ACADEMIC
PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Games and Economic Behavior 45 (2003) 73–96

GAMES and
Economic
Behavior

www.elsevier.com/locate/geb

Learning, hypothesis testing, and Nash equilibrium

Dean P. Foster [a] and H. Peyton Young [b,c,*]

[a] Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Economics, Johns Hopkins University, 3400 North Charles Street,
Baltimore, MD 21218-2685, USA
[c] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received 27 March 2001

Abstract

Consider a finite stage game G that is repeated infinitely often. At each time, the players have hypotheses about their opponents' repeated game strategies. They frequently test their hypotheses against the opponents' recent actions. When a hypothesis fails a test, a new one is adopted. Play is almost rational in the sense that, at each point in time, the players' strategies are ϵ-best replies to their beliefs. We show that, at least $1-\epsilon$ of the time t these hypothesis testing strategies constitute an ϵ-equilibrium of the repeated game from t on; in fact the strategies are close to being subgame perfect for long stretches of time. This approach solves the problem of learning to play equilibrium with no prior knowledge (even probabilistic knowledge) of the opponents' strategies or their payoffs.

JEL classification: C72; C12

Keywords: Repeated game; Nash equilibrium; Subgame perfect equilibrium; Hypothesis test

1. Statement of the problem

Consider a group of players who are engaged in a repeated game and are trying to learn the behavior of their opponents. At every time they play optimally, or almost optimally, given their beliefs. Their beliefs are generated by a *learning process*, that is, a procedure that maps past play to predictions about future play. Are there learning strategies that come close to equilibrium play of the repeated game *without* assuming any prior knowledge of the opponents' strategies or payoffs?

[*] Corresponding author.
E-mail address: pyoung@brook.edu (H.P. Young).

0899-8256/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0899-8256(03)00025-3

There have been several lines of attack on this learning problem, but none of them is robust in the above sense. The oldest branch of the literature is built on fictitious play (Brown, 1951). This simple learning process converges to Nash equilibrium for special classes of games, such as zero-sum games, dominance solvable games, games with strategic complementarities and diminishing returns, and potential games (Robinson, 1951; Milgrom and Roberts, 1991; Krishna, 1992; Monderer and Shapley, 1996). However, there are many examples in which it does not converge to Nash equilibrium (Shapley, 1964; Jordan, 1993; Foster and Young, 1998a). Further, even in some situations where it does converge, such as zero-sum games with only mixed equilibria, the convergence is not of the type we desire: namely, the players' forecasts are not close to the actual next round probabilities of play, nor are their strategies close to being in equilibrium.

A second branch of the literature explores conditions under which Bayesian rational players can learn to play Nash when the distribution of payoff-types is common knowledge. One answer is provided by Jordan (1991), who shows that if the players' prior beliefs about the others' strategies constitutes a sophisticated Bayesian equilibrium, then every accumulation point of the posterior beliefs is, with probability one, a Nash equilibrium in beliefs (see also Jordan, 1992, 1995). In our view this merely pushes the problem of learning an equilibrium onto another level. Moreover, these results do not solve our version of the problem because the learning process does not necessarily result in equilibrium or good prediction for the *realized* payoff-types.

A third approach to the learning problem was pioneered by Kalai and Lehrer (1993). They show that if the players' prior beliefs attach positive probability to all events that have positive probability under the players' actual strategies, then with probability one, play will eventually come ϵ-close to an ϵ-equilibrium of the repeated game. Their result would solve our problem *if* one could identify a robust procedure for constructing priors on the opponents' strategies such that the players' best-response strategies are absolutely continuous with respect to their beliefs. They do not give an example of such a rule; moreover, the subsequent literature suggests that it may be very difficult to do so (Nachbar, 1997, 1999, 2001; Miller and Sanchirico, 1997, 1999; Foster and Young, 2001).

A fourth branch of the literature focuses on learning procedures that are based on backward rather than forward-looking criteria of performance. For example, given a complete history of play, one might ask whether it *would have been* better, on average, to play x instead of y in all of those situations where one actually did play y. If no such substitution would have resulted in a higher average payoff, the history *minimizes conditional regret*. There exist a variety of quite simple learning rules that have this property. Moreover, when every player minimizes conditional regret, the empirical distribution of play converges with probability one to the set of correlated equilibria of the game (Foster and Vohra, 1997, 1998, 1999; Foster, 1999; Fudenberg and Levine, 1995, 1999a, 1999b; Hart and Mas-Colell, 2000, 2001).

Like fictitious play, these learning rules are quite simple computationally, and presume nothing about the players' prior information about the opponents' payoffs. Indeed, they are more satisfactory than fictitious play in the sense that they work for all finite games. But they are less satisfactory in the sense that they rely on backward-looking criteria of performance, rather than forward-looking ones. Furthermore, convergence is to the set of correlated equilibria, which is much larger than the set of Nash equilibria.[1]

Before proceeding, let us remark that there is no difficulty in constructing dynamic *algorithms* that converge to Nash equilibrium of the stage game. Indeed, given a small $\epsilon > 0$, simply search the space of mixed strategies until one finds an ϵ-equilibrium of the stage game, then search the space until one finds an $\epsilon/2$-equilibrium, and so forth ad infinitum.

There are several reasons why this is not a satisfactory solution to the learning problem. First, it is not a decentralized learning process; it requires coordinated search. Second, it requires that the payoffs be common knowledge to all players, hence it is not robust. Third, it ignores the desire of the players to maximize their payoffs as the learning proceeds.

In this paper we exhibit a natural class of learning procedures, based on the classical notion of hypothesis testing, that avoids these difficulties. There is no coordination, no common knowledge, and no special assumption about the players' priors. In fact, not only is there no *common* knowledge, there is *no* knowledge of either the opponents' payoffs or their distribution. Nevertheless, these hypothesis testing strategies come close to equilibrium in the following sense: at least $1 - \epsilon$ of the time t they are ϵ-equilibria of the repeated game from t onward, and in fact they are ϵ-close to being subgame perfect for long stretches of time.

2. Statement of the main result

We begin by recalling the elements of hypothesis testing. A hypothesis test has four basic ingredients:

(i) first one observes sample outcomes of a process;
(ii) one then compares the observations to predictions under a null hypothesis;
(iii) one rejects the null hypothesis if it is improbable that the observations could have occurred given that the null hypothesis is true;
(iv) one then selects a new hypothesis, and the process is repeated.

In the context of a repeated game, the data consist of the actions taken in previous periods, which we assume are publicly observed. A "hypothesis" is a model or forecast of how one's opponents are going to act in the future (conditional on play so far) together with one's own behavioral response. In other words, it is a hypothetical probability distribution over future actions, conditioned on each and every history that could unfold.

There is, however, a complication in testing hypotheses about people that does not crop up when testing hypotheses about nature. While holding a hypothesis, the hypothesis tester is forced to take actions that his opponent can observe. If the hypothesis tester is rational, these actions must bear a particular relationship to the hypothesis. This means that the opponents can make inferences about the tester's current hypothesis, and hence his future

[1] Forward-looking learning processes that lead to correlated equilibria are discussed by Nyarko (1994, 1997).

actions, which may cause them to alter *their* future actions, thus perhaps invalidating the tester's hypothesis. In short, there is a feedback loop from the model to the object being modeled. The challenge is to design a hypothesis testing rule that solves the associated fixed-point problem, yet is simple to implement and requires no coordination among the players.

We begin by considering the hypotheses that agents entertain about the strategies of their opponents, and the hypothesis tests they employ. Say that a repeated game strategy has *memory m* if it conditions only on the previous m periods of history. A player's *hypothesis* has *memory m* if it attributes memory m strategies to each of the opponents, and the player's response also has memory m.[2] Notice that this definition is consistent with rationality, because if an agent's model of the opponents' behavior has memory m, then there exists at least one best-response behavioral strategy that has memory m.

Any strategy of memory m for player i can be represented by a point in a finite-dimensional Euclidean space. Specifically, let X_i denote i's finite action space, and let Δ_i be the set of probability distributions on X_i. A *strategy of memory m* consists of a choice in Δ_i for each length-m history of play. Thus the set $\mathcal{A}_i$ of i's memory-m strategies is a product of simplexes, one for each length-m history on which the strategy next period can be conditioned. A *hypothesis of memory m* for player i can therefore be represented by a point in $\mathcal{A} \equiv \prod_j \mathcal{A}_j$. Note that player i is "hypothesizing" about the stochastic process generating the sequence of actions taken, including i's own actions. This simplifies notation since everyone's hypotheses lie in the same space.

We hasten to point out that while the players' *hypotheses* have bounded memory, typically their repeated game *strategies* do not. The reason is that hypotheses change from time to time due to the failure of hypothesis tests. Hence the players' repeated game strategies are much more complex than their models, and typically have unbounded memory. The same holds for the process generating the players' updated beliefs. Thus the assumption of bounded memory for the hypotheses should be interpreted as a form of bounded rationality in which players construct simplified models of the world around them. As Theorem 1 will show, this simplification "works" in the sense that actual behavior over long periods of time will in fact be very close to the simplified model. (To avoid misunderstandings in what follows, we shall reserve the term "strategy" to mean repeated-game strategy, and "belief" to mean a probability distribution over the opponents' strategies. The adoption and rejection of hypotheses define the players' beliefs, and the behavioral responses to the changing hypotheses define their repeated-game strategies.)

Next let us consider what kinds of hypothesis tests the agent should employ. The objective is to identify the strategies that one's opponents are currently using, not what they did in the distant past. It therefore makes sense to conduct a test only on *recent data*, say plays in the last s_i periods, where s_i is agent i's *sample size*. In addition, one needs to use a test with sufficient power. If the hypothesis is sufficiently far from the truth, the test should reject it with probability close to one: the probability of a type-II error should be small. Similarly, if the hypothesis is sufficiently close to the truth, the probability of

[2] This approach contrasts with Jehiel (1995, 1998) who examines learning processes in which the players look forward only a finite number of periods.

rejecting it should be close to zero: the probability of a type-I error should be small. As we shall see later, it is easy to design tests such that type-I and type-II errors go to zero exponentially fast as s_i increases. Such a family of tests is said to be *powerful*.

Next we turn to the question of how a tester chooses a new hypothesis after he rejects an old one. Classical hypothesis testing is silent on the question of how alternative hypotheses are chosen. One could, for example, use a maximum likelihood approach, but this is only one of many possibilities. We shall therefore remain agnostic about the exact mechanism that leads to the choice of new hypotheses. The remarkable fact is that virtually *any* mechanism works as long as no hypothesis is "excluded" in a sense made precise below.

Specifically, let us suppose that when agent i's test fails in period t he chooses a new hypothesis according to some density $f_i(\cdot|\overline{\omega}^t)$ on the space $\mathcal{A}$. (Note that the choice of new hypothesis may depend on the entire history of play, not just the most recent test data.) If we wish, we can view i's choice of new hypothesis at time t as being determined by the outcome of a random variable $\hat{\theta}_i^t$ that conveys private information to i. With this interpretation, i's new hypothesis can simply be viewed as the updating of i's belief given the new information. Since the realized values θ_i^t play no role in the proof, however, we shall suppress them in what follows. All that matters are the conditional distributions f_i that govern the choice of new hypotheses when a hypothesis test fails.

We make two assumptions about the f_i. First, each $f_i(\cdot|\overline{\omega}^t)$ is uniformly bounded away from zero for all $\overline{\omega}^t$.[3] Second, the new hypothesis lies within λ_i of the old hypothesis with probability at least $1-\lambda_i$, where λ_i is positive and close to zero. These assumptions amount to saying that the tester adopts "radical" hypotheses with non-negligible probability no matter what the history tells him, but that for the most part he chooses new hypotheses that are close to the old ones. Such a hypothesis tester is said to be *flexible* and *conservative*. (Actually, conservatism is not needed to prove convergence, but it is needed to show that hypothesis testing is almost rational as a repeated game strategy.)

Lastly, we need to specify how a tester chooses his own response given his model of the others' behavioral responses. Assume that for each player i there is a family $\{A_i^{\sigma_i}\}$ of response functions, indexed by a *response parameter* $\sigma_i > 0$, such that the following conditions hold. First, each $A_i^{\sigma_i}$ is a continuous mapping from i's model space $\mathcal{B}_i$ to i's response space $\mathcal{A}_i$ and is also continuous in the payoffs $u_i(x)$. Second, if σ_i is small, the response is *almost rational* in the sense that its expected payoff is within σ_i of the payoff from an optimal strategy for some small $\sigma_i > 0$. Third, the response is *diffuse* in the sense that i plays every action in every period with positive probability. We shall say that $\{A_i^{\sigma_i}\}$ is a family of *smoothed best response functions* if it satisfies these three conditions. This type of response behavior can be viewed as a form of bounded rationality or, if we wish, as strictly optimizing behavior under small, unobserved utility shocks.

A player who never changes response—no matter what model he holds—has no motivation to predict accurately, because accuracy in prediction does not affect his payoff. (An example is a player who only considers memory zero models and has a strictly dominant strategy.) We say that *prediction does not matter* for such a player. More

[3] In the proof we will actually use a weaker condition, namely, we shall assume that f_i is *diffuse* in the sense that, for each $\tau_0 > 0$, the f_i-measure of any τ_0-ball of hypotheses is bounded below by a strictly positive number $f_*(\tau_0) > 0$. This allows us to consider discrete distributions as well as continuous ones.

generally, say that *prediction matters by at most ϵ* if a player has a strategy that is an ϵ-best response to every memory-m model of the opponents; otherwise prediction matters by more than ϵ.

Player i is an *ϵ-good predictor* if i's prediction of the opponents' behavior differs from their planned next period behavior by at most ϵ at least $1-\epsilon$ of the time. Equivalently, the average of i's absolute errors in prediction up to time t is bounded above by a small number for all sufficiently large t.

In general let G^∞ be an infinitely repeated n-person game, where the stage game G is finite. An n-tuple of repeated game strategies is an *ϵ-equilibrium* at time t if, given any realization to time t, the continuation strategy for each player i has an expected discounted payoff that is within ϵ of the maximum expected discounted payoff among all continuation strategies.

Let $a_i \in \mathcal{A}_i$ be a memory-m strategy for each player i. Suppose that player i plans to play a_i in each period $t \geqslant m+1$. Then we obtain a repeated game strategy, which we also denote by a_i, that has memory m and is time homogeneous. Note that the n-tuple of strategies $\vec{a} \in \mathcal{A}$ is a subgame perfect equilibrium if, for each $\overline{\omega}^t$, playing a_i in each period from $t+1$ on is an optimal response given that every player $j \neq i$ plays a_j from $t+1$ on.

Given $\epsilon > 0$, we say that a repeated game strategy-tuple $(S_1, \ldots, S_n)$ is *ϵ-close to being a subgame perfect equilibrium at time t* if there exists a subgame perfect equilibrium $\vec{a} \in \mathcal{A}$ such that for at least $1/\epsilon$ periods beginning at t the conditional distribution on X_i induced by S_i is ϵ-close to the conditional distribution induced by a_i, where closeness is measured by Euclidean distance in Δ_i. Notice that being ϵ-close to subgame perfect equilibrium requires not only that behavior is close in a *given* period, but that it stays close for a long period of time.

We can now state our principal result (Theorem 2 below).

Let G be a finite, normal-form, n-person game and let $\epsilon > 0$. If the players are almost rational, use sufficiently powerful hypothesis tests with comparable amounts of data, and are flexible in their adoption of new hypotheses, then at least $1-\epsilon$ of the time:
(i) *their repeated-game strategies are ϵ-close to subgame perfect equilibrium,*
(ii) *all players for whom prediction matters are ϵ-good predictors.*

3. A simple example

A particularly simple example of this set-up is the following. At each time t, agent i's model b_i^t is that the opponents are using a memory-zero strategy, that is, b_i^t generates the same distribution over actions x_{-i} in every period $t' > t$ irrespective of the history. Let $U_i(x_i, b_i^t)$ be i's expected utility, discounted to time t, from playing x_i in each period from t on, given the model b_i^t. Without loss of generality we may choose $U_i(\cdot)$ so that the maximum expected payoff is one and the minimum is zero. Let

$$p_i\left(x_i \middle| b_i^t\right) = \mathrm{e}^{U_i(x_i, b_i^t)/\sigma_i} \Big/ \sum_{y_i \in X_i} \mathrm{e}^{U_i(y_i, b_i^t)/\sigma_i}. \tag{1}$$

This is the *logit response function* introduced by McKelvey and Palfrey (1995). The smaller σ_i is, the closer the response is to being a best reply. Specifically, given any $\epsilon > 0$, if σ_i is sufficiently small, the utility from this rule is within ϵ of the maximum utility over all rules, given the prediction b_i^t about the behavior of the opponents in all future periods.

At the end of each period t in which agent i is *not* presently conducting a hypothesis test, he initiates a test with probability $1/s_i$. After s_i periods have elapsed, he applies the "distance test" with tolerance τ_i, that is, he rejects his current hypothesis if and only if the observed empirical frequency distribution of plays in periods $t+1, \ldots, t+s_i$ is more than τ_i away from his hypothesis as measured by the Euclidean norm. Assume that, if the hypothesis is rejected, then with probability $1-\lambda_i$ he retains it, and with probability $\lambda_i > 0$ he draws a new hypothesis at random via the uniform distribution. We call this the *logit response and distance test*.

It is governed by four parameters: the degree of smoothing σ_i, the tolerance τ_i between the model and empirical distribution, the amount of data collected, s_i, and the degree of conservatism λ_i. Our result implies that, given any small $\epsilon > 0$, if all σ_i are sufficiently small, all τ_i are sufficiently small (given the σ_i), and all s_i are sufficiently large (given the σ_i and τ_i), then all players for whom prediction matters are ϵ-good predictors. Further, at least $1-\epsilon$ of the time the strategies constitute an ϵ-equilibrium of the repeated game and in fact are ϵ-close to being subgame perfect. (Note that these two statements do not depend on the size of the λ_i.) Finally, if the λ_i's are sufficiently small, then at all times the repeated game *strategies* generated by this hypothesis testing procedure are ϵ-optimal relative to the players' beliefs, that is, to the conditional probabilities they assign to future play paths taking into account their own future changes of hypothesis.

4. Models and hypotheses

Let G be a finite, n-person game with players $i = 1, 2, \ldots, n$. The action space of player i is denoted by X_i, and the utility function by $u_i : X \to R$, where $X = \prod X_i$. The game G is played infinitely often, and the actions in each period are publicly observed. A *history of play* is denoted by ω, and the set of all possible histories by Ω. Given a history ω, $\omega^t = (\omega_1^t, \ldots, \omega_n^t) \in X$ denotes the actions taken in period t. Let $\overline{\omega}^t = (\omega^1, \omega^2, \ldots, \omega^t)$ denote the sequence of actions taken in periods 1 through t inclusive. Finally, the set of all continuations of the initial history $\overline{\omega}^t$ is denoted by $\Omega(\overline{\omega}^t) = \{\alpha \in \Omega \mid \bar{\alpha}^t = \overline{\omega}^t\}$.

The basic building block of our analysis is the concept of a model. A *model* for a given player is a forecast of the player's opponents' one-step-ahead behaviors, conditional on every possible initial history. In other words, a model b_i for player i specifies a conditional probability distribution $p_i^t(x_{-i}|\overline{\omega}^{t-1}, b_i)$ for every initial history $\overline{\omega}^{t-1}$. We will assume that a model attributes conditionally independent plays to the opponents, though in fact all theorems and proofs go through without this assumption.

A model b_i has *memory at most m* if the conditional distributions satisfy

$$p_i^t\left(x_{-i}\middle|\overline{\omega}^{t-1}, b_i\right) = p_i^t\left(x_{-i}\middle|\omega^{t-m}, \ldots, \omega^{t-1}, b_i\right) \quad \text{for all } t > m.$$

If a model b_i has memory at most m, there are $M = |X|^m$ distinct objects on which the conditional probabilities $p_i^t \in \Delta_{-i}$ depend. (The distributions $p_i^t(x_{-i}|\overline{\omega}^{t-1}, b_i)$ for

$1 \leqslant t \leqslant m$ can be chosen arbitrarily and will be ignored in what follows.) A list of these conditional probabilities determines the model, that is, we can identify the model with a point in the Euclidean space $\mathcal{B}_i = \prod_{j \neq i} \Delta_j^M$. Thus $\mathcal{B}_i$ is the space of all models having memory at most m. Let $|\mathcal{B}_i|$ denote the dimension of this space.

5. Behavioral responses

Next we turn to the strategic behaviors that players adopt in response to their models. In general, a *behavioral response* a_i for player i defines the conditional probability $q_i^t(x_i|\overline{\omega}^{t-1}, a_i)$ that i plays action x_i in period t, given the history $\overline{\omega}^{t-1}$. A response a_i has *memory at most m* if

$$q_i^t\bigl(x_i\big|\overline{\omega}^{t-1}, a_i\bigr) = q_i^t\bigl(x_i\big|\omega^{t-m}, \ldots, \omega^{t-1}, a_i\bigr) \quad \text{for all } t > m.$$

(The responses $q_i^t(x_i|\omega^{t-1}, a_i)$ for $1 \leqslant t \leqslant m$ can be chosen arbitrarily and will be ignored in what follows.)

As before, if a behavioral response a_i has memory at most m, there are $M = |X|^m$ distinct objects on which the conditional probabilities can depend. For each of these, the conditional probability can be thought of as a point in Δ_i. So a_i can be thought of as a point in $\mathcal{A}_i = \Delta_i^M$, which is a subset of a Euclidean space. Note that the model space and the response space are related by

$$(\forall i) \quad \mathcal{B}_i = \prod_{j \neq i} \mathcal{A}_j.$$

Let $\mathcal{A} = \prod \mathcal{A}_i$ and let $\vec{a} = (a_1, \ldots, a_n) \in \mathcal{A}$ be the players' responses at a given point in time. Then the correct model for player i to hold is $\prod_{j \neq i} a_j$. The mapping from any response vector $\vec{a}$ to the correct model for i will be denoted by $B_i : \mathcal{A} \to \mathcal{B}_i$, where $B_i(a) = \prod_{j \neq i} a_j$.

Let $\rho_i < 1$ be i's discount factor, so that i's utility from time period 1 on is

$$U_i^1(\omega) = (1 - \rho_i) \sum_{t=1}^{\infty} \rho_i^{t-1} u_i\bigl(\omega^t\bigr).$$

Here we have normalized by the factor $1 - \rho_i$ so that ϵ-deviations from best replies are comparable among players with different discount factors. For the same reason, we normalize the utilities so that, for each player, the maximum utility of any stage-game strategy is one and the minimum utility is zero. (Thus the same upper and lower bounds hold for $U_i^1(\omega)$.)

Let ν_{a_i, b_i} be the probability measure over infinite histories induced by the response a_i and the model b_i. Then i's expected utility from the pair (a_i, b_i) is

$$E\bigl(U_i^1(\omega)\big|a_i, b_i\bigr) = \int U_i^1(\omega)\, \mathrm{d}\nu_{a_i, b_i}.$$

Similarly, if ϖ^{t-1} is an initial history, the future stream of utilities discounted to period t is given by

$$U_i^t(\omega) = (1-\rho_i)\sum_{t'=t}^{\infty}\rho_i^{t'-t}u_i\bigl(\omega^{t'}\bigr).$$

(If $\rho_i = 0$ we let $U_i^t(\omega^t) = u_i(\omega^t)$.) Player i's expected utility at time t over all continuation histories $\Omega(\varpi^{t-1})$ is

$$E\bigl(U_i^t(\omega)\big|a_i, b_i, \varpi^{t-1}\bigr) = \int\limits_{\Omega(\varpi^{t-1})} U_i^t(\omega)\,\mathrm{d}\nu_{a_i,b_i} \Big/ \int\limits_{\Omega(\varpi^{t-1})} \mathrm{d}\nu_{a_i,b_i}.$$

To simplify the notation we shall henceforth write

$$U_i^t(a_i, b_i) \equiv E\bigl(U_i^t(\omega)\big|a_i, b_i, \varpi^{t-1}\bigr).$$

Thus $U_i^t(a_i, b_i)$ is a random variable where we have suppressed the dependence on ϖ^{t-1}. We will say that a_i is an *optimal response to* b_i if for all other responses a_i'

$$(\forall t)\quad U_i^t(a_i, b_i) \geqslant U_i^t\bigl(a_i', b_i\bigr).$$

Given a small $\sigma_i > 0$, a_i is a σ_i*-optimal response to* b_i if

$$(\forall t)\bigl(\forall a_i'\bigr)\quad U_i^t(a_i, b_i) \geqslant U_i^t\bigl(a_i', b_i\bigr) - \sigma_i.$$

Consider an arbitrary response a_i and a model b_i which has memory at most m. In general, a σ_i-optimal response a_i need not have memory m or less, but there exists at least one σ_i-optimal response that does. Indeed, given any σ_i-optimal response a_i, we can construct such a σ_i-optimal response as follows. For each ϖ^{t-1}, define a response a_i' as follows: a_i' is the arithmetical average of the numbers $p_i^t(x_i|\eta^{t-1}, a_i)$ where η^{t-1} ranges over all length-$(t-1)$ histories whose last m entries agree with ϖ^{t-1}. Then a_i' is σ_i-optimal given ϖ^{t-1} and b_i, and it depends only on the last m entries of ϖ^{t-1}. Thus for each player i there is a σ_i*-optimal response function* from models to strategies of the form

$$A_i^{\sigma_i}: \mathcal{B}_i \to \mathcal{A}_i.$$

In addition to σ_i-optimality we shall assume that these response functions are *continuous* in b_i, continuous in each payoff $u_i(x)$, and are *diffuse* in the sense that each action is played with positive probability. We shall call such an $A_i^{\sigma_i}$ a σ_i*-smoothed best response function* and $\{A_i^{\sigma_i}\colon \sigma_i > 0\}$ is a *family* of smoothed best response functions. The dependence of $A_i^{\sigma_i}$ on U_i will not be explicitly included in the notation but will be used in the theorem.

5.1. Hypothesis tests

In each period t, let b_i^t denote the model that player i has at that time. Every so often i subjects his current model to a hypothesis test. A *null hypothesis* H_0 is a statement of the form "the real process generating the actions from time t on is described by the pair $(A_i^{\sigma_i}(b_i^t), b_i^t)$." Since $b_i \in \mathcal{B}_i = \prod_{j\neq i}\mathcal{A}_j$, we see that a hypothesis lies in the space

$\mathcal{A} \equiv \prod_i \mathcal{A}_i$. An *alternative hypothesis* H_1 is another probability distribution in $\mathcal{A}$. The hypothesis H_0 is rejected if the data fall into the rejection region determined by the test.

If i is not conducting a test at the beginning of period t_0, he begins a new test with probability $1/s_i$.[4] Over the next s_i periods he collects data on how the process evolves. During this *test phase*, he continues to play $A_i^{\sigma_i}(b_i^{t_0})$, because his model has not yet changed. At the end of period $t_0 + s_i$ he conducts the test, and either accepts the hypothesis or rejects it. If it is accepted, the model does not change next period, that is, $b_i^{t_0+s_i+1} = b_i^{t_0}$. If the hypothesis is rejected, player i chooses a new model $b_i^{t_0+s_i+1}$ according to a probability measure $f_i^{t_0+s_i+1}(b_i|\overline{\varpi}^{t_0+s_i})$. This new model will be held until the next time that player i conducts a test and the hypothesis is rejected.

We can think of the choice of new hypothesis as being governed by the realization of a random variable $\hat{\theta}_i(\overline{\varpi}^{t_0+s_i})$, which we suppress for notational simplicity. The only assumption we make about f_i is that the measure of every τ_0-ball is bounded below by some $f_*(\tau_0)$ that is positive for all $\tau_0 > 0$. This allows for discrete as well as continuous densities.

5.2. Powerful tests

A standard criterion for being a good test is that it accept the null with high probability when the null is correct, and reject with high probability when the null is not correct. To define these ideas formally, let $b_i^{t_0}$ be i's current model at time t_0, and let $A_i^{\sigma_i}(b_i^{t_0})$ be i's response. Together, $b_i^{t_0}$ and $A_i^{\sigma_i}(b_i^{t_0})$ generate a distribution of play paths that we can view as the *null distribution* $\nu_i^{t_0} = (A_i^{\sigma_i}(b_i^{t_0}), b_i^{t_0})$. This is the distribution that i thinks is generating the actions, both his own and others'. Given the data $\overline{\varpi}^{t_0-1}$, let $\alpha_{i,s_i}(\nu_i^{t_0}, \overline{\varpi}^{t_0-1})$ denote the probability of rejecting the null when it is correct, that is, of making a type-I error. Similarly, given an alternative distribution $\nu \neq \nu_i^{t_0}$, let $\beta_{i,s_i}(\nu, \nu_i^{t_0}, \overline{\varpi}^{t_0-1})$ denote the probability of accepting $\nu_i^{t_0}$ when in fact ν is correct. That is, $\beta_{i,s_i}(\nu, \nu_i^{t_0}, \overline{\varpi}^{t_0-1})$ is the probability of making a type-II error, and $1 - \beta_{i,s_i}(\nu, \nu_i^{t_0}, \overline{\varpi}^{t_0-1})$ is the *power* of the test.

It is useful to define a variant of β that discriminates between hypotheses that are good approximations of the truth and those that are not. Recall that a null distribution ν_0 and any alternative distribution ν can be viewed as points in the space Δ^M, and thus we can speak of the "distance" between the two distributions. Given a small *tolerance level* $\tau > 0$, define

$$\beta_{i,s_i,\tau} = \sup_{\overline{\varpi},\, t,\, \nu_0} \; \sup_{\nu: |\nu-\nu_0| > \tau} \beta_{i,s_i}\big(\nu, \nu_0, \overline{\varpi}^{t-1}\big).$$

This is the least upper bound on making a type-II error when the true distribution ν is more than τ away from the null ν_0. Likewise we can define

$$\alpha_{i,s_i} = \sup_{\overline{\varpi},\, t,\, \nu_0} \alpha_{i,s_i}\big(\nu_0, \overline{\varpi}^{t-1}\big).$$

[4] In fact it suffices to assume that the probability of commencing a test is bounded above by $1 - c_i/s_i$ and bounded below by c_i/s_i for some constant $c_i > 0$.

A family of tests is *powerful* if there exist functions $k_i(\tau) > 0$ and $r_i(\tau) > 0$ such that, for each tolerance $\tau > 0$, there is a test in the family such that

$$\alpha_{i,s_i} \leqslant k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i} \quad \text{and} \quad \beta_{i,s_i,\tau} \leqslant k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i}. \tag{2}$$

Powerful families are easy to generate. For example, take any test which has $\alpha_{i,s_0} < 0.5$ and $\beta_{i,s_0,\tau} < 0.5$ for some sample size s_0. For any s rounds of data actually collected, divide it up into $\lfloor s/s_0 \rfloor$ disjoint subsets, each containing s_0 elements, and ignore the remainder. Further, if $\lfloor s/s_0 \rfloor$ is even, throw out the last subset of size s_0. Now compute the result of the test on each of the k remaining subsets of size s_0, where k is odd. Reject the null if and only if the null is rejected by the original test on a majority of the subsets. Letting $\alpha = \alpha_{i,s_0}$, it follows that the probability of incorrectly rejecting the null is at most

$$\begin{aligned}\sum_{j>k/2} \binom{k}{j} \alpha^j (1-\alpha)^{k-j} &\leqslant \alpha \sum_{j>k/2} \binom{k}{j} \bigl(\alpha(1-\alpha)\bigr)^{\lfloor k/2 \rfloor} \\ &= 2\alpha\bigl[4\alpha(1-\alpha)\bigr]^{\lfloor k/2 \rfloor} \sum_{j>k/2} \binom{k}{j} (1/2)^k \\ &\leqslant \alpha\bigl[4\alpha(1-\alpha)\bigr]^{\lfloor k/2 \rfloor}.\end{aligned}$$

By construction, $\lfloor k/2 \rfloor \geqslant s/(2s_0) - 2$. Further, $4\alpha(1-\alpha) < 1$ because $\alpha < 1/2$. It follows that

$$\alpha_{i,s} \leqslant k'\mathrm{e}^{-r's},$$

where

$$k' = \alpha\bigl[4\alpha(1-\alpha)\bigr]^{-2} \quad \text{and} \quad r' = \ln\bigl[4\alpha(1-\alpha)\bigr]\big/2s_0.$$

Similarly, letting $\beta = \beta_{i,s_0,\tau} < 1/2$, we have

$$\beta_{i,s,\tau} \leqslant k''\mathrm{e}^{-r''s},$$

where

$$k'' = \beta\bigl[4\beta(1-\beta)\bigr]^{-2} \quad \text{and} \quad \ln\bigl[4\beta(1-\beta)\bigr]\big/2s_0.$$

Thus,

$$\alpha_{i,s} \leqslant k\mathrm{e}^{-rs} \quad \text{and} \quad \beta_{i,s,\tau} \leqslant k\mathrm{e}^{-rs},$$

where

$$k = \max\{k', k''\} \quad \text{and} \quad r = \min\{r', r''\}.$$

Analogously with $\beta_{i,s_i,\tau}$, let us define $\alpha_{i,s_i,\tau}$ to be the least upper bound on rejecting the null when the true distribution is within τ of the null. Note that $\beta_\tau \leqslant \beta_{\tau'}$ whenever $\tau < \tau'$ and $\alpha_\tau \leqslant \alpha_{\tau'}$ whenever $\tau < \tau'$.

Lemma 1. *Suppose that under the null hypothesis the conditional probability of observing any given combination of actions in any period is at least ρ. Then for every $\tau > 0$,*

$$\alpha_{i,s_i,\tau} \leqslant \left(1 + \frac{\tau}{\rho}\right)^{s_i} \alpha_{i,s_i}.$$

Proof. To simplify the notation we consider first the situation where the test is applied to data generated in the first s_i periods, namely the data $\bar{\omega}^{s_i}$. Denote the null distribution by $p(\bar{\omega}^{s_i})$, and let R be the rejection set, that is, the set of $\bar{\omega}^{s_i}$ such that the null is rejected where $\bar{\omega}^{s_i}$ is observed. Let $q(\bar{\omega}^{s_i})$, be any other distribution that lies within τ of p, that is, $\|q - p\| \leqslant \tau$ in the Euclidean norm. Then we certainly have

$$\left|q\left(\omega^t \middle| \bar{\omega}^{t-1}\right) - p\left(\omega^t \middle| \bar{\omega}^{t-1}\right)\right| \leqslant \tau \quad \text{for all } t \leqslant s_i.$$

We wish to show that if $p(R) \leqslant \alpha$, then

$$q(R) \leqslant \left(1 + \frac{\tau}{\rho}\right)^{s_i} \alpha.$$

Since $p(\omega^t|\bar{\omega}^{t-1}) \geqslant \rho$ for each $t \leqslant s_i$, we see that

$$\frac{q(\omega^t|\bar{\omega}^{t-1})}{p(\omega^t|\bar{\omega}^{t-1})} \leqslant 1 + \frac{\tau}{\rho}.$$

So,

$$q(R) = \sum_{\bar{\omega}^{s_i} \in R} q\left(\bar{\omega}^t\right) = \sum_{\bar{\omega}^{s_i} \in R} \prod_{t=1}^{s_i} q\left(\omega^t \middle| \bar{\omega}^{t-1}\right) \leqslant \sum_{\bar{\omega}^{s_i} \in R} \prod_{t=1}^{s_i} \left(1 + \frac{\tau}{\rho}\right) p\left(\omega^t \middle| \bar{\omega}^{t-1}\right).$$

The right-hand side can be expressed as:

$$\begin{aligned} \sum_{\bar{\omega}^{s_i} \in R} \prod_{t=1}^{s_i} \left(1 + \frac{\tau}{\rho}\right) p\left(\omega^t \middle| \bar{\omega}^{t-1}\right) &= \left(1 + \frac{\tau}{\rho}\right)^{s_i} \sum_{\bar{\omega}^{s_i} \in R} \prod_{t=1}^{s_i} p\left(\omega^t \middle| \bar{\omega}^{t-1}\right) \\ &= \left(1 + \frac{\tau}{\rho}\right)^{s_i} p(R), \end{aligned}$$

from which the desired conclusion follows. In the situation where the test is applied to data generated in periods $t, \ldots, t + s_i$ for some $t > 0$ we replace $p(\bar{\omega}^{s_i})$ by the probability of observing the data conditional on $\bar{\omega}^{t-1}$, and the argument proceeds in the same way. □

Next we show that, for each tolerance τ, there is a value $c_i(\tau) \leqslant \tau$ such that

$$\alpha_{i,s_i,c_i(\tau)}, \beta_{i,s_i,\tau} \leqslant k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i/2}. \tag{3}$$

In other words, when the null is sufficiently far from the truth (more than τ), then one accepts with exponentially small probability, whereas when the null is sufficiently close to the truth (closer than $c_i(\tau)$) then one rejects with exponentially small probability.

Denote the minimum probability that i plays each action in each period by $h_i(\sigma_i)$, which is positive because the response function is diffuse. Every combination of actions is played with probability at least $\prod h_i(\sigma_i)$. Let

$$c_i(\tau) = \left[\prod_j h_j(\sigma_j)\right] r_i(\tau)/2. \tag{4}$$

By the previous lemma,

$$\alpha_{i,s_i,c_i(\tau)} \leqslant \left(1 + \frac{c_i(\tau)}{\prod h_i(\sigma_i)}\right)^{s_i} \alpha_{i,s_i}. \tag{5}$$

Since i employs a powerful family of tests, $\alpha_{i,s_i} \leqslant k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i}$. Thus

$$\alpha_{i,s_i,c_i(\tau)} \leqslant \left(1+\frac{c_i(\tau)}{\prod h_i(\sigma_i)}\right)^{s_i} k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i} \leqslant \left(1+\frac{r_i(\tau)}{2}\right)^{s_i} k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i}$$
$$\leqslant \mathrm{e}^{r_i(\tau)s_i/2}k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i} = k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i/2}.$$

Thus the inequality (3) holds for $\alpha_{i,s_i,c_i(\tau)}$. It also holds for $\beta_{i,s_i,\tau}$ because

$$\beta_{i,s_i,\tau} \leqslant k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i} \leqslant k_i(\tau)\mathrm{e}^{-r_i(\tau)s_i/2}.$$

Next we construct a uniform bound analogous to (3) that holds for all i. Namely, let

$$c(\tau) \equiv \min_i c_i(\tau), \qquad r(\tau) \equiv \min_i r_i(\tau)/2, \quad \text{and} \quad k(\tau) \equiv \max_i k_i(\tau).$$

Then

$$(\forall i) \quad \alpha_{i,s_i,c(\tau)}, \beta_{i,s_i,\tau} \leqslant k(\tau)\mathrm{e}^{-r(\tau)s_i}. \tag{6}$$

We shall say that two agents i and j *use comparable amounts of data* if there is a p, $1 < p < 2$, such that for every i and j, $s_i \leqslant s_j^p$. Note that this is a very weak condition since, if s_i and s_j are sufficiently large (which is the case of interest), the ratio of any two agents' sample sizes can be made arbitrarily large, and so can the ratio of the power of their tests.

If we define $s^* = \max_i s_i$, and $s_* = \min_i s_i$ then we can combine (3) and (6) to obtain the uniform bound

$$\alpha_{i,s_i,c(\tau)}, \beta_{i,s_i,\tau} \leqslant k(\tau)\mathrm{e}^{-r(\tau)s_*}. \tag{7}$$

5.3. Prediction

We shall say that player i is an *ϵ-good predictor* if the mean square error of his predictions is almost surely bounded above by ϵ as $t \to \infty$. Specifically, player i is an ϵ-good predictor if

$$\lim_{T\to\infty} \sup \frac{1}{T}\sum_{t=1}^{T}\left(b_i^t - B_i\left(\vec{a}^t\right)\right)^2 \leqslant \epsilon \quad \text{a.s.} \tag{8}$$

This condition implies that, for any set of actions x_{-i}, and any realization of the random process to date, the conditional probability that b_i^t assigns to the occurrence of x_{-i} in period t comes within ϵ of the actual probability of its occurrence, except possibly for a sparse set of times. (We remark that this condition is weaker than the one used by Kalai and Lehrer (1993) who require that, for every $\epsilon > 0$ there is a time t_ϵ such that the conditional probabilities come within ϵ of the actual behavior for all times $t \geqslant t_\epsilon$.)

6. Proof of the main result

Let $X = \prod X_i$ be a finite state space and write $G(\vec{u}, X)$ for the stage game on X defined by the utility functions $\vec{u} = \{u_1, \ldots, u_n\}$, $u_i : X \to \Re$. We shall say that a property of a learning process is *robust* if it holds independently of the payoff functions $\vec{u}$. Our first theorem gives some robust properties of hypothesis testers.

Theorem 1. *Suppose that the players adopt hypotheses with finite memory, have σ_i-smoothed best response functions, employ powerful hypothesis tests with comparable amounts of data, and are flexible in the adoption of new hypotheses. Given any $\epsilon > 0$, if the σ_i are small (given ϵ), if the test tolerances τ_i are sufficiently fine (given ϵ and σ_i) and if the amounts of data collected, s_i, are sufficiently large (given ϵ, σ_i and τ) then:*

(1) *The repeated-game strategies are ϵ-equilibria of the repeated game $G^{\infty}(\vec{u}, X)$ at least $1-\epsilon$ of the time.*
(2) *All players for whom prediction matters by at least ϵ are ϵ-good predictors.*

Thus the property of being an ϵ-equilibrium most of the time does not depend on the utility functions—it suffices that the responses are sufficiently sharp, the test tolerances are sufficiently fine, and the sample sizes are sufficiently large. Similarly, the property of being an ϵ-good predictor is robust for those players who care enough about prediction.

In preparation for proving Theorem 1, we recall some earlier notation. Let $\mathcal{A}_i$ be the set of responses for player i that have memory at most m. Let $\mathcal{B}_i \equiv \prod_{j \neq i} \mathcal{A}_j$ be the set of models for player i that have memory at most m. In general, $a_i \in \mathcal{A}_i$ will denote a response for i, whereas $b_i \in \mathcal{B}_i$ will denote a model. A vector of responses will be denoted by $\vec{a} \equiv (a_1, \dots, a_n)$, and the space of such response-tuples will be denoted $\mathcal{A} \equiv \prod_i \mathcal{A}_i$. Similarly, a vector of models will be denoted by $\vec{b} \equiv (b_1, \dots, b_n)$, and its associated space by $\mathcal{B} \equiv \prod_i \mathcal{B}_i$. A *hypothesis* for i is a pair $\nu_i = (a_i, b_i)$, where $a_i = A_i^{\sigma_i}(b_i)$.

If the players hold models $\vec{b}$ at time t, then the distributions generating the process are

$$A^{\vec{\sigma}} : \mathcal{B} \mapsto \mathcal{A},$$
$$A^{\vec{\sigma}}(\vec{b}) = \bigl(A_1^{\sigma_1}(b_1), \dots, A_i^{\sigma_i}(b_i), \dots, A_n^{\sigma_n}(b_n)\bigr).$$

Given a vector of responses $\vec{a}$, the *correct models* are given by the function

$$B : \mathcal{A} \mapsto \mathcal{B},$$
$$B(\vec{a}) = \left(\prod_{j \neq 1} a_j, \dots, \prod_{j \neq i} a_j, \dots, \prod_{j \neq n} a_j\right).$$

A *model fixed point* is a solution $\vec{b}$ of

$$B\bigl(A^{\vec{\sigma}}(\vec{b})\bigr) = \vec{b},$$

whereas a *response fixed point* is a solution of

$$A^{\vec{\sigma}}\bigl(B(\vec{a})\bigr) = \vec{a}.$$

Of course these are equivalent in the sense that $\vec{b}$ is a model fixed point if and only if $A^{\vec{\sigma}}(\vec{b})$ is a response fixed point, and $\vec{a}$ is a response fixed point if and only if $B(\vec{a})$ is a model fixed point. The existence of such fixed points follows from the fact that both maps are continuous functions from compact convex spaces to themselves.

Fix a small tolerance level $\tau > 0$. Given a model vector $\vec{b}$, say that $\vec{b}$ is *good for i* if b_i is within τ of being correct, that is,

$$\left|b_i - B_i\left(A^{\vec{\sigma}}(\vec{b})\right)\right| \leqslant \tau. \tag{9}$$

Otherwise $\vec{b}$ is *bad for i*.

The gist of the proof goes as follows. Choose a specific fixed point $\vec{b}^*$. We divide the players into two classes: the "unresponsive" players do not care much about prediction because their payoffs remain more or less the same independently of the model they have of their opponents; all other players are "responsive." Suppose that $\vec{b}$ is bad for some responsive player. Call this player the *leader* and without loss of generality, assume that his index number is 1. The leader rejects his current hypothesis with high probability after the next test (assuming that the others' models stayed at or close to $\vec{b}$ during his test phase). Since the leader is responsive, he has a model $w_1 = w_1(\vec{b})$ that invalidates all the other players' models of himself, namely,

$$(\forall j \neq 1) \quad \left|A_1^{\sigma_1}(w_1) - (b_j)_1\right| > \tau. \tag{10}$$

(The proof will have to establish the existence of such a "wrong" model w_1.) Suppose that after rejecting the model b_1, the leader chooses a new model in the vicinity of w_1. This causes every player $j \neq 1$ to reject his model b_j. When they reject, there is a positive probability that their new models will all lie very close to the chosen fixed point $\vec{b}^*$. We can set things up so that this situation invalidates 1's model w_1. Thus with positive probability, after 1's next test he too will adopt a model that is very close to b^*. If the new vector of models lies close enough to $\vec{b}^*$, then it is good. Further, if it is close enough to $\vec{b}^*$, there is an exponentially small probability that any player's model will be rejected after any given test. It follows that the model vector $\vec{b}^t$ is good for all responsive players a large fraction of the time. By choosing the test tolerances τ_i and the response parameters σ_i small enough, we can conclude that the process is an ϵ-equilibrium at least $1 - \epsilon$ of the time.

We now turn to the details. In particular we are going to show that we can bound the learning parameters such that for *every* game G on X the following statements hold at least $1 - \epsilon$ of the time:

(i) the responses are close to being a fixed point;
(ii) the responses are ϵ-optimal given the models;
(iii) the models are within ϵ of being correct for all those players who care about prediction by at least ϵ.

More precisely we shall establish the following.

Lemma 2. *Fix a finite action space $X = \prod_{i=1}^{n} X_i$. Given any $\epsilon > 0$, and any finite memory m, there exist functions $\sigma(\epsilon)$, $\tau(\epsilon, \sigma)$, $s(\epsilon, \sigma, \tau)$ such that if these functions bound the parameters with the same names, then at least $1 - \epsilon$ of the time t,*

(1) $|a^t - A^{\vec{\sigma}}(B(\vec{a}^t))| \leqslant \epsilon/2$.
(2) $|U_i^t(a_i^t, B_i(\vec{a}^t)) - \max_{a_i'} U_i^t(a_i', B_i(\vec{a}^t))| \leqslant \epsilon$ *for all i.*
(3) $|b_i^t - B_i(A^{\sigma_i}(\vec{b}^t))| \leqslant \epsilon$ *for every player i for whom prediction matters by at least ϵ.*

Proof. We will arbitrarily pick one fixed point of the mapping $A^{\sigma} \circ B$ and designate it by $\vec{a}^*$. It follows that $\vec{b}^* = B(\vec{a}^*)$ is a fixed point of $B \circ A^{\sigma}$.

Let us choose all σ_i such that

$$(\forall i) \quad \sigma_i \leqslant \epsilon/2. \tag{11}$$

Since each $A_i^{\sigma_i}(\cdot)$ is continuous as a function of b_i and as a function of the stage game payoffs u_i, both of which lie in compact domains, there is a $\delta > 0$ and such that

$$(\forall \vec{u}, t, i) \quad \left(\forall b_i, b_i'\right) \quad \left|b_i - b_i'\right| \leqslant \delta \quad \Rightarrow \quad \left|A_i^{\sigma_i}(b_i) - A_i^{\sigma_i}\left(b_i'\right)\right| \leqslant \frac{\epsilon}{2n} \tag{12}$$

and

$$\delta < \epsilon/2n. \tag{13}$$

Let $d_i > 0$ be the diameter of the image of $A_i^{\sigma_i}(\mathcal{B}_i)$ in the space of i's responses, $\mathcal{A}_i$. If

$$d_i > \delta, \tag{14}$$

we will say that player i is *responsive*; otherwise i is unresponsive. Notice that d_i may depend on the smoothing parameter, σ_i, and the stage game utility, u_i, hence responsiveness also depends on σ_i and u_i. The maximum change in payoff that an unresponsive player i can induce, either for himself or anyone else, is bounded above by δ because the payoffs $U_i^t(\cdot, \cdot)$ lie between 0 and 1. It follows that if prediction matters to a player by more than ϵ (which is greater than δ) then that player must be responsive.

To establish Lemma 2 we shall consider two cases.

Case 1. All players are unresponsive.

Proof. Then every possible response of every player i lies within $\delta < \epsilon/2n$ of the optimal response, so statement (2) of the lemma holds. Further, each player's utility varies by more at most $\epsilon/2n < \epsilon$, so all actions are ϵ-close to optimal and statement (2) holds. Finally, statement (2) holds vacuously because there are no responsive players.

This concludes the proof of Case 1. □

Case 2. At least one player is responsive.

Proof. Fix a small tolerance level $\tau > 0$ such that

$$\tau \leqslant \frac{\delta}{2(n+1)}. \tag{15}$$

As in the preamble to Eq. (9) we say that a model vector $\vec{b}$ is *good* for i if

$$\left|b_i - B_i\left(A^{\vec{\sigma}}(\vec{b})\right)\right| \leqslant \tau. \tag{16}$$

Otherwise $\vec{b}$ is *bad* for i. A model vector $\vec{b}$ is *all good* if it is good for all players. It is *fairly good* if it is good for all responsive players, and it is *bad* if it is not fairly good. The proof of this case is a consequence of the following two claims.

Claim 1. *If the model vector $\vec{b}^t$ is fairly good at least $1-\epsilon$ of the time, then all three statements of Lemma* 2 *follow.*

Claim 2. *The model vector $\vec{b}^t$ is fairly good at least $1-\epsilon$ of the time.*

Proof of Claim 1. Equations (15) and (13) imply that $\tau < \delta$ and $\delta < \epsilon/2n$, hence $\tau < \epsilon$. When $\vec{b}^t$ is fairly good, (16) implies that $|b_i - B_i(A^{\vec{\sigma}}(\vec{b}))| \leqslant \epsilon$ for responsive players. Responsive players include all players for whom prediction matters by more than δ and hence all players for whom prediction matters by at least ϵ (since $\epsilon > \delta$). This establishes statement (2) of the lemma.

Next we establish statement (2). For every responsive player i, (16) implies that $|b_i - B_i(A^{\vec{\sigma}}(\vec{b}))| \leqslant \tau < \delta$. It follows from (12) that $|A_i^{\sigma_i}(b_i) - A_i^{\sigma_i}(B_i(A^{\vec{\sigma}}(\vec{b})))| \leqslant \epsilon/2n$ for each responsive player, that is $|a_i - A_i^{\sigma_i}(B_i(a_i))| \leqslant \epsilon/2n$. For each unresponsive player, $|a_i - A_i^{\sigma_i}(B_i(a_i))| \leqslant \delta \leqslant \epsilon/2n$. Putting these together we get $|\vec{a} - A^{\sigma}(B(a))| \leqslant \epsilon/2$.

It remains to show statement (2). For each i, the response $A_i^{\sigma_i}(B_i(\vec{a}^t))$ involves at most a loss of $\sigma_i \leqslant \epsilon/2$ in utility as compared to a best response to $B_i(\vec{a}^t)$, say a_i^*. By statement (2) we know that $|a_i^t - A_i^{\sigma_i}(B(\vec{a}^t))| \leqslant \epsilon/2$. Hence the payoff difference between a_i^t and $A_i^{\sigma_i}(B_i(\vec{a}^t))$ is at most $\epsilon/2$, because the utility functions are bounded between 0 and 1. Hence $|U_i^t(a_i^t, B_i(\vec{a}^t)) - U_i^t(a_i^*, B_i(\vec{a}^t))| \leqslant \epsilon/2 + \epsilon/2 = \epsilon$. □

Proof of Claim 2. Recall from (7) that, for each τ, there is a real number $0 < c(\tau) < \tau$ such that players reject with exponentially small probability whenever their models are within $c(\tau)$ of the truth and they reject with probability close to one when their models are at least τ away from the truth. In particular, there exist functions $k(\tau)$ and $r(\tau)$ such that whenever a player's model is within $c(\tau)$ of the correct model, he rejects with probability at most

$$k(\tau)\mathrm{e}^{-r(\tau)s_*}.$$

Next, we claim that there is a value $\gamma > 0$ such that, for our chosen model fixed point $\vec{b}^* = B(\vec{a}^*)$,

$$(\forall i) \quad \left|b_i - b_i^*\right| < \gamma \quad \Rightarrow \quad \left|b_i - B_i\bigl(A^{\vec{\sigma}}(\vec{b})\bigr)\right| < c(\tau). \tag{17}$$

Indeed, for each i, $A_i^{\sigma_i} : \mathcal{B}_i \to \mathcal{A}_i$ is continuous on a compact domain so it is uniformly continuous. Hence there exists $\gamma > 0$ such that

$$(\forall i) \quad \bigl(\forall b_i, b_i' \in \mathcal{B}_i\bigr) \quad \left|b_i - b_i'\right| < \gamma \quad \Rightarrow \quad \left|A_i^{\sigma_i}(b_i) - A_i^{\sigma_i}\bigl(b_i'\bigr)\right| < \frac{c(\tau)}{2n}. \tag{18}$$

Let us also choose $\gamma < c(\tau)/2$. It follows from the preceding that

$$\begin{aligned}(\forall b_i \in \mathcal{B}_i) \quad \left|B_i\bigl(A_i^{\sigma_i}(b_i)\bigr) - B_i\bigl(A_i^{\sigma_i}(b_i^*)\bigr)\right| &= \left|B_i\bigl(A_i^{\sigma_i}(b_i)\bigr) - b_i^*\right| \\ &< \frac{(n-1)c(\tau)}{2n}.\end{aligned} \tag{19}$$

Hence

$$\left|B_i\bigl(A_i^{\sigma_i}(b_i)\bigr) - b_i\right| \leqslant \left|B_i\bigl(A_i^{\sigma_i}(b_i)\bigr) - b_i^*\right| + \left|b_i^* - b_i\right| < c(\tau). \tag{20}$$

This establishes (17).

The values of δ, τ and γ defined in (13), (15), and (17) will remain fixed for the rest of the proof.

Next we establish the existence of a "wrong" model for each responsive player i. Let $\vec{b}$ be a model vector. For each responsive player i, we can choose $n+1$ points in the image set $A_i^{\sigma_i}(\mathcal{B}_i)$ such that no two points are closer than 2τ where τ is defined in (15). Hence there exist at least two of these points, say a_i' and a_i'' such that

$$(\forall j \neq i) \quad \left|(b_j)_i - a_i'\right| \geqslant \tau \quad \text{and} \quad \left|(b_j)_i - a_i''\right| \geqslant \tau.$$

Further, one or both of a_i', a_i'' is at least τ from $(b_j^*)_i$. Say without loss of generality that $|(b_j^*)_i - a_i'| \geqslant \tau$. Then we define $w_i(\vec{b})$ to be any model b_i' such that $A_i^{\sigma_i}(b_i') = a_i'$ and call this a *wrong model for i given $\vec{b}$*.

Define a *great state* to be a state such that, for every player i, i's model is within γ of b_i^* and no player is currently in a test phase. Beginning at a time t such that the vector $\vec{b}^t$ is bad, consider the following sequence of events leading to a great state.

Step 1. A responsive player whose model is bad, say 1, starts a new test phase as soon as all current tests (including possibly his own) are completed. During 1's new test phase no other player starts a test. After 1's new test phase is completed, he rejects his current hypothesis and adopts a model within γ of $\omega_1(\vec{b}^t)$. (Duration: at most $2s^*$ periods.)

Step 2. In succession, players $2, 3, \ldots, n$ conduct tests on non-overlapping sets of data. (Player 1 does not test during this period.) At the end of each of these $n-1$ tests, the tester rejects and adopts a model that is within γ of his part of b^*. (Duration: at most $(n-1)s^*$ periods.)

Step 3. Player 1 starts a test phase and no other players test while it is in progress. He rejects this test and jumps to a point within γ of $\vec{b}_1^*$. (Duration: at most s^* periods).

Step 4. If the number of periods in Steps 1–3 is $T < (n+2)s^*$, no player begins a test for the next $(n+2)s^* - T$ periods.

The duration of the whole sequence is exactly $(n+2)s^*$.

We now need to compute the probability of a special sequence of this type occurring conditional on being in a bad state at the start of the sequence. Ignoring player 1's first (partial) test phase, each of the subsequent $n+1$ test phases ends with a rejection. We can choose the test parameters so that each rejection occurs with probability at least $1/2$. After rejection, a target of radius γ must be hit within the rejector's model space. These $n+1$ events have probability at least $(f_*/2)^{n+1}$, where $f_* = f_*(\gamma)$. In addition, each of the $n+1$ test phases must begin at a specific time and no other players can be testing during another's phase. The probability of this is at least

$$\left((1/s^*)(1-1/s_*)^{(n-1)s^*}\right)^{n+1}. \tag{21}$$

We can assume that all $s_i \geqslant 2$, so $(1-1/s_*)^{s_*} \geqslant 1/4$. Then (21) is bounded below by

$$\left((1/s^*)(1/4)^{(n-1)s^*/s_*}\right)^{n+1}. \tag{22}$$

Finally, all players must cease testing until $(n+2)s^*$ periods have elapsed from the start of the sequence. This event has probability at least

$$(1 - 1/s_*)^{n(n+2)s^*} \geqslant 1/4^{n(n+2)s^*/s_*}. \tag{23}$$

Putting all of this together there is a positive integer N such that the probability of Steps 1–4 is at least

$$4^{-Ns^*/s_*}(f_*/2s^*)^{n+1}. \tag{24}$$

By assumption $s^* \leqslant s_*^p$ for some $p \in (1,2)$, so $s^*/s_* \leqslant s_*^q$ where $q = p - 1 > 0$. Thus there are constants $\alpha, \beta > 0$ such that the probability of Steps 1–4 is at least $\alpha s_*^{(n+1)q} \mathrm{e}^{-\beta s_*^q}$. This establishes the following fact.

If the model vector $\vec{b}^t$ is bad, the probability of being in a great state at time $t + (n+2)s^*$ is at least

$$\eta \equiv \alpha s_*^{(n+1)q} \mathrm{e}^{-\beta s_*^q}. \tag{25}$$

Continuing with our proof of Lemma 2, suppose that the process is in a great state at time t. Let T be a large positive integer. We will compute the probability that, over the next T periods, no player rejects a test. By construction, each null hypothesis (a_i^t, b_i^t) is within $\tau_0 = c(\tau)$ of the truth. Hence the probability that a given player i rejects a test is at most

$$\alpha_{i,s_i,\tau_0} \leqslant k_0 \mathrm{e}^{-r_0 s_i} \leqslant k_0 \mathrm{e}^{-r_0 s_*}, \tag{26}$$

where $k_0 = k(\tau_0) > 0$ and $r_0 = r(\tau_0) > 0$.

Over the course of the next T periods at most nT/s_* tests are concluded. Hence the probability that any player rejects a hypothesis during periods $t+1, \ldots, t+T$ is bounded above by

$$(nT/s_*)k_0 \mathrm{e}^{r_0 s_*} < T\mathrm{e}^{-4rs_*} \tag{27}$$

where the inequality holds for all sufficiently large s_* and some $r > 0$.

We will now use what we know about bad states and great states to show that the fraction of times that the process is in a bad state is very small. Recall that, conditional on being in a bad state at time t, the probability of being in a great state by time $t + (n+2)s^*$ is at least η by expression (25). We also know that, conditional on being in a great state at time t', the probability of staying in good states (states where the model vector is all good) for at least T periods is at least $T\mathrm{e}^{-4rs_*}$. Letting $T = \mathrm{e}^{3rs_*}$, the probability of leaving a great state is at most

$$\mathrm{e}^{-rs_*}. \tag{28}$$

Starting from a given time t, let $\mathcal{E}$ be the event "the realized states in at least ϵT of the periods $t+1, \ldots, t+T$ are bad." Let $\mathcal{E}'$ be the sub-event of $\mathcal{E}$ in which no great state is realized before the last bad state, and let $\mathcal{E}'' = \mathcal{E} - \mathcal{E}'$. We shall bound the conditional probabilities of $\mathcal{E}'$ and $\mathcal{E}''$ from above independently of the state at time t.

If $\mathcal{E}'$ occurs, there are at least $\lfloor \epsilon T/(n+2)s^* \rfloor = k$ distinct times $t < t_1 < \cdots < t_k \leqslant t + T$ such that the following hold:

- $t_{j+1} - t_j \geqslant (n+2)s^*$ for $1 \leqslant j < k$,
- the state at time t_j is bad for $1 \leqslant j < k$,
- no great state occurs from t_1 to t_k.

By the preceding, the probability of this event is at most $(1-\eta)^{k-1} < \mathrm{e}^{-\eta(k-1)}$.

Note that since ϵ and n are fixed, for all sufficiently large s_* we have $k - 1 > \mathrm{e}^{2rs_*}$.

Since β is fixed and $q < 1$, (25) implies that $\eta > \mathrm{e}^{-rs_*}$ for all sufficiently large s_*. Thus,

$$P(\mathcal{E}') \leqslant (1-\eta)^{k-1} \leqslant \mathrm{e}^{-\eta(k-1)} \leqslant \mathrm{e}^{-\mathrm{e}^{rs_*}},$$

which can be made as small as we wish when s_* is large. In particular it can be made less than $\epsilon/2$.

If $\mathcal{E}''$ occurs, the process does *not* stay in good states for at least T periods after a great state, so from (28)

$$P(\mathcal{E}'') \leqslant \mathrm{e}^{-rs_*}. \tag{29}$$

This can also be made less than $\epsilon/2$ when s_* is sufficiently large. Putting all of this together we conclude that, for all sufficiently large s_*,

$$P(\mathcal{E}) = P(\mathcal{E}') + P(\mathcal{E}'') \leqslant \epsilon.$$

Now divide all times t into disjoint blocks of length T, and let Z_k be the fraction of bad times in the kth block. We have just shown that $P(Z_k \geqslant \epsilon) \leqslant \epsilon$ for all k. Hence

$$E(Z_k) \leqslant P(Z_k \geqslant \epsilon) \cdot 1 + P(Z_k < \epsilon) \cdot \epsilon \leqslant 2\epsilon.$$

It follows that the proportion of times that the process is in a bad state is almost surely less than 2ϵ. Rerunning the entire argument with $\epsilon/2$ yields the desired conclusion, namely, that $\vec{b}^t$ is fairly good at least $1-\epsilon$ of the time. This establishes Claim 2, from which Lemma 2 follows. □

Proof of Theorem 1. The first statement of Theorem 1 follows from statement (2) of Lemma 2. Likewise the second statement follows from statement (3) of Lemma 2. □

If we allow the learning parameters to depend on the payoff functions defining the stage game G, we can strengthen the conclusions of Theorem 1 as follows.

Theorem 2. *Let G be an n-person game on the finite state space X and let G^∞ be the infinitely repeated game of G. Given $\epsilon > 0$, there exist values of the learning parameters $\{\sigma_i, \tau_i, s_i\}$ such that*:

(1) *at least $1-\epsilon$ of the time the players' strategies are ϵ-close to being a subgame perfect equilibrium*;

(2) *all players for whom prediction matters are ϵ-good predictors.*

Proof. Let G be the stage game with utility functions $\vec{u}$. Fix $\epsilon > 0$. There clearly exists $\alpha > 0$ such that every player who cares about prediction cares about it by more than α. Let $\epsilon' = \min\{\alpha, \epsilon\}$. Lemma 2 (with ϵ' in place of ϵ) says that the learning parameters can be chosen so that everyone for whom prediction matters more than ϵ' is an ϵ'-good predictor, and hence an ϵ-good predictor. Since this includes every player for whom prediction matters at all, statement (2) of Theorem 2 follows at once.

We now establish the first statement of Theorem 2, namely, that play is close to being subgame perfect a large fraction of the time. Specifically, we are going to show that, at least $1-\epsilon$ of the time t, $\vec{a}^t$ is within ϵ of a memory-m subgame perfect equilibrium and $\vec{a}^t$ does not change for at least $1/\epsilon$ periods beginning at time t.

Let $\vec{A^0}(\vec{b})$ be the best response correspondence (a set valued function), that is, $\vec{a} \in \vec{A^0}(\vec{b})$ if and only if for every player i, a_i is a best response to b_i. Let $\mathcal{S}$ denote the set of fixed points of $\vec{A^0} \circ B$, that is, $\mathcal{S} = \{\vec{a}\colon \vec{a} \in \vec{A^0}(B(\vec{a}))\}$. Every $\vec{a} \in \mathcal{S}$ generates a subgame perfect equilibrium of the repeated game.

From statement (2) of Lemma 2, we know that $|a_i^t - A_i^{\sigma_i}(B(\vec{a}^t))| \leqslant \epsilon$ for all i at least $1-\epsilon$ of the time t. We also know that $A^{\vec{\sigma}} \to \vec{A^0}$ as $\sigma_i \to 0$, that is, every limit point of $A^{\sigma}(\vec{b})$ is in $\vec{A^0}(\vec{b})$. It follows that, by choosing all σ_i to be sufficiently small, we can ensure that $\vec{a}^t$ is within $2n\epsilon$ of $\mathcal{S}$ at least $1-\epsilon$ of the time t. (Note that this upper bound on the σ_i may be smaller than the bound $\sigma(\epsilon')$ established in Lemma 2, and in fact it may depend on the payoffs defining G.) By carrying out the argument with $\epsilon/2n$ in place of ϵ, we conclude that whenever $|a_i^t - A_i^{\sigma_i}(B(\vec{a}^t))| \leqslant \epsilon/2n$, $\vec{a}^t$ is within ϵ of the set $\mathcal{S}$ of subgame perfect equilibria at time t. It remains to be shown that $\vec{a}^t$ stays unchanged for at least $1/\epsilon$ periods.

In general call a time t *ϵ-steady* if no player changes his model for the next $1/\epsilon$ rounds following t. □

Claim 3. *If $s_* \equiv \min s_i > 2n/\epsilon^2$ then at least $1-\epsilon$ of the times are steady.*

Proof. The number of tests completed between t and $t+T$ is at most $n + T/s_* \leqslant n + T\epsilon^2/2$. By choosing T to be sufficiently large, the number of tests is less than $T\epsilon^2$. This means that at least $1-\epsilon$ of the times between t and $t+T$ must be followed immediately by at least $1/\epsilon$ times in which no player completes a test, and hence no player changes her model or her strategy. Thus at least $1-\epsilon$ of the times between t and $t+T$ must be steady, and so at least $1-\epsilon$ of all times are steady. This establishes the claim.

We have therefore shown that if all σ_i are sufficiently small, and all s_i are sufficiently large, the players will be playing close to an element of $\mathcal{S}$ at least $1-\epsilon$ of the time and be steady at least $1-\epsilon$ of the time. Hence they are close to subgame perfect and steady at least $1-2\epsilon$ of the time. By redoing the argument with ϵ cut in half, we obtain the desired result. □

7. Beliefs

Our results so far have been stated in terms of "models" and "responses." These are simplified versions of what is really going on in the repeated game. In particular, models are naive because they make no allowance for the possibility they will change when rejected by a test. Let us therefore define the *belief* of player i at time t_0 to be the true conditional probability that i assigns to the opponents' actions for all $t \geqslant t_0$, including conditional probabilities of future changes in i's own model. Notice that these conditional probabilities need not have memory m because the model a player holds in a given period may depend on the models he held in previous periods. The question is whether the players are behaving rationally, or almost rationally, with respect to their beliefs.

Conditional on player i's model not changing from period t to $t+1$, his one-step-ahead conditional probability distribution is identical with the one-step-ahead conditional probability distribution implied by his model. A player's model cannot change from t to $t+1$ unless he is at the end of a test phase and he rejects. Thus there are many times when player i's one-step-ahead conditional forecast is the same under both his model and under his beliefs. Furthermore, when a rejection occurs player i adopts a new hypothesis that is within λ_i of his previous hypothesis with probability at least $1-\lambda_i$. For any given discount factors, we can choose the λ_i small enough so that the difference in the expected discounted payoffs between the models and the beliefs are within $\epsilon/2$ of each other. (This uses the fact that the stage-game payoffs are bounded.) Thus, if the level of conservatism is high enough, a response that is an $\epsilon/2$-best reply to one of the distributions will be an ϵ-best reply to the other. It follows that at every time t each player's choice of response is an ϵ-best reply to his belief.

Corollary 1. *If the players are sufficiently conservative, have sufficiently sharp best responses and employ sufficiently powerful hypothesis tests with sufficiently fine tolerances, then at all times the hypothesis testing strategies are ϵ-best responses to their beliefs.*

8. Convergence in probability

In conclusion, we remark that we can obtain even sharper convergence results by allowing the parameters to change over time as in simulated annealing. Specifically, for a given game G and a given $\epsilon > 0$, Theorem 2 guarantees levels of the parameters $\sigma_i(\epsilon)$, $\tau_i(\epsilon)$, and $s_i(\epsilon)$ such that the strategies are ϵ-close to a subgame perfect equilibrium at least $1-\epsilon$ of the time. Now assume that at the end of each period, we tighten the parameters to the values $\sigma_i(\epsilon/2)$, $\tau_i(\epsilon/2)$, and $s_i(\epsilon/2)$ with some small probability $p(\epsilon) > 0$. Then we tighten the parameters again, using an even smaller probability $p(\epsilon/2)$ and so forth. These tightening probabilities can be chosen so that, after the kth tightening, the process has time to come within $\epsilon/2^k$ of the set of subgame perfect equilibria with probability at least $1-\epsilon/2^k$ before the next tightening occurs. In this manner we can construct a learning process such that the strategies come arbitrarily close to the set of subgame perfect equilibria an arbitrarily large fraction of the time from arbitrary initial conditions.

Theorem 3. *There exists a learning process based on hypothesis testing such that, at each point in time, the players' strategies are almost optimal given their beliefs and the strategies converge in probability to the set of subgame perfect equilibria.*

Acknowledgments

An earlier version of this paper appeared under the title "Learning with Hazy Beliefs" (1998b) The presentation greatly benefited from comments of an anonymous referee. We are also indebted to seminar participants at the University of Chicago, Northwestern University, Hebrew University, the University of Siena, Stanford, the Stockholm School of Economics, University College London and the NSF/NBER Decentralization Conference at Georgetown University. The second author's research was supported by NSF grant 9818975.

References

Brown, G.W., 1951. Iterative solutions of games by fictitious play. In: Koopmans, T.C. (Ed.), Activity Analysis of Production and Allocation. Wiley, New York, pp. 374–376.

Foster, D.P., 1999. A proof of calibration via Blackwell's approachability theorem. Games Econ. Behav. 29, 73–78.

Foster, D.P., Vohra, R.V., 1997. Calibrated learning and correlated equilibrium. Games Econ. Behav. 21, 40–55.

Foster, D.P., Vohra, R.V., 1998. Asymptotic calibration. Biometrika 85, 379–390.

Foster, D.P., Vohra, R.V., 1999. Regret in the on-line decision problem. Games Econ. Behav. 29, 7–35.

Foster, D.P., Young, H.P., 1998a. On the nonconvergence of fictitious play in coordination games. Games Econ. Behav. 25, 79–91.

Foster, D.P., Young, H.P., 1998b. Learning with hazy beliefs. In: Leinfellner, W., Koehler, E. (Eds.), Game Theory, Experience, Rationality. Kluwer, Amsterdam.

Foster, D.P., Young, H.P., 2001. On the impossibility of predicting the behavior of rational agents. Proc. Nat. Acad. Sci. USA 22, 12848–12853.

Fudenberg, D., Levine, D.K., 1995. Universal consistency and cautious fictitious play. J. Econ. Dynam. Control 19, 1065–1089.

Fudenberg, D., Levine, D.K., 1999a. Conditional universal consistency. Games Econ. Behav. 29, 104–130.

Fudenberg, D., Levine, D.K., 1999b. An easier way to calibrate. Games Econ. Behav. 29, 131–138.

Hart, S., Mas-Colell, A., 2000. A simple adaptive procedure leading to correlated equilibrium. Econometrica 68, 1127–1150.

Hart, S., Mas-Colell, A., 2001. A general class of adaptive strategies. J. Econ. Theory 98, 26–54.

Jehiel, P., 1995. Limited horizon forecast in repeated alternate games. J. Econ. Theory 67, 497–519.

Jehiel, P., 1998. Learning to play limited foresight equilibria. Games Econ. Behav. 22, 274–298.

Jordan, J.S., 1991. Bayesian learning in repeated games. Games Econ. Behav. 3, 60–91.

Jordan, J.S., 1992. Bayesian learning in games: a non-Bayesian perspective. Preprint. University of Minnesota.

Jordan, J.S., 1993. Three problems in learning mixed-strategy equilibria. Games Econ. Behav. 5, 368–386.

Jordan, J.S., 1995. Bayesian learning in repeated games. Games Econ. Behav. 9, 8–20.

Kalai, E., Lehrer, E., 1993. Rational learning leads to Nash equilibrium. Econometrica 61, 1019–1045.

Krishna, V., 1992. Learning in games with strategic complementarities. Mimeo. Pennsylvania State Univ., University Park, PA.

McKelvey, R., Palfrey, T., 1995. Quantal response equilibria for normal form games. Games Econ. Behav. 10, 6–38.

Milgrom, P., Roberts, J., 1991. Adaptive and sophisticated learning in normal form games. Games Econ. Behav. 3, 82–100.

Miller, R.I., Sanchirico, C.W., 1997. Almost everybody disagrees almost all of the time: the genericity of weakly-merging nowhere. Department of Economics, Columbia University. Working paper, 9697-25.

Miller, R.I., Sanchirico, C.W., 1999. The role of absolute continuity in "Merging of Opinions" and "rational learning." Games Econ. Behav. 29, 170–190.

Monderer, D., Shapley, L., 1996. Potential games. Games Econ. Behav. 14, 124–143.

Nachbar, J.H., 1997. Prediction, optimization, and learning in games. Econometrica 65, 275–309.

Nachbar, J.H. 1999. Rational Bayesian learning in repeated games. Working paper. Department of Economics, Washington University, St. Louis.

Nachbar, J.H., 2001. Bayesian learning in repeated games of incomplete information. Soc. Choice Welfare 18, 303–326.

Nyarko, Y., 1994. Bayesian learning leads to correlated equilibria in normal form games. Econ. Theory 4, 821–841.

Nyarko, Y., 1997. Savage-Bayesian agents play a repeated game. In: Bicchieri, C., Jeffrey, R., Skyrms, B. (Eds.), The Dynamics of Norms. Cambridge Univ. Press, Cambridge, UK.

Robinson, J., 1951. An iterative method of solving a game. Ann. Math. 54, 296–301.

Shapley, L.S., 1964. Some topics in two-person games. In: Dresher, M., Shapley, L.S., Tucker, A.W. (Eds.), Advances in Game Theory. Princeton Univ. Press, Princeton, NJ, pp. 1–28.