

Regret Testing: A Simple Payoff-Based Procedure for Learning Nash Equilibrium*

DEAN P. FOSTER

Department of Statistics, Wharton School,
University of Pennsylvania

H. PEYTON YOUNG

Department of Economics, Johns Hopkins University
The Santa Fe Institute

October, 2003

This version: April 12, 2005

*The authors thank Andrew Felton, Ben Klemens, and several anonymous referees for constructive comments on an earlier draft.

Abstract

A learning rule is *uncoupled* if a player does not condition his strategy on the opponent's payoffs. It is *radically uncoupled* if the player does not condition his strategy on the opponent's actions or payoffs. We demonstrate a simple class of radically uncoupled learning rules, patterned after aspiration learning models, whose period-by-period behavior comes arbitrarily close to Nash equilibrium behavior in any finite two-person game.

1 Payoff-based learning rules

In this paper we propose a class of simple, adaptive learning rules that depend only on players' *realized payoffs*, such that when two players employ a rule from this class their period-by-period strategic behavior approximates Nash equilibrium behavior. Like reinforcement and aspiration models, this type of rule depends only on summary statistics that are derived from the players' received payoffs;¹ indeed the players do not even need to know they are involved in a game for them to learn equilibrium eventually.

To position our contribution with respect to the recent literature, we need to consider three separate issues: i) the amount of information needed to implement a learning rule; ii) the type of equilibrium to which the learning process tends (Nash, correlated, etc.); iii) the sense in which the process can be said to "approximate" the type of equilibrium behavior in question. (For a further discussion of these issues see Young, 2004)

Consider, for example, the recently discovered regret matching rules of Hart and Mas-Colell (2000, 2001). The essential idea is that players randomize among actions in proportion to their regrets from not having played those actions in the past. Like the regret-testing rules we introduce here,

¹See, for example, Bush and Mosteler, 1955, Erev and Roth, 1998; Karandikar, Mookherjee, Ray, and Vega-Redondo, 1998; Börgers and Sarin, 2000; and Bendor, Mookherjee, and Ray, 2001.

regret matching can be set up in such a way that players use only their realized payoffs to estimate the regrets, hence they need have no knowledge of their opponents' payoffs or their actions. However, these rules approximate equilibrium behavior in a rather weak sense: although the joint empirical distribution converges to the set of correlated equilibria, nothing guarantees that period-by-period behavior comes close to Nash equilibrium or even to correlated equilibrium.²

A second class of learning procedures that are closely related to the present proposal are the *hypothesis testing rules* introduced by Foster and Young (2003). In this approach, players act like classical statisticians, testing alternative hypotheses about their opponents' behavior against data, and rejecting if the behaviors are improbable given the hypothesis. When a rejection occurs, a new hypothesis is chosen at random from a suitable space of hypotheses, say those that attribute stationary strategies of bounded recall to the opponents. These procedures lead to period-by-period behavior that approximates Nash equilibrium behavior in the following sense: the parameters can be chosen so that at least $1 - \epsilon$ of the time the players' strategies constitute an ϵ -equilibrium of the stage game (Foster and Young, 2003). However, the behaviors do not necessarily converge to an ϵ -equilibrium, or even to the set of ϵ -equilibria; they are close to equilibrium a large fraction of the time. More importantly for the purposes of this paper, hypothesis testing assumes that the opponents' actions are observable, because this is how a player tests whether a hypothesis about the opponent's behavior is likely to be valid. In other words, hypothesis testing is an uncoupled process but not a radically uncoupled one.

It is, in fact, quite difficult to design decentralized learning rules that lead to Nash equilibrium behavior without making some compromises about the form of learning or the form of convergence (or both). Hart and Mas-

²There are many different rules that cause the joint empirical frequency distribution to converge to the set of correlated equilibria; see for example Foster and Vohra, 1999; Fudenberg and Levine, 1995, 1998; and Cahn, 2004.

Colell (2003) have shown, for example, that there exists no *deterministic* adjustment process that (like fictitious play) depends only on the empirical frequency distribution of past play and not in any way on the opponent's payoffs, and that causes the empirical frequencies to converge to Nash equilibrium in every finite game. More recently they have shown that there is no way to achieve almost sure convergence to Nash equilibrium (or even to ϵ -equilibrium, for small enough ϵ) if the players use stochastic adjustment rules that are stationary, uncoupled, and of bounded recall (Hart and Mas-Colell, 2004).

These negative results apply to a wide class of learning procedures that are, in one way or another, boundedly rational. What happens when players are perfectly rational, that is, they update their beliefs according to Bayes' rule and always choose optimal actions given their beliefs as in Kalai and Lehrer (1993)? Unfortunately this leads to difficulties when the payoff functions of the opponent are unknown. In fact, one can construct games of incomplete information such that, with probability one, rational Bayesian players have period-by-period behaviors that are *far* from Nash equilibrium a very high proportion of the time. Furthermore this can happen even when the priors about the unknown payoffs are correct ex ante (Jordan, 1991, 1993; Foster and Young, 2001).³

The aim of this paper is to demonstrate a very simple class of learning procedures that fit into the gap between these positive and negative results. In particular, they are radically uncoupled and approximate Nash equilibrium behavior, though they need not converge to any single Nash equilibrium. When the parameters are tightened at a suitable rate, one obtains convergence in probability to the set of Nash equilibria.

³For another critique of Bayesian rational learning in repeated games see Nachbar (1997, 2003).

2 Regret testing

We shall first give an informal description of the approach that emphasizes its computational simplicity and complete lack of dependence on the existence of other players. In the next section we state the main result.

Consider an individual who lives alone. At regular intervals he chooses an action from a finite set X of m possible actions. The names of the actions are written on slips of paper that are stored in a hat containing h papers. Since a given action can be written on multiple papers, the hat is a device for generating probability distributions over actions.

Step 1. Once each period (say once a minute) he reaches into the hat, draws a slip, and takes the action prescribed. He then returns the slip to the hat.

Step 2. At random times this routine is interrupted by telephone calls, where the probability of receiving a call in any given period is $\lambda \in (0, 1)$. Calls occur independently among periods. During a call he absent-mindedly chooses an action uniformly at random instead of reaching into the hat.

Step 3. Every time he takes an action he receives a payoff. At the end of day t he checks on how he is doing. Specifically, he first tallies the average payoff, $\hat{\alpha}_t$, he received per action taken over the course of the day whenever he was not on the phone. For each of his actions, $1 \leq j \leq m$, he compares $\hat{\alpha}_t$ with the average payoff, $\hat{\alpha}_{j,t}$, he received *when he chose j and was on the phone*. In effect, $\hat{\alpha}_t$ represents an endogenous *aspiration level*, and the payoffs $\hat{\alpha}_{j,t}$ are estimates of the payoff that arise when he departs from his current probabilistic strategy.

Step 4. If all of the differences $\hat{\alpha}_{j,t} - \hat{\alpha}_t$ are small in the sense that

$$(\forall j) \quad \hat{\alpha}_{j,t} - \hat{\alpha}_t \leq \tau$$

for some small *tolerance level* $\tau > 0$, he keeps the hat for the next day. Otherwise he puts the hat on the shelf and takes down another hat, also containing h papers that represent some distribution over actions. We assume

that every possible distribution of the m actions on h papers is represented by exactly one hat, and that each hat is drawn with equal probability.

Although our protagonist is unaware of it, there is someone next door using the same method, though not necessarily with the same parameters. Furthermore, the payoffs in each period depend on their joint decisions. Thus they are unwittingly engaged in a game. We claim that, *given any $\epsilon > 0$, if the τ 's and λ 's are small enough, and the hat sizes and number of periods in the day are large enough, they will be playing an ϵ -equilibrium of the stage game at least $1 - \epsilon$ of the time.*

A procedure of this form will be called a *regret testing rule*. The reason is that $\hat{\alpha}_{j,t}^i$ amounts to an estimate of the payoff on day t that player i would have resulted from playing action j all day long, hence the difference $\hat{r}_{j,t}^i = \hat{\alpha}_{j,t}^i - \hat{\alpha}_t^i$ is the estimated regret from not having done so. (Recall that the regrets cannot be evaluated directly because the opponent's actions are not observed.) The logic is that if some regret $\hat{r}_{j,t}^i$ is larger than a given tolerance level τ , the player becomes dissatisfied and chooses a new strategy, i.e., a new hat from the shelf. Otherwise, out of inertia, he sticks with his current strategy.

We hasten to say that this rule is intended to be a contribution to learning theory and should not to be interpreted literally as an empirical model of behavior, any more than fictitious play should be. It is, however, similar in structure to a well-known class of learning procedures called “aspiration models” (see, for example, Karandkikar, Mookherjee, Ray, and Vega-Redondo, 1998; Börgers and Sarin, 2000; Bendor, Mookherjee, and Ray, 2001). In this type of model, a subject switches action with some probability if the realized payoff from taking the action falls short of some payoff level that the subject hopes to achieve. This *aspiration level* may be generated in a variety of ways, but it is not uncommon to assume that it tracks the average realized payoff from past plays. Our model is broadly similar in structure, but it differs from the usual set-up in two key respects. First, dissatisfaction is triggered when a particular test statistic—the payoff from randomly chosen actions—

is noticeably better than the current aspiration level, which is the average payoff during the day (not over all previous periods). Second, the player switches probabilistically to a different *probability distribution* over actions (a different hat), whereas in a standard aspiration model the player chooses a new action with a probability that is *determined* by the current aspiration level and the current payoff. These features, when properly combined, lead to approximate Nash equilibrium play in general finite two-person games.

With some further refinements of the approach one can achieve almost sure convergence and the results can be extended to the n-person case (Germano and Lugosi, 2004). Here we shall focus on the two-person case where sharper results are obtainable and the underlying mechanics are more transparent.

3 The main result

A *regret testing rule* for individual i is defined as in the preceding section and involves four parameters: a tolerance toward payoff shortfalls, $\tau_i > 0$; the size of i 's hat, h_i (a positive integer), a probability $\lambda_i > 0$ that a telephone call occurs in any given period of time (independently among periods), and the number s of plays that occur per day. Note that s is the same for both players. The number of i 's hats, H_i , is determined by our assumption that there is one hat for each distribution on m_i actions that is representable by integer multiples of $1/h_i$.

Let G be a finite two-person game with m_i actions for player i , $i = 1, 2$. A pair of mixed strategies (p, q) constitutes an ϵ -*equilibrium* of G if neither player can increase his expected payoff by more than ϵ through a unilateral change of strategy. To simplify the computation, we shall assume (without loss of generality) that each player's utility function has been normalized so that the smallest payoff is zero and the largest payoff is one.

Theorem 1 *Let G be a finite two-person game played by regret testers and let $\epsilon > 0$. The learning parameters can be chosen so that the players' joint*

behaviors constitute an ϵ -equilibrium of game G with probability at least $1 - \epsilon$ at all sufficiently large times.

Corollary 1 *In Theorem 1 the limiting fraction of times that the players' joint behaviors constitute an ϵ -equilibrium is at least $1 - \epsilon$.*

Later we shall give explicit bounds on the parameters under which the theorem holds (see section 3.3 below). First we make several general remarks and then provide an intuitive sketch of the proof.

Remark 1

Notice that we do not claim that the learning process *converges* to an ϵ -equilibrium of G ; rather, the behaviors are very close to equilibrium *a very large fraction of the time*. Thinking of the limit as an empirical frequency, the theorem shows that the proportion of times $t' \leq t$ such that the process is in an ϵ -equilibrium at t' has greatest lower bound $1 - \epsilon$ as t goes to infinity. This is akin to convergence in probability rather than deterministic convergence. Nevertheless it is a strong claim: it says that if we were to take a “snapshot” of the players' behaviors at a random point in time, they would be very close to Nash equilibrium behavior with very high probability. In the concluding section we shall show that by annealing the learning parameters at a suitable rate, we can achieve convergence in probability to the set of Nash equilibria. With some further refinements of the approach one can actually achieve almost sure convergence, as shown by Germano and Lugosi (2004).

Remark 2

There is a structural resemblance between regret-testing and hypothesis-testing that merits some further discussion. One key similarity is that players exhibit considerable inertia. Another is that changes in behavior are random variables that allow the whole space of behaviors to be searched eventually. The major difference between the two approaches is that they operate with

different information. Under hypothesis testing, players observe the pattern of past play in order to test their hypotheses, and they optimize (subject to some smoothing) based on their hypotheses. Under regret testing there are no hypotheses and no optimization; in fact, players need not even know that an opponent exists. In short, under regret testing a player has less information than under hypothesis testing; nevertheless it is enough so that even a naive use of it can lead to equilibrium behavior.

Remark 3

It is not necessary to assume that the players revise their strategies simultaneously, that is, at the end of each day. For example, we could assume instead that if player i 's measured regrets exceed his tolerance at the end of the day, he revises his strategy with probability $\theta_i \in (0, 1)$, and with probability $1 - \theta_i$ continues to play his current strategy on the following day. This does not change the conclusion of theorem 1 or the structure of the argument in any significant way. Similarly, one can assume different s -values for the two players (within certain limits), but this would significantly complicate the proof without delivering much generality. Finally, one can refine the search phase of the process in a variety of ways that may significantly reduce the convergence time; we make no claim that the method proposed here is efficient.

3.1 Proof sketch

Before giving the proof of theorem 1 in detail, we shall sketch the intuitive idea of the argument. Consider the subset of times t that represent the beginning of a new day. At each such time we may think of the current *state* as the composition of the two hats; that is, a pair of probability distributions (p, q) for the two players. Since the hats have finite capacities h_1 and h_2 , there is a finite set Z of states, namely, all those pairs $z = (p, q)$ such that (ph_1, qh_2) is integer-valued in every component. This defines time-independent, transition probabilities $P(z \rightarrow z')$ that lead from any given state z to any other state

at the start of following day. Thus we obtain a stationary Markov process P on the finite state space Z .

A given state (p, q) is a Nash equilibrium of G if and only if the *expected* regrets are non-positive. A state is an ϵ -equilibrium if and only if the expected regrets are ϵ or smaller. Of course, some of the *realized* regrets may be much larger due to sample variability, but these realizations are unlikely. The same holds for the differences $\widehat{r}_{j,t}^i$, which are statistical estimates of the regrets. (We remark that these estimates, while good, may be slightly biased.) By contrast, if a given state is not an ϵ -Nash equilibrium, the regrets, and hence the differences $\widehat{r}_{j,t}^i$, will be large. Hence, if the τ_i 's are substantially smaller than ϵ , at least one of the players will be dissatisfied with fairly high probability. The dissatisfied player will then revise his strategy, and the revision can be any discrete probability distribution over the actions that is consistent with his hat size.

We would therefore like to argue that: (i) when the state is an ϵ -equilibrium it remains in place for a long time, and (ii) when the state is not an ϵ -equilibrium the process moves to an ϵ -equilibrium within a short period of time. To establish these points requires a fairly detailed argument.

To illustrate some of the issues that arise, suppose we could show that the process moves with positive probability from any given non- ϵ -equilibrium state a to some ϵ -equilibrium state e in one period. Suppose further that once in an ϵ -equilibrium state the process leaves with very small probability. Then the following lemma would establish our theorem.

Lemma 1 *Consider any stationary finite-state Markov chain with transition matrix P . Let π be any stationary distribution of P . Then, for any two states a and e :*

$$\pi_a \leq \frac{1 - P(e \rightarrow e)}{P(a \rightarrow e)} \quad (1)$$

The bound holds trivially if $P(a \rightarrow e) = 0$.

Proof: By standard properties of finite-state Markov chains, we know that the stationary distribution satisfies:

$$\begin{aligned}\pi_e &= \sum_x \pi_x P(x \rightarrow e) \\ &\geq \pi_a P(a \rightarrow e) + \pi_e P(e \rightarrow e).\end{aligned}$$

Hence,

$$\frac{\pi_a P(a \rightarrow e)}{1 - P(e \rightarrow e)} \leq \pi_e \leq 1,$$

which implies (1). □

Let \mathcal{E} be the set of all ϵ -equilibrium states. Suppose, as above, that $\max_{e \in \mathcal{E}} P(e \rightarrow e)$ can be made as small as we like, and that $\max_{e \in \mathcal{E}} \min_{a \notin \mathcal{E}} P(a \rightarrow e)$ can be bounded away from zero. It would then follow from Lemma 1 that $\sum_{a \notin \mathcal{E}} \pi_a$ can be made smaller than ϵ , from which it follows that $\sum_{e \in \mathcal{E}} \pi_e$ is larger than $1 - \epsilon$.

The difficulty is that neither of these premises may hold (though the lemma will still prove to be useful later on). First, one player may have regret while the other does not. As they revise their strategies, a cyclic pattern may emerge in which one revises while the second stays fixed, then the second revises while the first stays fixed, and so forth ad infinitum. Hence the process does not necessarily transit to an ϵ -equilibrium in one step. Assuming this hurdle can be surmounted, there is a second difficulty: there may exist ϵ -equilibria where the process does not stay for a long time. For example, if the state is just barely an ϵ -equilibrium, the process may move away again fairly quickly. In fact, even if the process lands on an equilibrium exactly, there is a problem if it is a mixed equilibrium. Namely, there is a possibility of “bad draws” in which the implementation of the strategy leads to realized differences $\hat{r}_{j,t}^i$ that are bigger than the required tolerance (due to sample variability). Hence there could be a non-negligible probability that, once near equilibrium, the players will move away again.

The essence of the proof is to show that, beginning in any non-equilibrium state, there exists some path leading to a *very sticky* ϵ -equilibrium state

such that the probability of following this path is much higher than the probability of leaving the target state once it is reached. The subtlety of the proof hinges on the fact that this path may be indirect, that is, the process may first move further away from equilibrium before moving towards it. In particular, we shall show that, if neither player has a weakly dominant strategy, then the process can transit from a non- ϵ -equilibrium state to a very sticky ϵ -equilibrium state in either one or two steps, possibly via an intermediate non-equilibrium state. (The weakly dominant case must be treated separately and is surprisingly non-trivial.) Once we establish this point, the following variant of lemma 1 will deliver the desired conclusion.

Lemma 2 *Consider any stationary finite-state Markov chain with transition matrix P . Let π be any stationary distribution of P . Then, for any three states a , b and e :*

$$\pi_a \leq \frac{1 - P(e \rightarrow e)^2}{P(a \rightarrow b \rightarrow e)} \quad (2)$$

The bound holds trivially if $P(a \rightarrow b \rightarrow e) = 0$.

Proof: Let $P^2(x \rightarrow y)$ denote the probability of transiting from x to y in exactly two time periods. It is clear that P^2 is a finite-state Markov chain; moreover, any stationary distribution π of P is also a stationary distribution of P^2 . Hence we can apply Lemma 1 to deduce that

$$\pi_a \leq \frac{1 - P^2(e \rightarrow e)}{P^2(a \rightarrow e)}.$$

Now (2) follows from the fact that $P(e \rightarrow e)^2 \leq P^2(e \rightarrow e)$ and $P(a \rightarrow b \rightarrow e) \leq P^2(a \rightarrow e)$. \square

3.2 Expected versus estimated regret

Next we turn our attention to the problem of estimating the escape probabilities from different kinds of states. This is complicated by the fact that the players cannot observe the current state; the only information they have

is their realized payoffs. Thus we need to distinguish between the current state—the pair of probability distributions generating the payoffs—and the realized payoffs that determine the players' estimated regrets.

Given a game G on $m_1 \times m_2$ action space \mathcal{A} , let $u_{j,k}^i$ be the payoff to i when the pair of actions (j, k) is played, $1 \leq j \leq m_1$, $1 \leq k \leq m_2$. To simplify later computations we shall assume, without loss of generality, that the von Neumann-Morgenstern payoffs have been chosen so that for all i, j, k , we have $0 \leq u_{j,k}^i \leq 1$. Define the *regret function* for player 1 by

$$R^1(p, q) \equiv \max_h \sum_{j,k} (u_{h,k}^1 - u_{j,k}^1) p_j q_k, \quad (3)$$

and for player 2 by

$$R^2(p, q) \equiv \max_h \sum_{j,k} (u_{j,h}^2 - u_{j,k}^2) p_j q_k. \quad (4)$$

The players' measured regrets are random variables that approximate $R^i(p_t, q_t)$ in a sense that we now make precise. Recalling the definitions of $\hat{\alpha}_{j,t}^i$ and $\hat{\alpha}_t^i$ from Step 3 of regret testing, let

$$\alpha_{j,t}^i \equiv E(\hat{\alpha}_{j,t}^i | (p_t, q_t)) \quad (5)$$

and

$$\alpha_t^i \equiv E(\hat{\alpha}_t^i | (p_t, q_t)) \quad (6)$$

For player 1 we have

$$\alpha_{j,t}^1 = \sum_k ((1 - \lambda_2)(q_t)_k + \lambda_2/m_2) u_{j,k}^1,$$

and

$$\alpha_t^1 = \sum_{j,k} (p_t)_j ((1 - \lambda_2)(q_t)_k + \lambda_2/m_2) u_{j,k}^1.$$

Similar expressions hold for $\alpha_{j,t}^2$ and α_t^2 . Define

$$r_t^i \equiv \max_j \alpha_{j,t}^i - \alpha_t^i. \quad (7)$$

Then $r_t^1 = R(p_t, (1 - \lambda_2)q_t + \lambda_2\vec{1}/m_2)$, and $r_t^2 = R((1 - \lambda_1)p_t + \lambda_1\vec{1}/m_1, q_t)$ where $\vec{1}$ is a vector of 1's.

Since $E(\hat{\alpha}_{j,t}^i | (p_t, q_t)) = \alpha_{j,t}^i$ and $E(\hat{\alpha}_t^i | (p_t, q_t)) = \alpha_t^i$ we can think of the difference:

$$\hat{r}_t^i = \max_j \hat{\alpha}_{j,t}^i - \hat{\alpha}_t^i$$

as being an estimator of r_t^i . (Note, however, that it is not unbiased.)

Define the *estimation error in state* (p_t, q_t) to be

$$|\hat{r}_t^i - r_t^i|. \quad (8)$$

Recalling that the payoffs were normalized to lie between zero and one, it follows from (3) – (6) that

$$|r_t^i - R(p_t, q_t)| \leq 2\lambda_i. \quad (9)$$

In particular, when the r_t^i 's are small, so are the values of the regret function $R(\cdot, \cdot)$.

Define $\tilde{p}_t = (1 - \lambda_1)p_t + \lambda_1\vec{1}/m_1$ and $\tilde{q}_t = (1 - \lambda_2)q_t + \lambda_2\vec{1}/m_2$. Then the actual play probabilities at time t are $(\tilde{p}_t, \tilde{q}_t)$. Further,

$$|r_t^i - R(\tilde{p}_t, \tilde{q}_t)| \leq 2(\lambda_1 \vee \lambda_2) < 2(\lambda_1 + \lambda_2). \quad (10)$$

This shows that actual play will be an ϵ -equilibrium whenever $r_t^1, r_t^2 \leq \epsilon - 2(\lambda_1 + \lambda_2)$.

Next we estimate the distribution of the realized regret estimates \hat{r}_t^i .

Lemma 3 *If $\lambda_i \leq 1/4$, then for all $\delta \leq \frac{1}{3m_i}$,*

$$P\left(|\hat{r}_t^i - r_t^i| > \delta\right) \leq 6m_i e^{-\frac{s\lambda_i\delta^2}{16m_i}}.$$

Proof: Fix a player i and let $m = m_i$, $\lambda = \lambda_i$. Let (p_t, q_t) be the state at date t . Let N_j be the number of times action j is played on day t while player i is on the telephone. The average payoff during these times, $\hat{\alpha}_j^i$, is

an average of N_j items, each of which is bounded between zero and one. By Azuma's inequality (1967),

$$P(|\hat{\alpha}_{j,t}^i - \alpha_{j,t}^i| > \delta \mid (p_t, q_t), N_j) \leq 2e^{-N_j\delta^2/2}. \quad (11)$$

Letting $N = \sum_j N_j$, $s - N$ is the number of times i was not on the phone, hence

$$P(|\hat{\alpha}_t^i - \alpha_t^i| > \delta \mid (p_t, q_t), N) \leq 2e^{-(s-N)\delta^2/2}. \quad (12)$$

Using Bonferroni's inequality (1936), it follows that

$$P(|\hat{r}_t^i - r_t^i| \geq 2\delta \mid (p_t, q_t), N_1, N_2, \dots, N_m) \leq 2 \sum_{i=1}^m e^{-N_j\delta^2/2} + 2e^{-(s-N)\delta^2/2}. \quad (13)$$

Define the events

$$\mathcal{A} \equiv |\hat{r}_t^i - r_t^i| > 2\delta$$

and

$$\mathcal{B} \equiv \cap_j \{|N_j - \lambda s/m| \leq \lambda s/2m\}.$$

From our assumption that $\lambda \leq 1/4$ we know that if \mathcal{B} occurs then $s - N \geq s\lambda$, hence

$$P(\mathcal{A}|\mathcal{B}) \leq 2me^{-s\lambda\delta^2/4m} + 2e^{-s\lambda\delta^2/2} = 2(m+1)e^{-s\lambda\delta^2/4m}. \quad (14)$$

Since in general $P(\mathcal{A}) \leq P(\mathcal{A}|\mathcal{B}) + P(\mathcal{B}^c)$, we only need to bound $P(\mathcal{B}^c)$ to complete the argument.

We know that N_j is binomially distributed $B(\lambda/m, s)$. It follows from Bennett's inequality (1962)⁴

$$P\left(|N_j - \frac{s\lambda}{m}| \geq \frac{s\lambda}{2m}\right) \leq 2e^{-s\lambda/20m}. \quad (15)$$

⁴Bennett's bound (1962) is usually stated for a bounded collection of n independent random variables U_1, \dots, U_n with $\sup |U_i| < M$, $EU_i = 0$, and $\sum_i EU_i^2 = 1$. Then for every $\tau > 0$,

$$P\left(\sum_i U_i \geq \tau\right) \leq \exp\left(\frac{\tau}{M} - \left(\frac{\tau}{M} + \frac{1}{M^2}\right) \log(1 + M\tau)\right).$$

We will consider the special case of n IID random variables X_1, \dots, X_n , that are bounded in absolute value by 1, with $\text{Var}(X_i) = \sigma^2$. Letting $\tau = (\gamma/\sigma)\sqrt{n}$ and $M = 1/\sigma\sqrt{n}$,

Hence $P(\mathcal{B}^c) \leq 2me^{-s\lambda/20m}$ by Bonferroni (1936). Thus we have

$$P(|\hat{r}_t^i - r_t^i| > 2\delta) \leq 2me^{-\frac{s\lambda}{20m}} + 2(m+1)e^{-\frac{s\lambda\delta^2}{4m}}. \quad (16)$$

Changing from δ to $\delta/2$ we obtain

$$P(|\hat{r}_t^i - r_t^i| > \delta) \leq 2me^{-\frac{s\lambda}{20m}} + 2(m+1)e^{-\frac{s\lambda\delta^2}{16m}}. \quad (17)$$

By assumption, $\delta^2 \leq 1/9m_i^2 \leq 1/9$. Hence $\delta^2/16 \leq 1/20$ and $e^{-s\lambda/20m} \leq e^{-s\lambda\delta^2/16m}$, so (17) implies

$$P(|\hat{r}_t^i - r_t^i| > \delta) \leq (4m+2)e^{-\frac{s\lambda\delta^2}{16m}} \leq 6me^{-\frac{s\lambda\delta^2}{16m}}. \quad (18)$$

□

3.3 Bounding the learning parameters

In this section we shall provide specific bounds on the parameters τ_i , λ_i , h_i , s such that theorem 1 holds for a given G and $\epsilon > 0$.

Let Δ_i be the simplex of all probability distributions on m_i actions, and let $\Delta_i(h_i)$ be the finite subset of those distributions that can be represented by integer multiples of $1/h_i$. Then $\Delta_i(h_i)$ approximates Δ_i in the sense that

$$\forall p \in \Delta_i, \exists p' \in \Delta_i(h_i), \|p' - p\| \leq \sqrt{m_i}/h_i. \quad (19)$$

Bennett's bound can be rewritten as:

$$P(\bar{X} - EX \geq \gamma) \leq \exp(n\gamma - n(\gamma + \sigma^2) \log(1 + \gamma/\sigma^2)).$$

If we take $\gamma = \sigma^2/2$ and use the fact that $\log 3/2 \geq .4$,

$$P(\bar{X} - EX \geq \sigma^2/2) \leq \exp(-n\sigma^2/10).$$

When the X_i 's are binomial (p, n) with $0 < p < .5$, this implies

$$P(|\bar{X} - p| \geq p/2) \leq 2e^{-np/20}.$$

In the following we shall make frequent use of the concept of δ -best reply, by which we mean a strategy that cannot be improved upon (in expectation) by more than $\delta \geq 0$, given the strategy of one's opponent. A strategy is δ -dominant for player i if it is a δ -best reply against any strategy of the opponent. Define $d(G)$ to be the smallest $\delta \geq 0$ such that at least one of the players has a δ -dominant strategy. In particular, $d(G) = 0$ if and only if someone has a weakly dominant strategy.

Let (δ_1, δ_2) be a pair of nonnegative real numbers. We shall say that a state $z = (p, q)$ is a (δ_1, δ_2) -equilibrium if player i cannot improve his expected payoff by more than δ_i given the strategy of his opponent ($i = 1, 2$). (Using our notation from section 3.2, a state $z = (p, q)$ is a (δ_1, δ_2) -equilibrium if $R^i(p, q) \leq \delta_i$ for $i = (1, 2)$.) When $\delta_1 = \delta_2 = \delta$ the terms δ -equilibrium and (δ_1, δ_2) -equilibrium will be used interchangeably.

Lemma 4 *Suppose that $\tau_i \leq 1$ and $\lambda_i \leq \tau_i/8$ for both players. Let $z_t = (p_t, q_t)$ be the state at time t .*

1. *If state $z_t = (p_t, q_t)$ is a $(\tau_1/2, \tau_2/2)$ -equilibrium, a revision occurs at the end of period t with probability at most ae^{-bs} for all s , where $a = 12 \max_i m_i$ and $b = \min_i \left\{ \frac{\lambda_i \tau_i^2}{256 m_i} \right\}$.*
2. *If z_t is not a $(2\tau_1, 2\tau_2)$ -equilibrium, a revision occurs at the end of period t with probability at least .5 provided that $s \geq 800 \max_i \left\{ \frac{m_i^2}{\lambda_i \tau_i^2} \right\}$.*

Proof: In a $(\tau_1/2, \tau_2/2)$ -equilibrium, $r_t^i \leq \tau_i/2 + 2\lambda_i \leq 3\tau_i/4$. Hence, in order for a rejection to occur, we must have $|\hat{r}_t^i - r_t^i| > \tau_i/4$. By lemma 3 we know that for player i the probability of this occurring is less than $6m_i e^{\frac{-s\lambda_i\tau_i^2}{256m_i}}$. Thus the probability that one or both players reject is less than

$$\sum_{i=1}^2 6m_i e^{\frac{-s\lambda_i\tau_i^2}{256m_i}},$$

which implies part one.

If we are not in a $(2\tau_1, 2\tau_2)$ -equilibrium, $r_t^i \geq 2\tau_i - 2\lambda_i \geq 7\tau_i/4$ for at least one of the players. Call this player i' . For that player, a rejection will occur unless $\hat{r}_t^{i'} \leq \tau_{i'}$, which implies $|\hat{r}_t^{i'} - r_t^{i'}| > 3\tau_{i'}/4$. By lemma 3 we know that the probability of this is less than $6m_{i'}e^{-\frac{s\lambda_{i'}\tau_{i'}^2}{64m_{i'}}$. We wish to compute a value of s such that

$$6m_{i'}e^{-\frac{s\lambda_{i'}\tau_{i'}^2}{64m_{i'}}} < .5.$$

This holds if

$$s > \frac{64m_{i'}}{\lambda_{i'}\tau_{i'}^2} \log_e(12m_{i'}).$$

Noting that $\log x \leq x$ for $x \geq 1$, this simplifies to

$$s > \frac{768m_{i'}^2}{\lambda_{i'}\tau_{i'}^2},$$

which implies part two. □

A useful consequence of lemma 4 is the following

Lemma 5 *Suppose $\lambda_i \leq \tau_i/8 \leq 1/8$ and $s \geq 2000 \max_i\{((m_1 + m_2)^2)/(\epsilon\lambda_i\tau_i^2)\}$, for $i = 1, 2$. Then*

1. *If the state $z_t = (p_t, q_t)$ is a $(\tau_1/2, \tau_2/2)$ -equilibrium, a revision occurs at the end of period t with probability at most ϵ .*
2. *If the state $z_t = (p_t, q_t)$ is not a $(2\tau_1, 2\tau_2)$ -equilibrium, a revision occurs at the end of period t with probability at least .5.*

Proof: The first part follows if $ae^{-bs} \leq \epsilon$ which holds if $s \geq \log\{a/\epsilon\}/b$. Using the inequality $\log(x) \leq x$, we see that it is sufficient for $s \geq a/b\epsilon = 12 \max_i m_i / \min_i\{\lambda_i\tau_i^2/(128\epsilon m_i)\}$. Hence it is sufficient that $s \geq 2000 \max_i\{((m_1 + m_2)^2)/(\epsilon\lambda_i\tau_i^2)\}$.

The second part follows because the restriction given for s is tighter than in part two of the previous lemma. □

Fix an $m_1 \times m_2$ action space A . Let G be a game on A and let $\epsilon > 0$. Choose τ_i, λ_i, h_i and s such that

$$\tau_i \leq \epsilon^2/48 \tag{20}$$

$$\lambda_i \leq \tau_i/16 \tag{21}$$

$$h_i \geq 8\sqrt{m_i}/(\tau_1 \wedge \tau_2) \tag{22}$$

$$s \geq 8000(H_1 + H_2)^3(m_1 + m_2)^2 \max_i \left\{ \frac{1}{\epsilon \lambda_i \tau_i^2} \right\} \tag{23}$$

(Recall that H_i is the number of i 's hats and is determined by h_i .) In particular, these assumptions imply that i 's grid error is at most $\tau_i/8$ (see inequality (19) and (22)).

Theorem 1 (restatement) *Let G be a two-person game on A played by regret testers, and let $\epsilon > 0$. Whenever the parameters satisfy (20)–(23) and $d(G) \notin (0, \sqrt{48(\tau_1 \vee \tau_2)})$, the players' joint behaviors at time t constitute an ϵ -equilibrium of G with probability at least $1 - \epsilon$ as $t \rightarrow \infty$.*

This result implies theorem 1, because *given* G and $\epsilon > 0$, we can always choose τ_i small enough so that the conditions of Theorem 1 are satisfied. (Either $d(G) = 0$, in which case (20)–(23) suffice, or $d(G) > 0$, in which case we can choose τ small enough that (21)–(23) hold and also $d(G) \geq \sqrt{48(\tau_1 \vee \tau_2)}$.) We have stated theorem 1 in this way in order to call attention to the point that once the parameters are chosen to satisfy (20)–(23) for a given $\epsilon > 0$, the conclusion holds for all games G on A except possibly for games in which no one has a weakly dominant strategy but someone almost does, that is, $d(G) \in (0, \sqrt{48(\tau_1 \vee \tau_2)})$ for some i . (Under various natural assumptions about the distribution of payoffs in G , this excluded set will be small when the τ_i are small and hence when ϵ is small.) The trouble with such a game is that the payoff differences may not be large enough (relative to the tolerances) for the players to stick with an ϵ -equilibrium for very long once they find it. If however $d(G) = 0$, then someone has a weakly dominant strategy that he can lock into for long periods, which allows the other player time to adjust to ϵ -equilibrium behavior too. In the following section we shall

show that we can circumvent this difficulty by annealing the parameters at a suitable rate. In this case we obtain a variant of regret testing that guarantees convergence in probability to the set of Nash equilibria for all games G on a given action space A .

Proof of theorem 1. Let \mathcal{E} be the set of states $z = (p, q)$ on the grid such that the players' behavior in state z constitutes an ϵ -equilibrium of G . Recall that 1's actual behavior is given by the probability distribution $\tilde{p} \equiv (1 - \lambda_1)p + \lambda_1 \mathbf{1}/m_1$, and player 2's actual behavior is given by $\tilde{q} \equiv (1 - \lambda_2)q + \lambda_2 \mathbf{1}/m_2$. Thus it could happen that the strategies (p, q) constitute an ϵ -equilibrium of G but the behaviors do not. From equation (10) we see that when the λ_i are small, however, the actual behaviors in state z are very close to (p, q) ; indeed we can say the following:

If $\lambda_1, \lambda_2 \leq \epsilon/4$ and $z = (p, q)$ is an $\epsilon/2$ -equilibrium of G , then the actual behaviors in state z constitute an ϵ -equilibrium of G .

Define \mathcal{E}^* to be the set of all states z that are $\epsilon/2$ -equilibria of G . To prove Theorem 1, it suffices by (10) to show that the long-run probability of \mathcal{E}^* is at least $1 - \epsilon$. In other words, it suffices to show that for any stationary distribution π of the process,

$$\sum_{a \notin \mathcal{E}^*} \pi_a \leq \epsilon.$$

We need to consider two separate cases.

Case 1. $d(G) > 0$: neither player has a weakly dominant strategy.

In this case we have the additional hypothesis that $d(G) \geq \sqrt{48(\tau_1 \vee \tau_2)}$. Choose a Nash equilibrium $e \in \Delta_1 \times \Delta_2$. Since the grid errors are at most $\tau_i/8$ and the payoffs are bounded between 0 and 1, there exists a state $e^* = (p^*, q^*)$ on the grid that is a $(\tau_1/8, \tau_2/8)$ -equilibrium. We shall fix this state for the remainder of the proof of Case 1.

Letting $\epsilon' = \epsilon/4(H_1 + H_2)^3$ and applying Lemma 5 (with ϵ' instead of ϵ) we obtain

$$P(e^* \rightarrow e^*) \geq 1 - \epsilon' \geq 1 - \epsilon/4(H_1 + H_2)^3 \text{ for all } s. \quad (24)$$

The next step is to show that for all $a \notin \mathcal{E}^*$, the process moves from a to e^* (in one or two periods) with a sufficiently high probability.

Case 1a. $a \notin \mathcal{E}^*$ and each player can increase his payoff by more than $\epsilon/2$.

Suppose that $z_t = a = (p, q)$. Since each player i can increase his payoff by more than $\epsilon/2$, he can certainly increase it by more than $2\tau_i$. (Recall our assumption that $\tau_i \leq \epsilon^2/48$.) It follows from Lemma 5, part two, that the probability is at least $1/4$ that both players revise at the end of day t . Conditional on both rejecting, the probability is at least $1/H_1H_2$ that player 1 chooses p^* and player 2 chooses q^* in period $t + 1$. Hence

$$P(a \rightarrow e^*) \geq \frac{1}{4H_1H_2}.$$

Case 1b. $a \notin \mathcal{E}^*$ and only one of the players can improve his payoff by more than $\epsilon/2$.

This case requires a two-step argument: we shall show that the process can transit from state a to some intermediate state b with the property that in state b each player i can increase his payoff by more than $2\tau_i$. As in the proof of Case 1a, we then conclude that $P(b \rightarrow e^*) \geq \frac{1}{4H_1H_2}$.

Assume without loss of generality that in state $a = (p, q)$, player 1 can increase his payoff by more than $\epsilon/2$, that is, there exists $p' \in \Delta_1$ such that

$$u^1(p', q) - u^1(p, q) \geq \epsilon/2. \quad (25)$$

Let $\delta = d(G)$: then neither player has a δ' -dominant strategy for any $\delta' < \delta$. In particular, q is not $\delta/2$ -dominant for player 2. Hence there exists $p^* \in \Delta_1$ and $q' \in \Delta_2$ such that

$$u^2(p^*, q') - u^2(p^*, q) > \delta/2.$$

Consider the strategy

$$p'' = (\delta/4)p + (1 - \delta/4)p^*. \quad (26)$$

Since p' is a best response to q , we know that $u^1(p', q) - u^1(p^*, q) \geq 0$. It follows from (25) and (26) that

$$\begin{aligned} u^1(p', q) - u^1(p'', q) &= (\delta/4)[u^1(p', q') - u^1(p, q)] \\ &\quad + (1 - \delta/4)[u^1(p', q') - u^1(p^*, q)] \\ &\geq (\delta/4)[u^1(p', q') - u^1(p, q)] \\ &> \delta\epsilon/8. \end{aligned} \quad (27)$$

By assumption, $\delta \geq \sqrt{48(\tau_1 \vee \tau_2)}$ and $\epsilon \geq \sqrt{48(\tau_1 \vee \tau_2)}$, hence $\delta\epsilon/8 > 6\tau_1$. From this and (27) it follows that at (p'', q) player 1 can increase his payoff by more than $6\tau_1$.

Similar calculations show that at (p'', q) player 2 can increase his payoff by at least $\delta/2 - \delta/4 = \delta/4$, which by assumption is greater than $\sqrt{48(\tau_1 \vee \tau_2)}/4 \geq 3\tau_2$.

Although q is on player 2's grid, the definition of p'' in (26) does not guarantee that it is on player 1's grid. We know from (22), however, that there exists a grid point (p''', q) such that $|p''' - p''| \leq \sqrt{m_1}/h_1 \leq \tau_1 \wedge \tau_2$. Since the payoffs lie between zero and one, it follows from the preceding that $b = (p''', q)$ is on the grid and is not a $(5\tau_1, 2\tau_2)$ -equilibrium.

As in the proof of Case 1a, it follows that $P(b \rightarrow e^*) \geq 1/4H_1H_2$. Further, the process moves from state a to state b with probability at least $\frac{1}{H_1 \vee H_2}$, because only one player needs to revise to p''' . Thus

$$P(a \rightarrow b \rightarrow e^*) \geq \frac{1}{4(H_1 + H_2)^3}.$$

By (24) we know that

$$1 - P(e^* \rightarrow e^*) \leq \frac{\epsilon}{4(H_1 + H_2)^3}.$$

Applying Lemma 2, we conclude that the probability of being in \mathcal{E}^* , and hence in an ϵ -equilibrium state, is at least $1 - \epsilon$.

Case 2: $d_G = 0$: Weakly dominant strategies exist.

We can assume without loss of generality that player 1 has a weakly dominant strategy, in which case player 1 has a pure strategy that is weakly dominant.

The idea of the proof is to show that player 1 does not change strategy for very long stretches of time, and that during any such stretch player 2 does not change strategy very often either. Thus both players change strategies infrequently. Assuming that $\tau_i < \epsilon/2$, this implies that they must be in an ϵ -equilibrium a large fraction of the time (because in any period where they are not in an ϵ -equilibrium, one or both will reject and switch strategies with high probability, by Lemma 5.) An interesting feature of the argument is that it does not imply that player 1 is using his weakly dominant strategy for long stretches of time. (For example, he might be playing his part of a strict pure strategy equilibrium that does not involve his weakly dominant strategy.) It is the *presence* of a weakly dominant strategy that allows us to say that strategy changes are infrequent. We then use the ergodicity of the process to deduce that it is in an ϵ -equilibrium with high probability.

Define

$$T \equiv 2(H_1 + 1)(H_2 + 1)/\epsilon. \quad (28)$$

Assume for the sake of concreteness that action 1 is weakly dominant for player 1. There is some hat that consists entirely of papers labelled with this action. Any time player 1 picks a new hat off the shelf, the probability is $1/H_1$ that he grabs this hat. Once he starts using it, lemma 3 (and weak dominance) imply he will continue to use it for $e^{s\lambda_1\tau_1^2/16m_1}$ days in expectation. By assumption, equation (23) shows that

$$s\lambda_1\tau_1^2/16m_1 \geq 8000(H_1 + H_2)^3m_1/16\epsilon \geq T.$$

Hence player 1 will continue playing his weakly dominant strategy for at least T days in expectation.

Now let us estimate the frequency with which player 2 revises. Suppose that L days elapse between two successive revisions by player 1. If player 2

revises during this interval, there is a $1/H_2$ chance that she will pick a hat representing a pure strategy best reply to player 1's fixed behavior. Once this happens, the expected waiting time (in days) until she revises again is at least $\exp(s\lambda_2\tau_2^2/16m_2)$, which by construction is greater than T . If $L > T$ (i.e., L is long), the expected number of revisions by player 2 during this interval is at most H_2L/T . If $L \leq T$ (i.e., L is short), the expected number of revisions by player 2 is at most H_2 . Given any time $t > T$, it follows that the expected number of revisions by player 2 over all short intervals up to t is at most H_1H_2t/T , and the expected number of revisions over all long intervals is at most H_2t/T . Thus the expected number of times up to t when player 1 or player 2 revises is at most $H_1H_2t/T + H_2t/T + H_1t/T$ which is less than $(H_1 + 1)(H_2 + 1)t/T$.

If at some time t the players are not in an ϵ -equilibrium, then at least one of them (say i) has an expected regret that exceeds $2\tau_i$. Thus, by lemma 5, the probability is at least one-half that a revision occurs at t . It follows that the expected number of times through $t > T$ that the process is not in an ϵ -equilibrium is at most twice the expected number of revisions, that is, at most $2(H_1 + 1)(H_2 + 1)t/T$, which is at most ϵt . By the ergodic theorem for acyclic finite Markov chains, the long run frequency of being in an ϵ -equilibrium is also the limiting probability of being in an ϵ -equilibrium at any given time t as t goes to infinity, which by the preceding is at least $1 - \epsilon$.

□

4 Convergence in probability

Theorem 1 shows that regret-testing induces ϵ -equilibrium behavior with probability at least $1 - \epsilon$ provided that $d(G)$ is not in the excluded range $(0, \epsilon)$. If we think of G as a vector of payoffs in $\mathfrak{R}^{2m_1m_2}$ (one payoff for each player under each pair of actions), the excluded set will be small relative to Lebesgue measure when ϵ is small. By letting ϵ approach zero at a suitable rate, and tightening the parameters in accordance with the bounds (20)-(23),

the process will eventually capture all games G in the “net,” that is there will be no excluded cases. Moreover the process will converge in probability to the set of Nash equilibria, not just the ϵ -equilibria.

Fix finite action spaces A_1 and A_2 for players 1 and 2 respectively, where $|A_1| = m_1$ and $|A_2| = m_2$. Next, let $(\tau_1(\epsilon), \tau_2(\epsilon), \lambda_1(\epsilon), \lambda_2(\epsilon), h_1(\epsilon), h_2(\epsilon), s(\epsilon))$ be a system of parameters satisfying the bounds (20)—(23) for every $\epsilon > 0$.⁵ Let $P_G(\epsilon)$ denote the finite-state Markov process determined by G and the parameters $(\tau_1(\epsilon), \dots, s(\epsilon))$. Let $\mathcal{E}_G(\epsilon)$ be the finite subset of states consisting of ϵ -equilibria of G .

Definition 1 *Let P be an acyclic, finite Markov process and \mathcal{A} a subset of states. For each $\epsilon > 0$, let $T(P, \mathcal{A}, \epsilon)$ be the first time (if any) such that, for all $t > T(P, \mathcal{A}, \epsilon)$ and all initial states the probability is at least $1 - \epsilon$ that the process is in \mathcal{A} at time t .*

Since $P_G(\epsilon)$ is acyclic, it follows from theorem 1 that $T(P_G(\epsilon), \mathcal{E}(G), \epsilon)$ is finite. In particular, for all $t \geq T(P_G(\epsilon), \mathcal{E}_G(\epsilon), \epsilon)$, the probability is at least $1 - \epsilon$ that the process is in an ϵ -equilibrium of G at time t .

The time $T(P_G(\epsilon), \mathcal{E}_G(\epsilon), \epsilon)$ may depend on the payoffs, because these affect the details of the transition probabilities and the states that are ϵ -equilibria of G . We claim, however, that for every $\epsilon > 0$ there is a time $T(\epsilon)$ such that $T(\epsilon) \geq T(P_G(\epsilon), \mathcal{E}_G(\epsilon), \epsilon)$ for all G such that $\delta_G \notin (0, \epsilon)$.

To see why this is so, consider the realization of plays on any given day. A realization is a sequence of $s(\epsilon)$ actions for each player and a sequence of $s(\epsilon)$

⁵For example, we could choose

$$\begin{aligned} \tau_i(\epsilon) &= \epsilon^2/50, \\ \lambda_i(\epsilon) &= \frac{\epsilon^2}{800}, \\ h_i(\epsilon) &= \left\lceil \frac{400\sqrt{m_1 + m_2}}{\epsilon^2} \right\rceil, \\ s(\epsilon) &= \lceil (30/\epsilon)^7 (H_1 + H_2)^3 (m_1 + m_2)^2 \rceil \end{aligned}$$

where $\lceil x \rceil$ refers to the least integer greater than or equal to x .

binary outcomes (say 0 or 1) that indicate whether a given action was taken while on the phone or not. Hence there are $(4m_1m_2)^{s(\epsilon)}$ possible realizations. We may partition them into four disjoint classes: sequences that are rejected by both players, sequences that are rejected by player 1 but not player 2, sequences that are rejected by player 2 but not player 1, and sequences that are accepted by both. (Notice that this partition does not depend on the day t or on the strategies (p_t, q_t) in force during that day, but it does depend on the game G .) However, a player's response given a rejection does not depend on the sequence at all: it leads the player to choose a new strategy with uniform probability over all distributions consistent with his hat size.

The number of length- $s(\epsilon)$ realizations is finite, and there is a finite number of ways of partitioning them into four classes. Further, the probability that each sequence will be realized on a given day t is determined by the state (p_t, q_t) , and there are only a finite number of possible states. Hence, over all G , there can be only a finite number of Markov transition matrices $P_G(\epsilon)$. Further, there are only a finite number of different subsets of states that can be used to define $\mathcal{E}_G(\epsilon)$. Let us enumerate all of these possible pairs as follows $(P_1, \mathcal{E}_1), \dots, (P_k, \mathcal{E}_k)$. Now define $T(\epsilon) = \max_{1 \leq j \leq k} T(P_j, \mathcal{E}_j, \epsilon)$. Then $T(\epsilon)$ has the property that for all G such that $d(G) \notin (0, \epsilon)$, and for all $t \geq T(\epsilon)$, the process is at an ϵ -equilibrium at time t with probability at least $1 - \epsilon$.

Definition 2 (annealed regret testing) *Consider a positive sequence $\epsilon_1 > \epsilon_2 > \epsilon_3 > \dots$ decreasing to zero. The annealed regret testing procedure at stage k is the regret testing procedure with parameters $(\tau_1(\epsilon_k), \tau_2(\epsilon_k), \lambda_1(\epsilon_k), \lambda_2(\epsilon_k), h_1(\epsilon_k), h_2(\epsilon_k), s(\epsilon_k))$. Each day that the process is in stage k , the probability of moving to stage $k + 1$ on the following day is*

$$p_k \equiv \frac{\epsilon_{k+1}^2}{2k^2 T(\epsilon_{k+1})} \quad (29)$$

Theorem 1 *Fix an $m_1 \times m_2$ action space \mathcal{A} . The annealed regret testing process defined above has the property that for every game G on \mathcal{A} the process converges in probability to the set of Nash equilibria of G .*

Proof: We will show that, for all $\epsilon > 0$, the probability that at time t the process is in an ϵ -equilibrium converges to one as t goes to infinity. This implies that the process converges in probability to the set of Nash equilibria of G .

The definition of p_k means that we can recursively define a sequence of random variables N_t as follows:

$$\begin{aligned} N_1 &= 1 \\ N_{t+1} &= \begin{cases} N_t & \text{with probability } 1 - p_{N_t} \\ N_t + 1 & \text{with probability } p_{N_t} \end{cases} \end{aligned}$$

where the procedure uses parameters $(\tau_1(\epsilon_{N_t}), \tau_2(\epsilon_{N_t}), \lambda_1(\epsilon_{N_t}), \lambda_2(\epsilon_{N_t}), h_1(\epsilon_{N_t}), h_2(\epsilon_{N_t}), s(\epsilon_{N_t}))$ at time t .

Define $T_k \equiv \inf_t \{t : N_t \geq k\}$. In other words, T_k is the first time that the system shifts to the k th set of parameters. Now define $W_t \equiv t - T_{N_t}$. W_t is the length of time since the parameters were last changed. For any game G on \mathcal{A} , if $d(G) > 0$ then $d(G) \geq \epsilon_k$ for some k . The least such k is the *critical index* for G , denoted by k_G . In case $d(G) = 0$, we will take $k_G = 1$. Define $k_G^* = k_G \vee \min_k \{k \mid \epsilon_k < \epsilon/2\}$.

Consider two different cases at time t : $W_t \geq T(\epsilon_{N_t})$ and $W_t < T(\epsilon_{N_t})$. In the first case, by the Markov restart property and the definition of $T(\epsilon_{N_t})$ the process is in an ϵ_{N_t} -equilibrium with conditional probability greater than $1 - \epsilon_{N_t}$ where we have conditioned on N_t . If $N_t \geq k_G^*$ then $\epsilon_{N_t} \leq \epsilon/2$. When t is large enough, N_t is arbitrarily large, and hence the probability that $N_t > k_G^*$ can be made close to 1.

It remains to consider the second case, namely, $W_t < T(\epsilon_{N_t})$. We shall show that $P(W_t < T(\epsilon_{N_t}))$ converges to zero as $t \rightarrow \infty$, which will complete the proof.

We will say that episode k is “short” if it lasts at most $T(\epsilon_{k+1})/\epsilon_{k+1}^2$ periods. The probability of a short episode is less than $1/k^2$, which is summable. Hence if we pick $k_G^{**} \geq k_G^* \vee 4/\epsilon$, the chance of a short episode after stage k_G^{**}

is less than $\epsilon/2$. Define the set

$$A_t \equiv \{N_{t-T(\epsilon_{N_t})/\epsilon^2} \geq N_t - 1\}.$$

The event A_t occur if $N_t - 1$ is not a short episode. Pick t'' large enough so that the probability that $N_{t''} \geq k_G^{**} + 1$ is larger than $1 - \epsilon/4$. Then for any $t > t''$ we have $P(A_t^c) \leq \epsilon/2$.

We want to show that, for all $t > t''$,

$$P(W_t \geq T(\epsilon_{N_t})) \geq 1 - \epsilon.$$

For all $t > t''$,

$$\begin{aligned} P(W_t < T(\epsilon_{N_t})) &\leq P(W_t < T(\epsilon_{N_t})|A_t) + P(A_t^c) \\ &\leq P(W_t < T(\epsilon_{N_t})|A_t) + \epsilon/4 \end{aligned}$$

Thus we only need to bound $P(W_t < T(\epsilon_{N_t})|N_t, A_t)$ by $\epsilon/2$ and the proof will be complete. We have

$$\begin{aligned} P(W_t < T(\epsilon_{N_t})|A_t) &= \sum_n P(W_t < T(\epsilon_{N_t})|N_t = n, A_t)P(N_t = n|A_t) \\ &\leq \max_n P(W_t < T(\epsilon_{N_t})|N_t = n, A_t) \end{aligned}$$

Define $A_t^0 = \{N_{t-T(\epsilon_{N_t})/\epsilon^2} = N_t\}$, and $A_t^1 = \{N_{t-T(\epsilon_{N_t})/\epsilon^2} = N_t - 1\}$. Then $A_t = A_t^0 \cup A_t^1$ and

$$\begin{aligned} P(W_t < T(\epsilon_{N_t})|N_t = n, A_t) &= P(W_t < T(\epsilon_{N_t})|N_t = n, A_t^0)P(A_t^0|A_t) + \\ &\quad P(W_t < T(\epsilon_{N_t})|N_t = n, A_t^1)P(A_t^1|A_t). \end{aligned}$$

It is clear that $W_t = T(\epsilon_{N_t})/\epsilon^2 > T(\epsilon_{N_t})$ when A_t^0 occurs. Hence,

$$P(W_t < T(\epsilon_{N_t})|N_t = n, A_t) \leq P(W_t < T(\epsilon_{N_t})|N_t = n, A_t^1)P(A_t^1|A_t).$$

The probability that $W_t = w$ is proportional to $(p_k^w p_{k+1})^{T(\epsilon_{N_t})/\epsilon^2 - w}$ which is decreasing in t . Hence for all sufficiently large t , $P(W_t < T(\epsilon_{N_t})|N_t = n, A_t^1)P(A_t^1|A_t)$ is less than $\epsilon/2$.

□

References

1. Azuma, K., 1967. "Weighted sums of certain dependent random variables," *Tohoku Math. J.*, **19**, pp. 357 - 367.
2. Bendor, Jonathan, Dilip Mookherjee, and Debraj Ray, 2001. "Aspiration-based reinforcement learning in repeated interaction games: an overview," *International Journal of Game Theory*, **3**, 159-174.
3. Bennett, G. (1962), "Probability inequalities for the sum of independent random variables," *JASA*, **57**, 33-45.
4. Bonferroni, C.E. (1936). "Teoria statistica delle classi e calcolo delle probabilita'" *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3-62.
5. Börgers, Tilman, and Rajiv Sarin, 2000. "Näive reinforcement learning with endogenous aspirations," *International Economic Review*, **41**, 921-950.
6. Bush, R. R. and F. Mosteller, 1955. *Stochastic Models for Learning*. New York: John Wiley.
7. Cahn, A. 2004. "General procedures leading to correlated equilibria," *International Journal of Game Theory*.
8. Erev, Ido, and Alvin E. Roth, 1998. "Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, **88**, 848-881.
9. Foster, Dean P. and Rakesh Vohra, 1993. "A randomization rule for selecting forecasts," *Operations Research*, **41**, 704-709.
10. Foster, Dean P., and Rakesh Vohra, 1999. "Regret in the on-line decision problem," *Games and Economic Behavior*, **29**, 7-35.

11. Foster, Dean P., and H. Peyton Young, 2001. "On the impossibility of predicting the behavior of rational agents," *Proceedings of the National Academy of Sciences of the USA*, **98**, no.222, 12848-12853.
12. Foster, Dean P., and H. Peyton Young, 2003. "Learning, hypothesis testing, and Nash equilibrium," *Games and Economic Behavior*, **45**, 73-96.
13. Fudenberg, Drew, and David Levine, 1995. "Consistency and cautious fictitious play," *Journal of Economic Dynamics and Control*, **19**, 1065-90.
14. Fudenberg, Drew, and David Levine, 1998. *The Theory of Learning in Games*. Cambridge MA: MIT Press.
15. Germano, Fabrizio, and Gabo Lugosi, 2004. "Global convergence of Foster and Young's regret testing," Working paper, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona.
16. Hart, Sergiu, and Andreu Mas-Colell, 2000. "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, **68**, 1127-1150.
17. Hart, Sergiu, and Andreu Mas-Colell, 2001. "A general class of adaptive strategies," *Journal of Economic Theory*, **98**, 26-54.
18. Hart, Sergiu, and Andreu Mas-Colell, 2003. "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, **93**, 1830-1836.
19. Hart, Sergiu, and Andreu Mas-Colell, 2004. "Stochastic uncoupled dynamics and Nash equilibrium." Technical Report, Hebrew University of Jerusalem.
20. Jordan, James S., 1991. "Bayesian learning in normal form games," *Games and Economic Behavior*, **5**, 368-386.

21. Jordan, James S., 1993. "Three problems in learning mixed-strategy equilibria," *Games and Economic Behavior*, 5, 368-386.
22. Kalai, Ehud, and Ehud Lehrer, 1993. "Rational learning leads to Nash equilibrium," *Econometrica*, **61**, 1019-1045.
23. Karandikar, Rajeeva, Dilip Mookherjee, Debraj Ray, and Fernando Vega-Redondo, 1998. "Evolving aspirations and cooperation," *Journal of Economic Theory*, **80**, 292-331.
24. Nachbar, John H., 1997. "Prediction, optimization, and learning in games," *Econometrica*, **65**, 275-309.
25. Nachbar, John H., 2003, "Beliefs in repeated games," Working paper, Department of Economics, Washington University, St. Louis.
26. H. Peyton Young, 2004. *Strategic Learning and Its Limits*. Arne Ryde Memorial Lectures. Oxford, UK: Oxford University Press.