Learning with Hazy Beliefs

Dean Foster* and Peyton Young**

May, 1996

*Department of Statistics, Wharton School, Univ. of Pennsylvania, Philadelphia, PA 19104.
**Department of Economics, Johns Hopkins University, Baltimore, MD 21218.

Abstract

Players are rational if they always choose best replies given their beliefs. They are good predictors if the difference between their beliefs and the distribution of the others' actual strategies goes to zero over time. Learning is deterministic if beliefs are fully determined by the initial conditions and the observed data. (Bayesian updating is a particular example). If players are rational, good predictors, and learn deterministically, there are many games for which neither beliefs nor actions converge to a Nash equilibrium. We introduce an alternative approach to learning called *prospecting* in which players are rational and good predictors, but beliefs have a small random component. In any finite game, and from any initial conditions, prospecting players learn to play arbitrarily close to Nash equilibrium with probability one.

Learning with Hazy Beliefs

Can people learn to play equilibrium in a repeated game when they start with beliefs that are inconsistent with the facts? Kalai and Lehrer (1993) show that the answer is yes provided that players are rational, they update their beliefs by Bayes' rule, and their initial beliefs contain a "grain of truth" in the sense that they assign positive probability to the players' actual strategies. In this case each player learns to predict the others' strategies, and the empirical distribution of actions converges with probability one to an ε-Nash equilibrium.[1] In a different framework, Jordan (1991, 1995) argues that the answer to the learning question is "sort of. " Suppose the players have a common prior about the payoff functions and are sophisticated in the sense that in every period they play a Bayesian Nash equilibrium given the information revealed up to that point. Then, although neither the beliefs nor the empirical distribution of actions need converge, it is true that every limit point of the belief sequence is a Nash equilibrium of the game.[2]

On the opposite side of the question, Nachbar (1995) argues that learning by rational Bayesian players is highly problematical. Nachbar supposes that the payoff functions are common knowledge and the players' initial beliefs contain a "grain of reflection": if player i intends to use a strategy, then she should not rule out the possibility that player j will adopt a best reply to her strategy. In other words, the support of a player's beliefs should include all best replies to the player's own strategy. A player should also not rule out the possibility that another player will use a pure strategy when the latter is a best reply. In other words, players are not be forced to randomize. (Nachbar calls these kinds of beliefs "plausible.") Nachbar shows that for a large class of games this leads to an impossibility result: if players are rational, Bayesian,

---

[1] In fact, Kalai and Lehrer show that this result holds under the weaker assumption that players' beliefs assign positive probability to any measurable set of outcomes that have positive probability under the actual strategies. See Blume and Easley (1995) for a discussion of other sufficient conditions that lead to good prediction.

[2] For related work see Nyarko (1994).

and have plausible beliefs, then with probability one they will never come close to predicting the actual behavioral strategies.

In this paper we argue that there is an essential conflict between rationality, prediction, and learning even when players are *not* Bayesian. Specifically, if players always choose best replies to their beliefs, and if they are good predictors in the sense that beliefs eventually come close to predicting behavioral strategies, then *no matter how the beliefs are generated* there are many games in which the beliefs either converge to equilibrium in finite time, or they never converge. To put it differently, if the initial uncertainty about the others' strategies is not resolved in a finite period of time, then the players never come close to predicting the course of the game in an infinite period of time. This obviously puts a heavy burden on what the players must know at the outset of the game for them to "learn" to play an equilibrium.

In the second part of the paper we argue that this impossibility theorem arises because the world of a game is too complex. In Savage's terminology it is a large world rather than a small world (Savage, 1951; see also Binmore, 1991). In particular, players' strategies and beliefs live in a continuum, while the information revealed over the course of the game -- the discrete actions chosen by the players in each period -- lives in a countable set. Thus the history of play does not necessarily provide enough information to weed out false beliefs unless the beliefs are serendipitously chosen at the outset. It is like searching for a needle in an uncountable haystack by counting through the haystraws one-by-one: unless one has a pretty good idea where the needle is to begin with, the search will fail.

The way out of the difficulty is to smooth the search process by introducing additional random variation into the players' behavior. For example we could posit that players sometimes tremble or experiment with suboptimal actions given their beliefs.[3] This would require us to give up strict rationality unless we have a theory of optimal experimentation. A second solution is to retain rationality, but abandon the Bayesian idea that beliefs are determined with *certainty* given the data and the priors. This is the approach we shall

[3] Fudenberg and Kreps (1993) and Kaniovski and Young (1995) study variants of fictitious play with random experimentation in 2 x 2 games.

adopt here. We posit that beliefs are *hazy*, that is, they are random variables whose distribution is determined by the initial conditions and revealed information. By thus allowing uncertainty at two levels -- in the choice of action and in the formation of beliefs -- we can construct a learning rule that comes arbitrarily close to equilibrium with probability one in any finite game.

2. An impossibility theorem under deterministic learning

In this section we show why any learning rule based on deterministic beliefs is fragile. Let G be an n-person game with finite action spaces Xl, $X_2, \ldots, X_n$ and payoff functions $u_i: X \to R$, where $X = \prod X_i$. A one-period *outcome* is an n-tuple of actions $x \in X$. All actions are publicly observed. Let Δ_i denote the set of probability distributions over Xi, and let A denote the product set of mixed strategies: $\Delta = \prod \Delta_i$. A *belief* of player i will be denoted by $p_{-i} \in \Delta_{-i} = \prod_{j \neq i} \Delta_j$. A (mixed) *strategy* of i will be denoted by $q_i \in \Delta_i$, and a vector of strategies by $q \in A$. If we are given a vector $q \in A$ and wish to consider all components except i, we shall write $(q)_{-i}$.

Let us first consider the situation if G is played exactly once. Before the game is played, the state can be described by a family of n pairs $\{(p_{-i}, q_i)\}_{1 \leq i \leq n}$ where p_{-i} is player i's belief and q_i is i's strategy. *Rationality* says that each player chooses only best replies given his beliefs, that is,

$$\forall i, \qquad q_i(x_i) > 0 \Rightarrow x_i \in BR(p_{-i}). \tag{R^0}$$

By itself this does not take us very far -- some further behavioral assumption is necessary to obtain an equilibrium. This is provided by the idea of prediction: i is a *good predictor* if his beliefs about the others' coincide with their actual plans:

$$\forall i, \qquad | p_{-i} - (q)_{-i} | = 0. \tag{P^0}$$

Any family $\{(p_{-i}, q_i)\}_{1 \leq i \leq n}$ that satisfies both (R^0) and (P^0) corresponds to a Nash equilibrium of the game; conversely, any Nash equilibrium can be described in this manner.

With these ideas in hand, let us now consider the dynamic case. Let G^∞ be the infinitely repeated game of G. A *history of length* $t \geq 0$ is a sequence of t outcomes drawn from X: $h^t = (x^1, x^2, \ldots, x^t)$. We let h^0 denote the null history. Let H^t be the set of all length-t histories, and let $H = \cup H^t$. A *behavioral strategy* for player i is a function from histories to strategies. Since the strategy also depends on the payoff function, we shall write $g_i: U \times H \to \Delta_i$, where U is the set of all payoff functions on X. Let Γ_i denote the set of behavioral strategies for i. A *belief* by player i is a probability distribution on $\prod_{j \neq i} \Gamma_j$. Any such belief can be represented by a function $f_i: U \times H \to \Delta_{-i}$, where $(f_i(h^t))_j$ is the probability that i assigns to j's actions in period t + 1, given the history h^t up to that point. Note that the beliefs are built on the assumption that the other players' strategies are independent of one another. This assumption is justified because the strategies are independent in fact.

Suppose now that the players are Bayesian. Let π^0_i be i's prior at the beginning of the game about the behavioral strategies the others are going to use. As above we can represent π^0_i by a function $f^0_i: U \times H \to A_{-i}$. Consider the situation at the end of period t, when i has observed the history h^t. What is i's posterior belief about j's behavior in period t + 1? The answer is $f^0_i(h^t)$ -- the product probability distribution that i *originally believed* would describe the others' actions in period t + 1, conditional on the history h^t actually occurring (see Kalai and Lehrer, 1993). As we shall soon see, the essential point here is not that players use Bayes' rule to update their beliefs, but that their beliefs at each period are *fully determined* by the observed history and the initial conditions.

Following Jordan (1992) we say that a *learning process* (with deterministic beliefs) is a family of 2n functions $(f_i, g_i\}_{1 \leq i \leq n}$ such that

$$f_i: U \times H \to \Delta_{-i} \text{ and } g_i: U \times H \to \Delta_i. \qquad (D)$$

A *learning path* is an infinite sequence $\{(h^t, p^t_{-i}, q^t_i): t = 1, 2, \ldots\}$ where $p^t_{-i} = f_i(h^{t-1})$ is player i's belief at time t and $q^t_i = g_i(h^{t-1})$ is i's strategy at time t. Note that a learning path is not fully observed, only the histories h^t are.

We assume that the players are *myopically rational:* in each period they choose strategies that maximize their expected payoff given their beliefs about the others' strategies in that period.[4] Thus every learning path $\{(h^t, p^t_i, p^t_{-i})\}$ satisfies

$$\forall t,\ q^t_i(x_i) > 0 \Rightarrow x_i \in BR_i(p^t_{-i}). \tag{R}$$

We say that player i "learns to predict" the behavior of j if his forecast of j's probability distribution over actions becomes better and better over time. A strong way of stating this condition is to require that for all distinct i and j, $\lim_{t\to\infty} | (p^t_{-i})_j - q_j{}^t | = 0$.[5] A somewhat weaker version -- which is still strong enough for our purposes -- is that the mean square error in prediction go to zero:[6]

$$\forall i, \qquad \lim_{T\to\infty} \sum_{t=1}^{T} (p^t_{-i} - (q^t)_{-i})^2/T = 0. \tag{P}$$

A simple example of a learning process is fictitious play. Here the belief of player i at time t is simply the empirical frequency distribution of actions taken by the others up through time t - 1, and i's strategy is a distribution over the set of best replies to these beliefs. Hence the process is by definition rational. For many games, fictitious play does not converge either in beliefs or in strategies (Shapley, 1964; Jordan, 1993; Foster and Young, 1995). Moreover, even if the beliefs converge, they need not be a good predictor of the behavioral strategies being used by the players at any given time. (An exception occurs when play piles up on some pure Nash equilibrium.) We are going to argue that one or more of these problems is going to arise with *any* deterministic learning rule.

---

[4] An alternative model is that players maximize their expected discounted payoffs over the future course of the game, given their beliefs about the others' behavior in each future period. The justification for myopic optimization is that the players are learning to play a one-shot game without knowing in advance that they are in fact going to play it infinitely often.

[5] Kalai and Lehrer (1993) and Nachbar (1995) use a similar condition in the context of discounted payoffs.

[6] In fact, propositions 1 and 2 below hold under the still weaker assumption that $\liminf_{T\to\infty} \Sigma(p^t_{-i} - (q^t)_{-i})^2/T = 0$.

Suppose that, at the beginning of the process, the payoffs are private information and are drawn from a density μ that is common knowledge to the players. We can think of each payoff function u_i as a point in $R^{|X|}$, and the vector of payoff functions u as a point in $R^{n|X|}$; thus μ is a density on $R^{n|X|}$. We shall suppose that the payoffs are independently distributed, $\mu = \prod \mu_i$, and that they are absolutely continuous with respect to Lebesgue measure.

Assume now that the players are Bayesian, and that μ is their common prior. Since the distributions μ_i are independent, i's actual payoff function gives i no new information about the payoffs of anyone else. What then will player i forecast about j's behavior in period 1? It seems implausible that i would condition his forecast on information (namely u_i) that i knows j cannot infer from his current information except through the prior μ. Moreover, i cannot condition his forecast on information (namely $u_j, j \neq i$) that he himself cannot infer except from μ. This means that i's forecast p^1_{-i} depends only on μ. The same reasoning extends to every player in every subsequent period t: i will base his forecast of j's behavior only on the publicly observed data h^{t-1} and the common prior μ. Note that this is a plausibility assumption only: a Bayesian player *could* make his forecast of others' behavior contingent on his own payoff; we are only saying that it does not seem reasonable to do so under the circumstances.

A learning process $\{\mu, f_i, g_i\}$ satisfies *independence* (I) if, whenever the payoffs are independently distributed, then the beliefs are independent of the realized value of the payoffs, that is, $f_i(u, h^{t-1}) = f_i(h^{t-1})$ for all initial histories h^{t-1}. (Note that this condition is not restricted to Bayesian learning.) We say that beliefs *converge* along a learning path $\{h^t, p^t_{-i}, q^t_i\}$ if there exists $p \in A$ such that $p^t_{-i} \to (p)_{-i}$ for all i.

PROPOSITION 1. *Let* $\{\mu, f_i, g_i\}$ *be a learning process such that* $\mu = \prod \mu_i$ *is absolutely continuous with respect to Lebesgue measure. Assume that the process is predictive (P), rational* (R), *independent t* (I), *and deterministic in beliefs* (D). *With p-probability one, every learning path either converges in strategies to a pure Nash equilibrium, or does not converge in either beliefs or strategies.*

COROLLARY. *Let* $\{\mu, f_i, g_i\}$ *be a process satisfying the conditions of theorem 1, and suppose that every game in the support of* μ *has only mixed strategy equilibria. Then* no *learning path* converges in *either* beliefs *or strategies.*

Proof. Let $\{(\mu, f_i, g_i)\}$ be a PRID process with payoffs distributed according to $\mu = \prod \mu_i$, where μ (and hence each Pi) is absolutely continuous with respect to Lebesgue measure. By (I) the functions f_i may depend on $\mu(u)$, but they do not depend on the realized value of the random variable u.

Fix a player i. Let X_{-i} denote the set of all action-combinations that players other than i can take. For each $p_{-i} \in \Delta_{-i}$ and $x_{-i} \in X_{-i}$, write $p_{-i}(x_{-i}) = \prod_{j \neq i}(p_{-i})_j((x_{-i})_j)$. Given the payoff functions u, $u_i(x)$ is the payoff to i associated with the action tuple $x \in X$. Given u and $p_{-i} \in \Delta_{-i}$ we say that u_i *ties at* p_{-i} if there exist xi, $y_i \in X_i$ and $x_{-i} \in X_{-i}$ such that $x_i \neq y_i$ and

$$\sum_{x_{-i} \in X_{-i}} u_i(x_i, x_{-i}) p_{-i}(x_{-i}) - \sum_{x_{-i} \in X_{-i}} u_i(y_i, x_{-i}) p_{-i}(x_{-i}) = 0. \qquad (1)$$

Let $L(p_{-i})$ denote the set of all $u_i \in R^{|X|}$ that tie at p_{-i}. From (1) it is clear that $L(p_{-i})$ is a finite union of hyperplanes, one for each distinct pair xi, $y_i \in X_i$. Let

$$L_i = \bigcup_{p_{-i} \in \text{Range } (f_i)} L(p_{-i}). \qquad (2)$$

Since H is countable, the range of f_i is countable, so L_i is a countable union of hyperplanes in $R^{|X|}$. By assumption, μ_i is absolutely continuous with respect to Lebesgue measure, so $\mu_i(K) = 0$ for any hyperplane K in $R^{|X|}$. Thus L_i is a countable union of sets having μ_i-measure zero, hence L_i itself has μ_i-measure zero.

Let $U^0_i = (\text{supp } \mu_i) - L_i$ for each i, and let $U^0 = \prod U^0_i$. From the preceding we know that $\mu(U^0) = 1$.

A Nash equilibrium $p^* \in A$ is *isolated* if there exists an open neighborhood of p^* that contains no Nash equilibrium other than p^*. Let U^* be the set of all $u \in U$ such that all Nash equilibria of u are isolated. Clearly $\mu(U^*) = 1$.

Fix $u \in U^*$ and consider any learning path $\{h^t, p^t_{-i}, q^t_i\}$ generated by the process when the payoff functions are $u = (u_1, u_2, \ldots, u_n)$. By choice of u, there are no payoff ties produced by the players' beliefs at any time in the history. Rationality therefore implies that each player's strategy q^t_i puts probability one on the *unique* action $x^t_i \in X_i$ that constitutes i's best repy to p^t_{-i}. It follows that there exists at least one $x \in X$ such that

$$\limsup_{T \to \infty} I\{t \leq T: x^t = x\} I / T \geq 1 / |X 1. \tag{3}$$

Fix such an x and consider some player i. Let $\mathcal{T}$ be the set of all times at which x is played. The predictive property (P) implies that there is an infinite subset $\{t(1), t(2), \ldots\} \subset \mathcal{T}$ such that $(p^{t(k)}_{-i})_j \to 1_{x_i}$ for every $j \neq i$. In each of these periods i plays x_i, so x_i is a best reply by i to each of the vectors $p^{t(k)}_{-i}$. It follows from the continuity of the payoff function that x_i is a best reply by i to $\lim_{k\to\infty} p^{t(k)}_{-i}$, that is, x_i is a best reply by i to x_{-i}. Since this argument holds for all i, we conclude that x is a pure Nash equilibrium of the game, and that it is played often in the sense of (3).

Now suppose that the beliefs converge. By the preceding, $(p^t_{-i})_j$ must converge to 1_{x_j} for every distinct i and j. The action x^t_i of player i at time t is a best reply to the beliefs p^t_{-i}, which (by the preceding) converge to the pure strategy actions xi. By the continuity of the payoff function, there must exist some finite time T such that for all $t \geq T$, x^t_i is a best reply to x_{-i}. Since all Nash equilibria are isolated, it follows that $x^t_i = x_i$ for all $t \geq T$. Hence if the beliefs converge, then both strategies and beliefs converge to a pure Nash equilibrium x.

Suppose instead that the strategies converge. By choice of u, strategies are pure at all times, so we have $x^t = x^*$ for some $x^* \in X$ and all $t \geq t^*$. By the predictive property, there is a subsequence t(l), t(2), . . . such that $(p^{t(k)}_{-i})_j \to IX_j$ for every i and j as $k \to$ CO. From this and rationality it follows that x^* is a pure Nash equilibrium of G. Thus if the strategies converge, they converge to a Nash equilibrium of G.

To sum up: if the strategies converge along a learning path, they converge to a pure Nash equilibrium. If the strategies do not converge, the beliefs do not converge either. Thus, if the game has no pure Nash equilibrium then neither the strategies nor the beliefs converge along any learning path. This concludes the proof of proposition 1 and its corollary.

Proposition 1 is connected with previous results of Jordan (1992) and Nachbar (1995); indeed it can be viewed as a marriage of their approaches. Jordan considers the set of all finite n-person games defined on a given strategy space X, and *fixes* a learning process $\{(f_i, g_i)\}$ that is independent of the payoff functions u. Using a counting argument, he shows that, except for a set of games having Lebesgue measure zero, the best replies along any infinite history are uniquely determined for all players. It follows that, if the game has only mixed Nash equilibria, the actions cannot be converging to an equilibrium.

One limitation of this result is the assumption that the learning rule is fixed: why don't the players condition their learning on the payoff information they have at the beginning of the game? The answer we have suggested is independence: *if* payoffs are independently distributed ex ante, then such conditioning is not plausible. Moreover, Jordan's result leaves open the possibility that the *beliefs* could be converging to equilibrium. Proposition 1 shows, however, that this does not happen (except on a set of measure zero) when players predict and are rational. As in Nachbar (1995), it is the *combination* of these two properties that produces the impossibility theorem.

How does this negative result square with the positive results of Kalai-Lehrer and Jordan mentioned in the introduction? In the Kalai-Lehrer model players are rational and learn to predict, but the ability to do so hinges on a fortuitous alignment of their initial beliefs.[7] Jordan (1991, 1995) shows that if players are sophisticated Bayesians (which presupposes coordination on a Bayesian equilibrium in each period), then every convergent subsequence of beliefs converges to a Nash equilibrium. But the beliefs need not converge, and in general the players will fail to predict in our sense. (In fact the

[7]The Kalai-Lehrer model assumes discounted payoffs, but as we shall show in a companion paper, our results can be extended to this case.

empirical frequency distribution may not converge either.) The bottom line is that even if one assumes a high order of rationality, plus a common knowledge of rationality, Bayesian players may not be good learners unless their beliefs are reasonably consistent with each other to begin with.

3. A second impossibility result in learning.

Before turning to our main result (which is a positive one), let us consider a bit more carefully what is driving the negative result. Proposition 1 relies on four assumptions: prediction, rationality, independence, and deterministic beliefs. As we noted earlier, rationality and prediction are the keystones of equilibrium even in the one-shot case. To jettison either one of them undermines the whole idea of equilibrium behavior. Moreover, if players are not trying to predict (period by period) what the others are going to do, then the notion of belief does not make much sense. This leaves independence and determinism. Ultimately we are going to conclude that determinism is the culprit, and that we can get all of the other properties (prediction, rationality, and independence) plus learning to play Nash if only we give it up. First, however, we shall argue that independence is (like Nachbar's "grain of reflection") mainly an assumption of convenience and not the source of the difficulty.

An infinite sequence $p^t_{-i} \in \Delta_{-i}$ is *almost constant* if there exists a $p^*_{-i} \in \Delta_{-i}$ such that p^t_{-i} is almost always equal to p^*, that is,

$$\lim_{T \to \infty} (1/T)\, I\{t \leq T: p^t_{-i} \neq p^*_{-i}\}\, I = 0. \tag{4}$$

PROPOSITION 2. *Let G be a finite two-person game with known payoff functions* u *and a unique Nash equilibrium* $p^* = (p^*_1, p^*_2)$, *which has full support on* $X_1 \times X_2$. *Thus* $\mu = 1_u$ *is a point mass at* u. *Let* $\{\mu, f_i, g_i\}$ *be a learning process that is rational* (R), *predictive* (P), *and deterministic* (D). *Along any learning path* $\{h^t, p^t_{-i}, q^t_i\}$, *either the* beliefs *do not converge or they are almost constant with value* p^*.

The essence of this result is that if there is *any* initial uncertainty or misspecification about what strategies the other players are going to follow, it must be entirely resolved in a finite number of periods or else the players' beliefs will never converge. In particular they will not converge to a Nash equilibrium. This means that there cannot have been very much uncertainty or misspecification to begin with -- the players must have more or less guessed at the beginning of the game what the other players are going to do.

It might be objected that it is not very difficult for the players to guess the equilibrium in a game that has only one equilibrium. To this we answer that we have in mind a situation where there is some initial lack of information -- about the others' payoffs, or their behavioral strategies, or perhaps their rationality -- that we do not model explicitly and which makes it difficult for the players to guess exactly what the others are going to do. Moreover, even if they have full information, the fact that the equilibrium is mixed means that the players can never be sure at any finite time that the others actually are playing the equilibrium. The message of proposition 2 is that, whether the 'players are Bayesian or not, a learning process that is predictive, rational, and deterministic is only certain to converge to a Nash equilibrium if there is relatively little uncertainty at the outset.

Proof of Proposition 2. Let $p^* \in \Delta$ be the unique Nash equilibrium of G, and consider any $p \neq p^*$, $p = (p_1, p_2)$. Let $BR_1(p_2)$ denote the set of best replies by player 1 to p_2, and $BR_2(p_1)$ the set of best replies by player 2 to p_1. Suppose that $BR_1(p_2) = X_1$. By hypothesis p^* has full support, which implies that $BR_2(p^*_1) = X_2$. Thus (p^*_1, p_2) is a Nash equilibrium different from p^*, a contradiction. A similar argument holds for player 2. We therefore conclude that

$$p \neq p^* \Rightarrow BR_1(p_2) \neq X_1 \text{ and } BR_2(p_1) \neq X_2. \qquad (5)$$

Let $\{(h^t, p^t_{-i}, q^t_i)\}$ be a learning path and suppose that the beliefs converge. Thus there are vectors p_{-1}, p_{-2} such that $p^t_{-1} \to p_{-1}$ and $p^t_{-2} \to p_{-2}$. Assume that (4) does not hold for one or both of the players, say player 1. Then

$$\limsup_{T \to \infty} (1/T)\, I\{t < T: p^t_{-1} \neq p^*_{-1}\}\, I = r > 0.$$

From this and (5) it follows that there exists $x_1 \in X_1$ such that

$$\limsup_{T \to \infty} (1/T)\, I\{t < T: x_1 \notin BR_1(p^t_{-1})\} I \geq r / |X_1| > 0.$$

Rationality implies that 1's strategy put zero weight on x_1 at all of those times t such that $x_1 \notin BR_1(p^t_{-1})$. Good prediction implies that 2's forecast about 1's behavior puts increasingly small weight on x_1 at almost all of these times too. Since 2's beliefs converge as $t \to \infty$, we conclude that p_{-2} puts zero weight on x_1. Since p^* has full support, it follows that (p_{-2}, p_{-1}) is different from p^*. In particular, since p^* is the unique Nash equilibrium of the game, we know that (p_{-2}, p_{-1}) is *not* a Nash equilibrium. Hence one or both of the following hold:

$$\exists\, y_1 \in X_1 \text{ such that } p_{-2}(y_1) > 0 \text{ and } y_1 \notin BR_1(p_{-1}), \tag{6}$$

$$\exists\, y_2 \in X_2 \text{ such that } p_{-1}(y_2) > 0 \text{ and } y_2 \notin BR_2(p_{-2}). \tag{7}$$

Without loss of generality assume (6). Since 2 is a good predictor, 2's beliefs about player 1 converge in mean square to 1's actual strategy. Hence $q^t_1(y_1) > 0$ for almost all sufficiently large t. Since 1 is rational, $y_1 \in BR_1(p^t_{-1})$ for all t. This together with the continuity of 1's payoff function and the fact that $p^t_{-1} \to p_{-1}$, imply that $y_1 \in BR_1(p_{-1})$, which contradicts (6). A similar contradiction is obtained if (7) holds. This completes the proof of proposition 2.

We remark that this proposition fails when there are three or more players. To see why, consider the following three-person version of matching pennies due to Jordan (1993). "Player 1 seeks to match player 2, player 2 seeks to match player 3, and player 3 seeks to *mismatch* with player 1. In particular, each player i is concerned only with predicting the actions of player i + 1 (mod 3). It is easily seen that the unique Nash equilibrium requires each player to mix equally between *heads* and *tails.*" Jordan (1993, p. 374). Suppose that each player i predicts that player i + 1 (mod 3) will play heads and tails with probability 50-50. This leaves player i indifferent between heads and tails; let

us assume that player i decides to randomize 50-50. Suppose further that player i predicts that player i - 1 (mod 3) will play heads with probability 50 + 1/t and tails with probability 50- I/t. These are deterministic forecasts so (D) holds. The forecasts converge to 50-50 for each player, which by assumption is each player's behavioral strategy, so (1?) holds. Finally, each is playing a best reply, so (R) holds. But the beliefs are *not* almost constant, that is, the conclusion of proposition 2 fails.

4. Prospecting for equilibrium

We now propose a class of learning rules called *prospecting rules* that satisfy rationality and good prediction, and are capable of discovering a Nash equilibrium under arbitrary initial conditions. Quite apart from the fact that prospecting works, we believe it represents a plausible model of how people react to complex learning environments. The idea is best illustrated by an everyday example. Consider a fisherman who is deciding where to search for the day's catch. He begins with some generalized belief about the areas most likely to contain fish, and anchors his boat in one of them. He then casts his line about at random to see whether the fish are biting. If they are he continues fishing there, if not he moves on to another area.

This familiar type of search rule has three noteworthy elements. First, the fisherman's beliefs about the likelihood of catching fish at each particular *point* are rather hazy; hence he casts about at random within a local area. Second, if the results do not pass some reality test after a period of experimentation, he moves on. Third, he is ultimately willing to try every area even though initially he may have assigned it a low probability of being productive. (If he ruled out some areas altogether, this is surely where the fish would be.) A similar story could be told for people prospecting for oil, searching for lost items, and so forth. But fishing -- like playing a game -- has the complicating feature that the environment is constantly changing because of others' responses to one's own actions.

In general, let S be a compact subset of a metric space, and let v be a probability distribution on S that is continuous and has full support. This plays the role of the prospector's initial belief about the areas most likely to yield a high

payoff, where the notion of "payoff" depends on the context. The learning process proceeds in successive *rounds* r = 1, 2, 3, . . . that involve increasingly clear beliefs, but never perfectly clear ones.

1. Choose a point $x \in S$ using v. (This is the wide-area search phase.)

2. Choose a site at random from a neighborhood of x using a local probability distribution ω_r whose variance is σ_r^2. Do this for $1/\sigma_r^3$ periods in succession. In each period thereafter go to step 3 with probability σ_r, and with probability 1 -or repeat step 2. (This is the local-area search phase.)

3. Compute the average "payoff" or "yield" from the sites examined in the last $1/@$ periods. If it passes a threshold test (which may depend on r) go to step 4a; if it fails the threshold test go to step 4b. (This is the test phase.)

4a). With probability π_r increase r by one; otherwise keep r fixed. Go to step 2.
4b). With probability π_r increase r by one; otherwise keep r fixed. Go to step 1.

Any algorithm of this general form will be called a *prospecting rule*, where the parameters σ_r and π_r approach zero at rates that depend on the structure of the problem at hand.

We now specialize this idea to games, and show that with appropriately defined parameters (which do not depend on the game), prospecting players learn to play arbitrarily close to a Nash equilibrium with probability one in any finite game.

Given is an n-person game with payoff functions $u = (u_1, u_2, \ldots, u_n)$. Each person knows his own payoff function u_i and may or may not know the payoffs of the others. As we shall see, the learning process is not conditioned on the prior distribution of P(u), hence we shall drop it from the notation. (Further, the algorithm will converge for all games, not just for a subset of p-measure one.)

Fix a best reply mapping for each player, that is, a single-valued function $B_i : \Delta_{-i} \to \Delta_i$ from i's beliefs at any given time to a one-shot (possibly mixed)

strategy at that time, where the strategy puts positive weight only on actions that are best replies to the beliefs:

$$q_i = B_i(p_{-i}) \text{ and } q_i(x_i) > 0 \text{ implies } x_i \text{ maximizes } \sum_{x_{-i} \in X_{-i}} u_i(x_i, x_{-i}) p_{-i}(x_{-i}). \tag{8}$$

We can think of the "beliefs" of each player as having two components. One is the density ν_i on $\Delta_{-i} = \prod_{j \neq i} \Delta_j$ that. governs where i looks in the wide-area search phase; the other is the random belief generated in the local-area search phase. For the time being we shall fix ν_i; later we shall see how to modify ν_i to incorporate new information that arises during the learning process. It is assumed throughout that ν_i is continuous and has full support on Δ_{-i} for every i, $1 \leq i \leq n$.

The following terminology will be helpful. A *round* is a sequence of periods during which r does not change in step 4. During a given round r, each wide-area search by i yields a core *belief* p^*_{-i} surrounded by a *haze* that is represented by a local distribution $\omega_{i,r}(p_{-i} \mid p^*_{-i})$. (In fisherman's parlance, the beliefs are temporarily "anchored" at p^*_{-i}.) We assume that, for every $p^*_{-i} \in \Delta_{-i}$, the hazy beliefs have the following regularity properties:

i) for every $p^*_{-i} \in \Delta_{-i}$, $\omega_{i,}(p_{-i} \mid p^*_{-i})$ converges (as $r \to 00$) to the degenerate distribution with mass one at p^*_{-i};

ii) $O < \operatorname{var} \omega_{i,r} \leq \sigma_r^2$;

iii) $\omega_{i,r}$ is Lipshitz in the L_1-norm with a constant that is at most $O(1/\sigma_r)$, that is, for some constant L and every $p'_{-i}, p''_{-i} \in \Delta_{-i}$,

$$\|\omega_{i,r}(\cdot \mid p'_{-i}) - \omega_{i,r}(\cdot \mid p''_{-i})\|_1 = \int_{\Delta_{-i}} |\omega_{i,r}(p_{-i} \mid p'_{-i}) - \omega_{i,r}(p_{-i} \mid p''_{-i})| \, dp_{-i}$$
$$< (L/\sigma_r) \mid p'_{-i} - p''_{-i} \mid . \tag{9}$$

A canonical example is the uniform distribution on $S(p^*_{-i}) \cap \Delta_{-i}$, where $S(p^*_{-i})$ is the sphere of radius σ_r around p^*_{-i}. Note that we cannot expect the Lipshitz constant to be independent of σ_r, since $\operatorname{var} \omega_{i,r} \leq \sigma_r^2$. The bound

L / σ_r says that the density does not change too rapidly given that the variance of $\omega_{i,r}$ is at most σ_r^2.

Assume that each of the players employs a prospecting rule with the parameters

$$\sigma_r = 2^{-r} \text{ and } \pi_r = 2^{-(1/\sigma_r)}. \tag{10}$$

The learning proceeds independently across players (except insofar as they are reacting to the others' previous decisions). Let $r^t(i)$ denote the round that player i is in at time t. At the beginning of the process $r^0(i)= 1$ for every i. (In what follows we shall often suppress mention of t for notational convenience.) Each player i employs the following rule.

1. Pick p^*_{-i} at random from Δ_{-i} using the density ν_i. (This is the wide-area search.)

2. Pick p_{-i} at random from Δ_{-i} using the density $\omega_{i,r(i)}(p_{-i} \mid p^*_{-i})$, and choose a best reply according to the (mixed) strategy $B_i(p_{-i})$. Do this for $(1/\sigma_{r(i)})^3$ periods in succession, making independent draws from $\omega_{i,r}$ each time. Then go to step 3 with probability $\sigma_{r(i)}$, and repeat step 2 with probability $1 - \sigma_{r(i)}$. (This is the local-area search.)

3. Compute the empirical frequency distribution $Q_j(x_j)$ of actions taken by player j over the most recent $(1/\sigma_{r(i)})^3$ periods, and let $Q_{-i} = \prod_{j \neq i} Q_j$. If $\| Q_{-i} - p^*_{-i} \| \leq 2\sigma_{r(i)}$, go to step 4a; if $\| Q_{-i} - p^*_{-i} \| > 2\sigma_{r(i)}$, go to step 4b. (This is the test phase.)

4a. With probability $\pi_{r(i)} = 2^{-(1/\sigma_{r(i)})}$ increase r(i) by 1, with probability $1- \pi_{r(i)}$ keep r(i) fixed. Go to step 2.

4b. With probability $\pi_{r(i)} = 2^{-(1/\sigma_{r(i)})}$ increase r(i) by 1, with probability $1- \pi_{r(i)}$ keep r(i) fixed. Go to step 1.

The idea is that the player i goes back to a wide-area search every time his core beliefs fail the "reality test" in step 3. If they pass the test, he retains these

beliefs and returns to step 2. In either case, however, there is a small probability (in step 4) that he sharpens the focus of the local area-search by cutting σ_r in half (this also sharpens the test criterion). For convenience in counting time periods, we shall assume that time elapses only in step 2; the other steps in the algorithm are instantaneous.

THEOREM 1. *Let G be a finite n-person game. If all players learn by the prospecting rule, then they are rational, good predictors, and with arbitrarily high probability they eventually play arbitrarily close to a Nash equilibrium. That is, for every* $\varepsilon > 0$ *there exists a time* T_ε *such that for all learning paths* $\{h^t/p^t_{-i}, q^t_i\}$ *and all* $t \geq T_\varepsilon$, *the probability is at least* $1 - \varepsilon$ *that for some Nash equilibrium* q^*, $|p^t_{-i} - (q^*)_{-i}| < \varepsilon$ and $|(q^t)_i - (q^*)_i| < \varepsilon$ *for all* i.

Before proving this result several remarks are in order. First, unlike Bayesian learning, the process works even when players are completely naive and their initial conjectures about the other players do not contain even a "grain of truth." In particular, the players do not need to operate from a "model" of what the others are doing or why they are doing it. Second, like Bayesians, they *learn* in the sense that they observe what others are doing, and compare this with their current beliefs. Their response, however, is to change beliefs only if there is a large enough gap between the data and their beliefs; otherwise their beliefs stay put. The process gradually homes in on *equilibrium behavior* in the sense that the players eventually discover a set of core beliefs that mirror (fairly closely) what people are actually doing. Once this happens, they continue to play with these beliefs for a long period of time. The process does not necessarily *converge* to equilibrium because a player who happens to be predicting poorly will eventually try something radically different (loop to step 1). As time runs on, however, these large deviations become more and more infrequent. The result is that the chances become better and better that the players' beliefs (and actions) are arbitrarily close to a Nash equilibrium at any given time.

The proof proceeds by a series of lemmas. Let $r^t(i)$ be the round in which player i finds himself at time t, and let $\mathbf{r}^t = (r^t(1), ..., r^t(n))$ denote the corresponding vector of rounds.

LEMMA 1. The probability goes to one that all players are within one round of each other: for all $i \neq j$, $P(|r^t(i) - r^t(j)| \leq 1) \to 1$ as $t \to \infty$.

PROOF. Fix a player i. For each positive integer r, let T_r be a random variable representing the number of periods that player i spends in round r. The expected number of iterations of step 3 during round r is $1/\pi_r = 2^{1/\sigma_r}$ and the expected number of time periods between iterations is $(1/\sigma_r)^4$. Letting $\gamma_r = (1/\sigma_r)^4 (2^{(1/\sigma_r)})$, we have that

$$P(T_r > t) \approx e^{-t/\gamma_r}. \tag{11}$$

Since i is fixed we shall let $r^t(i) = r(t)$ for notational convenience. Thus

$$r(t) = \min \{ r' : \sum_{s=1}^{r'} T_s \geq t \}.$$

Since $E(T_r)$ grows exponentially, $\rho(t) = \log_2 \log_2(t)$ is a fairly good estimate of the value of r(t). In general, let $\lfloor x \rfloor$ denote the integer part of x. Define

$$r^-(t) = \lfloor \rho(t) - 1/2 \rfloor + 1,$$

and

$$r^+(t) = \lfloor \rho(t) \rfloor + 1.$$

Note that $0 \leq r^+(t) - r^-(t) \leq 1$ and in some cases we have $r^+(t) = r(t)$. We claim that

$$P(r^-(t) \leq r(t) \leq r^+(t)) \to 1 \text{ as } t \to +\infty. \tag{12}$$

It is clear from (11) and the definition of $r^-(t)$ that

$$P(r(t) \geq r^-(t)) = P\left(\sum_{s=1}^{r^-(t)-1} T_s < t \right) \to 1 \text{ as } t \to \infty.$$

Similarly,

$$P\left(\sum_{s=1}^{r+(t)} T_s \geq t\right) \geq P\left(T_{r^+(t)} \geq t\right) \to 1.$$

Hence $P(r(t) \leq r^+(t)) \to 1$ as $t \to \infty$, which establishes (12). Since this holds simultaneously for all players, all of them are within one round of each other with arbitrarily high probability as t becomes large. This completes the proof of Lemma 1.

Suppose now that player i is in round r, and is in the local search phase (step 2 of the algorithm) with random beliefs generated by the distribution $\omega_{i,r}(p_{-i} \mid p^*_{-i})$. Define the function $\mu_{i,r} : \Delta_{-i} + \Delta_i$ such that

$$q_i = \mu_{i,r}(p^*_{-i}) = \int_{A_{-i}} B_i(p_{-i})\omega_{i,r}(p_{-i} \mid p^*_{-i})dp_{-i}, \tag{13}$$

where B_i is the best-reply function for player i that was fixed earlier. Thus $\mu_{i,r}(p^*_{-i})$ is the distribution of player i's actions in each time period of round r when his core beliefs are p^*_{-i}. Note that player i has definite beliefs at any given time (though he may not know how he came to hold these beliefs), and he is rational in the sense that he chooses a best reply given his beliefs. From the others' perspective, i's beliefs are random and his actions are described by the probability distribution q_i.

Let $r = (r^t(1), \ldots, r^t(n))$ where $r^t(i)$ is the round that player i is in at time t. Because of Lemma 1 we shall usually restrict ourselves to those times t such that $r^t(i) - r^t(j) \leq 1$ for all i and j. Such an **r** will be called *regular*. (In what follows we shall often omit the t-superscripts for notational economy.)

LEMMA 2. For every **r** there exists a vector $q^r \in A$ such that

$$\forall i, \mu_{i,r(i)}[(q^r)_{-i}] = (q^r)_i .$$

PROOF. Define the function $F_r : A \to A$ such that

$$\forall q \in A, F_r(q_1, \ldots, q_n) = (\mu_{1,r(1)}[(q)_{-1}], \mu_{2,r(2)}[(q)_{-2}], \ldots, \mu_{n,r(n)}[(q)_{-n}]).$$

The hazy beliefs smooth the best replies as in (13), hence F_r is continuous on the compact, convex set A. It therefore has a fixed point q^r, which has the desired property. This completes the proof of Lemma 2.

The vector q^r is a mixed strategy n-tuple such that, when every player i has core beliefs $(q^r)_{-i}$ in round r(i), i's actions are distributed according to $(q^r)_i$. From here on we shall fix a particular q^r for each **r.**

LEMMA 3. Let r_k be a sequence such that $\min_i r_k(i) \to \infty$ as $k \to 00$. If $q^{r_k} \to q^*$, then q* is a Nash equilibrium of the game.

PROOF. For every $q \in A$, and every player i, define

$$b_i((q)_{-i}) = \max \{u_i(q'_i, (q)_{-i}) : q'_i \in \Delta_i\}. \tag{14}$$

Thus $b_i((q)_{-i})$ is the *payoff* to i from making a best reply to $(q)_{-i}$, whereas $B_i((q)_{-i})$ is a best reply strategy. It is easily checked that $b_i((q)_{-i})$ is continuous and bounded. Without loss of generality we can normalize the utilities so that $O \le b_i((q)_{-i}) \le 1$ for all q and all i.

Suppose now that $q^{r_k} \to q^*$. Fix a small $\varepsilon > 0$. There exists δ (depending on ε) such that

$$\forall i, \forall q \in N_\delta, b_i((q)_{-i}) \ge b_i((q^*)_{-i}) - \varepsilon/2, \tag{15}$$

where

$$N_\delta = \{q \in A: | q - q' | \le \delta\}.$$

Since $q^{r_k} \to q^*$ and var $\omega_{i,r_k(i)}(\cdot | (q^{r_k})_{-i}) \to O$ as $k \to$ CO, there exists an integer k_ε such that

$$\forall i, \ \forall k \ge k_\varepsilon, \quad \int_{N_\delta} \omega_{i,r_k(i)}(p_{-i} | (q^{r_k})_{-i}) dp_{-i} \ge 1 - \varepsilon/2. \tag{16}$$

By construction,

$$(q^{r_k})_i = \mu_{i,r_k}((q^{r_k})_{-i}) = \int_{\Delta_{-i}} B_i(p_{-i}) \omega_{i,r_k(i)}(p_{-i} | (q^{r_k})_{-i}) dp_{-i}. \tag{17}$$

From (15)-(17) and the normalization $O \le b_i((q)_{-i}) \le 1$ we conclude that

$$\forall i, \forall k \ge k(\varepsilon), \quad u_i((q^{r_k})_{-i}) \ge (b_i(q^*_{-i}) - \&/2)\ (1 - \varepsilon/2).$$

Since $q^{r_k} \to q^*$, $u_i(\cdot)$ is continuous, and ε is arbitrary, it follows that

$$Vi,\ u_i(q^*) \ge b_i(q^*_{-i}),$$

which proves that q^* is a Nash equilibrium of the game. This concludes the proof of Lemma 3.

Since A is compact, for every $\varepsilon > 0$ there is an integer r_ε such that

$$\forall \mathbf{r}, \min r_i \ge r_\varepsilon \text{ implies } q^{\mathbf{r}} \text{ is within } \varepsilon \text{ of some Nash equilibrium.} \tag{18}$$

A *cycle* is a maximal sequence of consecutive time periods in which no one's test fails and no one changes round. In particular, i's core beliefs p_{-i} are constant throughout a cycle, and so is the vector of rounds r. The cycle is ε-*close* to a vector q^0 if each player's core beliefs are close to q^0 and the expected actions $q_i = \mu_{i,r(i)}(p_{-i})$ are close to q^0 as well:

$$\forall i, \quad |p_{-i} - (q^0)_{-i}| < E \text{ and } |(q)_{-i} - (q^0)_{-i}| < 8. \tag{19}$$

Let $\Pi = (p_{-1}, p_{-2}, \ldots, p_{-n}) \in \Delta_{-1} \times \Delta_{-2} \times \ldots \times \Delta_{-n}$ be an n-tuple of core beliefs. Given a regular vector r, let $r = \min_i r(i)$. We say that Π is *r-virtuous* if

$$\forall i, |p_{-i} - (q^r)_{-i}| < (\sigma_r)^3. \tag{20}$$

LEMMA 4. Let r be regular, and let $r = \min r(i)$. There is an integer r^* such that, for all $r \ge r^*$, every r-virtuous belief vector initiates a cycle that is σ_{r+1}-close to q^r and has an expected length of at least e^{b/σ_r} periods, where b is a positive exponent independent of r.

PROOF. By the Lipschitz condition on $\omega_{i,r(i)}(\cdot)$,

$$\forall p', p'' \in \Delta_{-i}, \quad \|\omega_{i,r(i)}(p_{-i} \mid p'_{-i}) - \omega_{i,r(i)}(p_{-i} \mid p''_{-i})\|_1 < (k/\sigma_{r(i)}) \mid p'_{-i} - p''_{-i} \mid .$$

From the definition of $\mu_{i,r(i)}$ and the fact that $\mid B_i(\,) \mid \leq 1$, it follows that

$$\forall p', p'' \in \Delta_{-i}, \mid \mu_{i,r(i)}(p'_{-i}) - \mu_{i,r(i)}(p''_{-i}) \mid = 0(\mid p'_{-i} - p''_{-i} \mid / \sigma_{r(i)}). \tag{21}$$

If the beliefs satisfy (20), it follows from (21) that

$$\mid \mu_{i,r(i)}(p_{-i}) - \mu_{i,r(i)}(q^r_{-i}) \mid \leq O(\sigma_r^2).$$

But $\mu_{i,r(i)}(q^r)_{-i}) = (q^r)_i$ because q^r is a fixed point. Let $q_i = \mu_{i,r(i)}(p_{-i})$ for each i. Recalling that $r = \min_i r(i)$, it follows that for all sufficiently large r,

$$\mid (q)_{-i} - (q^r)_{-i} \mid < \sigma_r/2 = \sigma_{r+1} . \tag{22}$$

This proves that the cycle initiated by the beliefs is σ_{r+1}-close to q^r.

We estimate the expected number of periods in the cycle as follows. Suppose that some player i is just entering the test phase (step 3). For each $j \neq i$, let the random variable Qj denote the actual distribution of actions by j over the preceding $N = (1/\sigma_{r(i)})^3$ periods, and $Q_{-i} = \prod_{j \neq i} Q_j$. The test fails for player i if

$$\mid Q_{-i} - p_{-i} \mid > 2\sigma_{r(i)} \geq \sigma_r . \tag{23}$$

The preceding argument shows that for all sufficiently large r,

$$\mid (q)_{-i} - (q^r)_{-i} \mid < \sigma_r /2. \tag{24}$$

From this and (20) we have that, for all suitably large r,

$$\mid Q_{-i} - (q)_{-i} \mid > \sigma_r /4. \tag{25}$$

Now

$$P(\mid Q_{-i} - (q)_{-i} \mid > \sigma_r /4) \leq \text{Max}_{i \neq i} P(\mid Q_i - q_i \mid > \sigma_r /4(n-1)^{1/2}). \tag{26}$$

Fix $j \neq i$. For each $x_j \in X_j$, let $Q_j(x_j)$ be the empirical proportion of strategy x_j over the N periods in player i's current test phase. $Q_j(x_j)$ is a binomial random variable with expectation $q_j(x_j)$, and standard deviation $s = O((\sigma_r)^{3/2})$. Let $Y_j(x_j) = (Q_j(x_j) - q_j(x_j))/s$ and $y = \sigma_r/4(n-1)^{1/2}s$. The probability that the test fails is no larger than $I?(Y_j > y)$. By the theory of large deviations (Billingsley, Theorem 9.4),

$$P(Y_j > y) \leq e^{-O(y^2)} \text{ provided } y/N^{1/2} \text{ is small.} \tag{27}$$

Since $y/N^{1/2} = O(\sigma_r)$ and $\sigma_r \to O$ the latter condition is satisfied for all sufficiently large r. Since $y = O((1/\sigma_r)^{1/2})$, we conclude that there exist positive constants α and β such that, for all sufficiently large r,

$$P(|Q_{-i} - (q)_{-i}| > \sigma_r/4) \leq \alpha e^{-\beta/\sigma_r}. \tag{28}$$

This estimate applies to every player. The probability at a given time t of player i entering the test phase is at most $\sigma_r \geq \sigma_{r(i)}$, and the probability that he fails the test is at most $\alpha e^{-\beta/\sigma_r}$. Thus the probability of i looping to step 1 at the end of period t is at most $\alpha\sigma_r e^{-\beta/\sigma_r}$. The probability of anyone looping to step 1 at the end of period t is therefore at most $\alpha n\sigma_r e^{-\beta/\sigma_r}$. However, the cycle can only end by someone looping to step 1 or someone changing round in step 4, and the latter has probability at most $2^{-(1/\sigma_r)}$. We conclude that there exists a positive exponent b and a round r^* such that for all $r \geq r^*$, the expected number of periods in every r-virtuous cycle is at least e^{b/σ_r}. This completes the proof of Lemma 4.

Fix $E > 0$ and consider a cycle initiated in round r with core beliefs $\Pi^0 = (p\text{-}1, p\text{-}2, \ldots, p\text{-}n)$. The cycle is *bad in beliefs* if the core beliefs $p\text{-}i$ of at least one player are not ε-close to some Nash equilibrium. It is *bad in behaviors* if the expected actions q_i of at least one player are not ε-close to some Nash equilibrium. It is *good* if it is neither bad in beliefs nor bad in behaviors.

LEMMA 5. For each $\varepsilon > 0$ there is an r^*_ε, such that for all r satisfying $\min_i r(i) \geq r^*_\varepsilon$, the expected length of every bad cycle is at most $32/\sigma_r^4$.

PROOF. Let the core beliefs $\Pi^0 = (p_{-1}, p_{-2}, \ldots p_{-n})$ initiate a bad cycle.

Case i): The cycle is bad in beliefs. Then for every Nash equilibrium q^* we have $\sum |\Pi^0_i - (q^*)_{-i}| > \varepsilon$. Let $N_{\varepsilon/2}(\Pi^0)$ be the set of all beliefs Π such that $\sum |\Pi_i - \Pi^0_i| \leq \varepsilon/2$. In general, let $B_{-i}(\Pi)$ denote the vector of best reply strategies of all players other than i, given that their core beliefs are described by Π. Since no $\Pi \in N_{\varepsilon/2}(\Pi^0)$ is a Nash equilibrium, $\sum |\Pi_i - B_{-i}(\Pi)| > 0$. Indeed, since the set $N_{\varepsilon/2}(\Pi^0)$ is bounded away from the Nash equilibria, there exists $\delta_\varepsilon > 0$ such that

$$\inf_{\Pi \in N_{\varepsilon/2}(\Pi^0)} \sum |\Pi_i - B_{-i}(\Pi)| \geq \delta_\varepsilon > 0. \tag{30}$$

For all sufficiently large r, say all $r \geq r'_\varepsilon$, the distribution $\omega = \omega_{1,r(1)}(\cdot | p_{-1}) \times \ldots \times \omega_{n,r(n)}(\cdot | p_{-n})$ puts at least 50% of the probability on NE/L(n). The expected distribution of the players' actions is an average (using ω) of the best replies $B_i(\Pi_i)$. Letting q_i denote i's expected distribution, and recalling that $\Pi_i = p_{-i}$, it follows from (30) that

$$\forall\, \Pi \in N_{\varepsilon/2}(\Pi^0), \quad \sum |p_{-i} - (q)_{-i}| \geq \delta_\varepsilon/2. \tag{31}$$

Now player i's test fails in step 3 if $|p_{-i} - Q_{-i}| > 2\sigma_{r(i)} \geq \sigma_r$. As noted in the proof of Lemma 4, $|(q)_{-i} - Q_{-i}|$ has a standard deviation that is $O((\sigma_r)^{3/2})$. Thus the test is likely to fail if $\delta_\varepsilon/2$ is large compared with both σ_r and $(\sigma_r)^{3/2}$, which is the case if r is sufficiently large. In particular, there is an r''_ε such that for all $r \geq r^*_\varepsilon = \max\{r''_\varepsilon, r'_\varepsilon\}$, the test fails (for any given player) with probability one-half or greater. Since the expected number of periods in step 2 before entering a test phase is at most $1/\sigma_{r(i)}^4 \leq 16/\sigma_r^4$, we conclude that the expected length of the cycle is at most $32/\sigma_r^4$.

Case ii). The cycle is bad in behaviors. Let $q(r) \in A$ denote the vector of expected actions given the beliefs. By the preceding case, we may as well

suppose that the cycle is good in beliefs. Thus for every Nash equilibrium $q^* \in A$ and for all i,

$$I\, p_{-i}(q^*)_{-i} I < \&,$$

whereas for some i,

$$I\,(q(\mathbf{r}))_{-i} - (q^*)_{-i}\, I \geq \varepsilon.$$

Hence there exists $\delta > 0$ such that, for some i,

$$I\,(q(r))_{-i} - p_{-i}\, I \geq \delta.$$

If δ is sufficiently large relative to $\sigma_r^{3/2}$ (the standard deviation of the test statistic $I Q_{-i} - p_{-i} I$), the test of player i will fail with probability at least one-half. The expected number of periods until player i enters the test phase is at most $1/\sigma_{r(i)} \leq 16/\sigma_r^4$. Hence the expected number of periods in the cycle is at most $32/\sigma_r^4$. This concludes the proof of Lemma 5.

To prove the theorem, we need to show that for each $\varepsilon > 0$ there is a time T_ε such that for all $t \geq T_\varepsilon$, the strategies and beliefs of every player are ε-close to some Nash equilibrium with probability at least $1 - \varepsilon$. In the entire subsequent discussion, ε is fixed.

Let $\mathbf{r}^t$ be the random variable of rounds in which the players find themselves at time t. It is clear from the algorithm that each $r^t(i)$ goes to infinity with probability one:

$$P(r \leq \min\, r^t(i)) + 1 \text{ as } t \to \infty. \tag{32}$$

Moreover, Lemma 1 implies that

$$P(\mathbf{r}^t \text{ is regular}) \to 1 \text{ as } t \to \infty. \tag{33}$$

Thus for any integer r there exists a time T_r such that

$$\forall t \geq T_r, \quad P(\mathbf{r}^t \text{ is regular and all } r^t(i) \geq r) \geq 1 - \varepsilon/2. \tag{34}$$

Let $r \geq \max(r^*, r^*_\varepsilon, r\&)$ where r^* is defined in Lemma 4, r^*_ε is defined in Lemma 5, and r_ε is defined in (18). Fix some $t \geq T_r$. Let r be regular with $\min_i r_i \geq r$. The process can be in one of three "states": a$_n$ r-virtuous cycle (I), a good cycle that is not r-virtuous (2), and a bad cycle (3). Transitions from one state to another can only occur at a time when someone fails a test. (For purposes of this estimation we can ignore transitions to higher r that occur in step 4 because these are extremely improbable relative to the other transitions.) Let π_1, π_2, π_3 be the stationary probabilities of being in the three states conditional on no transitions to higher r occurring. We need to show that π_3 is smaller than $\varepsilon/2$ whenever r is sufficiently large.

Denote the transition probabilities by P_{ij}, $1 < i, j \leq 3$. Lemma 4 shows that the probability of leaving a virtuous cycle is bounded as follows:

$$P_{12} + P_{13} \leq e^{-b/\sigma_r}. \tag{35}$$

Suppose that the process is in a bad cycle: what is the probability of starting a virtuous cycle in the next period? We know from Lemma 5 that the probability of exiting from the cycle is at least $(\sigma_r^4)/32$. The probability that all players choose core beliefs that lie within σ_r^3 of q^r is bounded below by $c\sigma_r^d$ for some positive constants c and d (in fact we can take $d = 3n|X|$). Hence the probability of transiting from a bad cycle to a virtuous cycle is

$$P_{31} \geq (c/32)\sigma_r^{d+4}. \tag{36}$$

But the stationarity equations imply that

$$\pi_3 P_{31} \leq \pi_3(P_{31} + P_{32}) = \pi_1(P_{12} + P_{13}). \tag{37}$$

Hence

$$(\pi_1 + \pi_2)/\pi_3 \geq \pi_1/\pi_3 \geq P_{31}/(P_{12} + P_{13}) \geq (c/32)\sigma_r^{3d+4} e^{b/\sigma_r}. \tag{38}$$

Thus the probability of being in a good cycle at time $t \geq T_r$ can be made as close as we wish to unity by choosing r sufficiently large. This concludes the proof of the theorem.

Many variants of the above prospecting rule will also work; we briefly mention two of them here. First, the only reason to clamp down on $\sigma_{r(i)}$ in step 4 is to guarantee that the search process gets closer and closer to some Nash equilibrium, with increasingly high probability, as time runs on. Instead, we could fix each player's value $\sigma_{r(i)}$ (they need not be equal), and drop step 4. Then Lemma 1 is irrelevant to the proof, and the remainder of the argument shows that, given $\varepsilon > 0$, if all $\sigma_{r(i)}$ are small enough, then the process will be ε-close to some Nash equilibrium with probability at least $1 - \varepsilon$ after some time T&.

A second variation is to modify the wide-area search probability based on information that emerges during the learning process. Let $f_i: H \to \Delta_{-i}$ be any deterministic belief formation process as described in section 2. In particular, f_i could represent ordinary Bayesian updating of a prior on Δ_{-i}. Suppose that the prospecting process (for player i) loops to step 1 at time t, and that h^t is the history of actions up to that time. Let $p^t_{-i} = f_i(h^t)$ and let $\mathbf{1}(p^t_{-i})$ denote the degenerate distribution with mass 1 at $p^t_{-i} \in \Delta_{-i}$. We can think of p^t_{-i} as i's current estimate of the most likely spot at which to anchor his core beliefs. Suppose that player i conducts the wide-area search at time t according to the distribution

$$\nu^t_i(h^t) = \lambda^t \nu_i + (1 - \lambda^t)\mathbf{1}(p^t_{-i}), \tag{39}$$

where λ^t is an appropriately chosen weight. If λ^t does not go to zero too quickly with t, the above proof shows that the process will be arbitrarily close to a Nash equilibrium. Indeed, this will be the case if the expected waiting time until hitting an r-virtuous cycle is on the order of σ_r^{-k} for some positive exponent k. This clearly holds if $\lambda^t = O(\sigma_r) = O(2^{-r})$. Recalling that $r \approx \log_2\log_2 t$, we see that this holds if λ^t goes to zero no faster than $1/\log t$. In other words, the wide-area search can be modified by any deterministic updating process so long as the resulting wide-area forecast is surrounded by a "haze" $\lambda^t \nu_i$ that goes to zero no faster than $1/\log t$.

References

Billingsley, Patrick (1986): *Probability and Measure*, 2nd edition. New York: Wiley.

Binmore, Ken (1987): "Modelling Rational Players, Part I," *Economics and Philosophy, 3, 179-214.*

Binmore, Ken (1991): "DeBayesing Bame Theory," International Conferenc, on Game Theory, Florence, Italy.

Blume, Larry, and David Easley (1995): "Rational Expectations and Rational Learning," Preprint, Cornell University.

Foster, Dean, and Rakesh Vohra (1992): "Calibrated Learning and Correlated Equilibrium," Preprint, University of Pennsylvania.

Foster, Dean, and H. Peyton Young (1995): "On the Nonconvergence of Fictitious Play in Coordination Games," Working Paper, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Fudenberg, D. and D. Kreps (1993): "Learning Mixed Equilibria," *Games and Economic Behavior, 5, 320-367.*

Jordan, James S. (1991): "Bayesian Learning in Normal Form Games," *Games and Economic Behavior, 3, 60-91.*

Jordan, James S. (1992): "Bayesian Learning in Games: A NonBayesian Perspective," Preprint, University of Minnesota.

Jordan, James S. (1993): "Three Problems in Learning Mixed-Strategy Equilibria," *Games and Economic Behavior, 5, 368-386..*

Jordan, James S. (1995): "Bayesian Learning in Repeated Games," *Games and Economic Behavior, 9, 8-20.*

Kalai, Ehud, and Ehud Lehrer (1993): "Rational Learning Leads to Nash Equilibrium," *Econometric, 61, 1019-1045.*

Kaniovski, Yuri, and H. Peyton Young (1995): "Learning Dynamics in Games with Stochastic Perturb ations," *Games and Economic Behavior, 7, 330-363.*

Nachbar, John. H. (1995): "Prediction, Optimization, and Learning in Games," Preprint, Department of Economics, Washington University, St. Louis.

Nyarko, Y. (1994): "Bayesian Learning Leads to Correlated Equilibria in Normal Form Games," *Economic Theory, 4, 821-841.*

Savage, Leonard (1951): *The Foundations of Statistics. New* York: Wiley.

Shapley, LLoyd, (1964): "Some Topics in Two-Person Games," in *Advances in Game Theory, Annals of Mathematics Studies* vol. *52* (M. Dresher et al., eds.), *1-28.*