On the Impossibility of Predicting the Behavior of Rational Agents

Dean P. Foster
H. Peyton Young

SFI WORKING PAPER: 2001-08-039

SANTA FE INSTITUTE

On the Impossibility of Predicting the Behavior of Rational Agents

Dean P. Foster* and H. Peyton Young**

February, 1999
This version: June, 2001

Acknowledgments. This paper benefited from discussions at the Santa Fe Institute, and also from comments by Junfu Zhang and several anonymous referees. The research was supported in part by NSF grant SBR 9601743.

*Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104. dean@foster.net

**Department of Economics, Johns Hopkins University, Baltimore, MD 21218-2685. pyoung@jhu.edu

Abstract

A foundational assumption in economics is that people are rational -- they choose optimal plans of action given their predictions about future states of the world. In games of strategy this means that each players' strategy should be optimal given his or her prediction of the opponents' strategies. We demonstrate that there is an inherent tension between rationality and prediction when players are uncertain about their opponents' payoff functions. Specifically, there are games in which it is impossible for perfectly rational players to learn to predict the future behavior of their opponents (even approximately) no matter what learning rule they use. The reason is that, in trying to predict the next-period behavior of an opponent, a rational player must take an action this period that the opponent can observe. This observation may cause the opponent to alter his next-period behavior, thus invalidating the first player's prediction. The resulting feedback loop has the property that, in almost every time period, someone predicts that his opponent has a non-negligible probability of choosing one action, when in fact the opponent is *certain* to choose a different action. We conclude that there are strategic situations where it is impossible *in principle* for perfectly rational agents to learn to predict the future behavior of other perfectly rational agents, based solely on their observed actions.

Rationality vs predictability

A standard assumption in economics is that people are *rational*: they maximize their expected payoffs given their beliefs about future states of the world. This hypothesis plays a crucial role in game theory, where each player is assumed to choose an optimal strategy given his belief about the strategies of his opponents. In this setting, a belief amounts to a forecast or *prediction* of the opponents' future behavior, that is, of the probability with which the opponents will take various actions. The prediction is *good* if the forecasted probabilities are close to the actual probabilities. Together, prediction and rationality justify the central solution concept of the theory. Namely, if each player correctly predicts the opponents' strategies, and if each chooses an optimal strategy given his prediction, then the strategies form a Nash equilibrium of the repeated game. But under what circumstances will rational players actually learn to predict the behavior of others starting from out of equilibrium conditions?

In this paper we show that there are very simple games of incomplete information such that players almost never learn to predict their opponents' behavior even approximately, and they almost never come close to playing a Nash equilibrium. This impossibility result and its proof builds on the existing literature on learning in repeated games, including Jordan (1991, 1993, 1995), Kalai and Lehrer (1993), Lehrer and Smorodinsky (1997), Nachbar (1997, 1999, 2001), and Miller and Sanchirico (1997). It is also related to an earlier literature critiquing Bayesian learning more generally (Binmore, 1987, 1990, 1991; Diaconis and Freedman, 1986). The novelty of the present contribution is to demonstrate the incompatibility between rationality and prediction without placing any restrictions on the players' prior beliefs, their learning rules, or the degree to which they are forward-looking.

An example

We begin by illustrating the problem in a concrete case. Consider two individuals, A and B, who are playing the game of matching pennies. Simultaneously each turns a penny face up or face down. If the pennies match (both are Heads or both are Tails), then B buys a prize for A; if they do not match, A buys a prize for B. Assume first that the prize is one dollar, and that the utility of both players is linear in money. Then the game has a unique Nash equilibrium in which each player randomizes by choosing Heads (H) and Tails (T) with equal probability. If both adopt this strategy, then each is optimizing given the strategy of the other. Moreover, although neither can predict the

realized action of the opponent in any given period, each can predict his *strategy*, namely, the probabilities with which the actions will be taken. In this case no tension exists between rationality and prediction because the game has a unique equilibrium and the players know what it is.

Now change the situation by assuming that if both players choose Heads, then B buys an ice cream cone for A, whereas if both choose Tails then B buys A a milk shake. Similarly, if A chooses Heads and B chooses Tails then A buys B a coke, whereas if the opposite occurs then A buys B a bag of chips. Assume that the game is played once each day, the players' tastes do not change from one day to the next, and they have a fixed positive utility for each of the prizes and also for money. Unlike the previous situation, this is a game of incomplete information in which neither player knows the other's payoffs.

		B's action		B's action	
		H	T	H	T
A's action	H	eat cone	buy coke	buy cone	drink coke
	T	buy chips	drink shake	eat chips	buy shake
		Outcomes for A		**Outcomes for B**	

For expositional simplicity assume first that the players are myopic, that is, they do not worry about the effect of their actions on the future course of the game. Imagine that the following sequence of actions has occurred over the first ten periods

Period	1	2	3	4	5	6	7	8	9	10	11
A:	H	T	T	H	H	H	T	H	T	H	?
B:	T	H	T	H	T	H	T	H	T	H	?

The immediate problem for each player is to predict the *intention* of the opponent in period eleven, and to choose an optimal response. The opponent's intention might be to play Heads or Tails for sure, or it might be to randomize with some probability p for Heads and 1 – p for Tails. If the opponent's intention is to randomize, then obviously one cannot predict his realized action, but it does not seem too much to ask that one predict the approximate probability with which he intends to play each action. We claim, however, that this is essentially impossible.

To see why, let's put ourselves in A's shoes. The behavior of B suggests an alternating pattern, perhaps leading us to predict that B will play T next period. Since we are rational, we will (given our prediction) play T for sure next period. But if B is a good predictor, then she must be able to predict that, with high probability, we are in fact going to play T next period. This prediction by B leads her to play H next period, thus falsifying our original prediction that she is about to play T.

The point is that if either side makes a prediction that leads them to play H or T for sure, the other side must predict that they are going to do so with high probability, which means that they too will choose H or T for sure. But there is no pair of predictions such that both are approximately correct and the optimal responses are H or T for sure. It follows that, for both players to be good predictors of the opponent's next period behavior, at least one of them must be intending to play a mixed strategy next period, and the other must predict this.

Suppose, for example, that player B intends to play a mixed strategy in period eleven. Since B is rational, she only plays a mixed strategy if she is *exactly indifferent* between playing H and T given her predictions about A. (If there is a slight difference in payoff between the two actions, strict rationality requires that the one with higher payoff be chosen exclusively.) Now B's predictions about A's behavior in the eleventh period are based on the observed history of play in the first ten periods. Let's say that the particular history given above leads B to predict that A will play H with probability .127. Since B intends to play mixed, it must be the case that B's expected utility from playing H or T are identical, given B's utility function u_B for the various outcomes. In other words, it must be that

$.127\, u_B(\text{buy cone}) + .873 u_B(\text{eat chips}) = .127 u_B(\text{drink coke}) + .873 u_B(\text{buy shake})$.

But there is no reason to think that B's utilities actually do satisfy this equation exactly. More precisely, let us suppose that B's utility for each outcome could be any real number within a certain interval, and that B's actual utility (B's *type*) is the result of a random draw from among these possible values. (The draw occurs once and for all before the game begins.) Following Jordan (1993), we claim that the probability is zero that the above equation will be satisfied. The reason is that there are only a finite number of distinct predictions that B could make at this point in time, because B's prediction can only be based A's observed behavior (together with B's initial beliefs). Since this argument holds for every period, the probability is zero that B will ever be indifferent. From this and the preceding argument it follows that, in any given period, one or both players must be making a bad prediction. Moreover, they cannot be playing a Nash equilibrium in any given

period (or even close to a Nash equilibrium), because this would require them to play mixed strategies, which means that both must be indifferent.

Jordan (1993) was the first to employ this kind of argument to show that myopic players effectively cannot learn mixed equilibria no matter what their beliefs are. Moreover, as we have just seen, the same argument shows that at least one of them cannot learn to predict the behavior of the other. The limitation of Jordan's result is that it assumes players are completely myopic. Forward-looking behavior allows for a much richer repertoire of learning strategies, and more time to detect complex patterns in the behavior of one's opponent. Nevertheless, the incompatibility between rationality and prediction continues to hold even in this case, as we shall show below.

A second closely related body of work is due to Nachbar (1997, 1999, 2001). He was the first to argue that there is a fundamental tension between prediction and rationality in the context of Bayesian learning even when players are forward-looking. Nachbar's critique was prompted by an earlier paper by Kalai and Lehrer (1993), which laid out conditions under which Bayesian rational players would in fact be able to learn to predict the behavior of their opponents. Suppose that each player begins the game with a prior belief over the possible repeated game strategies that his opponents might use. Kalai and Lehrer show that, if these prior beliefs contain a "grain of truth," that is, they put positive probability (however small) on the *actual* repeated game strategies of the opponents, then players learn to predict with probability one.

As Nachbar points out, however, the grain of truth condition may be very difficult to satisfy in practice. To illustrate, consider the preceding example and suppose that the players are perfectly myopic. Then the unique equilibrium of the repeated game is for A to play Heads with some fixed probability p* each period, and for B to play Heads with some fixed probability q* each period. These values are not known to the players because p* depends on B's payoffs whereas q* depends on A's payoffs. Can they be learned through Bayesian updating of a diffuse prior? Suppose that each player begins with a belief that the other is playing an i.i.d. strategy with an unknown parameter (the probability of playing Heads), where the beliefs have full support on the interval [0, 1]. In any given period, the players will almost surely have updated beliefs that lead them to play H or T *with probability one* in that period because the expected payoffs from Heads and Tails are not exactly equal. However, their updated beliefs lead them to predict that their opponent is almost surely going to play a *mixed* strategy next period. Thus their predictions are almost certainly not close to their actual strategies. Furthermore, as the game proceeds, rationality causes them to play H for sure in some periods and T for sure in others. Hence their actual strategies are not i.i.d., and hence not in the support of their beliefs. More generally, Nachbar

(1997, 1999, 2001) argues that in games like this it is difficult to identify any plausible family of beliefs such that the players' best response strategies are in the support of their beliefs.[1]

In this paper we are agnostic about whether or not the players are Bayesian, and what the structure of their priors might be. Instead we show that no matter how players use the information revealed by repeated play, it does not suffice to learn to predict the opponents' behavior in some kinds of games. In other words, they must have additional information about the opponents' strategies or payoffs for prediction to be possible.

Before turning to a precise statement of our result, we should point out that it is *prediction by the players* that is problematical; to an *observer* the average behavior of the players may exhibit empirical regularities. For example, it could be that the cumulative frequency distribution of play approaches a Nash equilibrium of the game. In fact this will be the case for fictitious play, in which each player uses the empirical distribution of the opponent's play up through a given period to predict his next-period behavior, then chooses a best response given that prediction. In games like matching pennies, this simple learning rule induces long-run *average* behavior that converges to the mixed Nash equilibrium of the game (Miyasawa, 1961; Monderer and Shapley, 1996).[2] But this does not imply that the players themselves ever play Nash equilibrium strategies, or that they learn to predict.

The learning model

We now describe our impossibility result in detail. Consider an n-person game G with finite action space $X = \prod X_i$ and utility functions $u_i: X \to R$. We shall assume that the payoffs take the form $u_i(x) = u_i^0(x) + \omega_i(x)$, where the $u_i^0(x)$ are payoffs in a "benchmark" game G^0, and the $\omega_i(x)$ are i.i.d. random variables drawn from a continuous density $\nu(\omega)$ whose support is the interval $I_\lambda = [-\lambda/2, \lambda/2]$. The parameter $\lambda > 0$ is the *range of uncertainty* in the payoffs. We shall assume that the payoffs from the benchmark game G^0, as well as the error structure, are common knowledge. However, the realized payoff $u_i(x)$ is known only to player i. Errors are drawn once only before play begins, and the resulting realized game is played infinitely often. We shall call the realized stage game a ν-*perturbation of* G^0.

[1] Another paper in the same general spirit is due to Miller and Sanchirico (1997).

[2] There are other models in which average behavior in a population of heterogeneous players mimics Nash equilibrium from the observer's point of view (Harsanyi, 1973 ; Fudenberg and Kreps, 1993); indeed Nash (1950) himself suggested such an interpretation.

Each player takes an action once in each time period t = 1, 2, 3, ... The *outcome* in period t is an n-tuple of actions $x^t \in X$, where x_i^t is the action taken by i in period t. A *state* of the process at time t is a history of play up to t, that is, a sequence of outcomes $h^t = (x^1, x^2, \ldots, x^t)$. Let h^0 represent the null history, H^t the set of all length-t histories, and $H = \cup H^t$ the set of all finite histories, i.e., the set of all states. A *realization* of the process will be denoted by h, and the set of realizations (i.e., the set of infinite histories) by H^∞. Histories are publicly observed, that is, there is perfect monitoring.

The *discounted payoff* to player i from a realization $h = (x^1, x^2, \ldots, x^t, \ldots)$ is

$$U_i(h) = (1 - \delta_i) \sum_{t=1}^{\infty} \delta_i^{t-1} u_i(x^t), \qquad (1)$$

where δ_i is i's *discount factor*, $0 \quad \delta_i < 1$. (If $\delta_i = 0$, $U_i(h) = u_i(x^1)$.) Let Δ_i denote the set of probability distributions over X_i. Let $\Delta = \prod \Delta_i$ denote the product set of mixtures, and let $\Delta_{-i} = \prod_{j \ i} \Delta_j$ be the product set of mixtures by i's opponents. A *behavioral strategy* for player i specifies a conditional probability distribution over i's actions in each period, conditioned on the state in the previous period. Thus we can represent i's strategy by a function $q_i^t = g_i(h^{t-1}) \in \Delta_i$, where $q_i^t(x_i)$ is the probability that i plays x_i in period t given that h^{t-1} is the state in period t – 1. This is of course a function of i's realized utility function u_i, but we shall not write this dependence explicitly.

A *prior belief* of player i is a probability distribution over all possible combinations of the opponents' strategies. We can decompose any such belief into one-step-ahead forecasts of the opponents' behavior conditional on each possible state. Thus, if h^{t-1} is the state at time t - 1, i's forecast about the behavior of her opponents in period t can be represented by a probability distribution $p_{-i}^t = f_i(h^{t-1}) \in \Delta_{-i}$, where $p_{-i}^t(x_{-i})$ is the probability that i assigns to the others playing the combination x_{-i} in period t. The function $f_i: H \rightarrow \Delta_{-i}$ will be called i's *forecasting function*. Given any vector of forecasting functions $f = (f_1, f_2, \ldots, f_n)$, it can be shown that there exists a set of prior beliefs such that the f_i describe the one-step-ahead forecasts of players with these beliefs (see Kalai and Lehrer, 1993).

Consider the situation just after the players have been informed privately of their realized payoff functions u_i. Because of the independence of the draws among players, no one knows anything he did not already know about the others' payoffs, and this fact is common knowledge. This has an implication for the forecasting functions. Namely, at the beginning of each period t, i knows that j's information consists solely of the publicly observed history h^{t-1} and j's own payoff function u_j.

Player j's behavior cannot be conditioned on information that j does not have (namely u_{-j}), and player i's forecast of j's behavior cannot be conditioned on information that i does not have (namely, u_{-i}). Thus i's forecast $(f_i(h^{t-1}))_j$ about j's behavior in period t does not depend on the realization of the values u_k for *every* k, including k = i, j. Moreover, this holds for all t and all i. It follows that the functions f_i do not depend on the realized payoff functions $u_i(x)$, though they may depend on ν. Another way of saying this is that the prior beliefs must be consistent with the players' a priori knowledge of the information structure.

Following Jordan (1993), we shall say that a *learning process* is a pair $(f, g) = (f_1, \ldots, f_n; g_1, \ldots, g_n)$ where $f_i: H \to \Delta_{-i}$ and $g_i: H \to \Delta_i$ for each player i. Given a realization of the process h, we shall denote player i's forecast in period t by $p^t_{-i}(h) = f^t_i(h^{t-1})$, and i's behavioral strategy in period t by $q^t_i(h) = g_i(h^{t-1})$.

The pair (f_i, g_i) induces a probability measure on the set of all realizations H^∞. Similarly, for every state h^{t-1}, f_i and g_i induce a conditional probability distribution on all continuations of h^{t-1}. Denote this conditional distribution by $\mu_i(f_i, g_i \,|h^{t-1})$. We say that individual i is *rational* if, for every h^{t-1}, i's conditional strategy $g_i(\cdot|h^{t-1})$ optimizes i's expected utility from time t on, given i's conditional forecast $f_i(\cdot|h^{t-1})$. (This is also known as *sequential rationality*.) Specifically, for every alternative choice of strategy $g_i(\cdot|h^{t-1})$,

$$\int U_i(h)\, d\mu_i(f_i, g_i \mid h^{t-1}) \quad \int U_i(h)\, d\mu_i(f_i, g'_i|h^{t-1}). \qquad (2)$$

Prediction

Intuitively, player i "learns to predict" the behavior of his opponent(s) if i's forecast of their next-period behavior becomes closer and closer to their actual next-period strategies. This idea may be formalized as follows. Consider a learning process (f, g), and let $\mu(g)$ denote the probability measure induced on H^∞ by the strategies $g = (g_1, g_2, \ldots, g_n)$. We say that player i *learns to predict* if the mean-square error of i's next-period predictions goes to zero over almost all histories of play. In other words, for $\mu(g)$-almost all realizations h,

$$\lim_{T\to\infty} \sum_{t=1}^{T} (|\, p^t_{-i}(h) - q^t_{-i}(h)|)^2/T = 0. \qquad (3)$$

Similarly, we shall say that player i *never learns to predict* if the subset of histories for which (3) holds has μ-measure zero. Note that condition (3) permits players to make bad forecasts from time to time, so long as they do not occur too often.

An impossibility theorem

We now demonstrate a class of repeated games such that, with probability one, some player never learns to predict his opponent's behavior, and this holds for *all* prior beliefs. Since our result holds for all beliefs, it must hold for beliefs that are in some sense best possible. A reasonable candidate for "best possible beliefs" are rational expectations beliefs. These have the property that, at every point in time, each player's prediction of his opponent's future behavior is correctly conditioned on the posterior distribution of payoff-types revealed by play so far. These posterior distributions converge to the set of Nash equilibria of the game (Jordan, 1993; see also Nyarko, 1998). However, this does not imply that the posteriors lead to predictions that are close to being correct for a *given* opponent. Our result shows, in fact, that these rational expectations predictions are *not* close to being correct for almost all opponents.

This still leaves open the possibility that for some combinations of beliefs the players' *strategies* converge to Nash equilibrium even though their predictions do not. In a repeated game convergence to equilibrium can be given a variety of interpretations; we shall show that the process fails to converge to equilibrium in almost any reasonable sense. Let Q^N be the set of all one-period strategy tuples $q \in \Delta$ such that q occurs in *some* time period in *some* Nash equilibrium of the repeated game. For every $q \in \Delta$ let $d(q, Q^N)$ be the minimum Euclidean distance between q and the compact set Q^N. Given a learning process (f, g) and a specific history h, if the behavioral strategies come close to Nash equilibrium on h then at a minimum we would expect the following condition to hold,

$$\lim_{T\to\infty} [\sum_{t=1}^{T} d(q^t(h), Q^N)^2]/T = 0. \qquad (4)$$

This implies that, for every $\varepsilon > 0$, play is within ε of some Nash equilibrium at each point in time except possibly for a sparse set of times. We shall show that the process *fails to come close to Nash* in the sense that condition (4) fails to hold for *almost all* histories h.

Theorem. *Let* η *be a continuous density on* $[-\lambda/2, \lambda/2]$*, and let G be a* η*-perturbation of a finite, zero-sum, two-person game* G^0*, all of whose Nash equilibria have full support. Assume that the players are perfectly rational, have arbitrary discount factors less than unity, and that each updates his predictions of the opponents' future behavior by a learning rule that is based solely on observable actions. If* λ *is sufficiently small, then for* η*-almost all payoff realizations, the probability is one that someone never learns to predict and that play fails to come close to Nash.*

We remark that the set of games for which this impossibility result holds is actually much larger than the one stated in the theorem. Consider, for example, any two-person game G with strategy space $Y_1 \times Y_2$, such that $|Y_i| \quad 2$, all Nash equilibria have full support on $Y_1 \times Y_2$, and every action not in Y_i is strictly dominated by some action in Y_i. Then the theorem holds for perturbed versions of this game. Next let us extend G to an n-person game G* by adjoining n - 2 players as follows: each new player has a strictly dominant action, and G* is the two-person subgame that results when they play these actions. It follows that, for any finite action space $X = \quad X_i$, there exists an n-person game G* on X such that when the payoffs of G* are perturbed by small i.i.d random errors, good prediction fails to occur with probability one.

Now consider *any* n-person game G on the finite strategy space $X = \quad X_i$. Suppose that we perturb the payoffs of G by i.i.d. random errors drawn from a normal distribution, or in fact any distribution with a continuous density whose support is the whole real line. With positive probability the payoffs of the realized game will be close to the game G* constructed above. Thus as a corollary we obtain the following.

Corollary. *Let G be any finite n-person game whose payoffs are perturbed once by i.i.d normally distributed random errors. Assume that the players are perfectly rational, have arbitrary discount factors less than unity, and that each updates his predictions of the opponents' future behavior by a learning rule that is based solely on observable actions. For almost all payoff realizations, there is a positive probability that someone never learns to predict and that play fails to come close to Nash.*

Proof of the theorem

Since the proof is somewhat involved, we shall first explain why the argument given in the introduction for myopic players does not extend easily to the general case. One difficulty is that

patient players might interact through conditional strategies that involve no randomization, and these might be predictable at least some of the time. Eliminating this case requires a delicate probabilistic argument. The second difficulty is that, even when players randomize and are therefore indifferent among alternative strategies, this does not imply that the stage-game payoffs are solutions of a linear equation. Rather, they are the roots of a nonlinear function, and we must show that the roots of this function constitute a set of measure zero.

To increase the transparency of the proof, we shall give it for the game of matching pennies. It generalizes readily to any finite, zero-sum, two-person game whose stage-game Nash equilibria are all strictly interior in the space of mixed strategies. Fix a continuous density ν whose support is $[-\lambda/2, \lambda/2]$. To be concrete, we may think of ν as the uniform distribution. The perturbed game has payoff matrix

	1	2
1	$1 + \omega_{11}, -1 + \omega'_{11}$	$-1 + \omega_{12}, 1 + \omega'_{12}$
2	$-1 + \omega_{21}, 1 + \omega'_{21}$	$1 + \omega_{22}, -1 + \omega'_{22}$

where ω_{ij}, ω'_{ij} are i.i.d. random variables distributed according to ν.

Fix two rational players, 1 and 2, with discount factors $0 \quad \delta_1 \quad \delta_2 < 1$. Let their beliefs be f_1, f_2, and let their strategies be $g_1(\cdot|\mathbf{A})$, $g_2(\cdot|\mathbf{B})$, where **A** and **B** are the realized values of the players' payoff matrices. The functions f_1, f_2, g_1, g_2 will be fixed throughout the proof. All probability statements will be conditional on them without writing this dependence explicitly. Let H(**A**, **B**) be the set of all histories h such that good prediction (3) holds when the realized payoffs are (**A**, **B**). Let P be the set of pairs (**A**, **B**) such that good prediction holds with positive probability, that is, $\mu(H(\mathbf{A}, \mathbf{B})) > 0$. First we shall show that $\nu(P) = 0$, that is, there are almost no payoff realizations (**A**, **B**) such that both players learn to predict with positive probability. In the second part of the proof we shall show that for almost all (**A**, **B**) the process fails to come close to Nash.

Lemma 1. For every positive integer m, every $0 < \varepsilon' \quad \varepsilon < 1$, and every (**A**, **B**) $\in$ P, there exists a time T, possibly depending on m, ε, ε', **A**, **B**, such that, with μ-probability at least $1 - \varepsilon'$, each player forecasts the other's next-period strategy within ε in each of the periods $T + 1, \ldots, T + m$.

Proof. Let (**A**, **B**) $\in$ P and suppose there were no such time T. Then for *every* time T the μ-probability would be greater than $\varepsilon' > 0$ that at least one player misforecasts the opponent's behavior by more than ε in one or more of the periods T + 1, . . . , T + m. This would imply that condition (3) is violated for almost all histories, that is, $\mu(H(\mathbf{A}, \mathbf{B})) = 0$, which contradicts our assumption that (**A**, **B**) $\in$ P.

Lemma 2. For each (**A**, **B**) $\in$ P there exists a time T and a history h^T, possibly depending on **A**, **B**, such that, conditional on h^T, each player's expected future payoffs, discounted to T + 1, are bounded above by $c\lambda$ for some positive number c that depends only on the discount rates.

Proof. Given a small $\lambda > 0$, choose m 1 such that δ_2^m λ and $0 < \varepsilon'$ ε $\lambda/m4^m e^m$. As guaranteed by Lemma 1, let h^T be a history such that the μ-probability is at least $1 - \varepsilon'$ that each player forecasts the other's next-period strategy within ε in each of the periods T + 1, . . . , T + m. Let α^*_{T+1} and β^*_{T+1} be the payoffs that players 1 and 2 *expect* to get from period T + 1 on, discounted to period T + 1. We shall exhibit a positive constant c, depending only on the discount factors, such that $\alpha^*_{T+1}, \beta^*_{T+1}$ $c\lambda$. Note first that each player has the option of playing fifty-fifty in each period from T + 1 on, which has expected discounted payoff at least $-\lambda/2$. Since each player's strategy is optimal, it follows that $\alpha^*_{T+1}, \beta^*_{T+1}$ $-\lambda/2$.

For each j, 1 j m, let α_{T+j} be the player 1's expected *undiscounted* payoff in period T + j *as forecast* by player 1 at the end of period T. Define β_{T+j} similarly for player 2. Let H_{j,h^T} be the set of all continuations of h^T to time T + j – 1. Let $\phi_1(h^{T+j-1})$ denote player 1's probability assessment of $h^{T+j-1} \in H_{j,h^T}$ and similarly define $\phi_2(h^{T+j-1})$ for player 2. The true probability is $\mu_0(h^{T+j-1})$, where μ_0 is μ conditional on h^T. The set of continuations on which someone makes a bad forecast have μ_0-probability at most ε'. On the remaining *good* continuations, each player errs by at most ε in forecasting his opponent's stage-game behavior in each of j stages. Hence for every good continuation h^{T+j-1}, $|\phi_i(h^{T+j-1}) - \mu_0(h^{T+j-1})|$ $(1 + \varepsilon)^j - 1$ $(j\varepsilon)e^{j\varepsilon}$.

Each player's forecasted payoff in period T + j cannot differ from the actual payoff in period T + j by more than $2 + \lambda$ no matter how bad the forecast is. There are 4^j continuations to period T + j, including good and bad. Over all the good ones, player 1's forecasted expected payoff differs from his actual expected payoff by at most $4^j(j\varepsilon)e^{j\varepsilon}(2 + \lambda)$. Over all the bad ones the two differ by at most $\varepsilon'(2 + \lambda)$. Thus the difference between 1's forecasted expected payoff, α_{T+j}, and his actual expected payoff, $\bar{\alpha}_{T+j}$, is at most $(\varepsilon' + 4^j(j\varepsilon)e^{j\varepsilon})(2 + \lambda)$. By assumption, ε' ε $\lambda/m4^m e^m$ and j m, so $\varepsilon' + 4^j(j\varepsilon)e^{j\varepsilon}$ $\varepsilon + 4^m(m\varepsilon)e^{m\varepsilon}$ 2λ. Thus $|\alpha_{T+j} - \bar{\alpha}_{T+j}|$ $2\lambda(2 + \lambda)$ 6λ. Similarly $|\beta_{T+j} - \bar{\beta}_{T+j}|$ 6λ. The actual payoffs satisfy $|\bar{\alpha}_{T+j} + \bar{\beta}_{T+j}|$ λ, from which we conclude that

$$|\alpha_{T+j} + \beta_{T+j}| \quad 13\lambda \text{ for } 1 \quad j \quad m. \qquad (5)$$

For each j, 1 j m, let $\alpha^*_{T+j} = (1 - \delta_1)[\alpha_{T+j} + \delta_1\alpha_{T+j+1} + \delta_1^2\,\alpha_{T+j+2} + \ldots]$ be player 1's expected payoff from period T + j on, discounted to period T + j, as forecast at the end of period T. Similarly define $\beta^*_{T+j} = (1 - \delta_2)[\beta_{T+j} + \delta_2\beta_{T+j+1} + \delta_2^2\,\beta_{T+j+2} + \ldots]$. We claim that $\alpha^*_{T+j}, \beta^*_{T+j}$ $-\lambda/2$ for every j. If not, some player could switch his strategy to a fifty-fifty random mixture from period T + j on, thus increasing his expected payoff from that time on, which would contradict sequential rationality.

Beyond period T + m, the forecasts may no longer be good within ε. However, neither player expects to get more than $1 + \lambda/2$ in any period, so the sum of expected payoffs beyond period T + m, discounted to period T + 1, cannot be more than $(1 - \delta_2)\delta_2^m(1 + \lambda/2)$. By choice of m, δ_2^m λ, so the previous expression is at most 2λ when λ is small Putting this fact together with (5) it follows that, for 1 j m,

$$\beta^*_{T+j} \quad (1 - \delta_2)\,_{j=1,m}\delta_2^{j-1}\beta_{T+j} + 2\lambda \quad (1 - \delta_2)\,_{j=1,m}\delta_2^{j-1}(13\lambda - \alpha_{T+j}) + 2\lambda$$
$$15\lambda - (1 - \delta_2)\{\textstyle\sum_{j=1,m}\delta_2^{j-1}\alpha_{T+j}\}. \qquad (6)$$

The term $\{\sum_{j=1,m}\delta_2^{j-1}\alpha_{T+j}\}$ is similar in form to α^*_{T+j} except that the wrong discount factor is being used and the sum is truncated. Nevertheless we claim that if α^*_{T+j} is small, then so is the term in question. To see this, consider the identity $\alpha^*_{T+j} = \delta_1\alpha^*_{T+j+1} + (1 - \delta_1)\alpha_{T+j}$, which holds for all j. From this we obtain

$$\textstyle\sum_{j=1,m}\delta_2^{j-1}\alpha^*_{T+j} = \delta_1\,_{j=1,m}\,\delta_2^{j-1}\alpha^*_{T+j+1} + (1 - \delta_1)\sum_{j=1,m}\delta_2^{j-1}\alpha_{T+j,}$$

and after rearranging terms,

$$\textstyle\sum_{j=1,m}\delta_2^{j-1}\alpha_{T+j} = [1/(1 - \delta_1)][\alpha^*_{T+j} + (\delta_2 - \delta_1)\sum_{j=1,m-1}\delta_2^{j-1}\alpha^*_{T+j+1} - \delta_1\delta_2^{\,m-1}\alpha^*_{T+m+1}]. \quad (7)$$

All the terms $\alpha^*_{T+j+1}, \ldots \alpha^*_{T+j+m}$ are at least $-\lambda/2$, the term α^*_{T+m+1} is at most $1 + \lambda/2$, and $\delta_1\delta_2^{m-1}$ δ_2^m λ. Thus, the right-hand side of (7) is bounded above by $\alpha^*_{T+j}/(1 - \delta_1) + c'\lambda$, where $c' > 0$ depends only on the discount factors. The left-hand side of (7) is the term in curly brackets in (6). Substituting this expression into (6) we see that $\beta^*_{T+j} + [(1 - \delta_2)/(1 - \delta_1)]\alpha^*_{T+j}$ $15\lambda - (1 - \delta_2)c'\lambda$. Since $\alpha^*_{T+1}, \beta^*_{T+1}$ $-\lambda/2$ we conclude that both α^*_{T+1} and β^*_{T+1} are bounded above by $c\lambda$ for some c that depends only on the discount rates δ_1 and δ_2. This concludes the proof of Lemma 2.

Lemma 3. For every positive integer m and all sufficiently small $\lambda > 0$, if $(\mathbf{A}, \mathbf{B}) \in P$, then there exists a history h^T such that, conditional on h^T at time T, both players randomize in each of the periods $T + 1, \ldots, T + m$.

Proof. As in the proof of Lemma 2 choose m 1 such that $\delta_2^m \leq \lambda$ and let $0 < \varepsilon \leq \lambda/m4^m e^m$. Assume in addition that $\varepsilon' = \varepsilon^{4m}$. Now apply Lemma 1 with 2m instead of m: there is a time T such that, with probability at least $1 - \varepsilon'$, the next-period forecasts are within ε of being correct for the periods $T + 1, \ldots, T + 2m$.

For each h^{T+j}, $0 \leq j \leq 2m - 1$, say that h^{T+j} is *good* if both players' next-period forecasts are within ε of being correct; otherwise h^{T+j} is *bad*. Say that h^{T+j} is γ-good if it is good and, conditional on h^{T+j} occurring in period $T + j$, the probability is at most γ that someone makes a bad next-period forecast in any continuation of h^{T+j} through period $T + 2m - 1$.

By choice of T there is at least one state, h^T, that has positive probability under the strategies and is ε'-good. Lemma 2 implies that the expected discounted payoffs from T + 1 on are bounded above by $c\lambda$. We claim this implies that both players randomize in period T + 1, and in fact each of them chooses each action with probability at least ε. Suppose, to the contrary, that some player (say player 1) chooses action 1 with probability less than ε. Since h^T is good, player 2 forecasts that 1 will play action 2 with probability at least $1 - 2\varepsilon$. But then player 2 could obtain a higher expected payoff by mismatching (playing action 1) in period T + 1 and randomizing fifty-fifty in every period thereafter. (The expected payoff from the latter is at least $(1 - \delta_2)[(1 - \lambda/2) + \delta_2(-\lambda/2)]$, which is greater than $c\lambda$ for all sufficiently small λ.) This contradiction shows that player 1 chooses each action in period T + 1 with probability at least ε, and the same holds for player 2.

It follows that each of the four possible continuations of h^T to period T + 1 has probability at least ε^2. Since $\varepsilon^2 > \varepsilon'$ and h^T is ε'-good, none of these four continuations can be bad, and in fact each of them must be at least $(\varepsilon'/\varepsilon^2)$-good. Now apply Lemma 2 again (redefining ε' to be $\varepsilon^{4m}/\varepsilon^2$) and conclude that, for every continuation of h^T to some h^{T+2}, the conditional expected payoffs from period T + 2 forward are bounded above by $c\lambda$. As before we conclude that both players randomize in period T + 2, each putting at least ε on each action. Continuing in this manner, we deduce that both players randomize in every continuation of h^T to period T + m. This concludes the proof of Lemma 3.

The gist of the proof so far is that, if the payoff realizations (**A**, **B**) lead to good predictions with μ-positive probability, then for every sufficiently large positive integer m, there exists a state h^T that induces randomization by both players in each of the next m periods. We now show that this implies that the payoffs are zeroes of a function whose set of zeroes has ν-measure zero. This will show that good prediction occurs with ν-measure zero.

Let h^T be any state and let m be a positive integer. Suppose that player 1 plays action 1 in each of the periods T + 1 to T + m, after which he plays an optimal strategy given his beliefs. We can write his expected utility, discounted to time T + 1, as a function of his payoff matrix **A** as follows:

$$U_1(\mathbf{A}) = \theta_1 a_{11} + (1 - \delta_1^m - \theta_1)a_{12} + \delta_1^m R_1(\mathbf{A}).$$

Here θ_1 comes from player 2's randomization between actions 1 and 2, and the remainder term $R_1(\mathbf{A})$ is concave and bounded. In fact, $|R_1(\mathbf{A})| \leq (1 - \delta_1^m)(|a_{11}| + |a_{12}| + |a_{21}| + |a_{22}|)$. Likewise we can compute player 1's expected utility from playing action 2 in each of the next m periods and an optimal strategy thereafter:

$$U_2(\mathbf{A}) = \theta_2 a_{21} + (1 - \delta_1^m - \theta_2)a_{22} + \delta_1^m R_2(\mathbf{A}),$$

where $R_2(\mathbf{A})$ satisfies the same bound as $R_1(\mathbf{A})$. All of these functions depend on h^T, but we will suppress this dependence for the moment. It will be convenient to consider a one-dimensional subspace of the payoff matrices **A**. Namely, for every four real numbers w, x, y, z, let $\psi_{x,y,z}(w)$ be the 2 x 2 matrix with entries $a_{11} = w + x$, $a_{12} = w - x$, $a_{21} = y$, and $a_{22} = z$. Given x, y, and z, define the following function of w:

$$F_{x,y,z}(w) = U_1(\psi_{x,y,z}(w)) - U_2(\psi_{x,y,z}(w))$$

$$= K_{x,y,z} + (1 - \delta_1^m)w + \delta_1^m[R_1(\psi(w)) - R_2(\psi(w))],$$

where $K_{x,y,z}$ is a linear function of x, y, z and does not depend on w. The functions $R_i(\psi(w))$ are concave and bounded by the same bound as before. By choosing m to be sufficiently large, we can ensure that $F_{x,y,z}(w)$ is strictly monotone increasing in w. It follows that, for any triple x, y, z, there is at most one value of w such that $F_{x,y,z}(w) = 0$. Since x, y, z are drawn from the continuous density ν, we have $P[F_{x,y,z}(w) = 0|x, y, z)] = 0$. By the smoothing theorem (i.e., the law of iterated expectations), it follows that $P[\{(w, x, y, z) \in \Re^4 : F_{x,y,z}(w) = 0\}] = 0$.

To state this in terms of the matrix **A**, let $G(\mathbf{A}) = F_{x,y,z}(a_{11} + a_{12})/2)$ where $x = (a_{11} - a_{12})/2$, $y = a_{21}$, and $z = a_{22}$. The preceding implies that $P[\{\mathbf{A}: G(\mathbf{A}) = 0\}] = 0$. Recalling that F (and thus G) are conditional on a particular history h^T, we can write this as $P[\{\mathbf{A}: G(\mathbf{A}) = 0\}| h^T] = 0$. Hence

$$\sum_{h^T} P[\{\mathbf{A}: G(\mathbf{A}) = 0\}| h^T] P(h^T) = 0. \qquad (8)$$

Suppose now that (**A**, **B**) is a pair for which good prediction holds. Let h^T be a history as guaranteed by Lemma 3, where m is sufficiently large that F is strictly monotone increasing in w. By Lemma 3, player 1 randomizes in each of the periods T + 1, . . ., T + m. Hence he is indifferent between playing action 1 or action 2 in each of these periods. Hence for this **A** and this h^T, $U_1(\mathbf{A}| h^T) - U_2(\mathbf{A}| h^T) = 0$. Hence **A** lies in the set of ν-measure zero defined by (8). Thus we have proved the following: *there are* η*-almost no payoff realizations* (**A**, **B***) such that both players learn to predict with positive probability*. This establishes the first claim of the theorem.

For future reference we note that we have actually established the following fact.

Lemma 4. If m is large enough and λ is small enough, then the ν-probability is zero that there exists a state h^T such that, conditional on h^T at time T, both players randomize in each of the periods T + 1, . . . , T + m.

It remains to be shown that, for ν-almost all (**A**, **B**), play fails to come close to the set of Nash equilibria in the sense that (4) fails to hold for almost all histories h. The first step is to show that all Nash equilibria of the repeated game are sufficiently mixed in each time period provided that λ is sufficiently small.

Lemma 5. There exists $\varepsilon > 0$ and $\lambda' > 0$ such that, whenever $0 < \lambda \leq \lambda'$, every Nash equilibrium of the repeated game puts probability at least 2ε on each action in every time period.

The proof is similar to that of Lemma 2; in outline it runs as follows. First, each player can guarantee himself at least $-\lambda/2$ in every time period, so in equilibrium the expected payoff of each player must also satisfy this lower bound. Using this and the fact that the actual payoffs in each period sum to at most λ, we can bound each player's expected payoff from above by an expression of form $k\lambda$, where k is a positive constant that depends only on the discount factors. It

follows that no player puts probability less than 2ε on any action in any period, because otherwise the opponent could play a pure strategy in that period, playing a fifty-fifty mixed strategy thereafter, and obtain an expected payoff that is greater than $k\lambda$ (assuming ε and λ are sufficiently small).

Fix $\lambda \in (0, \lambda']$. For each pair of payoff matrices (**A**, **B**), let N(**A**, **B**) be the set of all histories h such that expression (4) holds, i.e., such that play comes close to Nash in a weak sense. We are going to show that there are μ-almost no such histories for ν-almost all (**A**, **B**). This is a consequence of the following.

Lemma 6. Let (**A**, **B**) be a pair of payoff matrices such that (4) holds with μ-positive probability. Then for every positive integer m and every sufficiently small λ, there exists a state h^T such that, conditional on h^T, each player randomizes in each of the periods $T + 1, \ldots, T + m$.

By choosing m large enough, it follows from Lemma 4 that there are ν-almost no payoff realizations (**A**, **B**) with this property. In other words, for ν-almost all payoff realizations play fails to come close to Nash. Thus, once we establish Lemma 6, we will have completed the proof of the theorem.

Proof of Lemma 6. Fix a pair (**A**, **B**) such that (4) holds with μ-positive probability. Choose ε and λ such that every element of Q^N puts probability at least 2ε on each action in each time period, as guaranteed by Lemma 5. Let m be a positive integer, and let $\varepsilon' = \varepsilon^{4m}$. There exists a time T such that, with μ-probability at least $1 - \varepsilon'$, $d(q^t(h^t), Q^N) \leq \varepsilon$ for every h^t in the interval $T \leq t \leq T + m$. (If this were not so, condition (4) would hold with μ-probability zero, contrary to our assumption.)

Say that a history h^t is *good* if $d(q^t(h^t), Q^N) \leq \varepsilon$. It is *very good* if it is good and all of its successors for the next m periods are good. If a history is good then each action is played in the next period with probability at least ε. Hence every continuation of a good history occurs with probability at least ε^2. If no history at time T is very good, then the μ-probability of a bad history occurring in the interval $T, T + 1, \ldots, T + m - 1$ is at least $\varepsilon^{2(m-1)} > \varepsilon'$, contrary to our assumption. Hence there exists h^T such that $d(q^t(h^t), Q^N) \leq \varepsilon$ for every continuation of h^T in the interval $T + 1 \leq t \leq T + m$, and hence both players randomize for m periods in succession. By Lemma 4 this happens with ν-probability zero. This concludes the proof of Lemma 6, and thereby the proof of the theorem.

References

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

---------- (1990): *Essays on Foundations of Game Theory*. Oxford: Basil Blackwell.

---------- (1991): "DeBayesing Game Theory," International Conference on Game Theory, Florence, Italy.

Diaconis, Persi, and David Freedman (19860; “On the Consistency of Bayes Estimates,” *Annals of Statistics*, 14, 1-26.

Fudenberg, Drew, and David Kreps, (1993): “Learning Mixed Equilibria,” *Games and Economic Behavior, 5, 320-367.*

Harsanyi, John (1973): “Games with Randomly Disturbed Payoffs; A New Rationale for Mixed-Strategy Equilibrium Points,” *International Journal of Game Theory*, 2, 1-23.

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

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

---------- (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," *Econometrica*, 61, 1019-1045.

Lehrer, Ehud, and Rann Smorodinsky (1997): “Repeated Large Games with Incomplete Information,” *Games and Economic Behavior*, 18, 116-134.

Miller, R. I. and Chris W. Sanchirico (1997): "Almost Everybody Disagrees Almost All the Time: The Genericity of Weakly-Merging Nowhere." Department of Economics, Columbia University.

Miyasawa, K. (1961): “On the Convergence of the Learning Process in a 2 x 2 Non-Zero-Sum Two-Person Game,” Economic Research Program, Research Memorandum no. 33, Princeton University, Princeton NJ.

Monderer, Dov, and Lloyd Shapley (1996): “Fictitious Play Property for Games with Identical Interests,” *Journal of Economic Theory* 68, 258-265.

Nachbar, John. H. (1997): "Prediction, Optimization, and Learning in Games," *Econometrica*, 65, 275-309.

---------- (1999): "Rational Bayesian Learning in Repeated Games." Working Paper, Department of Economics, Washington University, St. Louis.

---------- (2001): "Bayesian Learning in Repeated Games of Incomplete Information,” *Social Choice and Welfare*, 18, 303-326.

Nash, John (1950): *Non-cooperative Games*. Ph.D. Dissertation, Princeton University.

Nyarko, Yao (1998): “Bayesian Learning and Convergence to Nash Equilibria without Common Priors,” *Economic Theory*, 11, 643-655.