Theoretical Economics 1 (2006), 341–367 1555-7561/20060341

Regret testing: learning to play Nash equilibrium without knowing you have an opponent

DEAN P. FOSTER
Department of Statistics, Wharton School, University of Pennsylvania

H. PEYTON YOUNG
Department of Economics, Johns Hopkins University, and Department of Economics, University of Oxford

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

KEYWORDS. Learning, Nash equilibrium, regret, bounded rationality.
JEL CLASSIFICATION. C72, D83.

1. LEARNING EQUILIBRIUM

Although Nash equilibrium is the central solution concept in game theory, it has proved difficult to find adaptive learning rules that invariably lead to Nash equilibrium from out-of-equilibrium conditions. Of course, there exist particular rules, such as fictitious play, that work for particular classes of games, such as zero-sum games and potential games. And there exist sophisticated Bayesian updating procedures that lead to Nash equilibrium in any game provided that players' strategies and beliefs are sufficiently aligned at the outset.[1] The issue we consider here is whether there exist simple adaptive procedures that solve the "learning to play Nash" problem for general games without making large demands on the players' computational capacities and without imposing special initial conditions.

Dean P. Foster: dean@foster.net
H. Peyton Young: pyoung@jhu.edu

The authors thank Andrew Felton, Sham Kakade, Ben Klemens, Thomas Norman, Kislaya Prasad, and several anonymous referees for constructive comments. An earlier version of this paper appeared as Sante Fe Institute Working Paper 04-12-034 with the title "Regret Testing: A Simple Payoff-Based Procedure for Learning Nash Equilibrium."

[1] In a Bayesian framework, convergence to Nash equilibrium occurs with probability one provided that players' repeated-game strategies are optimal given their beliefs, and their beliefs put positive probability on all events that have positive probability under their strategies (Kalai and Lehrer 1993). Unfortunately it is difficult to satisfy the latter absolute continuity condition when players do not know their opponents' payoff functions (Jordan 1991, 1993, Foster and Young 2001, Nachbar 1997, 2005).

Copyright © 2006 Dean P. Foster and H. Peyton Young. Licensed under the Creative Commons Attribution-NonCommercial License 2.5. Available at http://econtheory.org.

To date, the main results on this problem have been negative. Consider, for example, the following criteria: i) a player's response rule may depend on the history of the game, but it should not depend on ex ante knowledge of the opponent's payoff function; ii) the rule should not depend on state variables of arbitrarily high dimensionality; iii) when all players use the rule, their period-by-period behaviors should converge (or at least come close) to Nash equilibrium behavior of the stage game or the repeated game.

A rule with the first property is said to be *uncoupled*. This is a reasonable requirement when the payoff structure of the game is not known precisely, which is often the case in practice. (Moreover, if coupled learning rules were allowed, one could simply "tailor" the learning rule to each payoff situation, and the program would amount to little more than a theory of equilibrium selection.) The second property expresses the idea that a rule should be simple to implement. One formulation of "simplicity"—admittedly rather restrictive—is that a player's behavioral response should depend only on histories of bounded length, or alternatively on a summary statistic of the whole history, such as the realized empirical frequency distribution (as in fictitious play). Such a rule is said to be *stationary* with respect to the state variable in question. The third property says that *period-by-period behaviors* should come close to Nash equilibrium; it is not enough that the cumulative empirical frequency of play come close. (The latter is the sense in which fictitious play converges in zero-sum games for example.)

A recent paper of Hart and Mas-Colell (2005) establishes the following impossibility theorem: when the relevant states are taken to be histories of bounded length, and convergence is defined as almost sure convergence of the period-by-period behavioral probabilities to an ε-equilibrium of the stage game, then for all sufficiently small $\varepsilon > 0$ there is no rule satisfying the above three properties on the set of finite two-person games.[2]

This impossibility result hinges crucially on a particular choice of state variable and a demanding notion of convergence. In an earlier paper, for example, we demonstrated a class of statistical learning procedures that are simple, uncoupled, and cause players' behaviors to converge in probability to the set of Nash equilibria in any finite game (Foster and Young 2003). In the simplest version of this approach, each player's state variable has three components: i) the empirical frequency distribution of the opponent's play over the past s periods, where s is finite; ii) a "hypothesis" about what frequency distribution the opponent is using during these periods, which is assumed to be unconditional on history; iii) a counting variable that tells whether the player is currently in "hypothesis testing" mode and how long he has been so. Once the count reaches s, a player conducts a hypothesis test, that is, he compares his current hypothesis with the observed behavior of the opponent over the last s periods. If the hypothesis is not too improbable given the data, he keeps the same hypothesis and eventually starts testing again. Otherwise he rejects his current hypothesis and chooses a new one at random from the finite-dimensional space of frequency distributions that the opponent could

[2] A related impossibility result states that if the state variable is the joint frequency distribution of play, there exists no uncoupled, deterministic, continuously differentiable adjustment dynamic such that the empirical frequency distribution converges to a stage-game Nash equilibrium in any finite two-person game (Hart and Mas-Colell 2003).

be using. Players are assumed to be boundedly rational in the sense that they choose smoothed best responses given their current hypotheses about the opponent. By annealing the parameters, it can be shown that, given any finite game, the players' behaviors converge in probability to the set of Nash equilibria of the stage game.

In this paper we introduce a new type of learning rule, called regret testing, that solves the "learning to play Nash" problem in an even simpler way. Unlike fictitious play, hypothesis testing, Kalai–Lehrer updating, and a host of other learning rules, regret testing does not depend on observation of the opponent's pattern of play or even on knowledge of the opponent's existence; *it depends only on summary statistics of a player's own realized payoffs*. In this sense it is similar in spirit to reinforcement and aspiration learning.[3] Response rules that depend only on a player's received payoffs are said to be *radically uncoupled*.

In the next section we define regret testing in detail; here we briefly outline how it works and why it avoids the impossibility theorems mentioned earlier. In each period a player has an *intended strategy*, that is, a probability mixture over actions that he plans to use in that period. With a small exogenous probability—independent among periods and players—a given player becomes distracted and uses an alternative strategy instead of his intended strategy. For simplicity assume that the alternative strategy involves choosing each action with equal probability. Periodically the player evaluates how his intended strategy is doing. He does this by comparing the average payoff he received when using his intended strategy with the average payoffs he received when distracted. If the latter payoffs are not markedly larger than the former, he continues as before. Otherwise he switches to a new (intended) strategy, where the choice of new strategy has a random component that assures that no region of his strategy space is completely excluded from consideration.

The random aspect of strategy switching is crucial because it allows for undirected search of the strategy space, and prevents the learning process from getting bogged down in disequilibrium mutual-adjustment cycles. It also side-steps the impossibility theorem of Hart and Mas-Colell mentioned at the outset: since behavior at a given point in time depends on the outcome of prior random variables (strategy switches), the learning process is *not stationary* with respect to any of the usual state variables such as history of play, history of payoffs, and so forth. Nevertheless, it is very simple and intuitive, and under an appropriate choice of the learning parameters, causes period-by-period behaviors to converge in probability to the set of stage-game Nash equilibria in any finite two-person game. The method can be extended to handle generic n-person games with finite action spaces (Germano and Lugosi 2004), but whether it works for all finite n-person games ($n \geq 3$) remains an open problem.

2. Regret testing

Consider an individual who lives alone. He has m possible actions, the names of which are written on "tickets" stored in "hats." Each hat contains $h \geq m$ tickets. Since a given

[3]Standard examples of reinforcement learning are given in Bush and Mosteller (1955) and Erev and Roth (1998). For models of aspiration learning see Karandikar et al. (1998), Börgers and Sarin (2000), Bendor et al. (2001), and Cho and Matsui (2005).

action can be written on multiple tickets, a hat is a device for generating probability distributions over actions. Every probability distribution that is expressible in integer multiples of $1/h$ is represented by exactly one hat. The larger is h, the more closely can any given distribution be approximated by one of these hats.

Step 1. A day consists of s periods, where s is large. Once each period, the player reaches into his current hat, draws a ticket, and takes the action prescribed. He then returns the ticket to the hat.

Step 2. At random times this routine is interrupted by telephone calls. 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 tallies the average payoff, $\widehat{\alpha}_t$, he received over the course of the day whenever he was not on the phone. For each action j, he compares $\widehat{\alpha}_t$ with the average payoff, $\widehat{\alpha}_{j,t}$, he received when he chose j and was on the phone.

Step 4. If at least one of the differences $\widehat{r}_{j,t} = \widehat{\alpha}_{j,t} - \widehat{\alpha}_t$ is greater than his tolerance level $\tau > 0$ he chooses a new hat, where each hat has a positive probability of being chosen. Otherwise he keeps his current hat and the process is repeated on day $t+1$.

Any procedure of this form is called a *regret testing rule*. The reason is that $\widehat{\alpha}_{j,t}$ amounts to a statistical estimate of the payoff on day t that the player would have received from playing action j all day long, hence the difference $\widehat{r}_{j,t} = \widehat{\alpha}_{j,t} - \widehat{\alpha}_t$ is the *estimated regret* from not having done so.[4] (Recall that the regrets cannot be evaluated directly because the opponent's actions are not observed.) The logic is simple: if one of the payoff-averages $\widehat{\alpha}_{j,t}$ during the experimental periods is significantly larger than the average payoff in the non-experimental periods, 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.

The revision process (Step 4) allows for many possibilities. The simplest is to choose each hat with *equal* probability, but this lacks behavioral plausibility. Instead, the player could exploit the information contained in the current payoffs, say by favoring strategies (hats) that put high probability on actions with high realized payoff $\widehat{\alpha}_{j,t}$. Consider, for example, the following revision rule: with probability $1-\varepsilon$ adopt the pure strategy that puts probability one on the action j that maximizes $\widehat{\alpha}_{j,t}$; and with probability ε choose a strategy at random. This is a trembled form of best response strategy revision, where the tremble is not in the *implementation* of the strategy but in the *choice* of strategy. In particular, a strategy that is far from being a best response strategy can be chosen by mistake, but the probability of such a mistake is small. While the use of recent payoff information may be sensible, however, we do not insist on it. The reason is that the process will eventually approximate Nash equilibrium behavior *irrespective* of the revision rule, as long as every hat is chosen with a probability that is uniformly bounded away from zero at all revision opportunities. This allows for a great deal of latitude in the specification of the learning process.

[4] A similar estimation device is used by Foster and Vohra (1993) and Hart and Mas-Colell (2000, 2001, 2005).

We hasten to say that this rule is intended to be a contribution to learning theory, and should not be interpreted literally as an empirical model of behavior, any more than fictitious play should be. Nevertheless it is composed of plausible elements that are found in other learning rules. One key element of regret testing is *inertia*: if there is no particular reason to change, play continues as before. In fact, inertia is built into the rule at two levels: there is no change of strategy while data is being collected over the course of a day, and change is implemented only if a *significant* improvement is possible—in other words, the alternative payoffs must exceed the current average payoff by more than some positive amount τ.

Inertia is an important aspect of aspiration learning as well as several other learning rules in the literature, including hypothesis testing (Foster and Young 2003) and regret matching (Hart and Mas-Colell 2000, 2001). In the latter procedure, a player continues to choose a given action with high probability from one period to the next. When change occurs, the probability of switching to each new action is proportional to its conditional regret relative to the current action.[5] Hart and Mas-Colell show that under this procedure the cumulative empirical frequencies converge almost surely to the set of correlated equilibria. (Note that this is quite different from saying that the period-by-period behaviors converge.)

A second key element of regret testing is that, when a change in strategy occurs, the choice of new strategy has a random component that allows for wide-area *search*. Except for hypothesis testing, this feature is not typical of other learning rules in the literature. For example, under regret matching, a player's strategy at any given time is either almost pure or involves switching probabilistically from one almost-pure strategy to another. Similarly, under aspiration learning, a player switches from one pure strategy to an alternative pure strategy when the former fails to deliver payoffs that meet a given aspiration level. In both of these situations there are probabilistic changes among particular classes of strategies, but not a wide-area search among strategies.

These two elements—inertia and search—play a key role in the learning process. Inertia stabilizes the players' behavior for long enough intervals that the players have a chance to learn something about their opponent's behavior. Search prevents the process from becoming trapped in adjustment cycles, such as the best response cycles that bedevil fictitious play in some settings. Intuitively, the way the process operates is that it discovers a (near) equilibrium through random search, then stays near equilibrium for a long time due to inertia. While it may seem obvious that this ought to work, it is a different matter to show that it actually does work. One difficulty is that the players' search episodes are not independent. Searches are linked via the history of play, so there is no guarantee that the joint strategy space will be searched systematically. A second difficulty is that, even when a search is successful and an equilibrium (or near equilibrium) has been found, the players do not know it. This is because they are ignorant of the opponent's payoff function, hence they cannot tell when a equilibrium is in hand, and may

[5]The *conditional regret* of action k relative to action j is the increase in average per-period payoff that would have resulted if k had been played whenever j actually was played. (The conditional regret is set equal to zero if k would have resulted in a lower average payoff than j.)

move away again. The essence of the proof is to show that, nevertheless, the expected time it takes to get close to equilibrium is much shorter than the expected time it takes to move away again.

3. Formal definitions and main result

Let G be a two-person game with finite action spaces X_1 and X_2 for players 1 and 2 respectively. Let $|X_i| = m_i$ and let $u^i : X_1 \times X_2 \to \mathbb{R}$ be i's utility function. In what follows, we assume (for computational convenience) that the von Neumann Morgenstern utility functions u^i are normalized so that all payoffs lie between zero and one:

$$\min_{x \in X_1 \times X_2} u^i(x) \geq 0 \qquad \text{and} \qquad \max_{x \in X_1 \times X_2} u^i(x) \leq 1. \tag{1}$$

Let Δ_i denote the set of probability mixtures over the m_i actions of player i. Let h_i be the uniform size of i's hats (a positive integer). The set of distributions in Δ_i that are representable as integer multiples of $1/h_i$ is denoted by P_i. Note that every strategy in Δ_i can be closely approximated by some strategy in P_i when h_i is sufficiently large. Let $\tau_i > 0$ denote i's *tolerance level*, let $\lambda_i \in (0,1)$ be the probability that a call is received by a player i during any given play of the game, and let s be the number of plays per day.

The *state space* is $Z = P_1 \times P_2$, which we sometimes refer to as the *probability grid*. The *state* of the learning process at the start of a given day t is $z_t = (p_t, q_t) \in P_1 \times P_2$. For each action j of player i, let $\widehat{\alpha}^i_{j,t} = \widehat{\alpha}^i_{j,t}(z_t)$ be the average payoff on day t in those periods when i played action j and was on the phone. Let $\widehat{\alpha}^i_t = \widehat{\alpha}^i_t(z_t)$ be i's average payoff on day t when *not* on the phone, and let $\widehat{\theta}^i_t = (\widehat{\alpha}^i_t, \widehat{\alpha}^i_{1,t}, \ldots, \widehat{\alpha}^i_{m_i,t})$. Note that $\widehat{\theta}^i_t$ contains enough information to implement a wide variety of updating rules, including trembled best response behavior and trembled better response behavior. Finally, let

$$\widehat{r}^i_t(z_t) = \max_{1 \leq j \leq m_i} \widehat{\alpha}^i_{j,t}(z_t) - \widehat{\alpha}^i_t(z_t).$$

A *regret-testing rule* for player 1 has the following form: there is a number $\gamma_1 > 0$ such that for every t and every state $z_t = (p_t, q_t)$,

$$\begin{aligned} &\widehat{r}^1_t(z_t) \leq \tau_1 \Rightarrow p_{t+1} = p_t \\ &\widehat{r}^1_t(z_t) > \tau_1 \Rightarrow P(p_{t+1} = p \mid p_t, \widehat{\theta}^1_t) \geq \gamma_1 \text{ for all } p \in P_1. \end{aligned} \tag{2}$$

The analogous definition holds for player 2. Note that we must have $\gamma_i \leq 1/|P_i|$ because the conditional probabilities in (2) sum to unity. The case $\gamma_i = 1/|P_i|$ corresponds to the uniform distribution, that is, all strategies in P_i are chosen with equal probability when a revision occurs. The class of regret testing rules is more general, however, because it allows for *any* conditional revision probabilities as long as they are uniformly bounded below by some positive constant.

A pair $(p,q) \in \Delta_1 \times \Delta_2$ is an *ε-equilibrium* of G if neither player can increase his payoff by more than ε through a unilateral change of strategy.

THEOREM 1. *Let G be a finite two-person game played by regret testers and let $\varepsilon > 0$. There are upper bounds on the tolerances τ_i and exploration rates λ_i, and lower bounds on the hat sizes h_i and frequency of play s, such that, at all sufficiently large times t, the players' joint behavior at t constitutes an ε-equilibrium of G with probability at least $1-\varepsilon$.*

Explicit bounds on the parameters are given in Section 5 below.

REMARK 1. 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 τ_i, he revises his strategy with probability $\theta_i \in (0,1)$ and with probability $1-\theta_i$ he continues to play his current strategy on the following day. We could also assume that the players use *different amounts of information*. Suppose, for example, that player i looks at the last k_i days of payoffs (k_i integer), and revises with probability $0 < \theta_i < 1$ whenever the estimated regrets exceed τ_i. With fixed values of k_i and θ_i this does not change the conclusion of Theorem 1 or the structure of the argument in any significant way.

REMARK 2. It is not necessary to assume that, when on the phone, a player chooses each of his actions with *equal* probability. Any fixed probability distribution that assigns positive probability to every action can be employed, but in this case the sample size may need to be larger than in the uniform case for the theorem to hold.

REMARK 3. Theorem 1 does not assert that the learning process *converges* to an ε-equilibrium of G; rather, it says that the players' period-by-period behaviors are *close to equilibrium with high probability* when t is large. By annealing the learning parameters at a suitable rate, one can achieve convergence in probability to the set of Nash equilibria, as we show in the concluding section. Moreover, with some further refinements of the approach one can actually achieve almost sure convergence, as shown by Germano and Lugosi (2004). Although these are probabilistic forms of convergence, the results are quite strong because they hold for the players' period-by-period behaviors. Regret matching, by contrast, only guarantees that the players' *time-average* behaviors converge, and then only to the set of correlated equilibria.[6]

Before giving the proof of Theorem 1 in detail, we give an overview of some of the technical issues that need to be dealt with. Regret testing defines one-step transition probabilities $P(z \to z')$ that lead from any given state z on day t to some other state z' on day $t+1$. Since these transition probabilities do not depend on t, they define a stationary Markov process P on the finite state space Z. A given state $z = (p,q)$ *induces a Nash equilibrium* in behaviors if and only if the *expected* regrets in that state are nonpositive. Similarly, (p,q) *induces an ε-equilibrium* in behaviors if and only if the expected regrets are ε or smaller. Note that this is not the same as saying that (p,q) itself is an

[6]Other rules whose long run average behavior converges to the correlated equilibrium set are discussed by Fudenberg and Levine (1995, 1998), Foster and Vohra (1999), and Cahn (2004). See Young (2004) for a general discussion of the convergence properties of learning rules.

ε-equilibrium, because the players' behaviors include experimentation, which distorts the probabilities slightly.

If a given state z does *not* induce an ε-equilibrium, the realized regrets $\widehat{r}^i_{j,t}$ are larger than ε with fairly high probability for at least one of the players. This player then revises his strategy. Since no strategy on his grid is excluded when he revises, there is a positive probability he hits upon a strategy that is close to being a best response to the opponent's current strategy. This is not good enough, however, because the new strategy pair does not necessarily induce an ε-equilibrium. What must be shown is that the players arrive *simultaneously* at strategies that induce an ε-equilibrium, a point that is not immediately obvious. For example, one player may revise while the second stays put, then the second may revise while the first stays put, and so forth.

Even if they do eventually arrive at an ε-equilibrium simultaneously, they must do so in a reasonably short period of time compared to the length of time they stay at the ε-equilibrium once they get there. Again this is not obvious. One difficulty is that the players do not know *when* they have arrived—they cannot see the opponent's strategy, or even his action, so they cannot determine when an ε-equilibrium is in hand. In particular, the realized regrets may be large (due to a series of bad draws) even though the state is close to equilibrium (or even *at* an equilibrium), in which case the players will mistakenly move away again. A second difficulty is that revisions by the two players are uncoupled, that is, they cannot coordinate the search process. In reality, however, their searches are linked because the regrets are generated by their joint actions. Thus, the fact that each player conducts a search of his own strategy space whenever he revises need not imply that the *joint* strategy space is searched systematically.

4. Entry and exit probabilities

The first step in proving Theorem 1 is to compare the probability of entering the set of ε-equilibrium states with the probability of leaving them. As a preliminary, we need to refine the concept of ε-equilibrium as follows. Given a pair of nonnegative real numbers $(\varepsilon_1,\varepsilon_2)$, say that a pair of strategies $(p,q)\in\Delta_1\times\Delta_2$ is an $(\varepsilon_1,\varepsilon_2)$-*equilibrium* if

$$\begin{aligned}\forall p'\in\Delta_1,\quad & u^1(p',q)-u^1(p,q)\le\varepsilon_1\\ \forall q'\in\Delta_2,\quad & u^2(p,q')-u^2(p,q)\le\varepsilon_2.\end{aligned}$$

When $\varepsilon_1=\varepsilon_2=\varepsilon$, we use the terms ε-equilibrium and $(\varepsilon_1,\varepsilon_2)$-equilibrium interchangeably. For any two real numbers x, y let $x\wedge y=\min\{x,y\}$ and $x\vee y=\max\{x,y\}$. Let us also recall that m_i denotes the number of actions available to player i.

Lemma 1. *Let $m=m_1\vee m_2$, $\tau=\tau_1\wedge\tau_2$, and $\lambda=\lambda_1\wedge\lambda_2$, and suppose that $0<\lambda_i\le\tau/8\le\frac{1}{8}$ for $i=1,2$. There exist positive constants a, b, and c such that, for all t,*

(i) 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.

(ii) If z_t is not a $(2\tau_1,2\tau_2)$-equilibrium and if $s\ge c$, then at least one player revises at the end of period t with probability greater than $\frac{1}{2}$; moreover if each player i is out

of equilibrium by at least $2\tau_i$, both revise at the end of period t with probability greater than $\frac{1}{4}$.

It suffices that $a = 12m$, $b = \lambda\tau^2/256m$, and $c = 10^3 m^2/\lambda\tau^2$.

REMARK 4. The proof shows, in addition, that if just *one* of the players, say i, can increase his payoff by more than $2\tau_i$, then i revises with probability greater than $\frac{1}{2}$ whenever $s \geq c$. Similarly, if one of the players i cannot increase his payoff by more than $\tau_i/2$, then i revises with probability at most ae^{-bs}. We sometimes use this unilateral version of Lemma 1 in what follows.

The proof of Lemma 1 involves a straightforward (but somewhat tedious) estimation of tail event probabilities, which is given in the Appendix. While it is a step in the right direction, however, it is not sufficient to establish Theorem 1. In particular, it is not enough to know that the process takes a long time (in expectation) to get *out* of a state that is very close to being an equilibrium; we also need to know how long it takes to get *into* such a state from somewhere else. What matters is the *ratio* between these entry and exit probabilities. This issue is addressed by the following general result on stationary, finite Markov chains.

LEMMA 2. *Consider a stationary Markov chain with transition probability function P on a finite state space Z. Suppose there exists a nonempty subset of states Z^0 and a state $w \notin Z^0$ such that:*

(i) in two periods the process moves from w into Z^0 with probability at least $\rho > 0$;

(ii) once in Z^0 the process stays there for at least one more period with probability at least $1-\theta$.

Then for any stationary distribution π of P, we have $\pi_w \leq 2\theta/\rho$.

PROOF. Let π be a stationary distribution of P. By definition $\pi P = \pi$, hence $\pi P^2 = \pi$; that is, π is also a stationary distribution of P^2. Condition (i) of the lemma says that

$$\sum_{z \in Z^0} P^2(w \to z) \geq \rho. \tag{3}$$

Condition (ii) implies that the probability of staying in Z^0 for at least two successive periods is at least $1-2\theta$, that is,

$$\forall y \in Z^0, \quad \sum_{z \in Z^0} P^2(y \to z) \geq 1 - 2\theta. \tag{4}$$

Since π is a stationary distribution of P^2, the stationarity equations imply that

$$\forall z \in Z^0, \quad \sum_{y \in Z^0} \pi_y P^2(y \to z) + \pi_w P^2(w \to z) \leq \pi_z. \tag{5}$$

Summing inequality (5) over all $z \in Z^0$ and using (3) and (4) we obtain

$$(1-2\theta)\sum_{y\in Z^0}\pi_y + \pi_w\rho \le \sum_{z\in Z^0}\pi_z.$$

Hence,

$$\pi_w\rho \le 2\theta\sum_{z\in Z^0}\pi_z \le 2\theta.$$

It follows that $\pi_w \le 2\theta/\rho$ as claimed. □

5. Proof of Theorem 1

We begin by restating Theorem 1, giving explicit bounds on the parameters. First we need some additional notation. Given $\delta \ge 0$, a strategy $p \in \Delta_1$ is δ-*subdominant* for player 1 if

$$\forall p' \in \Delta_1, \forall q \in \Delta_2, u^1(p',q) - u^1(p,q) \le \delta.$$

The analogous definition holds for player 2. A strategy is δ-subdominant if it is "almost" a weakly dominant strategy (assuming δ is small). A strategy is 0-*subdominant* if and only if it is a best reply irrespective of the opponent's strategy. (This is slightly weaker than weak dominance, because a 0-subdominant strategy is merely as good as any other strategy without necessarily ever being strictly better). Let $d(G)$ be the *least* $\delta \ge 0$ such that one or both players have a δ-subdominant strategy. Let $\tau = \tau_1 \wedge \tau_2$, $\lambda = \lambda_1 \wedge \lambda_2$, $\gamma = \gamma_1 \wedge \gamma_2$, and $m = m_1 \vee m_2$.

Theorem 1 (restatement). *Let G be a two-person game on the finite action space $X = X_1 \times X_2$ and let $\varepsilon > 0$. If the players use regret testing with strictly positive parameters satisfying the following bounds, then at all sufficiently large times t their joint behavior at t constitutes an ε-equilibrium of G with probability at least $1-\varepsilon$:*

$$\tau_i \le \varepsilon^2/48 \tag{6}$$

$$\tau_i \le d^2(G)/48 \quad \text{if } d(G) > 0 \tag{7}$$

$$\lambda_i \le \tau/16 \tag{8}$$

$$h_i \ge 8\sqrt{m}/\tau \tag{9}$$

$$\gamma_i \le 1/|P_i(h_i)| \tag{10}$$

$$s \ge (10^3 m^2/\lambda\tau^2)\ln(10^5 m/\varepsilon^2\gamma^7). \tag{11}$$

The need for some such bounds may be explained as follows. The tolerances τ_i must be sufficiently small relative to ε that the players reject with high probability when their behaviors are not an ε-equilibrium. The λ_i must be sufficiently small, relative to ε and τ, that the behaviors are close to equilibrium, and rejection is very unlikely, whenever the state (p,q) is sufficiently close to equilibrium. The h_i must be sufficiently large that the state space contains points that are close to equilibrium. The γ_i can be no larger than $1/|P_i(h_i)|$, where $|P_i(h_i)|$ is the number of probability distributions that can be accommodated by a hat of size h_i. The amount of information collected, s, must be

large enough that the probability of strategy revision is extremely small whenever the behaviors are sufficiently close to equilibrium. In addition, s must be large enough for Lemma 1 to hold, which is the case under assumption (11).

The most interesting, albeit somewhat mysterious, of these conditions is (7), which says that the tolerances must be small relative to $d(G)$. Since the tolerances are always positive, the condition states, in effect, that a given set of parameters will work for all games G except perhaps for those such that $d(G)$ is positive but very small, in particular, smaller than $(48\tau_i)^{1/2}$ for some player i.

To illustrate consider the following 2×2 coordination game:

	A	B
A	$1,1$	$1-\delta,0$
B	$0,1-\delta$	$1,1$

When δ is positive, action A is a δ-subdominant strategy for both players. When δ is negative, A is *strictly* dominant for both players. We have two different lines of argument for these situations.

When δ is negative, it can be shown that, once either player starts playing A (or a mixed strategy that puts very high probability on A), he will continue to play this strategy for a very long time. This gives the other player time to adjust and play a best response (which is also A). The same argument works when $\delta = 0$.

If δ is positive, however, we need a different argument. If the players are not very close to some equilibrium, we want to show that they will move simultaneously to, or close to, equilibrium within a relatively short period of time. This may not hold when one of the players has too high a tolerance relative to δ. The reason is that it may take him a rather long time to change strategy (i.e., to experience a regret that exceeds his tolerance), but he might not stay put long enough for the other player to adjust. Thus we cannot necessarily conclude that they arrive *simultaneously* at approximately mutual best responses in a short period of time. While we have not found an example showing that condition (7) is necessary for Theorem 1 to hold, we have also not been able to devise a method of proof that gets around it. For our purposes this does not particularly matter, because Theorem 2 exhibits an annealed version of the process that works for all two-person games on a given finite action space (with no excluded cases). It remains an open question whether Theorem 1 holds without imposing condition (7), in which case a fixed set of parameters would work for all two-person games on a given finite action space.

PROOF OF THEOREM 1. In state $z=(p,q)$, player 1 is playing the strategy $\widetilde{p}=(1-\lambda_1)p+(\lambda_1/m_1)\vec{1}_{m_1}$, where $\vec{1}_{m_1}$ is a length-m_1 vector of 1's. Similarly, player 2 is playing $\widetilde{q}=(1-\lambda_2)q+(\lambda_2/m_2)\vec{1}_{m_2}$. It follows that if (p,q) is an $\varepsilon/2$-equilibrium of G, then $(\widetilde{p},\widetilde{q})$ is an ε-equilibrium of G provided that the λ_i are sufficiently small. Since the payoffs lie between zero and one (see (1)), it suffices that $\lambda_1,\ \lambda_2 \le \varepsilon/4$. This holds because of assumptions (6) and (8).

Let E^* be the set of states in Z that are $\varepsilon/2$-equilibria of G (ignoring experimentation). We show first that, for every stationary distribution π of the process,

$$\sum_{z\notin E^*}\pi_z \le \varepsilon/2,$$

or equivalently,

$$\sum_{z\in E^*}\pi_z \ge 1-\varepsilon/2. \tag{12}$$

From this and the preceding remark it follows that the players' induced behaviors $(\widetilde{p},\widetilde{q})$ constitute an ε-equilibrium at least $1-\varepsilon/2$ of the time (and hence at least $1-\varepsilon$ of the time).

We need to show more however: namely, that the behaviors at time t constitute an ε-equilibrium with *probability* at least $1-\varepsilon$ for all sufficiently large times t. To see why this assertion holds, let P be the transition probability matrix of the process. If the process begins in state z_0, then the probability of being in state z at time t is $P^t(z_0\to z)$, where P^t is the t-fold product of P. We claim that P is acyclic; indeed this follows from the fact that for any state z, $P(z\to z)>0$. (Recall that, whenever a player revises, he chooses his previous strategy with positive probability.) It follows from standard results that the following limit exists

$$\forall z\in Z,\quad \lim_{t\to\infty}P^t(z_0\to z)=\pi_z,$$

and the limiting distribution π is a stationary distribution of P (Karlin and Taylor 1975, Theorem 1.2). From this and (12) it follows that

$$\lim_{t\to\infty}\sum_{z\notin E^*}P^t(z_0\to z)\le\varepsilon/2.$$

Hence

$$\exists T\quad \forall t\ge T,\quad \sum_{z\in E^*}P^t(z_0\to z)\ge 1-\varepsilon.$$

Thus, for all $t\ge T$, the probability is at least $1-\varepsilon$ that $z_t\in E^*$, in which case the induced behaviors at time t form an ε-equilibrium of G. This is precisely the desired conclusion. It therefore suffices to establish (12) to complete the proof of Theorem 1. We consider two cases: $d(G)>0$ and $d(G)=0$.

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

For every pair $(p,q)\in\Delta_1\times\Delta_2$, there exists $(p',q')\in Z$ such that

$$\left|p'-p\right|\le\sqrt{m_1}/h_1\quad\text{and}\quad\left|q'-q\right|\le\sqrt{m_2}/h_2.$$

By the lower bound (9) on the h_i, it follows that there is a point $(p',q')\in Z$ such that

$$\left|p'-p\right|\le\tau_1/8\quad\text{and}\quad\left|q'-q\right|\le\tau_2/8. \tag{13}$$

Now let (p,q) be a Nash equilibrium in the full space of mixed strategies, $\Delta_1 \times \Delta_2$. By (13) there is a state $e^* = (p^*, q^*) \in Z$ such that $\left|p^* - p\right| \leq \tau_1/8$ and $\left|q^* - q\right| \leq \tau_2/8$. Since all payoffs are bounded between zero and one, e^* is a $(\tau_1/8, \tau_2/8)$-equilibrium. In particular, $e^* \in E^*$, because by (6), $\tau_1/8 \leq \varepsilon/2$ and $\tau_2/8 \leq \varepsilon/2$. We fix $e^* = (p^*, q^*)$ for the remainder of the proof of Case 1.

It follows from Lemma 1, part (i), that

$$P(e^* \to e^*) \geq 1 - a e^{-bs}. \tag{14}$$

The next step is to show that for all $w \notin E^*$, the process enters E^* in two periods with fairly high probability; then we apply Lemma 2.

CASE 1A. $w \notin E^*$ and each player can, by a unilateral deviation, increase his payoff by more than $\varepsilon/2$.

Suppose that $z_t = w = (p,q)$. Since each player i can increase his payoff by more than $\varepsilon/2$, he can certainly increase it by more than $2\tau_i$ (because of the bound $\tau_i \leq \varepsilon^2/48$). It follows from Lemma 1, part (ii) that the probability is at least $\frac{1}{4}$ that both players revise at the end of day t.

Conditional on both revising, the probability is at least γ^2 that player 1 chooses p^* and player 2 chooses q^* in period $t+1$. Hence

$$P(w \to e^*) \geq \gamma^2/4,$$

so by (14),

$$P^2(w \to e^*) \geq (\gamma^2/4)(1 - a e^{-bs}).$$

CASE 1B. $w \notin E^*$ and only one of the players can improve his payoff by more than $\varepsilon/2$.

This case requires a two-step argument: we show that the process can transit from state w to some intermediate state x with the property that each player i can increase his payoff by more than $2\tau_i$. As in the proof of Case 1a, it follows that $P(x \to e^*) \geq \gamma^2/4$.

We now establish the existence of such an intermediate state. Assume without loss of generality that in state $w = (p,q)$, player 1 can increase his payoff by more than $\varepsilon/2$, whereas player 2 cannot. In particular, if $p' \in \Delta_1$ is a best response to q, then

$$u^1(p',q) - u^1(p,q) > \varepsilon/2. \tag{15}$$

Let $\delta = d(G)$: by definition neither player has a δ'-dominant strategy for any $\delta' < \delta$. In particular, q is not $\delta/2$-dominant for player 2. Hence there exist $p^0 \in \Delta_1$ and $q' \in \Delta_2$ such that

$$u^2(p^0,q') - u^2(p^0,q) > \delta/2. \tag{16}$$

Consider the strategy

$$p'' = (\delta/4)p + (1 - \delta/4)p^0. \tag{17}$$

By assumption, p' is a best response to q, so $u^1(p',q)-u^1(p^0,q)\geq 0$. It follows from (15) and (17) that

$$\begin{aligned} u^1(p',q)-u^1(p'',q) &= (\delta/4)[u^1(p',q)-u^1(p,q)] \\ &\quad +(1-\delta/4)[u^1(p',q)-u^1(p^0,q)] \\ &\geq (\delta/4)[u^1(p',q)-u^1(p,q)] \\ &> \delta\varepsilon/8. \end{aligned} \tag{18}$$

By assumptions (6) and (7), $\tau_1 \leq \delta^2/48$ and $\tau_1 \leq \varepsilon^2/48$, hence $48\tau_1 \leq \delta\varepsilon$, which implies $6\tau_1 < \delta\varepsilon/8$. From this and (18) we conclude that, given (p'',q), player 1 can deviate and increase his payoff by more than $6\tau_1$.

For player 2 we have, by definition of p'',

$$u^2(p'',q')-u^2(p'',q)=(\delta/4)[u^2(p,q')-u^2(p,q)]+(1-\delta/4)[u^2(p^0,q')-u^2(p^0,q)].$$

Since payoffs are bounded between zero and one, the first term on the right-hand side is at least $-\delta/4$. The second term is greater than $(1-\delta/4)(\delta/2)>3\delta/8$, by (16). Hence

$$u^2(p'',q')-u^2(p'',q)>\delta/8.$$

Since $\tau_2 \leq \delta^2/48 < \delta/48$, player 2 can deviate from (p'',q) and increase his payoff by more than $6\tau_2$. Hence (p'',q) is not a $(6\tau_1,6\tau_2)$-equilibrium.

Although q is on player 2's grid, the definition of p'' in (17) does not guarantee that it is on player 1's grid. We know, however, that there exists a grid point (p''',q) such that $|p'''-p''| \leq \sqrt{m_1}/h_1$. Since all payoffs lie between zero and one, the difference in payoff between (p''',q) and (p'',q) is at most $\sqrt{m_1}/h_1$ for *both* players. From (9) it follows that $\sqrt{m_1}/h_1 \leq \tau/8 \leq \tau_i/8$ for both players ($i=1,2$). Since (p'',q) is not a $(6\tau_1,6\tau_2)$-equilibrium, it follows that (p''',q) is not a $(5\tau_1,5\tau_2)$-equilibrium (and is on the grid).

Let $x=(p''',q)$. As in the proof of Case 1a, it follows that $P(x\to e^*)\geq\gamma^2/4$. Further, the process moves from state w to state x with probability at least $\gamma/2$, because only player 1 needs to revise: w and x differ only in the first coordinate. Hence,

$$P^2(w\to e^*)\geq\gamma^3/8.$$

In Case 1a we found that $P^2(w\to e^*)\geq(\gamma^2/4)(1-ae^{-bs})$, which is at least $\gamma^2/8$ provided that $ae^{-bs}\leq\frac{1}{2}$. This certainly holds under the assumptions in Lemma 1 on a, b, and s. Since $\gamma^2/8\geq\gamma^3/8$, it follows that in *both* cases

$$\forall w\notin E^*,\quad P^2(w\to e^*)\geq\gamma^3/8.$$

In both cases we also have $P(e^*\to e^*)\geq 1-ae^{-bs}$, by (14). Now apply Lemma 2 with $Z^0=\{e^*\}$, $\rho=\gamma^3/8$, and $\theta=ae^{-bs}$. We conclude that in *both* Case 1a and Case 1b, for every stationary distribution π of P,

$$\forall w\notin E^*,\quad \pi_w\leq\frac{2ae^{-bs}}{\gamma^3/8}=16ae^{-bs}/\gamma^3.$$

There are at most $1/\gamma^2$ states in Z altogether, so

$$\sum_{w \notin E^*} \pi_w \leq 16 a e^{-bs}/\gamma^5.$$

The right-hand side is at most $\varepsilon/2$ if $a e^{-bs} \leq \gamma^5 \varepsilon/32$, that is, if

$$s \geq (1/b)\ln(32a/\gamma^5\varepsilon).$$

By Lemma 1, we can take $a = 12m$ and $b = \lambda\tau^2/256m$. Thus it suffices that

$$s \geq \frac{256m}{\lambda\tau^2}\ln(384m/\gamma^5\varepsilon),$$

which is implied by the stronger bound in (11). This concludes the proof of Case 1.

CASE 2. $d(G) = 0$; some player has a 0-dominant strategy.

Fix a probability $0 < \beta < \frac{1}{2}$ that is *much smaller* than γ and *much larger* than $a e^{-bs}$; later we specify β and s more exactly. Define the following subset of states:

$$Z^\beta = \{z = (p,q) : \forall t, \forall q' \in P_2,\ P(p_{t+1} \neq p \mid z_t = (p,q')) \leq \beta\}.$$

In words, Z^β is the set of states such that the first player changes strategy with probability at most β no matter what strategy the second player is using on his grid.

Without loss of generality assume that player 1 has a 0-dominant strategy. Then he has a *pure* 0-dominant strategy, say p^*, which is in P_1. We fix p^* for the remainder of the proof.

Let Z^* be the set of states whose first coordinate is p^*. Then player 1 rejects with probability at most $a e^{-bs}$ (see Remark 4), that is,

$$P(z_{t+1} \in Z^* | z_t \in Z^*) \geq 1 - a e^{-bs}.$$

Hence $Z^* \subseteq Z^\beta$ provided that $a e^{-bs} \leq \beta$, which holds whenever s is sufficiently large (we assume henceforth that this is the case).

Let $w \in Z - E^*$. There are two possibilities: $w \notin Z^\beta$ and $w \in Z^\beta$.

CASE 2A. $w \notin E^*$ and $w \notin Z^\beta$.

Since $w = (p,q) \notin E^*, w$ is not an $\varepsilon/2$-equilibrium, and hence is not a $(2\tau_1, 2\tau_2)$-equilibrium. By Lemma 1 the probability is at least $\frac{1}{2}$ that there will be a revision next period by at least one of the players. If player 1 revises, a transition of form $(p,q) \to (p^*,\cdot) \in Z^*$ occurs with probability at least γ. After that, $(p^*,\cdot)$ stays in Z^* for one more period with probability at least $1 - a e^{-bs}$, which is at least $\frac{1}{2}$ because $a e^{-bs} < \beta < \frac{1}{2}$. Hence in this case

$$P^2(w \to Z^*) \geq \gamma/4.$$

If player 1 does not revise but player 2 does, then with probability at least γ we have a transition of form $(p,q) \to (p,q')$, where $q' \in P_2$ is a strategy for player 2 that makes

player 1 revise with probability greater than β. (There is such a q' because of our assumption that $w \notin Z^\beta$.) In the following period the transition $(p,q') \to (p^*,\cdot)$ occurs with probability greater than $\beta\gamma$. Hence in this case

$$P^2(w \to Z^*) \geq \beta\gamma^2/2.$$

Therefore, in either case,

$$P^2(w \to Z^*) \geq (\beta\gamma^2/2 \wedge \gamma/4) \geq \beta\gamma^2/4.$$

Now apply Lemma 2 with $Z^0 = Z^*$, $\theta = ae^{-bs}$, and $\rho = \beta\gamma^2/4$. Since $w \notin Z^*$ we conclude that

$$\pi_w \leq 2ae^{-bs}/(\beta\gamma^2/4) = 8ae^{-bs}/\beta\gamma^2. \tag{19}$$

CASE 2B. $w \notin E^*$ and $w \in Z^\beta$.

By definition of Z^β, player 1 revises with probability at most β, which by assumption is less than $\frac{1}{2}$. Since $w = (p,q) \notin E^*$, some player i can increase his payoff by at least $\varepsilon/2$, and hence by more than $2\tau_i$. This player will revise with probability greater than $\frac{1}{2}$ (see Remark 4), hence i cannot be player 1 (who revises with probability less than $\frac{1}{2}$). Therefore i must be player 2. By (9), there exists q'' on player 2's grid that is within $\tau_2/8$ of a best response to p. The probability is at least γ that 2 chooses q'' when he revises. Putting all of this together, we conclude that

$$P((p,q) \to (p,q'')) \geq \gamma/4. \tag{20}$$

By construction, state (p,q'') is a $(\cdot,\tau_2/8)$-equilibrium, hence player 2 revises with probability at most ae^{-bs} (see the remark after Lemma 1). By assumption, $(p,q) \in Z^\beta$, so player 1 revises with probability at most β against any strategy of player 2, including q''. Hence (p,q'') is also in Z^β, and

$$P((p,q'') \to (p,q'')) \geq (1-\beta)(1-ae^{-bs}) \geq (1-\beta)^2 > 1-2\beta.$$

From this and (20) we have

$$P^2((p,q) \to (p,q'')) \geq (\gamma/4)(1-\beta)^2 > \gamma/16,$$

the latter since $\beta < \frac{1}{2}$. Now apply Lemma 2 with $Z^0 = \{(p,q'')\}$, $\rho = \gamma/16$, and $\theta = 2\beta$. It follows that for every stationary distribution π of P,

$$\pi_w \leq 2(2\beta)/(\gamma/16) = 64\beta/\gamma. \tag{21}$$

Combining (19) and (21), it follows that in both Case 2a and Case 2b,

$$\forall w \notin E^*, \quad \pi_w \leq 64\beta/\gamma \vee 8ae^{-bs}/\beta\gamma^2. \tag{22}$$

The size of the state space is at least $1/\gamma^2$. Summing (22) over all $w \notin E^*$ it follows that

$$\pi(Z - E^*) \leq (1/\gamma^2)(64\beta/\gamma \vee 8ae^{-bs}/\beta\gamma^2).$$

We wish to show that this is at most $\varepsilon/2$. This will follow if we choose β and s so that $64\beta/\gamma^3 = \varepsilon/4$ and $8ae^{-bs}/\beta\gamma^4 \leq \varepsilon/4$. Specifically, it suffices that

$$\beta = \varepsilon\gamma^3/256$$

and

$$s \geq (1/b)\ln(8192a/\varepsilon^2\gamma^7).$$

By Lemma 1 we may choose $a = 12m$ and $b = \lambda\tau^2/256m$, hence it suffices that

$$s \geq (256m/\lambda\tau^2)\ln(98,304m/\varepsilon^2\gamma^7).$$

This certainly holds under (11), which states that $s \geq (10^3 m^2/\lambda\tau^2)\ln(10^5 m/\varepsilon^2\gamma^7)$. This concludes the proof of the theorem. □

6. Convergence in probability

Theorem 1 says that, for a given game G, regret testing induces an ε-equilibrium with high probability provided that the learning parameters satisfy the bounds given in (6)-(11). But it does not imply that, for a given set of parameters, an ε-equilibrium occurs with high probability for all games G. The difficulty is condition (7), which in effect requires that $d(G)$ not fall into the interval $(0, \sqrt{48(\tau_1 \vee \tau_2)})$. If we think of G as a vector of $2m_1m_2$ payoffs in Euclidean space, the excluded set is small relative to Lebesgue measure whenever the τ_i are small. Thus, if we tighten τ_1, τ_2 and the other parameters in tandem, the learning process eventually captures all games in the "net," that is, there are no excluded cases. In this section we show even more, namely, that by tightening the parameters sufficiently slowly, the players' period-by-period behavioral strategies converge in probability to the set of Nash equilibria of G.

Fix an $m_1 \times m_2$ action space $X = X_1 \times X_2$ and consider all games G on X with payoffs normalized to lie between zero and one. As before, let $m_1 \vee m_2$, $\lambda = \lambda_1 \wedge \lambda_2$, $\gamma = \gamma_1 \wedge \gamma_2$, and $\tau = \tau_1 \wedge \tau_2$. For each $\varepsilon > 0$, we choose particular values of these parameters that satisfy all the bounds except (7), namely,

$$\tau_i(\varepsilon) = \varepsilon^2/48 \tag{23}$$

$$\lambda_i(\varepsilon) = \tau(\varepsilon)/16 \tag{24}$$

$$h_i(\varepsilon) = \left\lceil 8\sqrt{m}/\tau(\varepsilon) \right\rceil \tag{25}$$

$$\gamma_i(\varepsilon) = 1/|P_i(h_i(\varepsilon))| \tag{26}$$

$$s(\varepsilon) = \left\lceil (10^3 m^2/\lambda(\varepsilon)\tau^2(\varepsilon)\ln(10^5 m/\varepsilon^2\gamma^7(\varepsilon)) \right\rceil. \tag{27}$$

Recall that $|P_i(h_i(\varepsilon))|$ is the number of distributions on i's grid when his hat size is $h_i(\varepsilon)$. Hence the players' grids become increasingly fine as ε becomes small. Note also

that (26) implies that each player chooses a new hat with *uniform* probability whenever a revision is called for. This proves to be analytically convenient in what follows, although more general assumptions could be made.

Let $P_G(\varepsilon)$ denote the finite-state Markov process determined by G and the parameters $(\tau_1(\varepsilon),\ldots,s(\varepsilon))$. Let $\mathscr{E}_G(\varepsilon)$ be the finite subset of states that induce an ε-equilibrium of G.

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

It follows from Theorem 1 that $T(P_G(\varepsilon),\mathscr{E}_G(\varepsilon),\varepsilon)$ is finite for all games G such that $d(G) \notin (0,\sqrt{48(\tau_1 \vee \tau_2)})$. By assumption (23), this holds whenever $d(G) \notin (0,\varepsilon)$. In this case, for all $t \geq T(P_G(\varepsilon),\mathscr{E}_G(\varepsilon),\varepsilon)$, the probability is at least $1-\varepsilon$ that the behavioral strategies constitute an ε-equilibrium of G at time t.

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

To see why this is so, consider the realization of plays on any given day. A realization is a sequence of $s(\varepsilon)$ action-outcome pairs, where an "outcome" is 0 or 1 depending on whether the action was taken by that player while on the phone or not. Hence there are $(4m_1m_2)^{s(\varepsilon)}$ possible realizations. We may partition them into four disjoint classes: sequences that cause both the players to reject (because the estimated regrets exceed their tolerances), 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. Moreover, a player's response *given a rejection* does not depend on other details of the sequence, because we are assuming that each player chooses a new strategy with uniform probability over all distributions on his grid.

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

DEFINITION 2 (annealed regret testing). Consider any positive sequence $\varepsilon_1 > \varepsilon_2 > \varepsilon_3 > \cdots$ decreasing to zero. The *annealed regret testing* procedure at stage k is the regret testing procedure with parameters $\tau_1(\varepsilon_k)$, $\tau_2(\varepsilon_k)$, $\lambda_1(\varepsilon_k)$, $\lambda_2(\varepsilon_k)$, $h_1(\varepsilon_k)$, $h_2(\varepsilon_k)$, $\gamma_1(\varepsilon_k)$,

$\gamma_2(\varepsilon_k)$, and $s(\varepsilon_k)$ as in (23)–(27). Each day that the process is in stage k, the probability of moving to stage $k+1$ on the following day is

$$p_k \equiv \frac{\varepsilon_{k+1}^2}{2k^2 T(\varepsilon_{k+1})}.$$

THEOREM 2. *Fix an $m_1 \times m_2$ action space $X = X_1 \times X_2$. Annealed regret testing has the property that, for every game G on X, the behavioral strategies converge in probability to the set of Nash equilibria of G.*

Although annealed regret testing seems to require that each player has an arbitrarily long memory, the process is actually of much lower dimension. To see why, let us fix a particular player i. Create one "payoff register" and one "counting register" for each of i's actions, plus one general payoff register and one general counting register for all of his actions together. Let k be a state variable that is common to all the players and indicates what stage the process is in (i.e., what parameters are currently in force). At each time t, i's general payoff register contains the running total of the payoffs he received when not on the phone, and the general counting register contains the number of times he was not on the phone. Similarly, each action-specific register contains the running total of the payoffs he received when he was on the phone and played that action, and the number of times the action was played while on the phone. When the sum over all counting registers reaches $s(\varepsilon_k)$, player i conducts a test using the k^{th} set of parameters, revises his strategy if this is called for, and empties all the registers. The process is then repeated. Thus player i needs to keep track of only $2m_i+3$ numbers—two for each of his actions, two in the general registers, and the current stage k. Thus the learning process requires very little memory or computational sophistication.

PROOF OF THEOREM 2. Given G, it suffices to show that, for every $\varepsilon > 0$, there is a finite time T_ε (possibly depending on G) such that for all $t \geq T_\varepsilon$, the probability is at least $1-\varepsilon$ that the behavioral strategies $(\widetilde{p}_t, \widetilde{q}_t)$ constitute an ε-equilibrium of G. Indeed, for every $\delta > 0$ there exists $0 < \varepsilon_\delta \leq \delta$ such that every ε_δ-equilibrium lies within δ of the compact set $\mathscr{N}_G$ of Nash equilibria of G. Hence, for all $t \geq T_{\varepsilon_\delta}$, the probability is at least $1-\varepsilon_\delta \geq 1-\delta$ that $(\widetilde{p}_t, \widetilde{q}_t)$ lies within δ of $\mathscr{N}_G$. Thus, the behavioral strategies converge in probability to the set of Nash equilibria of G.

To facilitate the proof we define three integer-valued random variables N_t, T_k, and W_k that describe the process as it transitions through stages. Let N_t be the stage that the process is in on day t. In other words, on day t the process is using the parameters $(\tau_1(\varepsilon_{N_t}), \tau_2(\varepsilon_{N_t}), \lambda_1(\varepsilon_{N_t}), \lambda_2(\varepsilon_{N_t}), h_1(\varepsilon_{N_t}), h_2(\varepsilon_{N_t}), \gamma_1(\varepsilon_{N_t}), \gamma_2(\varepsilon_{N_t}), s(\varepsilon_{N_t}))$. The distribution of the realizations of N_t depends on the transition probabilities as follows:

$$N_1 = 1$$

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

The first time that the system uses the k^{th} set of parameters is denoted by T_k, that is, $T_k \equiv \inf_t\{t : N_t \geq k\}$. Now define $W_t \equiv t - T_{N_t}$ to be the length of time since the parameters were last changed. In essence, the proof consists of establishing two facts about W_t. First we show that if W_t is "large" for a given t, the behavioral strategies are nearly a Nash equilibrium with high probability at time t. This follows by applying Theorem 1 to this setting. Second, we show that the probability that W_t is "large" converges to one as t converges to infinity. This follows from our assumption that the transition probabilities p_k are small. We now establish these points in detail.

For any game G on X, if $d(G) > 0$ then $d(G) \geq \varepsilon_k$ for all sufficiently large k, because the sequence $\{\varepsilon_k\}$ decreases to zero. The least such k is called the *critical index* of G, and denoted by k_G. In case $d(G) = 0$, we take $k_G = 1$. Fix $\varepsilon > 0$. Define $k_G^* = k_G \vee \min_k\{k \mid \varepsilon_k \leq \varepsilon/4\}$. It follows that if $N_t \geq k_G^*$ then $\varepsilon_{N_t} \leq \varepsilon/4$ and $d(G) \geq \varepsilon_{N_t}$.

Since $N_t \to \infty$ almost surely as $t \to \infty$, there is a time T^* such that, for all $t \geq T^*$, the probability is at least $1 - \varepsilon/4$ that $N_t \geq k_G^*$. From now on we consider only $t \geq T^*$.

Given $t \geq T^*$, consider two cases: $W_t \geq T(\varepsilon_{N_t})$ and $W_t < T(\varepsilon_{N_t})$. In the first case, the process is an ε_{N_t}-equilibrium with probability at least $1 - \varepsilon_{N_t}$. Since $t \geq T^*$, $N_t \geq k_G^*$ with probability at least $1 - \varepsilon/4$, in which case $\varepsilon_{N_t} \leq \varepsilon/4$. It follows that at time t the process is in an $\varepsilon/4$-equilibrium, and hence an ε-equilibrium, with probability at least $(1 - \varepsilon/4)(1 - \varepsilon/4) \geq 1 - \varepsilon/2$.

To complete the proof, it therefore suffices to show that, for all sufficiently large t, the second case occurs with probability at most $\varepsilon/2$, that is, there exists T^{**} such that

$$\forall\ t \geq T^{**}, \quad P(W_t < T(\varepsilon_{N_t})) \leq \varepsilon/2. \tag{28}$$

To establish (28) we proceed as follows. Recall that in the k^{th} stage of the process, the parameter values are $(\tau_1(\varepsilon_k), \ldots, s(\varepsilon_k))$. By choice of p_k, the k^{th} stage lasts for $2k^2 T(\varepsilon_{k+1})/\varepsilon_{k+1}^2$ periods in expectation. Say that the k^{th} stage is *short* if it lasts for at most $T(\varepsilon_{k+1})/\varepsilon_{k+1}^2$ periods, which is $1/2k^2$ times the expected number. This event has probability at most $1/k^2$. Hence, given any positive integer k_0, the probability that a short stage occurs at some time after the k_0^{th} stage is at most

$$\sum_{k > k_0} 1/k^2 \leq \int_{k_0}^{\infty} dx/x^2 = 1/k_0.$$

If we let $k_G^{**} = k_G^* \vee 16/\varepsilon$, it follows that *the probability is at most* $\varepsilon/16$ *that a short stage ever occurs after stage* k_G^{**}.

Now there exists a time T^{**} such that

$$\forall\ t \geq T^{**}, \quad P(N_t \geq k_G^{**} + 2) \geq 1 - \varepsilon/16.$$

We show that (28) holds for this value of T^{**}.

For each time $t \geq T^{**}$, define the event A_t to be the set of all realizations such that there is *at most one stage change between* $t - T(\varepsilon_{N_t})/\varepsilon_{N_t}^2$ *and* t, that is,

$$N_t \leq 1 + N_{t - T(\varepsilon_{N_t})/\varepsilon_{N_t}^2}.$$

Let A_t^c denote the complement of A_t. Since $t \geq T^{**}$, the probability is at least $1-\varepsilon/16$ that the process is at stage $k_G^{**}+2$ or higher at time t. Denote this event by B_t. If B_t and A_t^c both hold, then there were at least two stage changes between $t - T(\varepsilon_{N_t})/\varepsilon_{N_t}^2$ and t, hence the *previous* stage change (before the current stage) was short. But we already know that the probability of a short stage at any time beyond stage k_G^{**} is at most $\varepsilon/16$. Hence $P(A_t^c \mid B_t) \leq \varepsilon/16$ and $P(B_t^c) \leq \varepsilon/16$. Therefore

$$\forall\, t \geq T^{**}, \quad P(A_t^c) \leq P(A_t^c \mid B_t) + P(B_t^c) \leq 2(\varepsilon/16) = \varepsilon/8.$$

We now compute the probability that $W_t < T(\varepsilon_{N_t})$. By the preceding we know that

$$\begin{aligned} P(W_t < T(\varepsilon_{N_t})) &\leq P(W_t < T(\varepsilon_{N_t}) \mid A_t) + P(A_t^c) \\ &\leq P(W_t < T(\varepsilon_{N_t}) \mid A_t) + \varepsilon/8. \end{aligned}$$

Hence to establish (28) it suffices to show that

$$P(W_t < T(\varepsilon_{N_t}) \mid A_t) \leq 3\varepsilon/8.$$

Clearly,

$$\begin{aligned} P(W_t < T(\varepsilon_{N_t}) \mid A_t) &= \sum_k P(W_t < T(\varepsilon_{N_t}) \mid N_t = k, A_t) P(N_t = k \mid A_t) \\ &\leq \max_k P(W_t < T(\varepsilon_{N_t}) \mid N_t = k, A_t). \end{aligned}$$

Let $N_t = k$. The event A_t is the disjoint union of the event A_t^0 in which no stage change occurs between $t - T(\varepsilon_k)/\varepsilon_k^2$ and t, and the event A_t^1 in which exactly one stage change occurs.

When A_t^0 occurs, $W_t \geq T(\varepsilon_k)/\varepsilon_k^2 > T(\varepsilon_k)$, hence $P(W_t < T(\varepsilon_k) \mid A_t^0) = 0$. It remains only to show that $P(W_t < T(\varepsilon_k) \mid A_t^1) \leq 3\varepsilon/8$.

The conditional distribution of W_t is

$$f(w) \equiv P(W_t = w \mid N_t = k, A_t^1) = c_k (1-p_k)^{T(\varepsilon_k)/\varepsilon_k^2 - w} p_k (1-p_{k+1})^{w-1}, \tag{29}$$

where c_k is a positive constant and $1 \leq w \leq T(\varepsilon_k)/\varepsilon_k^2$. This follows because under A_t^1 a single stage change occurs during the interval, and it occurs exactly $W_t = w$ periods before period t. We may rewrite (29) in the form

$$f(w) = c_k' \left(\frac{1-p_{k+1}}{1-p_k} \right)^w$$

for some $c_k' > 0$. Since $p_k > p_{k+1}$, $f(w) \leq f(w+1)$. Hence for every T and w in the interval $1 \leq T$, $w \leq T(\varepsilon_k)/\varepsilon_k^2$,

$$\sum_{w<T} f(w) \leq T f(T) \quad \text{and} \quad \sum_{w \geq T} f(w) \geq (T(\varepsilon_k)/\varepsilon_k^2 - T) f(T).$$

In particular for $T = T(\varepsilon_k)$ we have

$$P(W_t < T(\varepsilon_k)) = \sum_{w<T(\varepsilon_k)} f(w)$$
$$= \frac{1}{1+\sum_{w\geq T(\varepsilon_k)} f(w)/\sum_{w<T(\varepsilon_k)} f(w)} \leq \frac{1}{1+\big(T(\varepsilon_k)/\varepsilon_k^2 - T(\varepsilon_k)\big)/T(\varepsilon_k)} = \varepsilon_k^2.$$

Since $t \geq T^{**}$, $\varepsilon_{N_t} = \varepsilon_k \leq \varepsilon/4$ with probability at least $1-\varepsilon/4$. Hence

$$P(W_t < T(\varepsilon_k)) \leq (\varepsilon/4)^2(1-\varepsilon/4)+\varepsilon/4 < 3\varepsilon/8.$$

This establishes (28) and completes the proof of Theorem 2. □

Appendix

Here we prove Lemma 1, which is restated for easy reference.

Lemma 1. *Let $m = m_1 \vee m_2$, $\tau = \tau_1 \wedge \tau_2$, and $\lambda = \lambda_1 \wedge \lambda_2$, and suppose that $0 < \lambda_i \leq \tau/8 \leq \frac{1}{8}$ for $i = 1,2$. There exist positive constants a, b, and c such that, for all t,*

(i) 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.

(ii) If z_t is not a $(2\tau_1, 2\tau_2)$-equilibrium and if $s \geq c$, then at least one player revises at the end of period t with probability greater than $\frac{1}{2}$; moreover if each player i is out of equilibrium by at least $2\tau_i$, both revise at the end of period t with probability greater than $\frac{1}{4}$.

It suffices that $a = 12m$, $b = \lambda\tau^2/256m$, and $c = 10^3 m^2/\lambda\tau^2$.

Proof. The player's strategy revisions are triggered by the size of their realized regrets $\widehat{r}_t^i$. Hence we need to estimate the distribution of $\widehat{r}_t^i$ conditional on the state at time t, namely, $z_t = (p_t, q_t)$. Recalling the definitions of $\widehat{\alpha}_{j,t}^i$ and $\widehat{\alpha}_t^i$ from Step 3 of regret testing, let

$$\alpha_{j,t}^i \equiv E(\widehat{\alpha}_{j,t}^i \,|\, (p_t, q_t)) \quad \text{and} \quad \alpha_t^i \equiv E(\widehat{\alpha}_t^i \,|\, (p_t, q_t)).$$

Recall that player 2 draws from his hat with probability $1-\lambda_2$, and plays an action uniformly at random with probability λ_2. (The uniform distribution over actions when experimenting contrasts with the possibly non-uniform distribution over hats when a rejection occurs.) Hence when player 1 chooses action j at time t, his expected payoff is

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

Similarly, 1's expected payoff at time t is

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

Similar expressions hold for $\alpha^2_{j,t}$ and α^2_t. Define

$$r^i_t \equiv \max_j \alpha^i_{j,t} - \alpha^i_t.$$

Since $E(\widehat{\alpha}^i_{j,t} \mid (p_t, q_t)) = \alpha^i_{j,t}$ and $E(\widehat{\alpha}^i_t \mid (p_t, q_t)) = \alpha^i_t$ we can think of the difference,

$$\widehat{r}^i_t = \max_j \widehat{\alpha}^i_{j,t} - \widehat{\alpha}^i_t,$$

as being an estimator of r^i_t.

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

$$|\widehat{r}^i_t - r^i_t|.$$

Next we estimate the distribution of the realized regret estimates $\widehat{r}^i_t$.

CLAIM. *If* $\lambda_i \le \frac{1}{3}$, *then for all* $\delta \le 1/\sqrt{2m_i}$, *and for all times* t,

$$P\left(|\widehat{r}^i_t - r^i_t| > \delta\right) \le 6m_i e^{-s\lambda_i \delta^2/16m_i}. \tag{A1}$$

PROOF. Fix a player i and let (p_t, q_t) be the state on day t. Let $N^i_{j,t}$ 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, $\widehat{\alpha}^i_{j,t}$, is an average of $N^i_{j,t}$ items, each of which is bounded between zero and one. By Azuma's inequality (Azuma 1967),

$$P(|\widehat{\alpha}^i_{j,t} - \alpha^i_{j,t}| > \delta \mid (p_t, q_t), N^i_{j,t}) \le 2e^{-N^i_{j,t}\delta^2/2}. \tag{A2}$$

Let $N^i_t = \sum_j N^i_{j,t}$. The number of times i was not on the phone on day t is $s - N^i_t$, hence again by Azuma's inequality

$$P(|\widehat{\alpha}^i_t - \alpha^i_t| > \delta \mid (p_t, q_t), N^i_t) \le 2e^{-(s-N^i_t)\delta^2/2}. \tag{A3}$$

Since for any two events $\mathscr{A}$ and $\mathscr{B}$, $P(\mathscr{A} \cup \mathscr{B}) \le P(\mathscr{A}) + P(\mathscr{B})$, it follows from (A2) and (A3) that

$$P(|\widehat{r}^i_t - r^i_t| > 2\delta \mid (p_t, q_t), N^i_{1,t}, N^i_{2,t}, \ldots, N^i_{m_i,t}) \le 2\sum_{j=1}^{m_i} e^{-N^i_{j,t}\delta^2/2} + 2e^{-(s-N^i_t)\delta^2/2}. \tag{A4}$$

The next step is to estimate the size of the tail of the random variable $N^i_{j,t}$, which is binomially distributed $B(\lambda_i/m_i, s)$. We claim that

$$P\left(\left|N^i_{j,t} - \frac{s\lambda_i}{m_i}\right| \ge \frac{s\lambda_i}{2m_i}\right) \le 2e^{-s\lambda_i/20m_i}. \tag{A5}$$

This can be derived from Bennett's inequality (Bennett 1962). Consider a collection of n independent random variables $U_1, \ldots, U_n$ with $\sup|U_i| \le M$, $E\,U_i = 0$, and $\sum_i E\,U_i^2 = 1$. Then for every $\tau > 0$,

$$P\left(\sum_i U_i \ge \tau\right) \le \exp\left(\frac{\tau}{M} - \left(\frac{\tau}{M} + \frac{1}{M^2}\right)\ln(1 + M\tau)\right). \tag{A6}$$

We apply this to the case of n i.i.d. random variables $X_1,\ldots,X_n$, with $\mathrm{Var}(X_i)=\sigma^2$ and $|X_i|\le 1$. Let $U_i=(1/\sigma\sqrt{n})(X_i-EX)$. Then $|U_i|<1/\sigma\sqrt{n}$, $EU_i=0$, and $\sum_{i=1}^n EU_i^2=1$. Letting $\tau=(\gamma/\sigma)\sqrt{n}$, $M=1/\sigma\sqrt{n}$, and $\overline{X}=\sum X_i/n$, it follows from (A6) that

$$P(\overline{X}-EX\ge\gamma)\le\exp\left(n\gamma-n(\gamma+\sigma^2)\ln(1+\gamma/\sigma^2)\right).$$

If we take $\gamma=\sigma^2/2$ and use the fact that $\ln(3/2)\ge.4$,

$$P(\overline{X}-EX\ge\sigma^2/2)\le\exp(-n\sigma^2/10).$$

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

$$P(\overline{X}-p\ge p/2)\le e^{-np/20}$$

and hence

$$P(|\overline{X}-p|\ge p/2)\le 2e^{-np/20},$$

from which (A5) follows immediately.

Consider the event $\mathscr{B}$ in which all of the $N^i_{j,t}$ lie within their expected value $\lambda_i s/m_i$ plus or minus half their expected value:

$$\mathscr{B}\equiv\bigcap\nolimits_j\{|N^i_{j,t}-\lambda_i s/m_i|\le\lambda_i s/2m_i\}.$$

From (A5) it follows that the probability of the complementary event $\mathscr{B}^c$ satisfies

$$P(\mathscr{B}^c)\le 2m_i e^{-s\lambda_i/20m_i}. \tag{A7}$$

Let $\mathscr{A}$ be the event $|\widehat{r}^i_t-r^i_t)|>2\delta$. From (A4) we have

$$P(\mathscr{A}\mid\mathscr{B})\le 2m_i e^{-\lambda_i s\delta^2/4}+2e^{-(s-N^i_t)\delta^2/2}.$$

But if $\mathscr{B}$ holds, then $s-N^i_t\ge s-3\lambda_i s/2=(1-3\lambda_i/2)s$. By hypothesis $\lambda_i\le\frac{1}{3}$, hence $s-N^i_t>s/2$ and

$$P(\mathscr{A}\mid\mathscr{B})\le 2m_i e^{-\lambda_i s\delta^2/4}+2e^{-s\delta^2/4}. \tag{A8}$$

Since $P(\mathscr{A})\le P(\mathscr{A}\mid\mathscr{B})+P(\mathscr{B}^c)$, it follows from (A7) and (A8) that

$$\begin{aligned}P(\mathscr{A})=P(|\widehat{r}^i_t-r^i_t|)>2\delta)&\le 2m_i e^{-s\lambda_i\delta^2/4}+2e^{-s\delta^2/4}+2m_i e^{-s\lambda_i/20m_i}\\&\le 2(m_i+1)e^{-s\lambda_i\delta^2/4}+2m_i e^{-s\lambda_i/20m_i}.\end{aligned}$$

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

$$P(|\widehat{r}^i_t-r^i_t|>\delta)\le 2(m_i+1)e^{-s\lambda_i\delta^2/16}+2m_i e^{-s\lambda_i/20m_i}. \tag{A9}$$

By assumption, $\delta\le 1\sqrt{2m_i}$; so $1/20m_i\ge\delta^2/16\ge\delta^2/16m_i$. It follows that $e^{-s\lambda_i\delta^2/16m_i}\ge e^{-s\lambda_i\delta^2/16}\ge e^{-s\lambda_i/20m_i}$, hence (A9) implies

$$P(|\widehat{r}^i_t-r^i_t|>\delta)\le(4m_i+2)e^{-s\lambda_i\delta^2/16m_i}\le 6m_i e^{-s\lambda_i\delta^2/16m_i}.$$

This establishes (A1) as claimed. □

Recall that in state (p,q) the behavioral probabilities are, for player 1,

$$\widetilde{p} = (1-\lambda_1)p + (\lambda_1/m_1)\vec{1}_{m_1},$$

and for player 2,

$$\widetilde{q} = (1-\lambda_2)q + (\lambda_2/m_2)\vec{1}_{m_2}. \tag{A10}$$

If (p_t,q_t) is an $(\varepsilon_1,\varepsilon_2)$-equilibrium, the expected regrets r_t^i in state (p_t,q_t) satisfy the bound

$$r_t^i \le \varepsilon_i + 2(\lambda_1 \vee \lambda_2). \tag{A11}$$

(This follows from (A10) and the assumption that the payoffs are bounded between 0 and 1.)

To prove Lemma 1, part (i), assume that $z_t = (p_t,q_t)$ is a $(\tau_1/2,\tau_2/2)$-equilibrium. Since by assumption, $\lambda_1, \lambda_2 \le \tau/8$, where $\tau = \tau_1 \wedge \tau_2$, it follows from (A11) that

$$r_t^i \le \tau_i/2 + \tau_i/4 = 3\tau_i/4.$$

In order for a rejection to occur, we must have $\widehat{r}_t^i > \tau_i$, which by the preceding implies that $|\widehat{r}_t^i - r_t^i| > \tau_i/4$. Letting $\delta = \tau_i/4$, it follows from (A1) that the probability of this occurring is less than $6m_i e^{-s\lambda_i\tau_i^2/256m_i}$. Thus the probability that one or both players reject is less than

$$\sum_{i=1}^{2} 6m_i e^{-s\lambda_i\tau_i^2/256m_i} \le 12m e^{-s\lambda\tau^2/256m}.$$

This establishes Lemma 1, part (i).

To prove part (ii) of the lemma, suppose that in state z_t at least one of the players, say i, can improve his payoff by more than $2\tau_i$. This implies $r_t^i > 2\tau_i - \tau_i/4 \ge 7\tau_i/4$. He rejects unless $\widehat{r}_t^i \le \tau_i$, which implies $|\widehat{r}_t^i - r_t^i| > 3\tau_i/4$. By (A1) we know that the probability of this is less than $6m_i e^{-9s\lambda_i\tau_i^2/256m_i} \le 6m_i e^{-s\lambda_i\tau_i^2/30m_i}$. Choose s large enough that

$$6m_i e^{-s\lambda_i\tau_i^2/30m_i} < \tfrac{1}{3}.$$

This holds if

$$s > \frac{30m_i}{\lambda_i\tau_i^2}\ln(18m_i).$$

Noting that $\ln x \le x$ for $x \ge 1$, this simplifies to

$$s > \frac{540m_i^2}{\lambda_i\tau_i^2}.$$

Recalling that $m = m_1 \vee m_2$, $\tau = \tau_1 \wedge \tau_2$, and $\lambda = \lambda_1 \wedge \lambda_2$, we see that this holds if $s \ge c = 10^3 m^2/\lambda\tau^2$, as posited in the lemma. We have therefore shown that, if player i is out of equilibrium by more than $2\tau_i$, then i accepts with probability at most $\frac{1}{3}$, and rejects with probability at least $\frac{2}{3}$, which is certainly greater than $\frac{1}{2}$. If *both* players are in this situation (as posited in part (ii) of the lemma), then each revises with probability at least $\frac{2}{3}$. Since the union of these two events has probability at most 1, their intersection has probability at least $\frac{1}{3}$ which is certainly greater than $\frac{1}{4}$. This concludes the proof of Lemma 1, part (ii). □

References

Azuma, Kazuoki (1967), "Weighted sums of certain dependent random variables." *Tôhoku Mathematical Journal*, 19, 357–367. [363]

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. [343]

Bennett, George (1962), "Probability inequalities for the sum of independent random variables." *Journal of the American Statistical Association*, 57, 33–45. [363]

Börgers, Tilman and Rajiv Sarin (2000), "Naive reinforcement learning with endogenous aspirations." *International Economic Review*, 41, 921–950. [343]

Bush, Robert R. and Frederick Mosteller (1955), *Stochastic Models for Learning*. John Wiley, New York. [343]

Cahn, Amotz (2004), "General procedures leading to correlated equilibria." *International Journal of Game Theory*, 33, 21–40. [347]

Cho, In-Koo and Akihiko Matsui (2005), "Learning aspiration in repeated games." *Journal of Economic Theory*, 124, 171–201. [343]

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. [343]

Foster, Dean P. and Rakesh Vohra (1993), "A randomization rule for selecting forecasts." *Operations Research*, 41, 704–709. [344]

Foster, Dean P. and Rakesh Vohra (1999), "Regret in the on-line decision problem." *Games and Economic Behavior*, 29, 7–35. [347]

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, 12848–12853. [341]

Foster, Dean P. and H. Peyton Young (2003), "Learning, hypothesis testing, and Nash equilibrium." *Games and Economic Behavior*, 45, 73–96. [342, 345]

Fudenberg, Drew and David K. Levine (1995), "Consistency and cautious fictitious play." *Journal of Economic Dynamics and Control*, 19, 1065–1090. [347]

Fudenberg, Drew and David K. Levine (1998), *The Theory of Learning in Games*. MIT Press, Cambridge, Massachusetts. [347]

Germano, Fabrizio and Gábor Lugosi (2004), "Global Nash convergence of Foster and Young's regret testing." Working Paper 788, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona. [343, 347]

Hart, Sergiu and Andreu Mas-Colell (2000), "A simple adaptive procedure leading to correlated equilibrium." *Econometrica*, 68, 1127–1150. [344, 345]

Hart, Sergiu and Andreu Mas-Colell (2001), "A general class of adaptive strategies." *Journal of Economic Theory*, 98, 26–54. [344, 345]

Hart, Sergiu and Andreu Mas-Colell (2003), "Uncoupled dynamics do not lead to nash equilibrium." *American Economic Review*, 93, 1830–1836. [342]

Hart, Sergiu and Andreu Mas-Colell (2005), "Stochastic uncoupled dynamics and Nash equilibrium." Technical Report DP-371, Center for Rationality, Hebrew University of Jerusalem. Forthcoming in *Games and Economic Behavior*. [342, 344]

Jordan, James S. (1991), "Bayesian learning in normal form games." *Games and Economic Behavior*, 3, 60–91. [341]

Jordan, James S. (1993), "Three problems in learning mixed-strategy equilibria." *Games and Economic Behavior*, 5, 368–386. [341]

Kalai, Ehud and Ehud Lehrer (1993), "Rational learning leads to Nash equilibrium." *Econometrica*, 61, 1019–1045. [341]

Karandikar, Rajeeva, Dilip Mookherjee, Debraj Ray, and Fernando Vega-Redondo (1998), "Evolving aspirations and cooperation." *Journal of Economic Theory*, 80, 292–331. [343]

Karlin, Samuel and Howard M. Taylor (1975), *A First Course in Stochastic Processes, 2nd edition*. Academic Press, New York. [352]

Nachbar, John H. (1997), "Prediction, optimization, and learning in games." *Econometrica*, 65, 275–309. [341]

Nachbar, John H. (2005), "Beliefs in repeated games." *Econometrica*, 73, 459–480. [341]

Young, H. Peyton (2004), *Strategic Learning and Its Limits*. Oxford University Press, Oxford, UK. [347]

Submitted 2006-2-7. Final version accepted 2006-5-5. Available online 2006-5-5.