RATIONAL LEARNING LEADS TO NASH EQUILIBRIUM

by

Ehud Kalai∗

and

Ehud Lehrer∗

March 1990
Final Revision, January 1993

∗Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management and Department of Mathematics, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208.
RATIONAL LEARNING LEADS TO NASH EQUILIBRIUM

by

Ehud Kalai

and

Ehud Lehrer

Forthcoming in Econometrica

March 1990

Final Revision, January 1993

*Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, and Department of Mathematics, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208.

The authors wish to thank Robert Aumann, Larry Blume, David Easley, Itzhak Gilboa, Sergiu Hart, Martin Hellwig, James Jordan, Dov Monderer, Dov Samet, Sylvain Sorin, Vernon Smith, Robert Wilson, and anonymous referees for helpful discussions and suggestions. This research was partly supported by Grants Nos. SES-9011790 and SES-9022305 from the National Science Foundation, Economics, and by the Department of Economics and the C. V. Starr Center for Applied Economics of New York University.
Abstract

Rational Learning Leads to Nash Equilibrium

by

Ehud Kalai and Ehud Lehrer

Each of \( n \) players, in an infinitely repeated game, starts with subjective beliefs about his opponents' strategies. If the individual beliefs are compatible with the true strategies chosen, then Bayesian updating will lead in the long run to accurate prediction of the future play of the game. It follows that individual players, who know their own payoff matrices and choose strategies to maximize their expected utility, must eventually play according to a Nash equilibrium of the repeated game. An immediate corollary is that, when playing a Harsanyi-Nash equilibrium of a repeated game of incomplete information about opponents' payoff matrices, players will eventually play a Nash equilibrium of the real game, as if they had complete information.
1. **Introduction**

The concept of Nash (1950) equilibrium has become central in game theory, economics, and other social sciences. Yet the process by which the players learn to play it, if they do, is not fully understood. This is not surprising for games played only once where players have no observations to guide them and learning theories are restricted to model thought processes. However, in repeated interaction, where the players do have enough time to observe the behavior of their opponents, one can hope to obtain a statistical learning theory that leads to Nash equilibrium.

While some experimental work (see, for example, Smith, 1990; McCabe et al., 1991; Linhart et al., 1989; Roth et al., 1991; and Prasnikar and Roth, 1992) supports the supposition that agents in repeated games do learn to play Nash equilibrium, no satisfactory theoretical explanation for this phenomenon exists. This is in spite of continuously growing game theoretic literature on repeated games with or without complete information (see Aumann, 1981, and Mertens, 1987, for surveys that are already outdated; see the forthcoming book by Mertens et al., 1990, for state-of-the-art knowledge on repeated games with and without complete information), and the interest in the topic of learning in economics (e.g., Blume et al., 1982; Jordan, 1985; Easley and Kiefer 1988; Bray and Kreps, 1987; McLennan, 1987; Grandmont-Laroque 1990; Woodford, 1990; and references therein).

The construction of processes that converge to Nash equilibrium is not a new topic to game theorists. Robinson (1951), Miyasawa (1961), and Shapley (1964) studied convergence properties of fictitious play. More recently, however, Fudenberg and Kreps (1988) recognized that such
mechanisms can be used as a basis to model learning by players with bounded rationality, and a large literature on the subject has developed. A sample of such papers includes: Selten, 1988; Crawford, 1989; Canning, 1992; Jordan, 1991, 1992; Brock et al., 1988; Milgrom and Roberts, 1991; Stanford, 1991; and Fudenberg and Levine, 1993.

Most of this literature, however, builds on assumptions not applicable to our subject of repeated play among a small number of subjectively rational agents. The dynamic models studied in this literature are often ones of fictitious play or of random matching in a large population, and the behavior of the players is often modeled to be "myopic" or "bounded" in other ways. For our players, this implies shortcomings of the following types.

1. In trying to predict future opponents' behavior, a boundedly rational player ignores the fact that his opponents are also engaged in a dynamic learning process.

2. A myopic player would not perform a costly experiment no matter how high the resulting expected future payoffs are.

3. A myopic player ignores strategic considerations regarding the future. For example, even if he believes his opponent in a repeated prisoners' dilemma game is playing a trigger strategy (see Example 2.1), his consideration of immediate payoff may lead him to a choice of a long run inferior action.

In order to overcome these types of flaws, we take a traditional decision theoretic approach to the problem. We assume that the players are engaged in a standard perfect-monitoring infinitely repeated game with discounting. Thus, each one possesses a fixed matrix which specifies his
payoff for every action combination taken by the group of players. In every period every player chooses his individual action, and the vector of chosen actions, through the individual payoff matrices, determines the period payoffs for all the players. Perfect monitoring means that before making the choice of a period's action, the player is informed of all the previous actions taken. Each player possesses a discount factor that he uses in evaluating future payoffs. His goal is to maximize the present value of his total expected payoff.

The players are assumed to be subjectively rational in the following sense. Each one starts with subjective beliefs about the individual strategies used by each of his opponents. He then uses these beliefs to compute his own optimal strategy. The strategies used for play, and for describing beliefs about an opponent's play, are behavior ones. Thus, they allow randomization in the choices of periods' actions (Section 3.4 elaborates on this topic and Kuhn's theorem, i.e., the fact that a probability distribution over many strategies can be replaced by a single behavior one). It is important to note that, unlike the stronger notion of rationalizable strategies (see Bernheim, 1986, and Pearce, 1984), the knowledge assumptions implicit in the definition of a subjectively rational strategy are weak. In order to choose one a player must only know his own payoff matrix and discount parameter, but needs not have any information about opponents' payoff matrices, nor assume anything about their rationality.

The main message of the paper is the following. If the players start with a vector of subjectively rational strategies, and if their individual subjective beliefs regarding opponents' strategies are "compatible with the
truly chosen strategies," then they must converge in finite time to play
according to an $\varepsilon$-Nash equilibrium of the repeated game, for arbitrarily
small $\varepsilon$. Moreover, their Bayes-updated posterior subjective beliefs
regarding future play of the game will become accurate with time. In other
words, they will learn to predict the future play of the game and to play
$\varepsilon$-Nash equilibrium for arbitrarily small $\varepsilon$. Some features and assumptions
of the model should be emphasized.

1. The players' objective is to maximize, relative to their
individual subjective beliefs, their long term expected discounted payoff.
Learning is not a goal in itself here but is, rather, a consequence of an
overall individual payoff maximization plans. Learning is acquired as the
real game progresses. In this sense it may be thought of as learning by
playing, paralleling the economic literature on "learning by doing" (see
Arrow (1962)).

2. Learning takes place thorough Bayesian updating of the individual
prior beliefs. This follows the traditional approach of games of incomplete
or imperfect information, e.g., Kuhn (1953), Harsanyi (1967), and Aumann and
Maschler (1967). However, since the use of Bayesian updating is a
consequence of expected utility maximization, (2) is already a consequence
of assumption (1).

3. We depart from the standard assumptions of game theory by not
requiring that the players have full knowledge of each others' strategies,
nor do they have commonly known prior distributions on the unknown
parameters of the game. (We do not prohibit such assumptions but they are
not necessary in our model.) Rather, we replace these assumptions by a
weaker one of compatibility of beliefs with the truth. This assumption
requires that players' subjective beliefs do not assign zero probability to events that can occur in the play of the game. In mathematical language, this means that on future play paths, the probability distribution induced by the chosen strategies must be absolutely continuous with respect to the probability distributions induced by the private beliefs of the players, i.e., any positive probability set of paths must be assigned some positive probability by each player. As an example, one may think of a situation where the beliefs about an opponent's strategy assign a small positive probability to the strategy actually chosen. In this case, we say that the beliefs contain a grain of truth, and compatibility of the beliefs with the truth is assured. Further discussion of these assumptions and their necessity will follow in subsequent sections.

An important corollary to the main result of this paper deals with Harsanyi-Nash equilibria of an n-person infinitely repeated game under discounting with n-sided incomplete information about opponents' payoff matrices. Assuming that the number of possible payoff matrices is finite or countable, the grain of truth condition stated above is satisfied. It follows that at such an equilibrium the players will eventually play according to an c-Nash equilibrium of the infinitely repeated realized game (the one with complete information) as if the uncertainties were not present. This corollary and its relation to Jordan's (1991) results will be discussed later in this paper.

As mentioned earlier, myopic theories of simultaneous learning involve fundamental difficulties due to the fact that what is being learned keeps changing. If players assume that their opponents' actions are fixed, yet the opponents, too, learn and change their own actions as a result of what
they learn, inconsistencies are likely to occur. Indeed, as is shown, for example, by Kirman (1983), and by Nyarko (1991), learning may never converge or, worse yet, it may converge to false beliefs. Dynamic approaches, like the one taken in this paper, have the potential to overcome this difficulty. They attempt to learn the strategies (or, more precisely, the reaction rules that guide the opponents), of the infinitely repeated game. These strategies, which do not change, already contain the fixed learning rules.

Also, existing results of game theory suggest learning to play Nash equilibrium in a repeated game should be easier than in a one shot game. Considering the extreme case with completely patient players, i.e., discount factor equals one, the folk theorem tells us that all feasible individually rational payoffs are Nash payoffs. This suggests that many, yet certainly not all, play paths are Nash paths. Thus, our result regarding convergence to Nash paths seems to be more meaningful for moderate or low discount parameters.

A second difficulty, associated with learning models, concerns experimentation. In order to avoid getting stuck at suboptimal solutions, a well designed process should occasionally try randomly generated experimentation. For example, the randomly generated mutants in the evolutionary models play such a role. However, as can be seen in Fudenberg-Kreps (1988), in a rational choice model the optimal determination of when and how to experiment is difficult. The subjectively rational approach suggested here overcomes this difficulty: every action in the current model, including experimentation, is evaluated according to its long run contribution to expected utility. And maximization, relative to posterior probability distribution regarding opponents' strategies, yields well
defined criterion for determining choices. Thus, a player, with a given
discount parameter, will experiment when he assesses that the information
gained will contribute positively to the present value of his expected
payoff. Given the strategic nature of the interaction in our model, for
some subjective beliefs regarding opponents' strategies, a player may find
it in his interest to choose randomly when and how to experiment. While, in
general, optimal experimentation computed against subjective beliefs does
not lead to an individually optimal solution (see the multi-arm bandit
example in the last section), it does so in the current paper due to the
special assumptions of perfect monitoring and knowledge of own payoff
matrices.

In addition to perfect monitoring, knowledge of own payoff matrices,
and compatibility of beliefs with the truth, our model contains several
additional restrictive assumptions. Independence of strategies is imposed
in two places. First, it is implicitly assumed that players' actual
strategies are chosen independently. In other words, players' choices
cannot depend on random events (unless they are completely private) since
such dependencies may lead to correlated strategies, which are assumed away
in the model. But this assumption of independence is also imposed on the
subjective beliefs of the players. In his beliefs regarding opponents'
strategies, a player assumes that the opponents choose their strategies
independently of each other.

Also, the assumption that players maximize their expected payoffs is
quite strong for infinitely repeated games. While this assumption is common
in game theory and economics, the solution of such a maximization problem,
in infinitely repeated games, may be very demanding.
Our preliminary studies regarding the relaxation of the various assumptions above, with the exception of truth-compatible beliefs, indicate that bounded learning will lead the players to correlated equilibrium (see Kalai-Lehrer, 1992) rather than Nash equilibrium. The need and possibilities of relaxing the truth compatibility assumption are discussed in several of the next sections.

Our proof of the convergence to playing an ε-Nash equilibrium is divided into three steps. The first establishes a general self-correcting property of Bayesian updating. This is a modified version of the seminal Blackwell and Dubins' (1962) result about merging of opinions. We give an independent easy proof of their result and an alternative characterization of their notion of merging.

When applied to our model, the self-correcting property shows that the probability distributions describing the players' beliefs about the future play of the game must converge to the true distribution. In other words, the beliefs and the real play become realization equivalent. At such time all learning possibilities have been exhausted. Remaining disagreement of the beliefs and the truth may only exist off the play path, and therefore will never be observed. We refer to such a situation of no further learning as subjective equilibrium. The notion of such an equilibrium and the fact that it may yield play different from Nash equilibrium were observed earlier in models of dynamic optimization, e.g., the multi-arm bandit literature (see, for example, Rothchild, 1974) and in a repeated game set-up by Fudenberg and Kreps (1988). Also, in a different learning model developed independently of ours, Fudenberg and Levine (1993a) developed and studied a
closely related notion called self-confirming equilibrium. We refer the 
reader to Battigalli et al. (1992) for a survey of the history of this 
concept.

The last step of the proof shows that in our model, the behavior 
induced by a subjective equilibrium, and even its perturbed versions, 
approximates the behavior of an ε-Nash equilibrium. Since this last step 
has independent interest of its own, and since proving it involves long 
computations not related to learning, we leave it to a companion paper (see 
Kalai and Lehrer, 1993).

Section 2 of this paper contains examples and additional elaborations 
on the approach taken and the assumptions made. The reader can skip it and 
move directly to Sections 3 and 4 which contain the formal presentation of 
the model and of the main results. Section 5 is devoted to the self 
correcting property of false priors by means of Bayesian updating. Section 
6 contains applications to repeated games with incomplete information and 
the relation to Jordan's (1991) results. Finally, in Section 7, we give 
further elaborations on some of the assumptions and possible extensions.

2. Examples and Elaboration

In the two person games that follow, we will refer to player 1, PI, as 
he, and player 2, PII, as she.

Example 2.1: Infinitely Repeated Prisoners' Dilemma Games.

As usual for these games, we will denote the possible actions for each 
of the two players in each stage of the game by A—to describe aggressive 
behavior, and by C—to describe cooperative behavior. The following matrix
represents the stage payoffs to PI as a function of pairs of action choices:

\[
\begin{array}{c|cc}
 & A & C \\
\hline
A & c & a \\
\hline
C & d & b \\
\end{array}
\]

As usual, we assume that \(a > b > c > d\). PI uses a discount parameter \(\lambda_1\) (\(0 < \lambda_1 < 1\)) to evaluate infinite streams of payoffs. We use the convention that \(\lambda_1\) close to 0 describes an impatient (myopic) player. PII has a similar payoff matrix and discount parameter but not necessarily with the same numerical values as the ones of PI. We assume here, as we do throughout this paper, that each player knows his own parameters, and that the game is played with perfect monitoring. That is, prior to making the choice in every stage, a player is informed of all the choices made by both players in all previous stages.

Departing from the traditional game theoretic approach, we do not explicitly model a player's knowledge about the parameters (payoff matrices, discount parameters, etc.) of his opponent. Instead, a player starts with prior subjective beliefs, described by a probability distribution, over the strategies his opponent will use. We assume that a player uses any specific knowledge he has about his opponent in creating these beliefs.

To illustrate such beliefs we consider a countable set of (pure) strategies \(g_t\) for \(t = 0, 1, 2, \ldots, \infty\), defined as follows. \(g_\infty\) is the well-known
(grim) trigger strategy. This strategy prescribes cooperation initially and after fully cooperative histories, but "triggers" to the aggressive action after every history that contains any aggression by either of the two players. For \( t < \infty \), \( g_t \) coincides with \( g_\infty \) at all histories shorter than \( t \) but prescribes the aggressive action \( A \) after all histories of length \( t \) or more. In other words, if not triggered earlier, \( g_t \) will prescribe unprovoked aggression starting from time \( t \) on. With this convention, \( g_0 \) is the constant aggressive strategy.

Suppose PII believes that PII is likely to cooperate by playing her grim trigger strategy; but he also believes there are positive probabilities that she will stop cooperating earlier for other reasons. More precisely, he will assign her strategies \( g_0, g_1, \ldots, g_\infty \) probabilities \( \beta = (\beta_0, \beta_1, \ldots, \beta_\infty) \) that sum to 1 and with each \( \beta_t > 0 \). Depending on his own parameters he chooses a best response strategy of the form \( g_{T_1} \) for some \( T_1 = 0, 1, \ldots \) or \( \infty \). PII holds similar beliefs, represented by a vector \( \alpha \), about PI's strategy, and chooses a strategy \( g_{T_2} \) as her best response. Now the game will really be played according to the two strategies \( (g_{T_1}, g_{T_2}) \).

It is easy to see that the beliefs are compatible with the chosen strategies. All positive probability events in the game, e.g., cooperation up to time \( t < \min(T_1, T_2) \), aggression in all times exceeding the min \( (T_1, T_2) \), are assigned positive probability by the original beliefs of both players. Thus, the results described earlier must hold.

Indeed, learning to predict the future play must occur. If, for instance, \( T_1 < T_2 \), then from time \( T_1 + 1 \) on, PII Bayesian updated beliefs regarding PI's choice will assign probability 1 to his choice of \( g_{T_1} \) and she will predict correctly the future noncooperative play. PI, on the other
hand, will never fully know her strategy since he would only know that $T_2 > T_1$. But he will still be able to infer the forthcoming noncooperative play. This should clarify the point that players do not learn the strategy of their opponent off the play path; they only learn to predict actual future play against the strategy they themselves use. Also notice that accuracy of the above predictions did not rely on $T_1$ and $T_2$ being optimal choices. It only relied on correct updating of the truth-compatible subjective beliefs.

A second point to emphasize is that players' beliefs will not necessarily coincide with the truth after a finite time; beliefs may only converge to the truth as time goes by without ever coinciding with it. Suppose, for example, that $T_1 = T_2 = \infty$. Now the only resulting play path is the totally cooperative one. After playing it for $t$ periods, PI, for example, will infer that she did not choose $g_0, g_1, \ldots, g_t$ and his Bayesian updated belief will assign probabilities $(\beta_{t+1}, \ldots, \beta_{\infty})/\sum_{t+1}^{\infty} \beta_r$ to her remaining strategies: $(g_{t+1}, \ldots, g_{\infty})$. Since $\beta_{\infty} > 0$, after sufficiently long time, his posterior probability $\beta_{\infty}/\sum_{t+1}^{\infty} \beta_r$ will be arbitrarily close to one and he will be almost certain that she chose $T_2 = \infty$.

The second main result, regarding convergence to Nash equilibrium play, is also easily seen in this example. The only two play paths that can survive after a long play are those generated by Nash equilibrium (and hence also by $\varepsilon$-Nash equilibrium). The totally aggressive path results from Nash equilibria regardless of the parameters of the game. But if the discount parameters are "generous," then the totally cooperative play path can also be obtained at a Nash equilibrium. Notice, however, that the overall play of these subjectively rational players can be of a type not generated by any Nash equilibrium, or even $\varepsilon$-Nash equilibrium for small $\varepsilon$. For example, the
path that is fully cooperative up to time 3 and not cooperative afterwards cannot be the outcome of any Nash, or small ε-Nash, equilibrium. But such a path will be generated by the players if their subjective individual beliefs assign high probability to the opponent not cooperating at time 4.

Nevertheless, as follows from our second main result, from a certain time on these players will follow an ε-Nash equilibrium path. Thus, if the true parameters of the game allow only the totally aggressive Nash equilibrium (and hence the only path compatible with arbitrarily small ε-Nash equilibrium is the totally aggressive one), then at least one of the $T_i$'s must be finite, and eventually constant mutual aggression must emerge. If, on the other hand, the game's parameters permit trigger strategies as a Nash equilibrium, then it is also possible that both $T_i$'s are infinite and the play path follows cooperation throughout.

It is easy to observe in this example that players who hold optimistic prior probabilities (high $\alpha_0$ and $\beta_0$) will follow a cooperative path while pessimistic players must eventually follow a noncooperative path. Thus, in the case of multiple equilibria, initial prior beliefs determine the final choice.

Example 2.2: Absolute Continuity and Grain of Truth Assumptions

In Example 2.1, each player's private beliefs assigned a strictly positive probability to the strategy actually chosen by the opponent. This condition, that beliefs regarding opponent's strategies contain a grain of truth, is stronger than needed.

Suppose, for instance, that in Example 2.1, PII's beliefs, given by the vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_\infty)$, had sufficiently low values of all $\alpha_i$'s with
t < \infty to allow the trigger strategy \( g_\infty \) as her best response. Then the well-known tit-for-tat (tft) strategy (where she starts by cooperation and then proceeds to mimic her opponent's last move) can also be chosen as a best response. If she actually chooses tft as her strategy, then PI's beliefs about her strategy do not contain a grain of truth, given that his beliefs assign probability zero to non-trigger strategies. Yet his beliefs regarding future play paths will contain a grain of truth. Consider the play paths \( z_0, z_1, \ldots, z_\infty \) with \( z_\infty \) describing the path in which both players cooperate up to time \( t \), he cooperates and she aggresses at time \( t \), and both aggress from time \( t + 1 \). If PI's beliefs about PII's strategy are described by the vector \( \beta \) as in Example 2.1, and in response he chooses \( g_\infty \) for himself, then his induced beliefs on the future play paths are given by a distribution \( \tilde{\mu}_1 \) which assigns probability \( \beta_1 \) to each of the paths \( z_\infty \). Given both of their choices, the true distribution on future play paths, \( \mu \), will assign probability one to the path \( z_\infty \). We now can write \( \tilde{\mu}_1 = \varepsilon \mu + (1 - \varepsilon) \hat{\mu} \) for some probability distribution \( \hat{\mu} \) and with \( \varepsilon = \beta_\infty > 0 \). When this is the case, i.e., when the belief distribution on future play paths assigns positive weight to the true distribution, we say that the beliefs on play paths contain a grain of truth. This last condition, which is weaker than the belief on strategies containing a grain of truth, is also sufficient for our main result.

The sufficient condition we end up using is weaker yet. We require that each player's belief distribution on play paths, \( \tilde{\mu}_1 \), not rule out positive probability events according to the real probability distribution, \( \mu \). That is, there should be no event in the play of the infinite game which can occur, i.e., has a \( \mu \) positive probability, yet be ruled out by the
beliefs of an individual player, i.e., has a zero probability according to \( \tilde{\mu}_1 \). In mathematical language, we require that \( \mu \) be absolutely continuous with respect to each \( \tilde{\mu}_1 \) (\( \mu \ll \tilde{\mu}_1 \)).

To understand the difference between the grain of truth conditions and absolute continuity it is useful to consider behavior and mixed strategies, i.e., ones that allow for randomization in the choice of actions and strategies (the next section contains a more detailed discussion of Kuhn's theorem and these notions). Suppose PI's beliefs about PII's strategy are as in Example 2.1 with \( \beta_t = (1/3)^{t+1} \) for \( t < \infty \) and \( \beta_\infty = 1/2 \). Suppose that his choice in response to these beliefs is to play \( g_\infty \). His induced beliefs on the future play paths, \( \tilde{\mu}_1 \), assign probability \((1/3)^{t+1}\) to each of the paths \( z_t \) with \( t < \infty \) and \( 1/2 \) to \( z_\infty \). If, unlike his beliefs, she chooses to randomize over the choices of \( g_t \) with probabilities \((1/2)^{t+1}\) (zero probability on \( g_\infty \)), then the real distribution on future play paths, \( \mu \), assigns probability \((1/2)^{t+1}\) to each of the paths \( z_t \) and zero to \( z_\infty \). It is easy to check that \( \tilde{\mu}_1 \) cannot be written as \( \varepsilon \mu + (1 - \varepsilon) \tilde{\mu} \) with a positive \( \varepsilon \) for any probability distribution \( \mu \). Thus, even his beliefs about play paths do not contain a grain of truth. Yet the absolute continuity condition holds. Every event in the play of the game that has \( \mu \)-positive probability, i.e., contains some paths \( z_t \) with finite \( t \), is assigned a positive probability by the belief distribution \( \tilde{\mu}_1 \). Thus, the results of this paper regarding learning and convergence to Nash equilibrium must hold.

In the above example, however, if PI assigned probability one to her playing \( g_\infty \), yet she randomized on \( g_t \) with probability \((1/2)^{t+1}\) and probability zero on \( g_\infty \), then the absolute continuity condition would fail. And, indeed, learning and convergence to Nash equilibrium will fail, too.
Example 2.3: On the Limitation of the Absolute Continuity Assumption:

Consider a repeated game with player one having to choose between $a$ and $r$ in every stage. Suppose player two believes that player one flips a coin to choose $a$ with probability $1/2$ and $r$ with probability $1/2$ after every history. This means that player two believes that future plays of player one are independent of his past actions and learning from the past is hopeless. Even if player one played the constant strategy $L$, always playing $a$, player two will not learn it since, given her initial beliefs, she will always dismiss long chains of $a$'s as random outcomes. Notice that this is a situation where the absolute continuity assumption is violated. The event "$a$ will be played forever" has probability one but is assigned probability zero in the beliefs of player two.

The discussion above shows that, without the absolute continuity assumption, or other assumptions that connect the future to the past, learning in general cannot take place. We know, however, that weaker assumptions suffice for approximate learning and for convergence of measures in a weak sense. Assume, for example, that player one plays a constant behavior strategy by which he randomizes with probability $\lambda$ on $a$ and $(1 - \lambda)$ on $r$ after every history of the game. Player two knows that this is the type of strategy player one uses but does not know the value of $\lambda$. She assumes that player one chose $\lambda$ according to a uniform distribution on the interval [0,1]. Now, player two's beliefs do not satisfy the absolute continuity assumption, which can be seen by considering the event that the long run average of $a$'s is $\lambda$ (it has probability one but is assigned probability zero by the diffused beliefs of player two). However, after
long enough play player two will be able to approximate the true $\lambda$ and have a fairly accurate prediction of player one's near future play. Section 7 contains discussion on weak learning and the possibility of weakening the absolute continuity assumption.

**Example 2.4: Learning and Teaching**

While this paper presents a theory of learning, it does not put the players in a passive state of learning. The following example shows that optimizers, who believe their opponents are open to learning, may find it in their own interest to act as teachers.

We consider a two person infinite symmetric version of a "chicken game" described as follows. Simultaneously, at the beginning and with perfect monitoring after every history, each player chooses to "yield" (Y) or "insist" (I). However, once a player yields (chooses Y) he has to continue yielding forever. The stage game payoffs are the following

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.0</td>
<td>1.2</td>
</tr>
<tr>
<td>I</td>
<td>2.1</td>
<td>-1,-1</td>
</tr>
</tbody>
</table>

and infinite payoff streams are evaluated with discounting. We denote the individual pure strategies of this game by $s_0, s_1, \ldots, s_\infty$ with $s_t$ indicating the one that prescribes the initial yielding at time $t$. 
Notice that this game differs from the prisoners' dilemma example in some important ways. First, the stage game has no dominant strategies and it contains two symmetric pure strategy equilibria. Also, the repeated game contains exactly two pure strategy Nash equilibria, the one where he yields immediately and she insists forever ($s_0, s_\infty$) and the symmetrically reversed one ($s_\infty, s_0$). (In addition, subgames following mutual simultaneous yield actions contain the mutual yield forever equilibrium of these subgames.) While technically this game is not an infinitely repeated one, due to the absorbing nature of the action $V$, the results of this paper still hold and offer interesting insights.

We assume as in the prisoner's dilemma example that PI's beliefs, about PII's strategy, are given by a vector $\beta = (\beta_0, \beta_1, \ldots, \beta_\infty)$ and, also, that PII's beliefs about PI are given by a similar vector, $\alpha$. Putting himself partially in her shoes, PI may think that she is equally likely to wait any number of the first $n$ periods before yielding, to see if he would yield first; or that she may insist forever with probability $\epsilon$ because, unlike him, she assigns a very large loss to ever yielding. Such thinking will lead him to a prior beliefs vector of the type $\beta = ((1 - \epsilon)/n, \ldots, (1 - \epsilon)/n, 0, 0, \ldots, \epsilon)$. If the future is important enough to PI, his best response to $\beta$ would be to wait $n$ periods in case she yields first, but if she does not, then yield himself at time $n + 1$. Interpreted according to the thought process that led him to the choice of vector $\beta$, he reasons that as long as she is willing to find out about him, he will try to convince her by his actions that he is indeed tough.

If both players adopt such reasoning, a pair of strategies $(s_{T_1}, s_{T_2})$ will be chosen. In cases of the type $T_1 = 0$ and $T_2 > 0$, there was no
attempt to teach on the part of PI and the resulting play is as in some Nash
equilibrium of the infinite game. But in cases of the form $0 < T_1 < T_2$, PI
failed in his attempt to teach her. The resulting play paths, with real
initial fighting segments and continuing with his yielding, could not be
justified by any Nash equilibrium or $\varepsilon$-Nash equilibrium with small $\varepsilon$.
Similarly, when $T_1 > T_2 > 0$, we obtain a non-Nash equilibrium path with him
winning. In any of the cases, however, as the main results of this paper
state, both players will learn to predict the future play and end up playing
a Nash equilibrium in sufficiently late subgames.

3. The Model and Assumptions

3.1 The Repeated Game

A group of $n$ players are about to play an infinitely repeated game.
The stage game is described by the following components.

1. $n$ finite sets $\Sigma_1, \Sigma_2, \ldots, \Sigma_n$ of actions with $\Sigma = \times_{i=1}^n \Sigma_i$ denoting
   the set of action combinations.

2. $n$ payoff functions $u_i : \Sigma \to \mathbb{R}$.

We let $H_t$ denote the set of histories of length $t$, $t = 0, 1, 2, \ldots$ (i.e.,
$H_t = \Sigma^t$), with $\Sigma^0$ being a singleton consisting of the null history. Denote
$H = \bigcup_t H_t$ the set of all (finite) histories. A (behavior) strategy of player
$i$ is a function $f_i : H \to \Delta(\Sigma_i)$ with $\Delta(\Sigma_i)$ denoting the set of probability
distributions on $\Sigma_i$. Thus, a strategy specifies how a player randomizes
over his choices of actions after every history.

We assume that each player knows his own payoff function and that the
game is played with perfect monitoring, i.e., the players are fully informed
about all realized past action combinations at each stage.
3.2 Infinite Play Paths

Let \( f = (f_1, \ldots, f_n) \) be a vector of behavior strategies. At the first stage player \( i \) plays \( f_i(h^0) \), where \( h^0 \) stands for the null history. Notice that \( f_i(h^0) \) is a probability distribution over his set of actions. Denote by \( z_i^1 (=z_i^1(f_i)) \), the realization of \( f_i(h^0) \) and by \( z^1 \) the realized action combination, i.e., \( z^1 = (z_1^1, \ldots, z_n^1) \). Player \( i \) is paid \( x_i^1 = u_i(z^1) \) and receives the datum \( z^1 \) (he is informed of the realized action combination).

At the second stage player \( i \) randomizes over his actions according to \( f_i(z^1) \). Denote by \( z_i^2 \) and by \( z^2 \) the realized action of player \( i \) at the second stage and the realized action combination, respectively. The payoff of player \( i \) is \( x_i^2 = u_i(z^2) \) and he is informed of \( z^2 \). The game proceeds in this fashion infinitely many times. The infinite vector \( (z^1, z^2, \ldots) \) of action combinations is the realized play path.

The procedure described above defines a probability distribution, \( \mu_f \), induced by the strategy vector \( f \), on the set of infinite play paths. First (with some abuse of notation), \( \mu_f \) is defined inductively for finite histories \( h \in \bar{H} \). \( \mu_f \) of the empty history is 1 and \( \mu_f(ha) = \mu_f(h) \times f_i(h)(a_i) \). In other words, the probability of the history \( h \) followed by an action vector \( a \) being played is the probability of \( h \) times the product of the \( a_i \)'s being selected by the individual \( f_i(h) \)'s.

In the set of infinite play paths, \( \mathcal{L}^\infty \), the event history \( h \) being played is described by the cylinder set \( C(h) \), consisting of all paths with initial segment \( h \). Thus \( f \) induces a probability \( \mu_f(C(h)) \) (the probability of the history \( h \)) to all such cylinder sets. Following the standard construction of probability theory, we let \( \mathcal{F}_T \) denote the \( \sigma \)-algebra generated by the
cylinder sets of the histories of length \( t \), and \( \mathcal{F} \), the \( \sigma \)-algebra used for \( \Sigma^\infty \), is the smallest one containing all \( \mathcal{F}_t \)'s. The probability distribution \( \mu_f \), defined on \( (\Sigma^\infty, \mathcal{F}) \), is the unique extension of \( \mu_f \) from the \( \mathcal{F}_t \)'s to \( \mathcal{F} \).

3.3 The Payoffs

Let \( \lambda_i, 0 < \lambda_i < 1 \), be the discount factor of player \( i \). Recall that \( x_i^t \) denotes player \( i \)'s payoff at stage \( t \). If the strategy vector \( f \) is played, then the payoff of player \( i \) in the repeated game is defined by

\[
U_i(f) = (1 - \lambda_i) \sum_{t=0}^{\infty} E_f(x_i^{t+1}) \lambda_i^t,
\]

where \( E_f \) denotes the expected value calculated with respect to the probability measure, \( \mu_f \), induced by \( f = (f_1, \ldots, f_n) \).

Notice that \( U_i(f) \) can be written also as \( (1 - \lambda_i) \int \sum x_i^{t+1} \lambda_i^t \, d\mu_f \).

3.4 Behavior and Beliefs

In order to play the infinite game, each player \( i \) chooses a behavior strategy \( f_i \). In addition, player \( i \) has a joint strategy \( f^i = (f^i_1, f^i_2, \ldots, f^i_n) \) describing his beliefs about the strategies adopted by his opponents. Thus, \( f^i_j \) denotes the behavior strategy that player \( i \) thinks player \( j \) will follow. We will assume throughout this paper that players know their own choice of strategies, i.e., \( f^i_i = f_i \).

As usual, we say that a strategy of player \( i \), \( f^i \), is a best response to \( f_{-i} = (f^i_1, \ldots, f^i_{i-1}, f^i_{i+1}, \ldots, f^i_n) \) if \( U_i(f^i_{-i}, \bar{f}_i) - U_i(f^i_{-i}, f_i) \leq 0 \) for all strategies \( \bar{f}_i \) of player \( i \). We say that \( f^i \) is an \( \varepsilon \)-best response (\( \varepsilon > 0 \)) if the same inequality holds with \( \varepsilon \) replacing 0.
Suppose that $f$ and $g$ are two vectors of individual behavior strategies in the repeated game, with $\mu_f$ and $\mu_g$ denoting the distributions over infinite play paths induced by $f$ and $g$, respectively. The measure $\mu_f$ is said to be absolutely continuous with respect to (w.r.t.) $\mu_g$ (denoted by $\mu_f \ll \mu_g$) if every event having a positive measure according to $\mu_f$ also has a positive measure according to $\mu_g$. Formally, $\mu_f(A) > 0$ implies $\mu_g(A) > 0$ for every measurable set $A \subseteq \Sigma^\infty$. If $\mu_f \ll \mu_g$ we also say that $f$ is absolutely continuous w.r.t. $g$.

It is important to expand on the assumption that the beliefs player $i$ holds regarding player $j$'s strategy are described by a single behavior strategy $f_{j,r}^i$. This represents no serious restriction because the well-known Kuhn's (1953) theorem (see also Selten, 1975) assures us that if player $i$'s beliefs were given by a probability distribution over behavior strategies of player $j$, then these beliefs could be replaced by an equivalent single behavior strategy. Since beliefs are a central topic of this paper, and since Kuhn's equivalent behavior strategies use in their construction Bayesian updating, another central topic to this paper, we briefly review this construction.

Suppose player $i$ believes that player $j$ will play the behavior strategy $f_{j,r}$ with probability $\lambda_r$, $r = 1, \ldots, L$. A Kuhn's equivalent behavior strategy $f_{j}^i$ will choose the action $a$ after the history $h$ with probability

$$f_{j}^i(h)(a) = \sum \lambda_r(h)f_{j,r}(h)(a)$$

with $\lambda_r|h$ being the posterior probability of $f_{j,r}$ having been chosen given the observed history $h$, i.e.,
\[ \lambda_r | h = \lambda_r \eta_{j,r} (h) / \sum_w \lambda_w \eta_{j,w} (h), \]

where \( \eta_{j,r} (h) \) denotes the probability of \( h \) being reached when all players other than \( j \) take the actions leading to \( h \) with probability one, and player \( j \) mixes according to \( f_{j,r} \). (In the case that \( \eta_{j,w} (h) = 0 \) for \( w = 1, \ldots, k \), \( f_{j,r}^i(h) \) can be chosen arbitrarily.)

Kuhn's equivalence is strong. Playing against the strategies \( (f_{j,r}^i) \) with the probabilities \( (\lambda_r) \) and playing against an equivalently constructed behavior strategy \( f_{j,r}^i \) generate identical probability distribution on the future play paths of the game (and hence also all positive probability subgames). We refer the reader to Aumann (1964) for the infinite version of Kuhn's theorem.

It is important to emphasize here an assumed restriction on the nature of the beliefs. Player \( i \) believes that different opponents, say, \( j \) and \( k \), are described by individual strategies \( f_{j,r}^i \) and \( f_{k,r}^i \). In evaluating the probabilities of potential histories, he uses the product of the probabilities induced by such strategies. In other words, he believes that player \( j \) and \( k \) choose their actions independently. This rules out important cases where player \( i \) believes that \( j \) and \( k \)'s strategy choices are correlated--for example, they both depend on the same random event whose outcome he himself does not know (e.g., they both went to school A or both went to school B and their strategies depend on the school they went to). Our results regarding future play prediction can be extended to these cases, but convergence to \( \varepsilon \)-Nash equilibrium may fail.
4. **Statement of the Main Results**

Recall that by a **path** we mean an infinite sequence of action combinations, i.e., an element of $\Sigma^\infty$. For any path $z$ and time $t \in \mathbb{N}$ we denote by $z(t)$ the $t$-prefix of $z$ (the element in $H_t$ consisting of the first $t$ action combinations of $z$).

**Definition 1:** Let $\varepsilon > 0$ and let $\mu$ and $\tilde{\mu}$ be two probability measures defined on the same space. We say that $\mu$ is $\varepsilon$-close to $\tilde{\mu}$ if there is a measurable set $Q$ satisfying:

1. $\mu(Q)$ and $\tilde{\mu}(Q)$ are greater than $1 - \varepsilon$; and
2. for every measurable set $A \subseteq Q$

$$(1 - \varepsilon)\tilde{\mu}(A) \leq \mu(A) \leq (1 + \varepsilon)\tilde{\mu}(A)$$

Notice that this notion of $\varepsilon$-closeness is strong. Unlike closeness measures that depend on differences (e.g., $|\mu(A) - \tilde{\mu}(A)| \leq \varepsilon$, where $\tilde{\mu}(A)$ can equal $2\mu(A)$ without violating the closeness requirement for small probability $A$), our definition requires that any two events in $Q$ can only differ by a small percentage. It also implies closeness of conditional probabilities. If $A, B \subseteq Q$ then $\mu$ being $\varepsilon$-close to $\tilde{\mu}$ in the above sense implies that

$$\tilde{\mu}(A|B)(1 - \varepsilon)/(1 + \varepsilon) \leq \mu(A|B) \leq \tilde{\mu}(A|B)(1 + \varepsilon)/(1 - \varepsilon).$$

Thus, in the sequel where $\mu$ represents true probabilities of events in the game and $\tilde{\mu}$ represents beliefs of a player, being $\varepsilon$-close would mean that not only does the player assess the future correctly, he even assesses
developments following small probability histories correctly, provided that he considers paths in the large set \( Q \). This is important since it implies no cumulative buildup of an error in his assessment of the future no matter how far.

Being close in our sense, on a large set \( Q \), and being close in the sense of differences, as mentioned above but without a restriction to a large set \( Q \), turn out, however, to be asymptotically equivalent notions, as we discuss in Section 5.

Let \( f \) and \( g \) be two joint strategies.

**Definition 2:** Let \( \varepsilon \geq 0 \). We say that \( f \) plays \( \varepsilon \)-like \( g \) if \( \mu_f \) is \( \varepsilon \)-close to \( \mu_g \).

**Definition 3:** Let \( f \) be a strategy, \( t \in \mathbb{N} \) and \( h \in H_t \). The induced strategy \( f_h \) is defined by

\[
f_h(h') = f(hh') \text{ for any } h' \in H_r.
\]

where \( hh' \) is the concatenation of \( h \) with \( h' \), i.e., the history of length \( t + r \) whose first \( t \) elements coincide with \( h \) followed by the \( r \) elements of \( h' \). If \( f = (f_1, \ldots, f_n) \) is a joint strategy, \( f_h \) denotes the joint strategy consisting of all the induced individual strategies.

The following theorem states that if the vector of strategies actually chosen is absolutely continuous w.r.t. the beliefs of a player, then the player will learn to accurately predict the future play of the game.
Theorem 1: Let $f$ and $f^i$ be two n-vectors of strategies, representing the ones actually chosen and the beliefs of player $i$, respectively. Assume that $f$ is absolutely continuous w.r.t. $f^i$. Then for every $\varepsilon > 0$ and for almost every play path $z$ (according to the measure induced by $f$) there is a time $T (=T(z, \varepsilon))$ such that for all $t \geq T$, $f^z_{\varepsilon}(t)$ plays $\varepsilon$-like $f^i_{\varepsilon}(t)$.

In other words, after the history $z(t)$, the real probability distribution over the future play of the game is $\varepsilon$-close to what player $i$ believes the distribution is. It implies that the real probability of any future history cannot differ from the beliefs of player $i$ by more than $\varepsilon$. But, as discussed earlier, it is substantially stronger. It implies closeness of probabilities for small events and for conditional probabilities.

Notice that, in Theorem 1, other than absolute continuity, no assumptions were made on $f$ and $f^i$. Thus, it is applicable to any strategies of interest and not just to those maximizing expected utility. For instance, if a player were following a minmax strategy and still conducting a Bayesian update, he would also learn to predict the future play. The theorem essentially states that Bayesian updating by itself will lead to a correct prediction of the important parts (those that determine the actual play) of other players' strategies. It does not state that a player would learn to predict other players' future randomization in response to actions that will not be taken.

Theorem 1, by itself, has immediate implications for theories dealing with payoff maximizing players. Suppose, as Theorem 1 implies, that $f^z_{\varepsilon}(t)$
plays $e$-like $f^i_{z(t)}$ for all $i = 1, \ldots, n$. Furthermore, assume that $f^i_1$ is a best response to $f^i$. Then, after sufficiently long time: (i) each player maximizes his payoff against his subjective beliefs and, moreover, (ii) these beliefs are almost (up to $e$) realization equivalent to the real strategies played. Thus, each player is maximizing relative to (possibly false) subjective beliefs which will never be contradicted by the play of the game (even statistically). The following solution concept captures these two elements.

**Definition 4:** An $n$-vector of strategies, $g$, is a subjective $e$-equilibrium if there is a matrix of strategies $(g^i_j)_{1 \leq i, j \leq n}$ with $g^i_1 = g^i_1$ s.t.

(i) $g^i_1$ is a best response to $g^i_{-1}$, $i = 1, \ldots, n$; and

(ii) $g$ plays $e$-like $g^i$, $i = 1, \ldots, n$.

**Corollary 1:** Let $f$ and $f^1, f^2, \ldots, f^n$ be vectors of strategies representing the actual choice and the beliefs of the players. Suppose that for every player $i$:

(i) $f^i_1$ is a best response to $f^i_{-1}$; and

(ii) $f$ is absolutely continuous w.r.t. $f^i$.

Then for every $e > 0$ and for almost every (w.r.t. $\mu_t$) path $z$ there is a time $T = T(z, e)$ such that for all $t \geq T$ $f^i_{z(t)}$ with $f^1_{z(t)}, \ldots, f^n_{z(t)}$ is a subjective $e$-equilibrium.

**Proof:** The corollary follows immediately from Theorem 1 when we recognize that maximizing expected discounted utility implies maximizing expected utility after every positive probability history relative to the posterior
distribution induced by the history.

Notice that if $g$ is a subjective 0-equilibrium (or just subjective equilibrium), then $\mu_g$, the distribution induced by $g$, is identical to $\mu_{g^1}$. Thus, $g$ and $g^1$ are realization equivalent. Despite the equivalence, a subjective 0-equilibrium does not necessarily induce the same behavior as a Nash equilibrium (the one person multi-arm bandit game is a well-known example).

However, under the assumptions of knowing own payoff matrices and perfect monitoring, or "observed-deviators" in the language of Fudenberg-Levine (1993a), it is easy to see that identical behavior is induced (see also Battigalli et al., 1992, for earlier versions of this observation). Clearly, every Nash equilibrium, being a subjective equilibrium, induces a subjective equilibrium behavior. Conversely, starting with a subjective equilibrium, one can modify the strategies used as follows. After histories that are in the support of all players' strategies leave the actions of all players' unchanged. In subgames that follow a one player deviation from his support, have all the players switch their actions to the ones attributed to them by the beliefs of the deviator. In subgames that follow a mutiperson deviations assign the players any actions. It is easy to observe that this modification yields a Nash equilibrium which is realization equivalent to the original subjective equilibrium.

When perturbations are introduced, to the accuracy of the beliefs in a subjective $\varepsilon$-equilibrium and to the accuracy of optimization in an $\varepsilon$-Nash equilibrium, the discrepancy between the two concepts is greater and the equivalence of behavior sketched above fails for obvious reasons (see Kalai
and Lehrer, 1993, for discussion and elaborations). Yet, for the family of games studied here, the two notions induce asymptotically identical behavior.

**Proposition 1:** For every \( \varepsilon > 0 \) there is \( \eta > 0 \) s.t. if \( g \) is a subjective \( \eta \)-equilibrium then there exists \( \tilde{f} \) s.t.

1. \( g \) plays \( \varepsilon \)-like \( \tilde{f} \), and
2. \( \tilde{f} \) is an \( \varepsilon \)-Nash equilibrium.

Theorem 1 will be proven in the next section. We refer the reader to Kalai and Lehrer (1993) for the proof of Proposition 1 and for a general discussion on subjective equilibrium. Together, however, Corollary 1 and Proposition 1 imply our main result.

**Theorem 2:** Let \( f \) and \( f^1, f^2, \ldots, f^n \) be strategy vectors representing respectively the one actually played and the beliefs of the players. Suppose that for every player \( i \):

1. \( f_i \) is a best response to \( f_{-i}^i \); and
2. \( f \) is absolutely continuous w.r.t. \( f^i \).

Then for every \( \varepsilon > 0 \) and for almost all (with respect to \( \mu_x \)) play paths \( z \) there is a time \( T = T(z, \varepsilon) \) such that for every \( t \geq T \) there exists an \( \varepsilon \)-equilibrium \( \tilde{f} \) of the repeated game satisfying \( f_{z(t)} \) plays \( \varepsilon \)-like \( \tilde{f} \).

In other words, given any \( \varepsilon > 0 \), with probability one there will be some time \( T \) after which the players will play \( \varepsilon \)-like an \( \varepsilon \)-Nash equilibrium. This means that if utility maximizing players start with individual
subjective beliefs, with respect to which the true strategies are absolutely continuous, then in the long run, their behavior must be essentially the same as a behavior described by an \( \epsilon \)-Nash equilibrium. In the last section of the paper, we show that by using a weaker version of closeness of behavior one can replace the \( \epsilon \)-Nash equilibrium in Theorem 2 by the usual Nash equilibrium.

5. **Bayesian Learning**

Our main result, Theorem 2, combines two issues: (i) Bayesian updating and (ii) payoff maximization. In this section, we concentrate on the first one and prove Theorem 1. In fact, the treatment of Bayesian updating, given here, is applicable to issues that lie beyond the scope of this paper. The reader is referred to Kalai and Lehrer (1990a,b and 1993a) and Monderer and Samet (1990).

Suppose that \( (\Omega, \mathcal{F}) \) is a measure space interpreted as the set of states of the world. Let \( \{\mathcal{P}_t\}_t \) be an increasing sequence of finite or countable partitions of \( \Omega \) (i.e., \( \mathcal{P}_{t+1} \) refines \( \mathcal{P}_t \)). \( \mathcal{P}_t \) is interpreted as the information available at time \( t \). In other words, at time \( t \) the agent is informed of the part \( P_t(\omega) \in \mathcal{P}_t \) that contains the prevailing state \( \omega \in \Omega \).

We assume that the \( \sigma \)-field \( \mathcal{F} \) is the smallest one that contains all the elements of all the \( \mathcal{P}_t \)'s.

The agent's initial belief about the distribution of states of nature is denoted by \( \tilde{\mu} \) (a \( \sigma \)-additive measure defined on \( (\Omega, \mathcal{F}) \)). However, the real distribution is given by a measure \( \mu \). Our task in this section is to show that the subjective probability (the belief) converges to the real one as information increases.
Denote the field generated by $\mathcal{F}_n$ by $\mathcal{F}_n$. The next theorem is a restatement of Theorem 1 but in the language of partitions. It is essentially equivalent to the Blackwell and Dubins (1962) theorem discussed later.

**Theorem 3:** Let $\mu \ll \tilde{\mu}$. With $\mu$-probability 1, for every $\varepsilon > 0$ there is a random variable time $r(\varepsilon)$ s.t. for all $r \geq r(\varepsilon)$, $\mu(\cdot|P_r(\omega))$ is $\varepsilon$-close to $\tilde{\mu}(\cdot|P_r(\omega))$.

**Proof:** Theorem 3 is a consequence of Proposition 2 and Lemma 1 (see also Monderer and Samet (1990)), which follow.

**Proposition 2:** Suppose that $\mu \ll \tilde{\mu}$ (i.e., $\mu(A) > 0$ implies $\tilde{\mu}(A) > 0$ for every $A \in \mathcal{F}$). With $\mu$-probability 1 for every $\varepsilon > 0$ there is a random variable $t(\varepsilon)$ s.t. for every $s \geq t \geq t(\varepsilon)$.

$$1 - \varepsilon \leq \frac{\mu(P_s(\omega)|P_t(\omega))}{\tilde{\mu}(P_s(\omega)|P_t(\omega))} \leq 1 + \varepsilon.$$  

**Proof:** Since $\mu \ll \tilde{\mu}$, by the Radon-Nikodym theorem, there is an $\mathcal{F}$-measurable function $\phi$ satisfying

$$\int_A \phi \, d\tilde{\mu} = \mu(A)$$  for every $A \in \mathcal{F}$.

By Levy's theorem (see Shiryayev, 1984), $E(\phi|\mathcal{F}_t) \rightarrow E(\phi|\mathcal{F}) = \phi$ $\mu$-almost surely (and therefore, $\mu$-a.s.). However, for $\tilde{\mu}$ almost all $\omega$
Moreover, by (2), $\phi > 0$ $\mu$-a.s. Thus, the right side of (3) tends $\mu$-a.s. to a positive number. In other words, there is a $t(\epsilon)$ such that for $\mu$-a.e. $\omega$ the following holds:

\[
1 - \epsilon \leq \frac{\mu(P_t(\omega))}{\tilde{\mu}(P_t(\omega))} \leq 1 + \epsilon \text{ for all } s \geq t \geq t(\epsilon).
\]

The middle term of (4) is equal to the middle one in (1). Since (1) holds for every $\epsilon > 0$ with $\mu$-probability 1, the proposition follows. //

**Lemma 1:** Let $\{W_t\}$ be an increasing sequence of events satisfying $\mu(W_t) \uparrow 1$. For every $\epsilon > 0$ there is a random time $t(\epsilon)$ such that any random $t \geq t(\epsilon)$ satisfies

\[
\mu(\omega; \mu(W_t \mid P_t(\omega)) \geq 1 - \epsilon) = 1.
\]

**Proof:** $\mu(W_t) \rightarrow_{t \rightarrow \infty} 1$. Thus, $\mu(C_t) \rightarrow_{t \rightarrow \infty} 0$, where $C_t = \Omega \setminus W_t$.

Suppose, to the contrary, that the lemma does not hold. Then there is a $\mu$-positive set $A$ and $\epsilon > 0$ such that for all $\omega \in A$, $\mu(W_t \mid P_t(\omega)) < 1 - \epsilon$ for infinitely many $t$'s.

Fix $s \in \mathbb{N}$ and define

\[
B_r = \{\omega \in A; r = \min(t \mid t \geq s \text{ and } \mu(W_t \mid P_t(\omega)) < 1 - \epsilon)\}.
\]
Observe that \( (B_r) \) are pairwise disjoint and, moreover, \( \{ \bigcup_{\omega \in B_r} P_r(\omega) \} \) are also pairwise disjoint. By the definition, \( A = \bigcup_{r \geq s} B_r. \)

Since \( C_s \supseteq C_t \) when \( t \geq s \), for all \( \omega \in A \), \( \mu(C_s | P_t(\omega)) > \epsilon \) for infinitely many \( t \). Thus, \( \mu(C_s | \bigcup_{\omega \in B_t} P_t(\omega)) > \epsilon \). Therefore, \( \mu(C_s) > \epsilon \mu(\bigcup_{t \geq s} B_t) = \epsilon \mu(A). \)

Hence, the sequence \( \{ \mu(C_s) \} \) is bounded away from zero, which is a contradiction. This concludes the proof of the lemma.

In order to apply the lemma set

\[
W_t = \{ \omega; |E(\phi | F_s)(\omega) / E(\phi | F_t)(\omega) - 1| < \epsilon \text{ for } \forall s \geq t \}.
\]

An immediate corollary is a version of the main result of Blackwell and Dubins (1962).

Corollary 2 (see Blackwell and Dubins, 1962): For \( \mu \)-a.e. \( \omega \) there is time \( t = t(\epsilon) \) such that for \( A \in F \) and \( s \geq t \)

\[
| \mu(A | P_s(\omega)) - \tilde{\mu}(A | P_s(\omega)) | < \epsilon.
\]

The converse statement, that the Blackwell-Dubins' result implies Theorem 3, is also true but not obvious. One can actually show that the topology generated by our notion of closeness is equivalent to the one generated by Blackwell and Dubins. That is a sequence of measures \( \mu_s \rightarrow \mu \) in one sense if and only if it does so in the other. That our topology is stronger is immediate. However, since the notion used by Blackwell and Dubins applies to all events, not just in large set \( Q \), it turns out to be as strong. See Kalai-Lehrer (1993a) for details.

Example: One biased coin with parameter \( p_i \) is selected with probability...
\( \alpha_i > 0 \) from a countable set of such coins. The chosen coin is tossed infinitely many times. An agent believes that the coin \( p_i \) is drawn with probability \( \beta_i \). Define \( \Omega \) to be the set of infinite sequences of 0's and 1's generated by the tosses of the coin. The probability measure on \( \Omega \), induced by \( \{\alpha_i\} \), say \( \mu \), is absolutely continuous with respect to the one induced by \( \{\beta_i\} \), say \( \tilde{\mu} \), if \( \beta_i > 0 \) for all \( i \). Theorem 3 states that, after sufficiently long time, the posterior probability of \( \mu \) will be arbitrarily close to the posterior one of \( \tilde{\mu} \).

**Remark:** For general probability measures, we say that \( \tilde{\mu} \) contains a "grain of truth" of \( \mu \), if \( \tilde{\mu} = \lambda \mu + (1 - \lambda)\tilde{\mu} \) for some probability measure \( \tilde{\mu} \) and \( \lambda > 0 \). It is equivalent to requiring that the Radon-Nikodym derivative, \( \phi = d\mu/d\tilde{\mu} \), is bounded.

Notice that in the previous example \( \tilde{\mu} \) contains a grain of truth if and only if \( \alpha_i/\beta_i \) are uniformly bounded.

6. **Repeated Games with Incomplete Information and Jordan's Results**

In this paper uncertainties regarding other players are captured by the individual beliefs a player holds about others' strategies. This is unlike traditional game theory where uncertainties are expressed by a commonly known prior distribution over the unknown parameters of the game (payoffs, discount parameters, feasible actions, etc.) with a commonly known signaling mechanism that gives different players different additional information. We proceed to show by example how the traditional approach can be viewed as a special case of the current paper. In particular, the equilibria of a large class of repeated games with incomplete information will satisfy the
assumptions of our main theorems, and the conclusions will yield interesting new insight.

Consider two players about to play an infinitely repeated game of the type described earlier, but with a randomly generated fixed size pair of payoff matrices \((A_i, B_j^{(i,j)})\) \((i,j) \in I \times J\). We assume that both \(I\) and \(J\) are finite or countable, and that the selection of the pair \((i,j)\) will be done according to a commonly known prior probability distribution \(\pi\) on \(I \times J\). After the selection, \(\Pi_1\) will be told the realized value \(i\) and \(\Pi_2\) will be told the realized value \(j\).

In order to play the game, \(\Pi_1\) chooses a vector \((f_i)_{i \in I}\) with each \(f_i\) being the infinite game strategy that he would follow if he is told that his realized payoff matrix is \(A_i\). \(\Pi_2\) chooses a similar vector of strategies \((g_j)_{j \in J}\). A pair of such vectors is a Harsanyi-Nash equilibrium if each \(f_i\) is a best response (in long term discounted utility) against the strategies \((g_j)_{j \in J}\) when mixed according to the conditional distribution on \(J\) given the realized value \(i\), \(\pi(j|i)\), and with the symmetric property holding for each \(g_j\).

To relate such an equilibrium to the current paper, assume that the random drawing of the payoff matrices has been performed and that positive probability \(\bar{i}\) and \(\bar{j}\) were selected. Thus, the real play of the game will follow the pair of strategies \((f_{\bar{i}}, g_{\bar{j}})\). Given his information, \(\Pi_1\) believes \(\hat{i|\bar{j}}\) \(\Pi_2\) will play \((g_j)_{j \in J}\) with probabilities \(\pi(j|\bar{i})\) and being at a Nash equilibrium his \(f_{\bar{i}}\) is actually a best response to this belief. Moreover, given the finiteness of \(J\), \(\Pi_1\)'s belief contains a grain of truth (i.e., assigns positive probability to \(g_{\bar{j}}\)). Similarly, \(\Pi_2\)'s choice of \(g_{\bar{j}}\) is a best response to the distribution \(\pi(i|\bar{j})\) on \((f_i)_{i \in I}\) and it also contains a
grain of truth.

In the set-up above, let \( f_i \) and \( g_j \) be the strategies realized and let \( \tilde{f} \)
and \( \tilde{g} \) be the induced beliefs over opponent's strategies, e.g., \( \tilde{g} \) is the
behavior strategy obtained by mixing the vector \( (g_j)_{j \in J} \) with probabilities \( \pi(j|i) \).

The analogues of Theorem 1, Proposition 1 and Theorem 2, when applied
to the Harsanyi-Nash equilibrium, follow as immediate corollaries.

**Theorem 1.1:** For every \( \varepsilon > 0 \) and almost every play path \( z \) (relative to the
distribution induced by \( f_i, g_j \)) there is a time \( T \) such that for all \( t \geq T \)
\( (f_i, g_j) z(t) \) plays \( \varepsilon \)-like \( (\hat{f}_i, \hat{g}_j) z(t) \).

In other words, at such a Harsanyi-Nash equilibrium the players
eventually predict the future play of the game accurately even if they do
not necessarily learn the payoff matrices of their opponents.

**Theorem 2.1:** For every \( \varepsilon > 0 \) and almost every play path we can find a time
\( T \) such that for all \( t \geq T \) there is an \( \varepsilon \)-Nash equilibrium of the realized
repeated game \( (A_i, B_j) \), \( (\hat{f}_i, \hat{g}_j) \), with \( (f_i, g_j) z(t) \) plays \( \varepsilon \)-like \( (\hat{f}_i, \hat{g}_j) \).

So even if the players do not learn the identity of the payoff matrices
actually played, they eventually play almost as \( \varepsilon \)-Nash players who do know
the identity of the payoff matrices.

Theorem 2.1 is related to an earlier result of Jordan (1991). His
players faced the same uncertainty about opponents' payoff matrices. His
prior distribution about such matrices, however, was more general since he did not restrict himself to a discrete set of possible matrices. On the other hand, his players played myopically. In each period they played a Harsanyi-Nash equilibrium, updated on all the information obtained earlier, as if each current period were the last one. He then studied the limit beliefs about opponents' next period actions as the number of periods became large. His main result was that all cluster points of the expectation sequence are Nash equilibria of the underlying realized stage game.

Our model can be made nearly myopic by letting the discount parameter approach zero. In general, when one totally discounts the future, Nash equilibria of the repeated game consist of repeated plays of Nash equilibria of the stage game. Thus, as a limit case when we let the discounted parameters approach zero, our result regarding convergence to Nash equilibria of the repeated game confirms Jordan's result of convergence to Nash equilibria of the stage game. (Of course, if the stage game had a multiplicity of equilibria, then one could see oscillation among them.)

Notice, though, that Jordan obtains convergence of the beliefs to Nash equilibrium, while we obtain convergence of the beliefs and the actual play to ε-Nash equilibrium.

When considering generalizations of Theorems 1.1 and 2.1 above to the n-player case with n > 2 we observe the following. Theorem 1.1 generalizes. Theorem 2.1 does not unless we impose an additional independence assumption on the prior distribution over payoff matrices. The condition needed is that the prior distribution, π(ι,ι',...,ι_n|ι_1), over payoff matrices should be independent over opponents for every realization of every player ι_j. For example, for player 1, π(ι_2,...,ι_n|ι_1) should be independent over ι_2.
through $i_n$.

The need for the above independence condition arises in the application of Proposition 1. As assumed in the definition of a subjective $\epsilon$-equilibrium, each player assigns independent beliefs to the strategies of his opponents. However, if the prior distribution $\pi$ did not satisfy the independence condition, one would not be able to replace the mixed combination of opponent's strategies by an equivalent product of behavior strategies, so the proposition would not hold. Indeed, convergence to $\epsilon$-Nash equilibrium will fail.

In order to correct for such dependencies, we would have to generalize the concept of subjective $\epsilon$-equilibrium to allow for correlated beliefs. This will yield a concept closer to the notion of self-confirming equilibrium developed by Fudenberg-Kreps (1988) and Fudenberg-Levine (1993a). The new concept will have to be defined for infinite games and will have to include a suitable notion of perturbation. The convergence of Theorem 2.1 to $\epsilon$-Nash equilibrium is then likely to be replaced by convergence to a correlated $\epsilon$-equilibrium.

7. Remarks

This section includes some additional remarks about assumptions of the model, the scope of the results, and possible extensions.

7.1 An Alternative Notion of Closeness

As discussed in this paper, the notions of one measure being close to another and of one strategy vector playing like another are strong. They guarantee that if $f$ plays $\epsilon$-like $g$, then with probability $1 - \epsilon$, $f$ and $g$
will assign close probabilities to future events and will continue to do so regardless of how long the game has been played. The result about learning to play \( \varepsilon \)-like \( \varepsilon \)-Nash gives a strong notion of proximity to the \( \varepsilon \)-Nash for the rest of the infinite game. We do not know, at this time, if the same theorem can be proven with only one \( \varepsilon \), i.e., learning to play \( \varepsilon \)-like a full Nash equilibrium. However, with a less demanding notion of being close, the players will learn to play a full, rather than \( \varepsilon \), Nash equilibrium.

For an \( \varepsilon > 0 \) and a positive integer \( k \) we say that \( \mu \) is \((\varepsilon, k)\)-close to \( \tilde{\mu} \) if for every history \( h \) of length \( k \) or less \(|\mu(h) - \tilde{\mu}(h)| \leq \varepsilon \). Similarly, \( f \) plays \((\varepsilon, k)\)-like \( g \) if the induced measure \( \mu_f \) is \((\varepsilon, k)\)-close to \( \mu_g \). Thus, playing \((\varepsilon, k)\)-like means playing \( \varepsilon \) the same up to a horizon of \( k \) periods.

It was shown in Kalai-Lehrer (1993) that for a given \( \varepsilon \) and \( k \), if \( g \) is a subjective \( \eta \)-equilibrium, with sufficiently small \( \eta \), then it must play \((\varepsilon, k)\)-like some Nash (rather than \( \varepsilon \)-Nash) equilibrium of the repeated game. So, taking this less ambitious notion of approximating behavior, we can obtain \( \varepsilon \)-closeness in finite time to a full Nash equilibrium. Therefore, Theorem 2 can be restated as:

**Theorem 2**. Let \( f \) and \( f^1, f^2, \ldots, f^n \) be strategy vectors representing the one actually played and the beliefs of the players. Suppose that for every play \( i \):

(i) \( f_i \) is best response to \( f_{-i}^i \); and

(ii) \( f \) is absolutely continuous w.r.t. \( f_i \).

Then for every \( \varepsilon > 0 \) and a positive integer \( k \) there is a time \( T = T(\varepsilon, k) \) such that for every \( t \geq T \) there is a Nash equilibrium \( \bar{f} \) of the repeated game satisfying \( f_z(t) \) plays \((\varepsilon, k)\)-like \( \bar{f} \).
The obvious identical modification can be applied to Theorem 2.1 as well.

7.2 Dispersed Beliefs and Weak Learning

Example 2.3 suggests that if the belief assigns a positive probability to any "neighborhood" of the real strategy, then a "weak" learning may take place. The following example shows that this is not an easy task, and that careful studies of the topologies and notions of learning involved have to be conducted.

The notion of "neighborhood" called for by Example 2.3 is the following. We say that a behavior strategy \( f' \) is in the \( \varepsilon \)-neighborhood of another behavior strategy \( f \) if the probabilities they assign to any action after every history are close to each other up to an \( \varepsilon \). In the following example we show a case where every neighborhood around the strategy denoted below by \( c_\infty \) is given positive probability and nevertheless no learning (even in a weak sense) takes place.

As before, PI has two actions: \( \& \) and \( r \). Let \( c_n \) be the stationary strategy that plays with probability \( 1 - 1/n \) the action \( \& \) and with probability \( 1/n \) the action \( r \), \( n = 1, 2, \ldots, \infty \). Let \( d_n \) be the strategy which plays \( n \) times \( \& \) and \( r \) thereafter. PII believes that \( c_n \) is played with probability \( \alpha_n > 0 \) (\( \sum \alpha_n \leq 1 \)) and \( d_n \) with \( \beta_n \), while in fact PI plays constantly \( \& \), i.e., \( c_\infty \), which is assigned zero probability by PII. It is clear then that PII ascribes positive probability to any neighborhood around \( c_\infty \). One might expect that \( c_n \)'s with large \( n \) will be assigned growing probabilities in the course of the infinite game. But we will show that whether or not this occurs depends on the sequences \( \{\alpha_n\} \) and \( \{\beta_n\} \). If the
sequence $\{\alpha_n\}$ tends to zero much faster than $\{\beta_n\}$ even "weak" learning fails.

After observing $t$ times $\lambda$, the posterior probability of $c_m$ being played is $\alpha_m (1 - 1/m)^t / (A_t + B_t)$ where $A_t = \sum_{m=1}^{\infty} \alpha_m (1 - 1/m)^t$ and $B_t = \sum_{m \geq t} \beta_m$. For sufficiently large $t$, $A_t$ can be bounded from above as follows.

$$A_t = \sum_{m=1}^{\lfloor t^{1/2} \rfloor} \alpha_m (1 - 1/m)^t + \sum_{\lfloor t^{1/2} \rfloor + 1}^{\infty} \alpha_m (1 - 1/m)^t.$$

The first term,

$$\sum_{m=1}^{\lfloor t^{1/2} \rfloor} \alpha_m (1 - 1/m)^t \leq (1 - 1/[t^{1/2}])^t \leq (1 - 1/[t^{1/2}])^{1/2} t^{1/2} \leq e^{-(t^{1/2})/2}.$$

The last inequality is obtained by the fact that $(1 - 1/[t^{1/2}])^{t^{1/2}} \to e^{-1}$ for large $t$'s. The second term,

$$\sum_{\lfloor t^{1/2} \rfloor + 1}^{\infty} \alpha_m (1 - 1/m)^t \leq \sum_{\lfloor t^{1/2} \rfloor + 1}^{\infty} \alpha_m.$$

If $\alpha_m = a2^{-m}$ and $a < 1$, then

$$\sum_{\lfloor t^{1/2} \rfloor + 1}^{\infty} \alpha_m < 2^{-[t^{1/2}]} < 2^{-(t^{1/2})/2}.$$

so that $A_t \leq 2 \cdot 2^{-(t^{1/2})/2}$. Suppose now that $\beta_m = a/m^2$, where the constant $a$ is chosen in such a way that $\sum_m (\alpha_m + \beta_m) = 1$ ($a < 1$). In this case, $B_t$ behaves asymptotically like $a/t$. Thus, $A_t/B_t$ approaches zero as $t$ goes to
infinity. We conclude that the probability assigned to the event that a $c_n$ is played approaches zero as more observations of $x$ arrive. Thus, the future event which consists of the infinite run of $x$'s is given smaller and smaller probability as time goes by. Since the strategy played by PI gives this infinite run of $x$'s probability 1, we do not have here learning in the sense discussed above. PII does not learn the future behavior of PI.

In this example, however, the updated belief assigns probability converging to 1 to the event that the next outcome will be $x$. Thus, PII learns to predict near future behavior of PI. In other words, PII learns in a weak sense the strategy of PI.

By defining the $\beta_n$'s a bit differently, we can construct an example in which every once in a while PII will expect the next outcome to be $r$ with probability close to 1/2, while the outcome will be always $x$. On an infinite set of integers, say, $M$, we set $\hat{\beta}_m = \sum_{k=m}^{\infty} \beta_k$, while if $m \notin M$ we define $\hat{\beta}_m = \beta_m$. Thus, if $M$ is very sparse, the series $\hat{\beta}_m$ converges and, moreover, $\hat{\beta}_m / \sum_{k=m+1}^{\infty} \beta_k$ tends to 1 as $m \to \infty$ if attention is restricted to $m \in M$. Now we define $\alpha_k = a2^{-n}$ and $\beta_m' = a\hat{\beta}_m$. Once again, $a$ is chosen in a way that $\sum (\alpha_m + \beta_m') = 1$. The calculation of $A_t$ in this case is the same as the calculation above. Defining $B_t' = \sum_{m \geq t} \beta_m'$ one gets $A_t / B_t' \leq A_t / B_t$. Therefore, as the game evolves the set of strategies $(c_n)$ is getting diminishing weight. After a long time, when $B_t'$ is very close to one and when $t = m$ for some $m \in M$, out of the probability $B_t'$ (assigned to all $d$'s) $B_t'$ is given to $d_m$, which plays $r$ at the $(m+1)$-th stage. But $\beta_m' / B_t'$ is close to 1/2. We conclude that after $m$ observations of $x$, the prediction of PII regarding the immediate play of PI is approximately 1/2 on $x$ and 1/2 on $r$, while in fact PI plays $x$ with probability 1. It means that PII does not
learn, even in the weak sense, the behavior of PI. In other words, Bayesian updating will not lead to accurate prediction even of events that take place in the near future.

7.3 The Necessity of Knowing Your Own Preferences

The assumption that each player knows his own preferences is crucial. For example, Blume and Easley (1992) show a repeated game with incomplete information where the players never converge to play an equilibrium of the complete information game. Thus, their results contradict Theorem 2.1. The difference lies in the fact that, in their example, players do not know their own payoff matrix.

More familiar, perhaps, are results regarding the multi-arm bandit problem (see, for example, Rothschild, 1974) where an optimal Bayesian learning strategy does not lead to an optimal strategy of the full information case. Since we can view the multi-arm bandit problem as a special case of the one person version of Corollary 1, optimal play corresponds to a Nash equilibrium. The discrepancy between the optimal Bayesian play and the optimal full information play contradicts the conclusion of Theorem 2.1, which states that the player should eventually behave ε-optimally as if the uncertainty were not present. The cause for this discrepancy lies in our assumption that the players know their own payoff matrix. In the multi-arm bandit problem this assumption requires that the player know the payoffs associated with the different actions, which is not true for that model.

The contrast with the multi-arm bandit problem illustrates an important point. The uncertainty in our model is regarding strategies of the
opponent. Unlike nature's uncertainties, opponents' actions will continue to be observed as long as the game lasts, and thus, perfect learning of them will take place.

7.4 The Need for Perfect Monitoring

Consider again the multi-arm bandit problem, but now view it as a two person game with the original player being player I and nature being player II. We let PII have a flat utility function, and his action set consists of choosing one of a few possible payoff distributions for each one of PI's activities. We assume that PII made his choices randomly at the beginning according to some fixed probability distribution \( \pi \), and then kept the same realized choices throughout the infinite game. We also assume that PI knows that PII followed the strategy described above but PI is not told the realized payoff choices.

It is easy to see that now we have modeled the multi-arm bandit problem as a two person infinite game. However, this game has imperfect monitoring. Nash equilibria with PII playing a strategy of the type discussed above exist, and they require that PI plays optimally against the realized choices of PII. Again, examples of optimal long term strategy in the multi-arm bandit problem violate the conclusion of our Theorem 2.1, but under this formulation, the discrepancy is explained by the failure of the perfect monitoring condition.

7.5 Extensions

Stochastic Games

The results of this paper should extend to the more general model of
stochastic games (see Shapley (1953)) under the informational assumptions that each player knows his own payoff matrices as well as the transition probabilities and the state realizations of the stochastic game. To what extent our results generalize when the realized states are not told to the players seem to be an interesting problem.

**General Continuous Payoff Functions**

Since the proof of Proposition 1 (in Kalai and Lehrer, 1993) relies only on the continuity of $U_i$, the proof of Theorem 1 applies also to repeated games with general payoff functions that are continuous with respect to the product topology: for example, when the discount factor changes with time.


Kalai, E. and E. Lehrer (1990a), "Bayesian Learning and Nash Equilibrium," Northwestern University.


Mertens, J. F., S. Sorin and S. Zamir (1990), Repeated Games, to be published.


