Learning by Trial and Error

H. Peyton Young

University of Oxford
The Brookings Institution

July 2, 2007

I am indebted to Thomas Norman, Tim Salmon, Christopher Wallace, and participants in the Game Theory Seminar at the Laboratoire d'Econometrie, Ecole Polytechnique, for helpful comments.

Abstract

A person *learns by trial and error* if he occasionally tries out new strategies, rejecting choices that are *erroneous* in the sense that they do not lead to higher payoffs. In a game, however, strategies can *become erroneous* due to a change of behavior by someone else. Such passive errors may also trigger a search for new and better strategies, but the nature of the search is different than when a player is actively engaged in experimentation. This paper introduces a simple version of this idea, called *interactive trial and error learning*, which has the property that it implements Nash equilibrium behavior in games that have at least one pure Nash equilibrium and no payoff ties. Unlike regret testing (Foster and Young, 2006), it requires no statistical estimation. Unlike a learning procedure proposed by Hart and Mas-Colell (2006), it requires no knowledge of the other players' actions: learning proceeds purely by responding to one's own payoff history.

1. Introduction

Consider a situation in which people interact, but they do not know how their interactions affect their payoffs. In other words, they are engaged in a game, but they do not know what the game is or who the players are. For example, commuters in a city can choose which routes to take to work. Their choices affect congestion on the roads, which determines the payoffs of other commuters. But no single commuter can be expected to know the others' commuting strategies or how their strategies influence his own commuting time. Similarly, in a market with many competing firms, no single firm is likely to know precisely what the other firms' marketing and pricing strategies are, or how these strategies affect its own profits (even though this assumption is routinely invoked in textbook models of competition). Likewise, traders in a financial market are typically unable to observe the strategies of the other traders, and probably do not even know the full set of players participating in the market.

In situations like these, one would like to have a learning procedure that does not depend on any knowledge of the others' actions or on their payoffs. Such a rule is said to be *payoff-based* or *radically uncoupled* (Foster and Young, 2006). Are there simple payoff-based learning rules such that, when used by everyone in a game, period-by-period play comes close to Nash equilibrium play a large proportion of the time? Several recent papers show that the answer is affirmative. Foster and Young (2006) introduced a learning procedure called *regret testing* that has this property for all finite, two-person games. Subsequently, Germano and Lugosi (2007) showed that regret testing leads to Nash equilibrium behavior in generic n-person games on a given finite action space.

More recently, Marden, Young, Arslan, and Shamma (2007), hereafter abbreviated MYAS, show that there are even simpler payoff-based learning rules that come close to pure Nash equilibrium behavior in the class of weakly acyclic games. These games have the property that from every joint action-tuple there exists a sequence of best replies -- one player moving at a time -- that ends at a pure Nash equilibrium. (Potential games and congestion games are special cases.) MYAS propose the following learning process: each player experiments in each period with very small probability, and adopts the experimental action if and only if his payoff increases. They prove that in any weakly acyclic game, this *simple experimentation procedure* implements Nash equilibrium in the sense that equilibrium behavior is observed in a very high proportion of all time periods.

A key feature of regret testing and the MYAS algorithm is that they cause period-by-period behavior to come close to equilibrium in a *probabilistic* sense, but behavior does not necessarily *converge* to equilibrium. Indeed, Hart and Mas-Colell (2003) have shown that there are severe limits to what can be achieved if one insists on convergence and the learning procedure is not, in a certain sense, 'rigged.' One definition of 'not rigged' is that each player's learning rule should be independent of the opponents' payoffs; such a rule is said to be *uncoupled*. Suppose further that each player's learning rule is deterministic and depends solely on the frequency distribution of past play (as in fictitious play). Hart and Mas-Colell (2003) show that there exists a large class of games for which no such rule, when used by all players, causes period-by-period behavior to converge to Nash equilibrium behavior. In a subsequent paper, they examine the situation where the learning procedure is stochastic, and is stationary with respect to histories of bounded length (Hart and Mas-Colell, 2006). In this case one can design simple, uncoupled rules that converge almost surely to Nash equilibrium behavior for games with a *pure* Nash equilibrium, but not for games in general.[1]

The results in the present paper differ from those of Hart and Mas-Colell in two key respects. First, we shall not insist on convergence to Nash equilibrium; it suffices that period-by-period play come close to Nash equilibrium quite often. Second, we shall show to achieve this by a learning process that *does not depend on the opponents' payoffs or their actions*. (The framework in Hart and Mas-Colell (2006) relies on the observability of others' actions; in other words their learning procedure is uncoupled but not radically uncoupled.) Unlike regret testing, the learning rule proposed here does not rely on statistical estimation; it is also intuitively more plausible as a behavioral model. Unlike the simple trial-and-error procedure of MYAS, the rule works for almost all games that possess at least one pure Nash equilibrium.[2]

A novel aspect of the approach is that a player's learning behavior depends on his mood, which can change if his recent payoffs are above or below his current expectations. Mood-driven learning has been suggested as an empirical phenomenon in a number of recent studies (Capra, 2004; Smith and Dickhaut, 2005; Kirchsteiger, Rigotti, and Rustichini, 2006), but to my knowledge the formal properties of such rules have not been previously investigated. In any event the rule proposed here is not intended to be an *empirical model* of mood-driven

[1] The rule operates as follows: if everyone played the same action over the past two periods, and if player *i*'s action is a best response to the others' actions, *i* plays that action again; otherwise *i* chooses an action uniformly at random.

[2] A game G on a finite action space A can be represented as a point in the Euclidean space $R^{n|A|}$. The subset of games with at least one pure Nash equilibrium has positive Lebesgue measure in $R^{n|A|}$, and a property holds for *almost all* such games if it holds except on a subset of Lebesgue measure zero.

learning, though it is composed of intuitively plausible elements that may turn out to have empirical validity. Rather, my intention is to show that rules of this type can be effective methods for learning equilibrium in situations where players have no knowledge of what other players are doing.

2. Interactive trial and error learning

We shall consider a learning rule in which each agent has one of four possible moods: content, discontent, watchful, and hopeful. When an agent is *content*, he occasionally experiments with new strategies, and switches if the new one is better than the old. When *discontent* he tries out new strategies frequently and at random, eventually becoming content with a probability that depends on how well his current strategy is doing. These are the main states, and reflect the idea that search can be of two kinds: careful and directed (when content), or flailing around (when discontent).

The other two states are transitional, and are triggered by changes in the behavior of *other* agents. Specifically, if an agent is currently content and does not experiment in a given period but his payoff changes anyway (because someone else changed strategy), then he becomes *hopeful* if his payoff went up and *watchful* if it went down. If he is hopeful and his payoff stays up for one more period, he becomes content again with a higher expectation about what his payoff should be. If he is watchful and his payoff stays down for one more period, he becomes discontent. [3]

I shall call this process *interactive trial and error learning*. It differs from ordinary trial and error learning, which involves trying new things and accepting them if

[3] The assumption of a one-period waiting time is purely for convenience; it could be any specified number of periods.

and only if they lead to higher payoffs. (This is the MYAS procedure.) In an interactive situation, however, "errors" can arise in two different ways: by trying something that turns out to be no better than what one was doing, or by continuing to do something that turns out to be worse than it used to be. The latter are *passive errors*, whereas the former are *active errors*. A key feature of ITE learning is that these two types of errors trigger different behavioral responses.

Let us now consider the states and transitions of the process in more detail. Let G be an n-person game with players $i = 1, 2, ..., n$, finite joint action space $A = \prod A_i$, and utility functions $u_i : A \to R$. A *state* of player i at a given point in time is a triple $z_i = (m_i, \overline{a}_i, \overline{u}_i)$, where m_i is i's current mood (content (c), discontent (d), hopeful (h), or watchful (w)), $\overline{a}_i$ is i's current benchmark action, and $\overline{u}_i$ is i's current benchmark payoff. A *state* z of the process specifies a state z_i for each player. We shall write this in the form $z = (m, \overline{a}, \overline{u})$, where each of the three components is an n-vector describing the players' moods, action benchmarks, and payoff benchmarks respectively. Let Z be the finite set of states corresponding to a given game G on A.

Given any state $z \in Z$, a joint action-tuple $a \in A$ is realized next period according to a conditional probability distribution $\psi(a | z)$. It will be useful to study the structure of these transitions without estimating the transition probabilities precisely (that will come later). In particular we shall examine how the state variable of each player shifts given the player's current state and the current realization of actions a. There are four cases to consider, depending on the player's current mood.

Content: $z_i = (c, \overline{a}_i, \overline{u}_i)$. Agent i chooses a_i next period, which differs from $\overline{a}_i$ if and only if i is experimenting. The possible transitions are:

$$\begin{tikzcd}
 & (c,\overline{a}_i,\overline{u}_i) & a_i \neq \overline{a}_i,\ u_i(a) \leq \overline{u}_i & (1) \\
 & (c,a_i,u_i(a)) & a_i \neq \overline{a}_i,\ u_i(a) > \overline{u}_i & (2) \\
z_i = (c,\overline{a}_i,\overline{u}_i) \arrow[uur] \arrow[ur] \arrow[r] \arrow[dr] \arrow[ddr] & (w,\overline{a}_i,\overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) < \overline{u}_i & (3) \\
 & (c,\overline{a}_i,\overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) = \overline{u}_i & (4) \\
 & (h,\overline{a}_i,\overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) > \overline{u}_i & (5)
\end{tikzcd}$$

The first case says that if i experiments and his payoff *does not increase*, then i keeps the previous benchmarks and remains content. The second case says that if i experiments and his payoff *does increas*e, he adjusts his benchmark payoff to the new higher level, takes the new strategy as his benchmark strategy, and remains content. The next three cases deal with the situation in which i does not experiment. He becomes watchful, content, or hopeful depending on whether the realized payoff was lower, the same, or higher than his benchmark.

Watchful: $z_i = (w,\overline{a}_i,\overline{u}_i)$. Agent i plays his benchmark strategy next period $(a_i = \overline{a}_i)$. If the realized payoff $u_i(a)$ is below his benchmark, he becomes discontent; if it equals his current benchmark he becomes content with the old benchmarks; if it is higher he becomes hopeful with the old benchmarks.

$$\begin{tikzcd}
 & (d,\overline{a}_i,\overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) < \overline{u}_i & (6) \\
z_i = (w,\overline{a}_i,\overline{u}_i) \arrow[ur] \arrow[r] \arrow[dr] & (c,\overline{a}_i,\overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) = \overline{u}_i & (7) \\
 & (h,\overline{a}_i,\overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) > \overline{u}_i & (8)
\end{tikzcd}$$

Hopeful: $z_i = (h, \overline{a}_i, \overline{u}_i)$. Agent i plays his benchmark strategy $(a_i = \overline{a}_i)$: if the realized payoff is lower than his benchmark he becomes watchful with the old benchmarks; if the realized payoff equals the benchmark he becomes content with the old benchmarks. If the realized payoff is higher, he becomes content with the realized payoff as the new benchmark.

$$\begin{tikzcd}[row sep=small]
 & (w, \overline{a}_i, \overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) < \overline{u}_i & (9) \\
z_i = (h, \overline{a}_i, \overline{u}_i) \arrow[ur] \arrow[r] \arrow[dr] & (c, \overline{a}_i, \overline{u}_i) & a_i = \overline{a}_i,\ u_i(a) = \overline{u}_i & (10) \\
 & (c, \overline{a}_i, u_i(a)) & a_i = \overline{a}_i,\ u_i(a) > \overline{u}_i & (11)
\end{tikzcd}$$

Discontent: $z_i = (d, \overline{a}_i, \overline{u}_i)$. In this case the agent's benchmark strategy and benchmark payoff do not matter: he plays a strategy a_i drawn uniformly at random from A_i. Spontaneously he becomes content with probability $\phi(u_i(a), \overline{u}_i)$, where the response function ϕ is bounded away from 0 and 1, that is, $\theta \le \phi(u_i, \overline{u}_i) \le 1 - \theta$ for some $\theta > 0$.[4] When agent i becomes content, his current strategy a_i and payoff level $u_i(a)$ serve as his new benchmarks; otherwise he continues to be discontent with the old benchmarks.

$$\begin{tikzcd}
 & (c, a_i, u_i(a)) & \textit{with prob } \phi(u_i(a)) & (12) \\
z_i = (d, \overline{a}_i, \overline{u}_i) \arrow[ur] \arrow[dr] & & & \\
 & (d, \overline{a}_i, \overline{u}_i) & \textit{with prob } 1 - \phi(u_i(a)) & (13)
\end{tikzcd}$$

The precise form of the response function ϕ is not important for our results, though from a behavioral standpoint it is natural to assume that it is *monotone increasing* in the realized payoff u_i and *monotone decreasing* in the benchmark $\overline{u}_i$:

[4] The response functions can differ among agents without changing the results; purely for notational convenience we shall assume that the same ϕ applies to everyone.

higher values of the former and lower values of the latter mean that the agent is more likely to become content again. Note, however, that there is no *guarantee* that the agent will become content no matter how high u_i is relative to $\bar{u}_i$; in particular he may remain discontent even if his previous benchmark is realized, and may become content even when it is not. In this sense the procedure differs from most forms of aspiration learning, where payoffs above or below the aspiration level determine the response more sharply (references).[5]

Definition. A game G is *interdependent* if any proper subset S of players can influence the payoff of at least one player not in S by some (joint) choice of actions. More precisely, G is *interdependent* if

$$\forall S, \varnothing \subset S \subset N, \forall a \in A, \exists i \notin S, \exists a'_S \neq a_S, \ u_i(a'_S, a_{-S}) \neq u_i(a_S, a_{-S}). \tag{14}$$

For a randomly generated game G on a finite strategy space A, interdependence holds *generically*, because it holds if there are no payoff ties. Notice, however, that interdependence is a considerably weaker condition: there can be many payoff ties so long as there is enough variation in payoffs that each subgroup can affect the payoff of *someone* not in the group by an appropriate choice of strategies.

Definition. Consider a stochastic process $\{X_t\}$ and suppose that each realization of X_t either does or does not have some property P. Given any realization of the process, let p_t be the proportion of times that property P holds in the first t

[5] An agent may forget what his earlier benchmark was. For example, he might remain discontent even though his realized payoff is higher than it ever was before, or he might become content even though his payoff is lower than the level that originally made him discontent. One is reminded of the rabbi who instructed the unhappy peasant to put a goat in his house: later he was delighted when the rabbi said he could take it out again.

periods. *Property* P *holds at least* r *of the time* if $\underline{\lim}\, p_t \geq r$ for almost all realizations.

Theorem 1. *Let* $\mathcal{G}_A^0$ *be the set of all n-person, interdependent games* G *on a finite joint action space A such that* G *possesses at least one pure Nash equilibrium. Suppose the players use ITE learning with experimentation probability* ε. *If* ε *is sufficiently small, then for all* $G \in \mathcal{G}_A^0$ *a pure Nash equilibrium is played at least* $1-\varepsilon$ *of the time.*

Notice that the theorem holds for *all* games in the given class provided that ε is small enough; in other words, the rate of experimentation does not have to be adjusted to the particular game in hand.

3. Discussion

Before proving theorem 1 formally let us briefly outline the argument. On the one hand, if the learning process is in a non-equilibrium state, it takes *only one* person to experiment with the 'right' action and the experiment will succeed (yield a higher payoff). Hence the process transits to a state having different benchmarks with probability at least $O(\varepsilon)$. On the other hand, if the process is in an equilibrium state, then *at least two* people must experiment together (or in close succession) for the experiments to succeed. Hence the process transits to a state with new benchmarks with probability at most $O(\varepsilon^2)$. Thus, when ε is very small, the process stays in the equilibrium states much longer than in the disequilibrium states. The key point to establish is that the process *enters* an equilibrium state with reasonably high probability starting from an arbitrary initial state. This requires a detailed argument and is the place where the interdependence property is used.

As we have already remarked, a much simpler version of this procedure works for potential games and the more general class of weakly acyclic games. A game G is *weakly acyclic* if, from every pure strategy-tuple there exists at least one sequence of strict better replies, one player moving at a time, that ends in a pure Nash equilibrium strategy-tuple. If all players use ordinary trial and error learning with sufficiently small experimentation probability ε, and if the game is weakly acyclic, they play a stage-game pure Nash equilibrium at least $1-\varepsilon$ of the time (Marden, Young, Arslan, and Shamma, 2007). As we have also pointed out, ITE learning is not the only payoff-based procedure that leads to Nash equilibrium with high probability. Regret testing also has this property, and works for generic games with pure or mixed equilibria (Foster and Young, 2006; Germano and Lugosi, 2007). Regret testing is more complex, however, because it relies on *statistical estimation*. Agents collect data and periodically compare the average payoffs generated by their current strategies with the average payoffs produced by occasional deviations. When the average payoff from deviating exceeds the average payoff from the current strategy by more than some tolerance level $\tau > 0$, the agent switches to a randomly chosen new strategy. In particular, the agent does not necessarily choose the strategy that actually did better when experimenting, hence the search has an undirected aspect. In ITE learning, by contrast, successful experiments are always implemented. Random search arises when payoffs got worse and the agent *did not experiment*; moreover even in this case the search is directed because higher realized payoffs lead the player to abandon the search with higher probability (assuming that ϕ is monotone increasing). This seems like a plausible behavioral hypothesis.

4. Proof of theorem 1: preliminaries

The proof uses the theory of perturbed Markov chains (Young, 1993). Suppose that all players in the game G use ITE learning with experimentation probability

ε and a given response function ϕ (which will be fixed throughout).[6] Let the probability transition matrix of this process be denoted by P^{ε}, where for every pair of states $z, z' \in Z$, $P^{\varepsilon}_{zz'}$ is the probability of transiting in one period from z to z'. We assert that $P^{\varepsilon}_{zz'}$ is a *polynomial* in ε (with coefficients depending on ϕ). To see why, suppose that z is the current state with benchmark strategies $\overline{a}$, and suppose that the vector a is realized next period, resulting in the state z'. If $a \neq \overline{a}$, some subset of k content players experimented. The probability of this event is $c\varepsilon^{k}(1-\varepsilon)^{n-k}$ where c is a constant. (The other $n-k$ players were either not content in z, or were content and did not experiment, but all of these events have probabilities that do not depend on ε.) If $a = \overline{a}$, no one experimented but someone's mood may have changed, the probability of such an event is independent of ε, which is trivially a polynomial in ε. Hence in all cases $P^{\varepsilon}_{zz'}$ is a polynomial in ε, possibly of degree zero.

Definition. The *resistance* of the transition $z \to z'$, written $r(z \to z')$, is the lowest exponent on ε among all nonzero terms in the polynomial describing $P_{zz'}$. (Note that r can be zero.)

Let $Z_1, Z_2, ..., Z_h$ be the distinct recurrence classes of the Markov chain P^{ε}. Starting from any initial state, the probability is one that the process eventually enters one of these classes and stays there ever after. To characterize the long-run behavior of P^{ε}, it therefore suffices to examine its long-run behavior when restricted to each of the classes Z_j. Let P^{ε}_{j} denote the process restricted to the recurrence class Z_j. This process is irreducible, and the resistances of its

[6] Players can have different experimentation probabilities provided they go to zero at the same rate. We could assume, for example, that each player i has an experimentation probability $\lambda_i \varepsilon > 0$, where the parameter ε is varied while the λ_i are held fixed. This complicates the notation unnecessarily, so in the proofs we shall assume a common rate ε.

transitions are defined just as for P^{ε}. Hence the restricted process is a regular, perturbed Markov chain (Young, 1993), and we can study its asymptotic behavior for small ε using the theory of large deviations.

Given a state $z \in Z_j$, a *tree rooted at* z, or z*-tree*, is a set of $|Z_j| - 1$ directed edges that span the vertex set Z_j, such that from every $z' \in Z_j - \{z\}$ there is a *unique directed path* from z' to z. Denote such a tree by $\mathscr{T}_z$. The *resistance* of $\mathscr{T}_z$ is defined to be the sum of the resistances of its edges:

$$r(\mathscr{T}_z) = \sum_{(z,z') \in \mathscr{T}_z} r(z \to z'). \qquad (15)$$

The *stochastic potential* of z is defined to be

$$\rho(z) = \min\{r(\mathscr{T}_z) : \mathscr{T}_z \text{ is a tree rooted at } z\}. \qquad (16)$$

Let Z^{-} be the subset of all states z that minimize $\rho(z)$. The following result follows from Young (1993, theorem 4).

For each class Z_j and each $z \in Z_j$, let $\mu_j^{\varepsilon}(z)$ be the long-run probability of z in the process P_j^{ε}. Then $\lim_{\varepsilon \to 0} \mu_j^{\varepsilon}(z) = \bar{\mu}_j(z)$ exists and $\bar{\mu}_j(z) > 0$ only if $z \in Z^{-}$. (17)

The states z such that $\bar{\mu}_j(z) > 0$ are said to be *stochastically stable* (Foster and Young, 1990). In effect, they are the only states that have nonvanishing probability when the parameter ε becomes arbitrarily small.

5. Proof of theorem 1.

The proof of theorem 1 amounts to showing that: i) every recurrence class Z_j contains at least one all-content state in which the benchmarks constitute a pure Nash equilibrium of G; ii) the stochastically stable states are all of this type.

Let Z^o be the subset of states $z = (m, \overline{a}, \overline{u})$ such that $\overline{u}_i = u_i(\overline{a})$ for all agents i. In other words, Z^o is the subset of states such that the agents' benchmark payoffs and benchmark actions are *aligned*. Let $C^o \subset Z^o$ be the subset of such states in which all agents are content. Let E^o be the subset of C^o in which the benchmark actions $\overline{a}$ form a pure Nash equilibrium of G. Finally, let $Z^o(\overline{a})$ denote the set of all states in Z^o such that the action benchmarks are some given $\overline{a}$.

Definition. A *path* in Z is a sequence of transitions $z^1 \to z^2 \to ... \to z^m$ such that all states on the path are distinct.

Claim 1. For every $z \notin C^o$ there exists a zero-resistance path of length at most three from z to some state in C^o.

Proof. Given any state $z = (m, \overline{a}, \overline{u}) \notin C^o$, I claim that the benchmark action-tuple $\overline{a}$ is played next period with probability $O(\varepsilon^0)$. Consider the cases: i) if in state z agent i is content, he plays $\overline{a}_i$ next period with probability $1-\varepsilon$; ii) if agent i is hopeful, he plays $\overline{a}_i$ again for sure and waits to see the payoff; iii) if agent i is watchful he plays $\overline{a}_i$ again for sure and waits to see the payoff; iv) if agent i is discontent, he plays $\overline{a}_i$ with probability $1/|A_i|$. Therefore $\overline{a}$ is played with probability $O(\varepsilon^0)$. Moreover, when $\overline{a}$ is played, each discontent agent i *spontaneously becomes content* with probability at least $q > 0$, in which case $i's$ new benchmark action is $\overline{a}_i$ and his new benchmark payoff is $u_i(\overline{a})$. Denote the resulting state by z'. By construction, z' has the benchmark actions $\overline{a}$; furthermore, all the content agents in z' have the corresponding payoff benchmarks $u_i(\overline{a})$. (If i just became content, i adopts the payoff from the preceding period as his benchmark, which was in fact $u_i(\overline{a})$.)

It could be that the transition $z \to z'$ caused some players to *become* hopeful, watchful, or discontent, so in state z' these players may have payoff benchmarks that are not aligned with the action benchmarks $\overline{a}$. In the *next period,* however, the same events occur with probability $O(\varepsilon^0)$: $\overline{a}$ will again be played and the discontent players will all become content with the appropriate payoff benchmarks. Call this state z''. Since $\overline{a}$ was played twice in succession on the path $z \to z' \to z''$, everyone's payoff stayed the same for one period. Hence every hopeful player in z' has now become content with payoff benchmark $u_i(\overline{a})$; furthermore every watchful player in z' has now become discontent. Meanwhile all the watchful players in z' have (by assumption) become content in z'' with the appropriate payoff benchmarks. Thus *in one more transition* of the same type, $\overline{a}$ will be played and everyone will become content with the payoff benchmarks $u_i(\overline{a})$. We have therefore shown that it takes at most three transitions, each having zero resistance, to go from any state not in C^o to some state in C^o.

Claim 2. If $e = (m, \overline{a}, \overline{u}) \in E^o$ and z has action benchmarks different from $\overline{a}$, then every path from e to z has resistance at least two.

Proof. Consider any path $e \to z^1 \to z^2 \to ... \to z^m = z$. By definition of E^o, everyone in e is content and they are playing a pure equilibrium, namely, $\bar{a}$. Hence $r(e \to z^1) \geq 1$, because at least one agent must experiment for the process to exit from e. If $r(e \to z^1) \geq 2$ we are done. Suppose therefore that $r(e \to z^1) = 1$, that is, the transition involves an experiment by *exactly one agent* (say i). Since $\bar{a}$ is an equilibrium, $i's$ experiment does not lead to a payoff improvement for i. Hence in state z^1 the benchmark actions are still $\bar{a}$, and the benchmark payoffs are still $\bar{u}$. (Note, however, that in z^1 some agents may have become hopeful or watchful, though none is yet discontent.)

Suppose that, in the transition $z^1 \to z^2$, none of the contented agents experiments. Then $\bar{a}$ is played, so in z^2 all the hopeful and watchful agents (if any) have *reverted* to a contented mood with benchmarks $\bar{a}, \bar{u}$. But this is the original state e, which contradicts the assumption that a path consists of *distinct* states. We conclude that at least one agent does experiment in the transition $z^1 \to z^2$, which implies that $r(z^1 \to z^2) \geq 1$. Hence the total resistance along the path is at least two, as claimed.

Definition. A transition from state z to another state is *easy* if it has the lowest resistance among all transitions out of z. A sequence of transitions $z^1 \to z^2 \to ... \to z^m$ is an *easy path* from z^1 to z^m if all states are distinct and all transitions are easy.

In particular, if $z^1 \to z^2 \to ... \to z^m$ is an easy path, then for every $k < m$,

$$z_k \in C^0 \Rightarrow r(z_k \to z_{k+1}) = 1 \text{ and } z_k \notin C^0 \Rightarrow r(z_k \to z_{k+1}) = 0. \qquad (19)$$

Claim 3. For every state not in E^o, there exists an easy path to some state in E^o.

Proof. Suppose that $z \notin E^o$. If also $z \notin C^o$, then by claim 1 there exists a zero-resistance path to some state $z^1 \in C^o$, which is obviously an easy path. If $z^1 \in E^o$ we are done. Otherwise it suffices to show that there exists an easy path from z_1 to some state in E^o. Let $(\overline{a}, \overline{u})$ be the benchmarks in state z^1, which are aligned in the sense that $\overline{u}_i = u_i(\overline{a})$ for all i, because $z^1 \in C^o$. Since $z^1 \in C^o - E^o$, there is an agent i and an action $a_i \neq \overline{a}_i$ such that $u_i(a_i, \overline{a}_{-i}) > u_i(\overline{a}_i, \overline{a}_{-i}) = \overline{u}_i$. The probability that $(a_i, \overline{a}_{-i})$ is realized next period is $(1-\varepsilon)^{n-1}\varepsilon/(|A_i|-1)$, which occurs when i experiments and chooses a_i, while the others do not experiment. This results in a state z^2 where i is content, $i's$ new benchmarks are a_i and $u_i(a_i, \overline{a}_{-i})$ respectively, and the others' benchmarks are as before (though their moods may have changed). Note that $i's$ payoff benchmark has *strictly increased*, while the others' payoff benchmarks have stayed the same. Note also that the lowest order term in $(1-\varepsilon)^{n-1}\varepsilon/(|A_i|-1)$ has order one (in ε), so $r(z^1 \to z^2) = 1$. Since all other transitions out of z^1 have resistance at least 1, $z^1 \to z^2$ is an easy path. As we have just noted, this is also a *monotone increasing path* in the sense that no one's benchmark payoff decreases and someone's strictly increases.

If $z^2 \in E^o$ we are done. Otherwise there are three possibilities to consider: i) everyone in z^2 is content; ii) some are hopeful and no one is watchful; iii) someone is watchful. (No one can be discontent at this stage, since it takes at least two periods of disappointing payoffs to become discontent.)

In the first case everyone is content, so evidently $i's$ change of action did not change anyone else's payoff. Hence $z^2 \in C^o$ and we can simply repeat the earlier argument to extend the path by one more transition, $z^2 \to z^3$, having resistance 1. This is an easy and monotone increasing continuation of the path. In the second case there is a zero-resistance (hence easy) transition to a state $z^3 \in C^o$ in which everyone becomes content, the benchmark payoffs for everyone are at least as high as they were in state z^2, and they are *strictly higher* for those who were hopeful (this happens when everyone in state z^2 plays his action benchmark). So again there is an easy and monotone increasing continuation of the path.

We shall consider the third case in a moment. Notice, however, that if the continuation of the path always involves cases i) and ii), then it will always be monotone increasing. Since the state space is finite, it must come to an end, which can only happen when it reaches some equilibrium state in E^o.

We now consider the other case, namely, the path reaches a first transition where some agent becomes *watchful*, but no one is yet discontent. Suppose this happens in the transition $z^k \to z^{k+1}$. Up to this point, transitions have either: i) involved a single contented agent making an experiment that led to a better payoff for himself; or ii) involved one or more hopeful agents playing their benchmark actions and becoming content with new higher benchmark payoffs (but not both i) and ii)). It follows that there are no hopeful agents in state z^k, because hopeful agents do not try new actions, so they cannot cause *someone else* to become watchful (which is what happened for the first time in the transition $z^k \to z^{k+1}$). Thus all agents in z^k are content, $z^k \in C^o$, and in the transition $z^k \to z^{k+1}$ there is exactly one agent, say i, who successfully experimented and caused the payoff of some other agent, say j, to go down.

Let $\overline{a}^k, \overline{u}^k$ be the benchmark actions and payoffs in state z^k; these are aligned because $z^k \in C^0$. Let $\overline{a}^{k+1}, \overline{u}^{k+1}$ be the benchmarks in state z^{k+1}. Note that only $i's$ benchmark action and payoff changed between the two states; agents who became watchful or hopeful in z^{k+1} have not changed their benchmarks yet (they will wait one more period). In the next period the probability is at least $(1-\varepsilon)^{n-1}$ that the current action benchmarks $\overline{a}^{k+1}$ will be played again. In this case all the watchful agents experience another disappointing payoff and become discontent, while all the other agents become content. Thus the process transits with zero resistance to a state z^{k+2} in which there is at least one discontent agent and there are no hopeful or watchful agents. In state z^{k+2} the benchmarks are still $\overline{a}^{k+1}, \overline{u}^{k+1}$, and they are partially aligned in the sense that $u_j(\overline{a}^{k+1}) = \overline{u}_j^{\,k+1}$ for all agents j who are not discontent.

Let D be the subset of discontent agents in z^{k+2}. To avoid notational clutter let us drop the superscripts on the current benchmarks and denote them by $(\overline{a}, \overline{u})$. By assumption G is interdependent, hence there exists an agent $j \notin D$ and an action-tuple a'_D such that $u_j(a'_D, \overline{a}_{N-D}) \neq u_j(\overline{a}_D, \overline{a}_{N-D}) = \overline{u}_j$. We claim that there is a sequence of four (or fewer) easy transitions that make all the agents in $D \cup \{j\}$ discontent.

Case 1. $u_j(a'_D, \overline{a}_{N-D}) > u_j(\overline{a}_D, \overline{a}_{N-D})$.

Consider the following sequence: in the first and second period the players in D play a'_D and in the third and fourth periods they revert to $\overline{a}_D$, all the while remaining discontent. (In each of these periods the players not in D keep playing $\overline{a}_{N-D}$.) This initially raises $j's$ expectations, which are later quashed (an inverted goat strategy). The sequence of transitions and play realizations looks like this:

$$\begin{tikzcd}
z^{k+2} \arrow[r, "(a'_D, \overline{a}_{N-D})"] & z^{k+3} \arrow[r, "(a'_D, \overline{a}_{N-D})"] & z^{k+4} \arrow[r, "(\overline{a}_D, \overline{a}_{N-D})"] & z^{k+5} \arrow[r, "(\overline{a}_D, \overline{a}_{N-D})"] & z^{k+6} \\
 & u_j \uparrow & \overline{u}_j \uparrow & u_j \downarrow & \\
 & j \textit{ hopeful} & j \textit{ content} & j \textit{ watchful} & j \textit{ discontent}
\end{tikzcd}$$

I claim that each of these transitions has zero resistance, so this is an easy path. Indeed, in each transition the players in D play their required actions *and stay discontent*, which has probability at least $(\theta / m)^{|D|}$, where $m = \max_i |A_i|$. Meanwhile each of the players $i \notin D$ continues playing his benchmark $\overline{a}_i$, which has probability $1-\varepsilon$ if content, probability at least θ / m if discontent, and probability 1 if watchful or hopeful. These probabilities are all bounded away from zero when ε is small, hence all the transitions have zero resistance. Thus by state z^{k+6}, and possibly earlier, the set of discontent agents has expanded from D to $D \cup \{j\}$ or more .

Case 2. $u_j(a'_D, \overline{a}_{N-D}) < u_j(\overline{a}_D, \overline{a}_{N-D})$

This case just involves two transitions: everyone in D plays a'_D and stays discontent, while the others play $\overline{a}_{N-D}$. This makes player j discontent in two steps.

Thus in both cases there is an easy path from z^{k+2} to a state $z*$ in which *all* agents are discontent. Given any $e \in E$, the probability is at least $(\theta / m)^n$ that $z* \to e$ in one period; indeed this happens if all n agents choose their part of the equilibrium specified by e and spontaneously become content.

We have therefore shown that, from any initial state $z^1 \notin E^o$ there exists an easy path to some state in E^o. This establishes claim 3.

Recall that $\rho(z)$ is defined to be the resistance of the least resistant tree(s) rooted at z. To establish theorem 1, it therefore suffices to show the following (see statement (17) in the preceding section).

Claim 4. $\forall z \notin E, \exists e \in E, \quad \rho(e) < \rho(z)$

Proof. Let z be in the recurrence class Z_j, and let $\mathscr{T}_z$ be a least-resistant tree that spans Z_j and is rooted at z. Suppose that $z \notin E$. By claim 3 there exists an easy path from z to some state $e \in E$. Denote this path by $z = z^0 \to ... \to z^k = e$, and let $\mathscr{P}$ be the set of its k directed edges. We shall construct a new tree that is rooted at e and has lower resistance than does $\mathscr{T}_z$.

In $\mathscr{T}_z$, each state $z' \neq z$ has a unique *successor state* $s(z')$; in other words, $z' \to s(z')$ is the unique edge exiting from z'. Adjoin the path $\mathscr{P}$ to $\mathscr{T}_z$, and remove each edge $z^j \to s(z^j)$ that is not in $\mathscr{P}$ (i.e., such that $s(z^j) \neq z^{j+1}$). Call the resulting set of edges $\mathscr{S}$. Since $\mathscr{P}$ is an easy path, each of its transitions $z^j \to z^{j+1}$ has least resistance among all transitions out of the state z^j, hence

$$r(z^j \to z^{j+1}) \leq r(z^j \to s(z^j)) \text{ for } 1 \leq j < k. \qquad (18)$$

Furthermore, $z \to z^1$ is an easy transition out of state z, so $r(z \to z^1) \leq 1$. Hence

$$r(\mathscr{S}) \leq r(\mathscr{T}_z) + 1. \qquad (19)$$

Next let $e \to w^1 \to w^2 \to ... \to w^j$ be the unique path in $\mathscr{T}_z$ (and $\mathscr{S}$) leading from e toward z, where w^j is the first state on the path such that e and w^j do *not* have the same benchmarks. From claim 2 we know that

$$r(e \to w^1) + r(w^1 \to w^2) + ... + r(w^{j-1} \to w^j) \geq 2\,. \qquad (20)$$

Remove each of these j edges from $\mathscr{S}$, and adjoin the $j-1$ edges

$$w^1 \to e, w^2 \to e, ..., w^{j-1} \to e\,. \qquad (21)$$

The result of all of these edge-exchanges is now a tree $\mathscr{T}_e$ rooted at e. (See figure 1 for an example.)

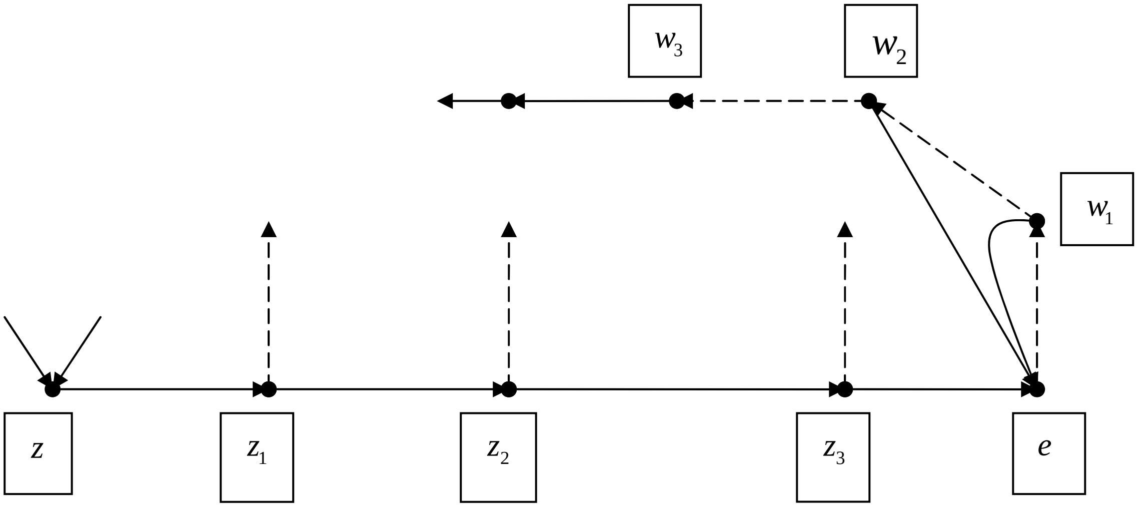

Figure 1. Construction of a tree rooted at e from a tree rooted at z by adding edges (solid) and subtracting edges (dashed).

By claim 1, each of the transitions in (21) has zero resistance, hence $r(\mathcal{T}_e) \le r(\mathcal{S}) - 2$. Combined with the previous result that $r(\mathcal{S}) \le r(\mathcal{T}_z) + 1$, we deduce that $r(\mathcal{T}_e) < r(\mathcal{T}_z)$. Hence $\rho(e) < \rho(z)$, because for any state w, $\rho(w)$ is the resistance of the least-resistant tree rooted at w. This completes the proof of theorem 1.

6. Non-generic payoffs

It is straightforward to construct games with non-generic payoffs such that ITE learning does not come close to Nash equilibrium behavior at *any* time, let alone *most* of the time. These examples require that the game have three or more players; ITE learning does come close to Nash equilibrium for all finite two-player games, as will be shown in Theorem 2.

First we demonstrate that some form of genericity is required when there are three or more players. Consider the three-person game in Figure 2, where each player has two actions. There is a unique pure equilibrium in the lower northeast corner, and a best response cycle on the top square. Suppose that the process starts in a state where player 3 is content. Since her payoffs are constant, no amount of experimenting will produce better results, and nothing the other players do will trigger a change in her mood. In short, once player 3 begins in a content state she remains content and never changes action. If she starts by playing the action corresponding to the top square, no combination of actions by the other two players constitutes a Nash equilibrium. Hence there exist initial states from which ITE learning never leads to a pure Nash equilibrium even though there is one. (By contrast, if the process begins in a state where player 3 chooses the action corresponding to the lower square, the pure equilibrium will eventually be played with probability one.)

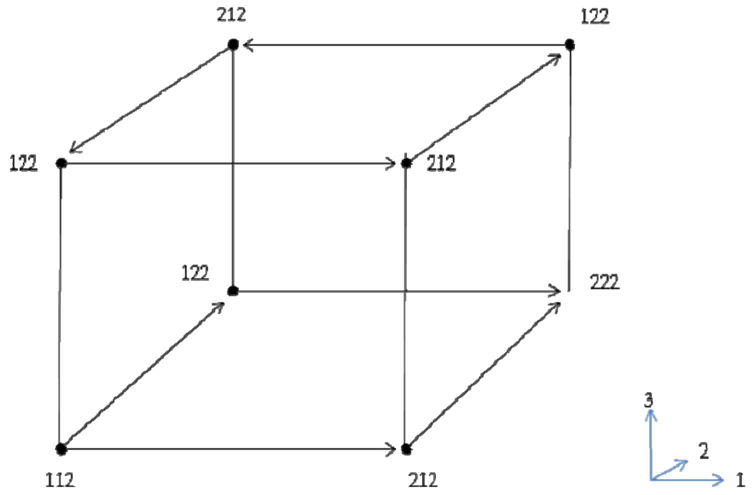

Figure 2. A three-person game with non-generic payoffs in which ITE learning does not necessarily lead to Nash equilibrium play. Arrows indicate best-response transitions.

Similar examples can be constructed when there are more than three players. This is not the case when there are only two players, as the next result shows.

Theorem 2. *Let $\mathcal{G}_A^*$ be the set of all two-person games G on a finite joint action space $A = A_1 \times A_2$ such that G possesses at least one pure Nash equilibrium. Suppose the players use ITE learning with experimentation probability ε. If ε is sufficiently small, then for all $G \in \mathcal{G}_A^*$ a pure Nash equilibrium is played at least $1-\varepsilon$ of the time.*

Proof. Consider a two-person game on a finite joint action space $A = A_1 \times A_2$, where the game possesses at least one pure Nash equilibrium. A *best response path* is a sequence of action-tuples $a^1 \rightarrow a^2 \rightarrow ... \rightarrow a^m$ such that the action-tuples are all distinct, and for each transition $a^k \to a^{k+1}$ there exists a unique player i such that $a^k_{-i} = a^{k+1}_{-i}$ and a^{k+1}_i is a *strict best response* by i to a^k_{-i} . The sequence is a *best response cycle* if all of these conditions hold except that $a^1 = a^m$.

The only part of the proof of theorem 1 that relied on the interdependence assumption was the proof of Claim 3. We shall show that this claim holds for two players without invoking interdependence, from which the theorem follows immediately.

Recall that E^o denotes the set of states such that everyone is content and the benchmarks correspond to a pure Nash equilibrium. (For other definitions and notation the reader is referred to the proof of theorem 1.)

Claim. For every state not in E^o, there exists an easy path to some state in E^o.

Proof. Suppose that $z \notin E^o$. If also $z \notin C^o$, then by claim 1 (in the proof of theorem 1) there exists a zero-resistance path to some state $z^1 \in C^o$, which is obviously an easy path. If $z^1 \in E^o$ we are done. Otherwise it suffices to show that there exists an easy path from z^1 to some state in E^o. Let $(\overline{a}^1, \overline{u}^1)$ be the benchmarks in state z^1, which are aligned by definition of C^o. We now distinguish two cases.

Case 1. There exists a best response path from $\overline{a}^1$ to a pure Nash equilibrium, say $\overline{a} = a^1 \to a^2 \to \ldots \to a^m$.

For $1 \le k \le m$, let z^k be the *state* that has action benchmarks a^k_i, payoff benchmarks $u_i(a^k)$, and everyone is content. We can construct an easy path to $z^m \in E^o$ by tracking the best response path as follows. In state z^1, let the relevant player experiment and choose a best reply to the others' current actions, while the others play these actions. Thus they play a^2, and the probability of this event is $O(\varepsilon)$. Of course this may cause some players to undergo a change of mood. In the next period, however, a^2 will be played again and everyone will become content, all with probability $O(\varepsilon^0)$. At this point the process has reached the state z^2 via easy transitions. Repeating the argument we conclude that there is an easy path to the target state $z^m \in E^o$, as claimed.

Case 2. There exists no best response path from $\overline{a}^1$ to a pure Nash equilibrium.

Given that there is no best response path from $\overline{a}^1$ to a pure Nash equilibrium, there must exist a best response path from $\overline{a}^1$ that leads to a best response cycle. Denote such a cycle by $a^j \rightarrow a^{j+1} \rightarrow ... \rightarrow a^m = a^j$ and the path to it by $\overline{a}^1 = a^1 \rightarrow a^2 \rightarrow ... \rightarrow a^j$. As in case 1 we can construct an easy path that mimics the best response path up to a^j; we need to show that it can be extended as an easy path to a Nash equilibrium.

Along the cycle the two players alternate in choosing best responses, say player 1 chooses a strict best response going from a^j to a^{j+1}, player 2 from a^{j+1} to a^{j+2}, and so forth. Since these are strict best responses and the process cycles, each player's payoff must at some stage *decrease*. Proceeding from a^j, let $a^k \rightarrow a^{k+1}$ be the first transition on the cycle such that some player's payoff strictly decreases, say player 2's. Since this is a best response cycle, player 1's payoff must strictly increase in the transition $a^k \rightarrow a^{k+1}$. Moreover in the *previous* transition on the cycle, $a^{k-1} \rightarrow a^k$, player 2's payoff must strictly increase because the players alternate in making best responses. We therefore know that

$$u_1(a^{k+1}) > u_1(a^k) \text{ and } u_2(a^k) > u_2(a^{k-1}). \qquad (22)$$

We now consider two possibilities.

Case 2a. $u_1(a^k) < u_1(a^{k-1})$.

By assumption $a^k \to a^{k+1}$ was the first transition after a^j such that any decrease occurred, so $j = k$. As in case 1 we can construct an easy path (in the full state space) that mimics the transitions along the path $\overline{a}^1 = a^1 \to a^2 \to ... \to a^k$ and then mimics the cycle beginning at a^k. Consider the situation when this path *first returns* to a^k, that is, the players are content with benchmarks a^{k-1} and in the next period they play a^k. This causes player 1's payoff to decrease, so player 1 becomes watchful, while player 2 remains content. In the next period the probability is $O(\varepsilon^0)$ that player 1 plays action a_1^k again and *becomes discontent*, while player 2 plays action a_2^k again and remains content. In the next period after that, the probability is $O(\varepsilon^0)$ that player 1 chooses action a_1^{k+1} and *remains discontent*, while player 2 does not experiment, chooses $a_2^k = a_2^{k+1}$ again, and remains content. By assumption, player 2's payoff decreases in this transition ($a^k \to a^{k+1}$). In the period after that, with probability $O(\varepsilon^0)$ they play a^{k+1} again, player 1 remains discontent, and player 2 becomes discontent. At this juncture *both* players are discontent. Hence in one more period they will jump to a pure Nash equilibrium, and in one period after that both will become content playing the Nash equilibrium, all with probability $O(\varepsilon^0)$. Thus in case 2a we have constructed an easy path to a state in E^o, that is, to an all-content Nash equilibrium.

Case 2b. $u_1(a^k) \geq u_1(a^{k-1})$.

In this case we construct an easy path (in the full state space) that mimics the transitions along the path $\bar{a}^1 = a^1 \to a^2 \to ... \to a^j$ and then mimics the cycle up to the point where a^{k+1} is first played. At this point player 2 becomes watchful while player 1 is content. In the next period a^{k+1} is played again with probability $O(\varepsilon^0)$, and player 2 becomes discontent while player 1 remains content. In the next period after that, player 2 plays a_2^{k-1} *and remains discontent*, while player 1 sticks with his current action a_1^{k+1}, all with probability $O(\varepsilon^0)$. Denote the resulting pair of actions by $\tilde{a} = (a_1^{k+1}, a_2^{k-1})$. Again we may distinguish two cases.

Case 2b′. $u_1(a^k) \geq u_1(a^{k-1})$ and $u_1(\tilde{a}) < u_1(a^{k+1})$.

In this case player 1 has become watchful while player 2 is discontent, so in one more period they will both be discontent with probability $O(\varepsilon^0)$. As we have already shown, this leads in two more easy steps to a Nash equilibrium, and we are done. It therefore only remains to consider the following.

Case 2b″. $u_1(a^k) \geq u_1(a^{k-1})$ and $u_1(\tilde{a}) \geq u_1(a^{k+1})$.

We claim that this case cannot occur. Recall that the players alternate in making best replies around the cycle. Since player 2 best responded in going from a^{k-1} to a^k, player 1 best responded in the previous move. It follows that a_1^{k-1} is 1's best response to a_2^{k-1}, from which we deduce that $u_1(a^{k-1}) \geq u_1(\tilde{a})$. Putting this together with the case 2b″ assumption we obtain $u_1(a^k) \geq u_1(a^{k-1}) \geq u_1(\tilde{a}) \geq u_1(a^{k+1})$, which implies that $u_1(a^k) \geq u_1(a^{k+1})$, contrary to (22). This concludes the proof of theorem 2.

7. Extensions

Interactive trial and error learning can be generalized in several ways. One is to assume that players react only to "sizable" changes in payoffs. Given a real number $\tau > 0$, define *ITE learning with payoff tolerance* τ to be the same as before except that: i) a player becomes *hopeful* only if the gain in payoff relative to the previous benchmark is strictly greater than τ ; ii) a player becomes *watchful* only if the loss in payoff relative to the previous benchmark is strictly greater than τ.

Say that a game is τ*-interdependent* if any proper subset S of players can -- by an appropriate choice of joint actions -- change the payoff of some player not in S by more than τ. An argument very similar to that of theorem 1 shows the following: *if a game has a* τ*-equilibrium and is* τ*-interdependent, ITE learning with tolerance* τ *and experimentation rate* ε *leads to* τ*-equilibrium play in at least* $1 - \varepsilon$ *of all time periods provided that* ε *is sufficiently small.*

Extensions of the approach to learning mixed equilibria are not quite as straightforward. The obvious modification to make in this case is to assume that each player computes the *average payoff over a large sample of plays* before changing mood or strategy. If the players are using mixed strategies, however, there is always a risk -- due to sample outcome variability -- that the realized average payoffs will differ substantially from their expected values, and hence that one or more players changes mood and strategy due to "measurement error" rather than fundamentals. Thus one needs to assume that players only react to *sizable changes in payoff* and that the *sample size is sufficiently large* that sizable changes occur with very low probability. Moreover, for our method of proof to work, one would need to know that the game is τ-interdependent for a suitable value of τ, but this does not necessarily hold for the mixed strategy version of the game when the underlying game is τ-interdependent. (Consider for example a 2 x 2 game in which every two payoffs differ by more than τ. Each player may nevertheless have a mixed strategy that equalizes his own payoffs for all strategies of the opponent, in which case the mixed-strategy version is certainly not τ-interdependent). Thus, while it may be possible to extend the approach to handle mixed equilibria, the result would be more complex and perhaps not as intuitively appealing as the version described here.

To sum up, interactive trial and error learning is a simple procedure for learning pure equilibria that does not rely on statistical estimation (like regret testing) and does not require observability of the opponents' actions (like the procedure of Hart and Mas-Colell). Even simpler procedures -- such as the MYAS experimentation rule -- work for weakly acyclic games, of which potential games are a special case. We conclude that there exist simple heuristics that allow players to learn equilibrium in a wide variety of strategic situations even when they know nothing about the structure of the game, who the other players are, or what strategies they are pursuing.

References

Capra, C. Monica (2004), "Mood-driven behavior in strategic interactions," *American Economic Review Papers and Proceedings*, 94, 367-372.

Foster, Dean P., and H. Peyton Young (2006), "Regret testing: learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, 1, 341-367.

Germano, Fabrizio, and Gabor Lugosi (2007), "Global convergence of Foster and Young's regret testing," *Games and Economic Behavior*, forthcoming.

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

Hart, Sergiu, and Andreu Mas-Colell (2006), "Stochastic uncoupled dynamics and Nash equilibrium," *Games and Economic Behavior*, 57, 286-303.

Kahneman, Daniel, and Amos Tversky (1979), "Prospect theory: an analysis of decision under risk," *Econometrica*, 47, 263-292.

Kirchsteiger, Georg, Luca Rigotti, and Aldo Rustichini (2006), "Your morals might be your moods," *Journal of Economic Behavior and Organization*.

Marden, Jason, H. Peyton Young, Gurdal Arslan, and Jeff Shamma (2007), "Payoff-based dynamics for multi-player weakly acyclic games," Working Paper, Department of Mechanical and Aerospace Engineering, University of California Los Angeles.

Smith, Kip, and John Dickhaut (2005), "Economics and emotion: institutions matter," *Games and Economic Behavior*, 52,316-335.

Young, H. Peyton (1993), "The evolution of conventions," *Econometrica,* 61, 57-84.