated such that at every period nature newly selects the types according to the given distribution. For every block of T periods, agents play fixed conditional actions, which are resampled if regrets over the previous T periods exceed the regret threshold and are kept unchanged otherwise (up to the experimentation probability λ). Given that the performance of the conditional actions is (unbiasedly) estimated during play, the present approach does not assume players to have any priors concerning nature's move, but rather to obtain them through repeated play. Players here are quite naive with respect to other players' actions and types, yet play converges to Bayesian Nash equilibria, in the different senses of Theorems 1 and 2 and of Corollary 2. This is to be contrasted with the belief-based learning approaches, such as, for example, Jordan [26, 27], Dekel, Fudenberg, and Levine [7], or also Kalai and Lehrer [29], Fudenberg and Levine [13], and Nachbar [31].

7 Proofs

Proof of Lemma 1. To see that the process is a Markov chain, note that at each $m = 0, 1, 2, \ldots$, σ_{mT} depends only on $\sigma_{(m-1)T}$ and the regrets $r^i_{(m-1)T,k}$ ($k = 1, \ldots, K_i$, $i \in N$). It is clearly L^1 since $\sigma_{mT,k} \in [0, 1]$ for all k, m, it is irreducible since at each $0, T, 2T, \ldots$, the probability of reaching some $\sigma'_{mT} \in A$ for any open set $A \subset \Sigma$ from any $\sigma_{(m-1)T} \in \Sigma$ is strictly positive when $\lambda > 0$, and it is recurrent since $\mathbb{E}[\sum_{m=0}^{\infty} \mathbf{1}_{\{\sigma_{mT} \in A\}} | \sigma_0 \in A] = \infty$ for all $\sigma_0 \in A$. The Doeblin condition follows simply from the presence of the "exploration parameter" λ in the definition of experimental regret testing. In particular, with probability λ^N every player chooses a mixed action randomly and, conditioned on this event, the distribution of σ_{mT} is uniform. □

Proof of Lemma 2. Harsanyi [17] shows that almost every game has a finite (and odd) number of Nash equilibria all of which are regular. Fix the number of players and actions and let $[0, 1]^{\kappa N}$ be the corresponding space of normal form games. Clearly, for any $S' \subset S$ we have that, for almost every $\gamma \in [0, 1]^{\kappa N}$, the associated pure subgame γ' of γ has finitely many equilibria, all regular. Since S is finite, there are finitely many $S' \subset S$ and hence finitely many pure subgames γ' of γ. Intersecting over all of these leaves almost all games in $[0, 1]^{\kappa N}$ with the property that all pure subgames have finitely many equilibria, all regular.

Next, we show that for almost every game $\gamma \in [0,1]^{\kappa N}$, given $J \subset N$, we have that for almost every profile $\sigma^J \in \Sigma_J$, the subgame γ_{σ^J} has all equilibria regular. (Notice that if all equilibria are regular then there can only be finitely many of them.) Moreover, since we can view γ as the pure subgame of another game, this will prove the general case as well. Fix $J \subset N$ and consider the map $\varphi_J : [0,1]^{\kappa N} \to \Sigma_J$ defined by

$$\varphi_J(\gamma) = \{\sigma^J \in \Sigma_J : \gamma_{\sigma^J} \text{ has nonregular Nash equilibria}\}.$$

Since checking whether an equilibrium is nonregular reduces to evaluating the Jacobian of an algebraic function, it is easy to see that this map is semi-algebraic (see Bochnak, Coste, and Roy [3, Prop. 2.2.4]). Therefore, its discontinuities lie on a closed lower-dimensional subset of $[0,1]^{\kappa N}$ such that there are finitely many connected components on which it is continuous (see Schanuel, Simon, and Zame [34] or Blume and Zame [2]). Moreover, if φ_J is semi-algebraic and takes a set of values E with $\mu(E) > 0$ at some point $\bar{\gamma}$ in the interior of a component on which it is continuous, then there must exist an open set $E_0 \subset E$ such that $E_0 \subset \varphi_J(\gamma)$ for any γ in an open neighborhood of $\bar{\gamma}$. In other words, for fixed $\sigma_J \in E_0$, the game γ_{σ^J} has nonregular Nash equilibria for any $\gamma \in G_0$, where $G_0 \subset [0,1]^{\kappa N}$ is an open neighborhood of $\bar{\gamma}$. But since we can view each game γ_{σ^J} as a game in $[0,1]^{\kappa_{Jc} N_{Jc}}$, and since, in particular, all games in an open neighborhood of $\bar{\gamma} \in [0,1]^{\kappa N}$ span a corresponding open neighborhood of games in $[0,1]^{\kappa_{Jc} N_{Jc}}$ around $\bar{\gamma}_{\sigma^J}$, (notice that $\sigma_J \in E_0$ is fixed), we would have that all games in such a neighborhood of $\bar{\gamma}_{\sigma^J}$ are degenerate, which is impossible. Hence, it must be the case that if φ_J takes a set of values with positive measure, it must be at a game where φ_J is discontinuous. But this can only happen on a lower dimensional set of measure zero and hence, for almost every game $\gamma \in [0,1]^{\kappa N}$, and for any $J \subset N$, we have that for almost every profile $\sigma^J \in \Sigma_J$, the subgame γ_{σ^J} has all Nash equilibria regular. □

Lemmas 4 and 5. The proof of Lemma 3 is based on two lemmas. Lemma 4 is the key in extending Foster and Young's results to the case of more than two players. It is concerned with the probabilities of moving from a situation where exactly $J < N$ agents have expected regret less than or equal to ρ (and are playing a profile that is not part of an ρ-Nash equilibrium of γ) to a situation where $J-1$ or less agents have expected regret less than or equal to ρ. Specifically, it shows that with positive probability, bounded away from zero, the $(N-J)$ agents with expected regret greater than ρ will select a action such that (at least) one of the agents in J will also have expected regret greater than ρ in the next period. This is expressed using the sets $C^J_\epsilon(\sigma^J)$ defined below.

Lemma 5 shows some basic properties of the volume and geometric structure of ϵ–Nash equilibria in generic games. Recall that for $J \subset N$, $\Sigma_J = \times_{i\in J}\Sigma_i$. Without loss we assume $K_i \geq 2, i \in N$.

Lemma 4 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic N-player normal form game with $K_i \geq 2, i \in N$, let $J \subset N$ with $J^c = N\backslash J \neq \emptyset$, and let*

$$C^J_\epsilon(\sigma^J) = \{\sigma^{J^c} \in \Sigma_{J^c} : (\sigma^J, \sigma^{J^c}) \in \cap_{i\in J} B^i_\epsilon\}$$

be the set of profiles in Σ_{J^c} to which $\sigma^J \in \Sigma_J$ is a joint ϵ–best reply by the players in J, $\epsilon \geq 0$. Then there exists $\delta(J) > 0$ and a positive number $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$,

$$\sup_{\sigma^J} \mu_{\Sigma_{J^c}}(C^J_\epsilon(\sigma^J)) \leq 1 - \delta(J) < 1,$$

where the supremum is taken over all $\sigma^J \in \Sigma_J$ that are not part of an ϵ–Nash equilibrium profile of γ.

Proof. For an arbitrary set $J \subset N$ and arbitrary mixed action profile $\sigma^J \in \Sigma_J$, let $\gamma_{\sigma^J} \in [0,1]^{\kappa_J(N-J)}$, where $\kappa_J = \Pi_{i\notin J} K_i$, denote the subgame where players in J play the fixed action σ^J. (Basically this reduces to a game between the players in J^c.)

First we show the statement for $\epsilon = 0$. To simplify notation, we drop the subscript ϵ whenever $\epsilon = 0$. Fix $J \subset N$ with $J^c \neq \emptyset$ and consider the correspondence $\eta(\sigma^{J^c})$ that maps σ^{J^c} to the set of Nash equilibria of the subgame $\gamma_{\sigma^{J^c}}$. This correspondence is semi-algebraic since it is the composition of two semi-algebraic maps, namely, the map mapping action profiles

$\sigma^{J^c} \in \Sigma_{J^c}$ to subgames $\gamma_{\sigma^{J^c}} \in [0,1]^{\kappa_{J^c J}}$ (this map is convex combinations of pure action payoffs) with the Nash correspondence $\mathcal{N}(\gamma_{\sigma^{J^c}})$ mapping subgames $\gamma_{\sigma^{J^c}}$ to Nash equilibria of $\gamma_{\sigma^{J^c}}$. Therefore its discontinuities lie on a closed lower-dimensional subset of Σ_{J^c} such that there are finitely many connected components on which it is continuous (see Schanuel, Simon and Zame [34] or Blume and Zame [2]). Moreover, by our genericity assumption it takes finitely many values for almost every profile $\sigma^{J^c} \in \Sigma_{J^c}$. This means that there exists a component $D \subset \Sigma_{J^c}$ and $\delta_0 > 0$ such that η is continuous on D, takes finitely many values on a dense subset of D, and $\mu_{\Sigma_{J^c}}(D) > \delta_0$.

To prove the lemma, suppose the claim is false. Suppose there exists a sequence of action profiles $\{\sigma^{J,n}\} \subset \Sigma_J$ such that

(i) for every n, $\sigma^{J,n}$ is not part of Nash profile of γ,

(ii) $\lim_{n\to\infty} \mu_{\Sigma_{J^c}}(C^J(\sigma^{J,n})) = 1$.

Because Σ_J is compact, there exists a convergent subsequence $\{\sigma^{J,n_k}\} \subset \Sigma_J$ such that (i) and (ii) hold for the corresponding elements. Let $\overline{\sigma}^J \in \Sigma_J$ be the limit of this subsequence, then $\mu_{\Sigma_{J^c}}(C^J(\overline{\sigma}^J)) = 1$. This means that for almost every $\sigma^{J^c} \in \Sigma_{J^c}$, $\overline{\sigma}^J \in \eta(\sigma_{J^c})$. Because η is semi-algbraic and upper hemi-continuous, (it is the composition of an upper hemi-continuous correspondence with a continuous map), if it takes the value $\overline{\sigma}^J$ almost everywhere on Σ_{J^c}, it must take it everywhere on Σ_{J^c}, i.e., $\overline{\sigma}_J \in \eta(\sigma_{J^c})$ for all $\sigma_{J^c} \in \Sigma_{J^c}$, in particular $\overline{\sigma}_J$ is part of a Nash profile of γ.

Hence, we may assume without loss that besides (i) and (ii), the sequence $\{\sigma^{J,n}\}$ also satisfies

(iii) $\lim_{n\to\infty} \sigma^{J,n} = \overline{\sigma}^J$,

(iv) for every n, $\mu_{\Sigma_{J^c}}(C^J(\sigma^{J,n})) < \mu_{\Sigma_{J^c}}(C^J(\sigma^{J,n+1})) < 1$.

This implies that there exists a sequence of subsets $\{E_n\} = C^J(\sigma^{J,n}) \subset \Sigma_{J^c}$ with $\mu_{\Sigma_{J^c}}(E_n) \uparrow 1$ such that, for every n, the correspondence η takes the value $\sigma^{J,n}$ on E_n, i.e., $\sigma^{J,n} \in \eta(\sigma^{J^c})$ for all $\sigma^{J^c} \in E_n$. But then there must exist a set E of positive measure such that η takes values arbitrarily close to $\overline{\sigma}^J$ on E (by property (iii) above). But this is impossible since on a set of measure one η is continuous and takes finitely many values of which $\overline{\sigma}^J$ is one of them.

Let now $\epsilon > 0$. Suppose that the statement is false, i.e., suppose that for any $\epsilon > 0$, $\sup_{\sigma^J} \mu_{\Sigma_{J^c}}(C^J_\epsilon(\sigma^J)) = 1$, where the supremum is taken over all $\sigma^J \in \Sigma_J$ that are not part of an ϵ–Nash equilibrium profile of γ. This implies that there is a set $E \subset \Sigma_{J^c}$ of strictly positive measure ($\geq \delta(J)$ from the case above with $\epsilon = 0$) such that for any $\sigma^{J^c} \in E$, $\sigma^J \in \eta_\epsilon(\sigma^{J^c})$ for any $\epsilon > 0$, and at the same time $\sigma^J \notin \eta(\sigma^{J^c})$. Again, this contradicts the fact that η_ϵ is semi-algebraic, upper hemi-continuous, and compact-valued. □

Lemma 5 *Let $\gamma \in [0,1]^{\kappa N}$ be a generic N-player normal form game. Then there exist positive constants $c_1, \ldots, c_8$ such that for all sufficiently small $\epsilon > 0$,*

(a) $D_\infty(\mathcal{N}, c_1\epsilon) \subset \mathcal{N}_\epsilon \subset D_\infty(\mathcal{N}, c_2\epsilon)$,

(b) $c_3\epsilon^{c_4} \leq \mu(\mathcal{N}_\epsilon) \leq c_5\epsilon^{c_4}$,

(c) if $\sigma \in \mathcal{N}_\epsilon$, then $D_\infty(\sigma, c_6\epsilon) \cap \mathcal{N} \neq \emptyset$,

(d) if $\rho > \epsilon$ and $\rho/\epsilon - 1$ is sufficiently small, then $\mu(\mathcal{N}_\rho \backslash \mathcal{N}_\epsilon) \leq c_7(\rho - \epsilon)^{c_8}$.

Proof. (a) Fix $\gamma \in [0,1]^{\kappa N}$ generic and let

$$\varphi^i(\sigma) = \max_{s^i_k \in S_i} \gamma^i(s^i_k, \sigma^{-i}) - \gamma^i(\sigma),$$

where $\gamma^i(\sigma) = \sum_{\nu \in S} \gamma^i_\nu \prod_{j \in N} \sigma^j_{\nu_j}$ denotes player i's payoff function. Notice that φ^i is semi-algbraic and Lipschitz continuous, where the Lipschitz constant depends only on parameters of the game. Recall $D_\infty(\mathcal{N}, \epsilon) = \cup_{\overline{\sigma} \in \mathcal{N}} D_\infty(\overline{\sigma}, \epsilon)$ and $\mathcal{N}_\epsilon = \{\sigma \in \Sigma : \varphi^i(\sigma) \leq \epsilon, i \in N\}$. By genericity of γ, the set $\mathcal{N}$ consists of a finite number of regular Nash equilibria, so that the set $\mathcal{N}_\epsilon$ can be written as the union of a finite number of neighborhoods, each of which is defined by a finite number of nicely behaved hypersurfaces. More precisely, there exists a positive number ϵ_0 such that for any $\epsilon < \epsilon_0$, we can write

$$\mathcal{N}_\epsilon = \cup_{\overline{\sigma} \in \mathcal{N}} U(\overline{\sigma}; \epsilon),$$

where the sets $U(\overline{\sigma}; \epsilon)$, $\overline{\sigma} \in \mathcal{N}$, satisfy

(i) $U(\overline{\sigma}; \epsilon) = \{\sigma \in \Sigma : \gamma^i(s^i_k, \sigma^{-i}) - \gamma^i(\sigma) \leq \epsilon$, for all $s^i_k \in \operatorname{supp}(\overline{\sigma})\}$,

(ii) the sets $U(\overline{\sigma};\epsilon)$ are pairwise disjoint and, are defined by a finite number of hypersurfaces (of dimension $K-2$; recall $\dim\Sigma = K-1$); moreover, except for the hypersurfaces defining Σ, which are fixed, all the others are parameterized by ϵ such that the Hausdorff distance $\mathrm{d}(\Sigma \setminus U(\overline{\sigma},\epsilon),\overline{\sigma})$ is strictly increasing in ϵ for ϵ small.

Because the equations $\gamma^i(s^i_k,\sigma^{-i}) - \gamma^i(\sigma) = \epsilon$, $s^i_k \in \mathrm{supp}(\overline{\sigma})$, that bound the sets $U(\overline{\sigma};\epsilon)$, vary smoothly with ϵ, it follows that $\mathrm{d}(\Sigma \setminus U(\overline{\sigma},\epsilon),\overline{\sigma})$ is increasing and Lipschitz continuous in ϵ. Moreover, the genericity assumption implies that the gradient of the functions $h_{s^i_k}(\sigma) = \gamma^i(s^i_k,\sigma^{-i}) - \gamma^i(\sigma)$, $s^i_k \in \mathrm{supp}(\overline{\sigma})$, is not the zero vector at $\overline{\sigma}$. Writing the distance (locally) between $\overline{\sigma}$ and the σ's satisfying $h_{s^i_k}(\sigma) = \epsilon$ as $\frac{\epsilon}{\|\nabla h_{s^i_k}(\overline{\sigma})\|_2}$, where $\|\cdot\|_2$ denotes the L_2 norm, we obtain that the slope of the Hausdorff distance $\mathrm{d}(\Sigma\setminus U(\overline{\sigma},\epsilon),\overline{\sigma})$ with respect to ϵ is positive and bounded away from zero. Thus there exist positive constants $C_1 < C_2$ such that $D_\infty(\overline{\sigma},C_1\epsilon) \subset U(\overline{\sigma},\epsilon) \subset D_\infty(\overline{\sigma},C_2\epsilon)$. Taking c_1, c_2 to be respectively the minimum and maximum over all such constants for the different Nash equilibria yields $D_\infty(\mathcal{N},c_1\epsilon) \subset \mathcal{N}_\epsilon \subset D_\infty(\mathcal{N},c_2\epsilon)$.

(b) This follows immediately given the statement and proof of (a). Since $\overline{\sigma}$ is a point in Σ, we have $\epsilon^{K-1} \leq \mu(D_\infty(\overline{\sigma},\epsilon)) \leq (2\epsilon)^{K-1}$ depending on whether $\overline{\sigma}$ is in the interior or on the boundary of Σ. In particular, we have,

$$(c_1\epsilon)^{K-1} \leq \mu(D_\infty(\mathcal{N},c_1\epsilon)) \leq \mu(\mathcal{N}_\epsilon) \leq \mu(D_\infty(\mathcal{N},c_2\epsilon)) \leq (2c_2\epsilon)^{K-1},$$

and we can take $c_3 = c_1^{K-1}, c_4 = K-1$, and $c_5 = (2c_2)^{K-1}$.

(c) From (a) we have for any $\epsilon > 0$ small, $D_\infty(\mathcal{N},c_1\epsilon) \subset \mathcal{N}_\epsilon \subset D_\infty(\mathcal{N},c_2\epsilon)$. Hence, if $\sigma \in \mathcal{N}_\epsilon$ then $\sigma \in D_\infty(\mathcal{N},c_2\epsilon)$. Taking $c_6 = 2c_2$ we have $D_\infty(\sigma,c_6\epsilon) \cap \mathcal{N} \neq \emptyset$.

(d) From (a) we have for any $\rho,\epsilon > 0$ small, $D_\infty(\mathcal{N},c_1\epsilon) \subset \mathcal{N}_\epsilon$ and $\mathcal{N}_\rho \subset D_\infty(\mathcal{N},c_2\rho)$, and hence, for $\rho > \epsilon$,

$$\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon \subset D_\infty(\mathcal{N},c_2\rho) \setminus D_\infty(\mathcal{N},c_1\epsilon),$$

where $c_2 \geq c_1$. For the volume we have,

$$\begin{aligned}
\mu(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon) &\leq \mu\left(D_\infty(\mathcal{N}, c_2\rho) \setminus D_\infty(\mathcal{N}, c_1\epsilon)\right) \\
&= \mu(D_\infty(\mathcal{N}, c_2\rho)) - \mu(D_\infty(\mathcal{N}, c_1\epsilon)) \\
&= (c_2(2\rho)^{K-1} - c_1(2\epsilon)^{K-1})(\#\mathcal{N}) \\
&\leq c_1 2^{K-1}(\#\mathcal{N})(\rho^{K-1} - \epsilon^{K-1}) \\
&\leq c_5(\rho - \epsilon),
\end{aligned}$$

where $c_5 = c_1 2^{K-1}(\#\mathcal{N}) < \infty$. The last inequality follows for $\rho/\epsilon - 1$ small. □

Proof of Lemma 3. Lemma 4 implies that, if there are exactly $J < N$ players who have regret less than ρ and are playing a profile $\sigma^J \in \Sigma_J$ that is not part of a ρ-Nash equilibrium profile, then there is a positive probability, bounded away from zero (uniformly for all possible subsets $J \subset N$; take $\min_{J \subset N} \frac{\delta(J)}{2}$), that the action profiles randomly chosen by the players in J^c will be such that all players in J^c and at least one player in J will have expected regret greater than ρ at the new action profile. For the remaining $J-1$ players, there are two possibilities: (a) their action profile is part of a ρ-Nash equilibrium, (b) their action profile is not part of a ρ-Nash equilibrium. Since we are looking for a lower bound for $P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho)$, it suffices to follow up on case (b). In case (b), Lemma 4 always applies, and repeatedly following up on those cases, one reaches a situation (after at most $N-1$ steps), where all N players randomly sample a new action. Applying Lemma 5 at this last step and combining this with the previous, we have that there exists $\delta > 0$ such that for every $\rho > 0$, $P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho) \geq \delta^{N-1} C_1 \rho^{C_2}$, for some positive constants C_1, C_2. In particular, there exist positive constants c_1, c_2 such that, for any $\rho > 0$, $P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho) \geq c_1 \rho^{c_2}$.

Proof of Proposition 1. First note that by Corollary 1,

$$P_M(\mathcal{N}_\epsilon^c) \leq \pi(\mathcal{N}_\epsilon^c) + (1 - \lambda^N)^M$$

so that it suffices to bound the measure of $\mathcal{N}_\epsilon^c$ under the stationary probability

π. Clearly,

$$\pi(\mathcal{N}_\rho) \;\; = \;\; \pi(\mathcal{N}_\rho^c) P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho) + \pi(\mathcal{N}_\rho) P^{(N)}(\mathcal{N}_\rho \to \mathcal{N}_\rho).$$

Writing $\pi(\mathcal{N}_\rho^c) = 1 - \pi(\mathcal{N}_\rho)$ and solving for $\pi(\mathcal{N}_\rho)$, we have

$$\pi(\mathcal{N}_\rho) = \frac{P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho)}{1 - P^{(N)}(\mathcal{N}_\rho \to \mathcal{N}_\rho) + P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho)}, \tag{3}$$

where

$$\begin{aligned} P^{(N)}(\mathcal{N}_\rho \to \mathcal{N}_\rho) \;\; &= \;\; \frac{\pi(\mathcal{N}_\epsilon) P^{(N)}(\mathcal{N}_\epsilon \to \mathcal{N}_\rho)}{\pi(\mathcal{N}_\rho)} \\ &\quad + \;\; \frac{\pi(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon) P^{(N)}(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon \to \mathcal{N}_\rho)}{\pi(\mathcal{N}_\rho)} \\ &\geq \;\; \frac{\pi(\mathcal{N}_\epsilon) P^{(N)}(\mathcal{N}_\epsilon \to \mathcal{N}_\rho)}{\pi(\mathcal{N}_\rho)}. \end{aligned} \tag{4}$$

To bound $P^{(N)}(\mathcal{N}_\epsilon \to \mathcal{N}_\rho)$ note that if $\sigma_{mT} \in \mathcal{N}_\epsilon$ then the expected regret of all players is at most ϵ. Since the regret estimates $r^i_{mT,k}$ are sums of T independent random variables taking values between 0 and 1 with mean at most ϵ, Hoeffding's inequality [25] implies that

$$\mathbb{P}\{r^i_{mT,k} \geq \rho\} \leq e^{-2T(\rho-\epsilon)^2} \;, \quad k = 1, \ldots, K_i, \quad i = 1, \ldots, N \;. \tag{5}$$

Then the probability that there is at least one player i and a action $k \leq K_i$ such that $r^i_{mT,k} \geq \rho$ is bounded by $\sum_{i=1}^N K_i e^{-2T(\rho-\epsilon)^2} = K e^{-2T(\rho-\epsilon)^2}$. Thus, with probability at least $(1-\lambda)^N (1 - K e^{-2T(\rho-\epsilon)^2})$, all players keep playing the same mixed action and therefore

$$P(\mathcal{N}_\epsilon \to \mathcal{N}_\epsilon) \geq (1-\lambda)^N (1 - K e^{-2T(\rho-\epsilon)^2}) \;.$$

Consequently, since $\rho > \epsilon$, we have $P(\mathcal{N}_\epsilon \to \mathcal{N}_\rho) \geq P(\mathcal{N}_\epsilon \to \mathcal{N}_\epsilon)$ and hence

$$P^{(N)}(\mathcal{N}_\epsilon \to \mathcal{N}_\rho) \geq (1-\lambda)^{N^2} (1 - K e^{-2T(\rho-\epsilon)^2})^N \geq 1 - N^2 \lambda - N K e^{-2T(\rho-\epsilon)^2}$$

(where we assumed $\lambda \leq 1$ and $K e^{-2T(\rho-\epsilon)^2} \leq 1$). Thus, using (4) and the obtained estimate, we have

$$P^{(N)}(\mathcal{N}_\rho \to \mathcal{N}_\rho) \geq (1 - N^2 \lambda - N K e^{-2T(\rho-\epsilon)^2}) \frac{\pi(\mathcal{N}_\epsilon)}{\pi(\mathcal{N}_\rho)} \;.$$

Next we need to show that, for proper choice of the parameters, $P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho)$ is sufficiently large. For generic games of N players, this follows from Lemma 3 which asserts that

$$P^{(N)}(\mathcal{N}_\rho^c \to \mathcal{N}_\rho) \geq C_1 \rho^{C_2}$$

for some positive constants C_1 and C_2 that depend on the game. Hence, from (3) we obtain

$$\pi(\mathcal{N}_\rho) \quad \geq \quad \frac{C_1\rho^{C_2}}{1-(1-N^2\lambda - NKe^{-2T(\rho-\epsilon)^2})\frac{\pi(\mathcal{N}_\epsilon)}{\pi(\mathcal{N}_\rho)} + C_1\rho^{C_2}}$$

It remains to estimate the measure $\pi(\mathcal{N}_\epsilon)/\pi(\mathcal{N}_\rho)$. To this end, observe that if $\rho - \epsilon$ is sufficiently small then the ratio $\pi(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon)/\pi(\mathcal{N}_\epsilon)$ is bounded by the ratio of the corresponding Lebesgue measures $\mu(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon)/\mu(\mathcal{N}_\epsilon)$. (Just note that the "density" of π decreases by moving away from a Nash equilibrium. More precisely, π may not be absolutely continuous with respect to the Lebesgue measure, but one can show that if $\sigma_1 \in \mathcal{N}_\rho \setminus \mathcal{N}_\epsilon$ and $\sigma_2 \in \mathcal{N}_\epsilon$ then for a sufficiently small $0 < \xi \ll \epsilon$ the L_∞ ball $D_\infty(\sigma_1, \xi)$ of radius ξ centered at σ_1 has a π-measure less than or equal to that of the same ball centered at σ_2.) The ratio of the volumes of $\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon$ and $\mathcal{N}_\epsilon$ may therefore be bounded by invoking parts (c) and (d) of Lemma 5. We obtain

$$\frac{\pi(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon)}{\pi(\mathcal{N}_\epsilon)} \leq \frac{C_3(\rho-\epsilon)^{C_4}}{C_5\epsilon^{C_6}}$$

so that

$$\frac{\pi(\mathcal{N}_\epsilon)}{\pi(\mathcal{N}_\rho)} = 1 - \frac{\pi(\mathcal{N}_\rho \setminus \mathcal{N}_\epsilon)}{\pi(\mathcal{N}_\rho)} \geq 1 - \frac{C_3(\rho-\epsilon)^{C_4}}{C_5\rho^{C_6}} \ .$$

In summary,

$$\begin{aligned}
&\pi(\mathcal{N}_\epsilon) \\
&\quad \geq \quad \pi(\mathcal{N}_\rho)\left(1 - \frac{C_3(\rho-\epsilon)^{C_4}}{C_5\rho^{C_6}}\right) \\
&\quad \geq \quad \left(1 - \frac{C_3(\rho-\epsilon)^{C_4}}{C_5\rho^{C_6}}\right)\frac{C_1\rho^{C_2}}{1-(1-N^2\lambda - NKe^{-2T(\rho-\epsilon)^2})(1-\frac{C_3(\rho-\epsilon)^{C_4}}{C_5\rho^{C_6}}) + C_1\rho^{C_2}}
\end{aligned}$$

for some positive constants $C_1, \ldots, C_6$. Substituting the choices of the parameters ρ, λ, T with sufficiently large constants $c_1, \ldots, c_6$ we have

$$\pi(\mathcal{N}_\epsilon^c) \leq \epsilon/2 \ .$$

If M is so large that $(1-\lambda^N)^M \leq \epsilon/2$, we have $P_M(\mathcal{N}_\epsilon^c) \leq \epsilon$ as desired. □

Proof of Corollary 2. Let $s^i(s) \in S_i$ denote player i's action in the action profile $s \in S$, and let $\sigma^i(s^i(s))$ denote the probability player i's mixed action σ^i assigns to the action profile s. We can then write the probability of action profile s occurring under mixed action profile σ as

$$P_s(\sigma) = \Pi_{i=1}^N \sigma^i(s^i(s)) \ , \quad s \in S, \sigma \in \Sigma \ .$$

Next, observe that by martingale convergence, for every $s \in S$,

$$\widehat{P}_{s,t} - \frac{1}{t}\sum_{\tau=1}^{t} P_s(\sigma_\tau) \to 0 \quad \text{almost surely.}$$

Therefore, it suffices to prove convergence of $\frac{1}{t}\sum_{\tau=1}^{t} P(\sigma_\tau)$. Since σ_τ is unchanged during periods of length T, we obviously have

$$\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} P(\sigma_\tau) = \lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} P(\sigma_{mT}) \ .$$

By Lemma 1 the process $\{\sigma_{mT}\}_{m=0}^\infty$ is a recurrent and irreducible Markov chain, so the ergodic theorem for Markov chains (see, e.g., [32]) implies that there exists a $\overline{\sigma} \in \Sigma$ such that

$$\lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} \sigma_{mT} = \overline{\sigma} \quad \text{almost surely,}$$

which implies that there exists a $\overline{P} \in \Delta(S)$ such that

$$\lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} P(\sigma_{mT}) = \overline{P} \quad \text{almost surely.}$$

It remains to show that $\overline{P} \in \text{co}(\mathcal{N}_\epsilon)$. By the ergodic theorem and continuity of P, in fact, $\overline{P} = \int_\Sigma P(\sigma) d\pi$, where π is the (unique) stationary distribution of the Markov process $\{\sigma_{mT}\}_{m=0}^\infty$ (on Σ).

Let $\epsilon' < \epsilon$ be a positive number such that

$$\{P \in \Delta(S) : \exists P' \in \mathrm{co}(\mathcal{N}_{\epsilon'}) \text{ such that } \|P - P'\|_1 < \epsilon'\} \subset \mathrm{co}(\mathcal{N}_\epsilon)$$

where $\| \cdot \|_1$ denotes the L_1 distance between probability measures in $\Delta(S)$. Observe that, for a generic game, such an ϵ' always exists by part (a) of Lemma 5. In fact, one may choose $\epsilon' = \epsilon/c_3$ for a sufficiently large positive constant c_3 (whose value depends on the game).

Now choose the parameters (T, ρ, λ) such that $\pi(\mathcal{N}_{\epsilon'}^c) < \epsilon'$. Proposition 1 guarantees the existence of such a choice.

Clearly,

$$\overline{P} = \int_\Sigma P(\sigma) d\pi = \int_{\mathcal{N}_{\epsilon'}} P(\sigma) d\pi + \int_{\mathcal{N}_{\epsilon'}^c} P(\sigma) d\pi \ .$$

Since $\int_{\mathcal{N}_{\epsilon'}} P(\sigma) d\pi \in \mathrm{co}(\mathcal{N}_{\epsilon'})$, we have that the L_1 distance of $\overline{P}$ and $\mathrm{co}(\mathcal{N}_{\epsilon'})$ satisfies

$$d_1(\overline{P}, \mathrm{co}(\mathcal{N}_{\epsilon'})) \leq \left\| \int_{\mathcal{N}_{\epsilon'}^c} P(\sigma) d\pi \right\|_1 \leq \int_{\mathcal{N}_{\epsilon'}^c} d\pi = \pi(\mathcal{N}_{\epsilon'}^c) < \epsilon' \ .$$

By the choice of ϵ' we indeed have $\overline{P} \in \mathrm{co}(\mathcal{N}_\epsilon)$. □

Proof of Theorem 1. The theorem follows from Proposition 1, Lemma 5, and the Borel-Cantelli lemma. First note that the parameters $(T_\ell, \rho_\ell, \lambda_\ell)$ are defined such that for all sufficiently large ℓ, they satisfy the conditions of Proposition 1 for $\epsilon = \epsilon_\ell$. Next, define the events

$$A_\ell = \{\sigma_{[\ell]} \in \mathcal{N}_{\epsilon_{\ell-1}}\} \text{ and } B_\ell = \{r^i_{mT_\ell} \leq \epsilon_\ell^{2/3}, \ \forall m \text{ in } \ell\text{-th regime}, \ \forall i \in N\},$$

where $\sigma_{[\ell]}$ is the mixed action profile played at the end of the $(\ell-1)$-st regime. We need to show that event A_ℓ occurs almost surely for all but finitely many regimes $\ell \in \mathbb{N}$. To see this, we show that the probability of event $A_{\ell+1}$ is high given event A_ℓ, and that, given event A_ℓ^c, the process almost surely reaches A_{ℓ_0} for some finite $\ell_0 > \ell$.

Fix the ℓ-th regime, $\ell \in \mathbb{N}$, and consider the events

$$C_\ell = \{r^i_{[\ell]} \geq \epsilon_{\ell-1}^{2/3}, \ \forall i \in N\} \text{ and } D_\ell = \{r^i_{[\ell]} < \epsilon_{\ell-1}^{2/3}, \ \forall i \in N\},$$

where $r^i_{[\ell]}$ is i's maximal average regret, among all players, at the end of the $\ell - 1$-st regime. Assuming event C_ℓ, annealed localized experimental regret testing is identical to the process where each player plays according to experimental regret testing with parameters $(T_\ell, \rho_\ell, \lambda_\ell)$ during M_ℓ periods of length T_ℓ (since, by (c1) and (c2), Σ_i will be the space from which agents sample throughout M_ℓ, given C_ℓ). Therefore, Proposition 1 applies directly and we have

$$\mathbb{P}(A_{\ell+1}|C_\ell) \geq 1 - \epsilon_\ell.$$

Next, consider the process where each player plays according to experimental regret testing with parameters $(T_\ell, \rho_\ell, \lambda_\ell)$ during M_ℓ periods of length T_ℓ with the only modification that in step (c) the set Σ_i is replaced by $D^i_\infty(\sigma^i_{[\ell]}, \sqrt{\epsilon_\ell})$. Assuming event A_ℓ, this process satisfies Proposition 1, since $D^i_\infty(\sigma^i_{[\ell]}, \sqrt{\epsilon_\ell}) \subset \Sigma_i$, moreover, by part (c) of Lemma 5, $D^i_\infty(\sigma^i_{[\ell]}, \sqrt{\epsilon_\ell}) \cap \mathcal{N} \neq \emptyset$. Assuming event D_ℓ, the above process differs from annealed experimental regret testing exactly on the event B^c_ℓ. The probability of this event, conditional on A_ℓ and D_ℓ, is no greater than $Ke^{-2T_\ell\left(\epsilon_\ell^{2/3} - \epsilon_{\ell-1}\right)^2}$, by Hoeffding's inequality. Therefore, we have

$$\begin{aligned}
\mathbb{P}(A_{\ell+1}|A_\ell, D_\ell) &\geq \mathbb{P}(A_{\ell+1} \cap B_\ell|A_\ell, D_\ell) \\
&= 1 - \mathbb{P}(A^c_{\ell+1}|A_\ell, D_\ell) - \mathbb{P}(A_{\ell+1} \cap B^c_\ell|A_\ell, D_\ell) \\
&\geq 1 - \epsilon_\ell - Ke^{-2T_\ell\left(\epsilon_\ell^{2/3} - \epsilon_{\ell-1}\right)^2}.
\end{aligned}$$

This shows that in the event $C_\ell \cup (A_\ell \cap D_\ell)$ with probability at least $1 - \epsilon_\ell - Ke^{-2T_\ell\left(\epsilon_\ell^{2/3} - \epsilon_{\ell-1}\right)^2}$, event $A_{\ell+1}$ occurs. It remains to show that for the cases where event $A^c_{\ell+1}$ does occur, the process is appropriately reinitialized almost surely after finitely many regimes, i.e., event $C_{\ell_0} \cup (A_{\ell_0} \cap D_{\ell_0})$ occurs, almost surely, after finitely many regimes, at $\ell_0 < \infty$. But this follows from the same reasoning as Proposition 1 and using Lemma 3, since in event $A^c_{\ell+1}$, with high probability, at least one agent will experience a large regret and so after few steps, the process will be either in $\mathcal{N}_{\epsilon_{\ell_0}}$ or have all agents simultaneously choosing from Σ_i. In either case, this eventually leads to event $C_{\ell_0} \cup (A_{\ell_0} \cap D_{\ell_0})$ occurring, with probability one, after finitely many regimes.

Putting together the probabilities and applying the Borel-Cantelli Lemma shows that event A_ℓ occurs almost surely for all but finitely many regimes. Since $\epsilon_\ell \to 0$, the process indeed converges to a Nash equilibrium with probability one. □

Proof of Theorem 3. The main step in proving the extension of Theorems 1 and 2 (as well as of Corollary 2) consists in showing that the estimated regrets (2) work in this case. For this, we need to establish an analog of inequality (5) for the deviations of the estimated regret. This is done in the next lemma.

Lemma 6 *Assume that in a certain period of length T, the expected regret $\mathbb{E}[r^i_{mT,k}|s_1,\ldots,s_{mT}]$ of player i is at most ϵ. Then, for a sufficiently small ϵ, with the choice of parameters of Proposition 1,*

$$\mathbb{P}\{\widehat{r}^i_{mT,k} \geq \rho\} \leq cT^{-1/3} + \exp\left(-T^{1/3}\left(\rho-\epsilon\right)^2\right) \ .$$

Proof. We show that, with large probability, $\widehat{r}^i_{mT,k}$ is close to $r^i_{mT,k}$. To this end, note first that

$$\left|\frac{1}{T-K_i n_i}\sum_{\tau=t+1}^{t+T}\gamma^i(s_\tau)\mathbb{I}_{U_{i,\tau}=0} - \frac{1}{T}\sum_{\tau=t+1}^{t+T}\gamma^i(s_\tau)\right| \leq 2\frac{\sum_{i=1}^N K_i n_i}{T} \ .$$

On the other hand, observe that, if there is no time instant τ for which $U_{i,\tau}=1$ and $U_{j,\tau}=1$ for some $j \neq i$, then,

$$\frac{1}{n_i}\sum_{\tau=t+1}^{t+T}\mathbb{I}_{U_{i,\tau}=k}\gamma^i(k,s_\tau^{-i})$$

is an unbiased estimate of $\frac{1}{T}\sum_{\tau=t+1}^{t+T}\gamma^i(k,s_\tau^{-i})$ obtained by random sampling. The probability that no two players sample at the same time is at most

$$TN^2 \max_{i,j\in N}\frac{K_i n_i}{T}\frac{K_j n_j}{T}$$

and by Hoeffding's inequality [25] for an average of a sample taken without replacement,

$$\widehat{\mathbb{P}}\left\{\left|\frac{1}{n_i}\sum_{\tau=t+1}^{t+T}\mathbb{I}_{U_{i,\tau}=k}\gamma^i(k,s_\tau^{-i}) - \frac{1}{T}\sum_{\tau=t+1}^{t+T}\gamma^i(k,s_\tau^{-i})\right| > \alpha\right\} \leq e^{-2n_i\alpha^2}$$

where $\widehat{\mathbb{P}}$ denotes the distribution induced by the random variables $U_{i,\tau}$. Putting everything together,

$$\mathbb{P}\{\widehat{r}^{\,i}_{mT,k} \geq \rho\} \leq TN^2 \max_{i,j \in N} \frac{K_i n_i}{T} \frac{K_j n_j}{T} + \exp\left(-2n_i \left(\rho - \epsilon - 2\frac{\sum_{i=1}^N K_i n_i}{T}\right)^2\right)$$

Choosing $n_i = O(T^{1/3})$, the first term on the right-hand side is of order $T^{-1/3}$ and $\sum_{i=1}^N K_i n_i/T = O(T^{-2/3})$ becomes negligible compared to $\rho - \epsilon$ which proves the statement. □

Thus, in the unknown game model, the estimate of inequality (5) can be replaced by that of Lemma 6. It is easy to see by inspecting the proofs that the rest of the arguments go through without modification. □

Acknowledgments. We thank Sergiu Hart, Sham Kakade, Andreu Mas-Colell, and Peyton Young for sharing their view with us on the subject, as well as Ehud Lehrer, Bill Zame, an associate editor and three referees for helpful comments. The first author acknowledges financial support from the Spanish Ministry of Science and Technology, grant SEJ2004-03619, and in form of a Ramon y Cajal fellowship. The second author acknowledges support by the PASCAL Network of Excellence under EC grant no. 506778. The work of the second author was supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2002.

[2] L.E. Blume and W.R. Zame. The algebraic geometry of perfect and sequential equilibrium. *Econometrica*, 62:783–794, 1994.

[3] J. Bochnak, M. Coste, and M.F. Roy. *Real Algebraic Geometry.* Springer-Verlag, Berlin, 1998.

[4] T. Börgers and R. Sarin. Naïve reinforcement learning with endogenous aspirations. *International Economic Review*, 41:921–950, 2000.

[5] A. Cahn. General procedures leading to correlated equilibria. *International Journal of Game Theory*, 33:21-40, 2004.

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

[7] E. Dekel, D. Fudenberg, and D. Levine. Learning to play Bayesian games. *Games and Economic Behavior*, 46:282–303, 2004.

[8] I. Erev, and A.E. Roth. Predicting how people play games: reinforcement learning in experimental games with unique mixed strategy equilibriu. *American Economic Review*, 88:848–881, 1998.

[9] D. Foster and R. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behaviour*, 21:40–55, 1997.

[10] D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–36, 1999.

[11] D.P. Foster and P.H. Young. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior*, 45:73–96, 2003.

[12] D.P. Foster and P.H. Young. Regret testing: A simple payoff-based procedure for learning Nash equilibrium. Mimeo, University of Pennsylvania and Johns Hopkins University, 2004.

[13] D. Fudenberg and D. Levine. Steady state learning and Nash equilibrium. *Econometrica*, 61:547–574, 1993.

[14] D. Fudenberg and D. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[15] D. Fudenberg and D. Levine. *The theory of learning in games.* MIT Press, Cambridge MA, 1998.

[16] D. Fudenberg and D. Levine. Universal conditional consistency. *Games and Economic Behavior*, 29:104–130, 1999.

[17] J. C. Harsanyi. Oddness of the number of equilibrium points: a new proof. *International Journal of Game Theory*, pages 235–250, 1973.

[18] S. Hart. Adaptive Heuristics. *Econometrica*, 73:1401–1430, 2005.

[19] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[20] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.

[21] S. Hart and A. Mas-Colell. A reinforcement procedure leading to correlated equilibrium. In G. Debreu, W. Neuefeind, and W. Trockel, editors, *Economic Essays: A Festschrift for Werner Hildenbrand*, pages 181–200. Srpinger, New York, 2002.

[22] S. Hart and A. Mas-Colell. Regret-based continuous-time dynamics. *Games and Economic Behavior*, 45:375–394, 2003.

[23] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93:1830–1836, 2003.

[24] S. Hart and A. Mas-Colell. Stochastic uncoupled dynamics and Nash equilibrium. Technical report, The Hebrew University of Jerusalem, 2005.

[25] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[26] J.S. Jordan. Bayesian learning in normal form games. *Games and Economic Behavior*, 3:60–81, 1991.

[27] J.S. Jordan. Bayesian learning in repeated games. *Games and Economic Behavior*, 9:8–20, 1995.

[28] S.M. Kakade and D.P. Foster. Deterministic calibration and Nash equilibrium. In *Proceedings of the 17th Annual Conference on Learning Theory*. Springer, 2004.

[29] E. Kalai and E. Lehrer Rational learning leads to Nash equilibrium. *Econometrica*, 61:1019–1045, 1993.

[30] M. Kandori, G. Mailath, and R. Rob. Learning, mutation and long run equilibria in games. *Econometrica*, 61:27–56, 1993.

[31] J.H. Nachbar Prediction, optimization, and learning in repeated games. *Econometrica*, 65:275–309, 1997.

[32] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability.* Springer-Verlag, London, 1993.

[33] K. Ritzberger. The theory of normal form games from the differentiable viewpoint. *International Journal of Game Theory*, 23:207–236, 1994.

[34] Schanuel S.H., L.K. Simon, and W.R. Zame. The algebraic geometry of games and the tracing procedure. In R. Selten, editor, *Game Equilibrium Models, II: Methods, Morals, and Markets.* Springer Verlag, Berlin, 1991.

[35] G. Stoltz and G. Lugosi. Learning correlated equilibria in games with compact sets of strategies. Technical report, Université Paris-Sud, Orsay, 2004.

[36] E. van Damme. *Stability and perfection of Nash equilibria.* Springer-Verlag, New York, 1991.

[37] P.H. Young. The evolution of conventions. *Econometrica*, 61:57–83, 1993.