ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

LEARNING BY TRIAL AND ERROR

H. Peyton Young

Number 384
January 2008

Manor Road Building, Oxford OX1 3UQ

Learning by Trial and Error

H. Peyton Young

University of Oxford
The Brookings Institution

Address: Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom.

Tel: 44 1865 271086
Fax: 44 1865 271094
Email: peyton.young@economics.ox.ac.uk

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 any game with generic payoffs and at least one pure Nash equilibrium. Unlike regret testing (Foster and Young, 2006), the method 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. The approach shows that there exist simple and intuitive rules for discovering equilibria in decentralized settings where players have no knowledge of the system in which they are embedded.

JEL Classification: C72, D83

Keywords: learning, adaptive dynamics, Markov process, Nash equilibrium, bounded rationality

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 finite n-person games with generic payoffs.

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 *stationary* with respect to histories of bounded length (Hart and Mas-Colell, 2006). They show that, 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, I shall not insist on convergence to Nash equilibrium; it suffices that period-by-period play come close to Nash equilibrium quite often. Second, I 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 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

[1] The rule operates as follows: if everyone played the same action over the last 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.

type can be effective methods for learning equilibrium in situations where players have no knowledge of what the other players are doing.

2. Interactive trial and error learning

I 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, but does not immediately change his payoff expectations.[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 and only if they lead to higher payoffs. (This is the MYAS procedure.) In an

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

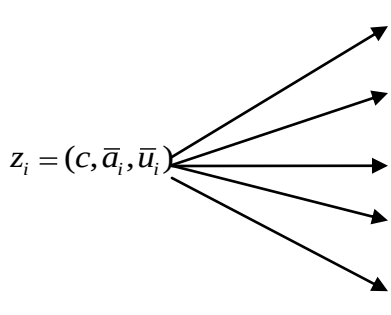
interactive situation, however, "errors" can arise in two different ways: by trying something that turns out to be no better than what one was doing, or by continuing to do something that turns out to be worse than it was before. The former are *active errors* whereas the latter are *passive errors*. These two types of errors trigger different behavioral responses. The key features of the learning process are that: i) each agent searches with small probability, and the search leads to a change of action only if there is payoff improvement (no active error); ii) when an agent experiences a passive error, he first waits to see if the error persists, and if it does he starts flailing around for a random number of periods until he eventually calms down again. The version of the rule proposed here is obviously not the only one with these properties, but this is the one we shall work with. Various extensions and variations will suggest themselves once the method of proof has been demonstrated.

We turn now to an analysis of the specific version of ITE learning defined above. 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 $z_i = (m_i, \bar{a}_i, \bar{u}_i)$, where m_i is i's current mood, $\bar{a}_i$ is i's current benchmark action, and $\bar{u}_i$ is i's current benchmark payoff. The four possible moods are content (c), discontent (d), hopeful (h), and watchful (w). A *state* z of the process specifies a state z_i for each player. We shall write this in the form $z = (m, \bar{a}, \bar{u})$, 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 $\psi(a \mid z)$. 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 realized actions a. There are four cases to consider, depending on the player's current mood.

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



$$(c, \bar{a}_i, \bar{u}_i) \qquad a_i \neq \bar{a}_i,\ u_i(a) \leq \bar{u}_i \qquad (1)$$

$$(c, a_i, u_i(a)) \qquad a_i \neq \bar{a}_i,\ u_i(a) > \bar{u}_i \qquad (2)$$

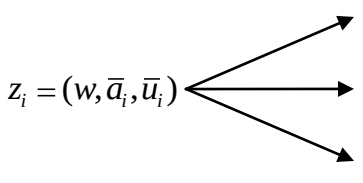
$$(w, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) < \bar{u}_i \qquad (3)$$

$$(c, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) = \bar{u}_i \qquad (4)$$

$$(h, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) > \bar{u}_i \qquad (5)$$

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: $z_i = (w, \bar{a}_i, \bar{u}_i)$. Agent i plays his benchmark strategy next period $(a_i = \bar{a}_i)$. If the realized payoff $u_i(a)$ is less than his payoff benchmark $\bar{u}_i$ he becomes discontent; if it equals $\bar{u}_i$ he becomes content with the old benchmarks; if it is greater than $\bar{u}_i$ he becomes hopeful with the old benchmarks. These possibilities are shown below:

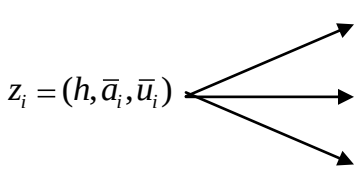


$$z_i = (w, \bar{a}_i, \bar{u}_i) \to (d, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) < \bar{u}_i \qquad (6)$$

$$z_i = (w, \bar{a}_i, \bar{u}_i) \to (c, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) = \bar{u}_i \qquad (7)$$

$$z_i = (w, \bar{a}_i, \bar{u}_i) \to (h, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) > \bar{u}_i \qquad (8)$$

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



$$z_i = (h, \bar{a}_i, \bar{u}_i) \to (w, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) < \bar{u}_i \qquad (9)$$

$$z_i = (h, \bar{a}_i, \bar{u}_i) \to (c, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) = \bar{u}_i \qquad (10)$$

$$z_i = (h, \bar{a}_i, \bar{u}_i) \to (c, \bar{a}_i, u_i(a)) \qquad a_i = \bar{a}_i,\ u_i(a) > \bar{u}_i \qquad (11)$$

Discontent: $z_i = (d, \bar{a}_i, \bar{u}_i)$. 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.[4] Spontaneously he becomes content with probability $\phi(u_i(a), \bar{u}_i)$, where the *response function* ϕ is bounded away from 0 and 1, that is, $\theta \le \phi(u_i, \bar{u}_i) \le 1 - \theta$ for some $\theta > 0$.[5] When agent i becomes content, his current

[4] The assumption of a uniform random draw is unimportant. It suffices that every action is chosen with a probability that is bounded away from zero over all possible states of the process.

[5] 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.

strategy a_i and payoff level $u_i(a)$ serve as his new benchmarks; otherwise he continues to be discontent with the old benchmarks.

$$z_i = (d, \bar{a}_i, \bar{u}_i) \begin{cases} \to (c, a_i, u_i(a)) & \text{with prob } \phi(u_i(a), \bar{u}_i) \quad (12) \\ \to (d, \bar{a}_i, \bar{u}_i) & \text{with prob } 1-\phi(u_i(a), \bar{u}_i) \quad (13) \end{cases}$$

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 $\bar{u}_i$: 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. Moods are not *determined* by the absolute level of one's payoffs, but moods can change when payoffs change.[6]

To state our main result we shall need two further definitions.

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, for every proper subset S and every action-tuple a,

[6] 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.

$$\exists i \notin S, \exists a'_S \neq a_S \text{ such that } u_i(a'_S, a_{-S}) \neq u_i(a_S, a_{-S}). \qquad (14)$$

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 periods. *Property P holds at least r of the time* if and only if $\liminf_t p_t \geq r$ for almost all realizations of the process.

Theorem 1. *Let G be an n-person game on a finite joint action space A such that G is interdependent and has at least one pure Nash equilibrium. If the players use ITE learning with experimentation probability ε and response function ϕ, then for all sufficiently small ε a pure Nash equilibrium is played at least $1-\varepsilon$ of the time.*

3. Proof of theorem 1: preliminaries

Before formally proving theorem 1 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 on the order of ε. On the other hand, if the process

is in an equilibrium state, there must be at least two experiments (together or in close succession) to exit from it more than temporarily, that is, for the benchmarks to change. Hence the process transits to a state with new benchmarks with probability on the order of ε^2 or less. It follows that, 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.

The proof uses the theory of perturbed Markov chains as developed in Young (1993), which builds on work of Freidlin and Wentzell (1984), Foster and Young (1990), and Kandori, Mailath, and Rob (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).[7] 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 if $P^\varepsilon_{zz'} > 0$, then $P^\varepsilon_{zz'}$ is of order ε^k for some non-negative integer k. To see why, suppose that z is the current state with benchmark strategies $\bar{a}$, and suppose that the vector a is realized next period, resulting in the state z'. If $a \neq \bar{a}$, some subset of $k \geq 1$ content players experimented. The probability of this event is $c\varepsilon^k(1-\varepsilon)^{n-k}$ where c depends on z' but not on ε. (The other $n-k$ players were either not content in z, or were content and did not experiment.) If $a = \bar{a}$, then no one experimented but someone's mood may have changed; the probability of this event is $c(1-\varepsilon)^n$ where again c depends on z' but not on ε.

[7] 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.

Hence in all cases $P^{\varepsilon}_{zz'}$ is of order ε^k for some integer $k \geq 0$. (In general we shall say that $P^{\varepsilon}_{zz'}$ *is of order* ε^k, written $P^{\varepsilon}_{zz'} \approx \varepsilon^k$, if $0 < \lim_{\varepsilon \to 0} P^{\varepsilon}_{zz'} / \varepsilon^k < \infty$.)

Definition. If the transition $z \to z'$ occurs with positive probability ($P^{\varepsilon}_{zz'} > 0$), the *resistance* of the transition , written $r(z \to z')$, is the unique integer $k \geq 0$ such that $P^{\varepsilon}_{zz'} \approx \varepsilon^k$.

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 forever. 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 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 $\mathcal{T}_z$. The *resistance* of $\mathcal{T}_z$ is defined to be the sum of the resistances of its edges:

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

The *stochastic potential* of z is defined to be

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

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

For each recurrence class Z_j and every $\varepsilon > 0$, let μ_j^ε be the unique stationary distribution of the process P_j^ε. Then for every $z \in Z_j$, $\lim_{\varepsilon \to 0} \mu_j^\varepsilon(z) = \bar{\mu}_j(z)$ exists and the support of $\bar{\mu}_j$ is contained in Z_j^-. (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.

4. 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 action benchmarks constitute a pure Nash equilibrium of G; ii) the stochastically stable states are all of this form.

Let Z^o be the subset of states $z = (m, \bar{a}, \bar{u})$ such that $\bar{u}_i = u_i(\bar{a})$ 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 $\bar{a}$ form a pure Nash equilibrium of G. The first step in the proof (claim 1 below) will be to show that the only candidates for stochastic stability are states in which everyone is content and benchmarks are aligned (states in C^o). The

remainder of the proof will establish that, in fact, the only candidates for stochastic stability are states in E^o.

Definition. A *path* in Z is a sequence of transitions $z^1 \to z^2 \to ... \to z^m$ such that all states 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. If state $z = (m, \bar{a}, \bar{u}) \notin C^o$, then someone is not content and/or someone's benchmark payoff is not aligned with the benchmark actions, that is, $\bar{u}_i \neq u_i(\bar{a})$ for some player i. I claim that the benchmark action-tuple $\bar{a}$ is played next period with probability $\approx \varepsilon^0$, that is, with a probability that is bounded away from zero for all small ε. Consider the cases: i) if in state z agent i is content, he plays $\bar{a}_i$ next period with probability $1-\varepsilon$; ii) if agent i is hopeful, he plays $\bar{a}_i$ again for sure and waits to see the payoff; iii) if agent i is watchful he plays $\bar{a}_i$ again for sure and waits to see the payoff; iv) if agent i is discontent, he plays $\bar{a}_i$ with probability $1/|A_i|$. Therefore $\bar{a}$ is played with probability $\approx \varepsilon^0$.

Notice that, if $\bar{a}$ is played, each discontent agent i *spontaneously becomes content* (and his benchmarks are $\bar{a}_i$, $u_i(\bar{a})$) with probability θ. Assume that this occurs for all discontent agents, and denote the resulting state by z'. Notice that if some player i was hopeful or watchful in z and becomes content in z', then i's new payoff benchmark is $u_i(\bar{a})$. We have therefore shown that, with probability $\approx \varepsilon^0$, $z \to z'$ where z' has action benchmark vector $\bar{a}$, and every content agent has a payoff benchmark that is aligned with $\bar{a}$.

We shall now show that in two more plays of $\bar{a}$, the process reaches a state in C^o. Let us observe first that the transition $z \to z'$ may have caused some players to *become* hopeful, watchful, or discontent, so we cannot assert that $z' \in C^o$. In the *next period,* however, the probability is $\approx \varepsilon^0$ that $\bar{a}$ will again be played and the previously discontent players (if any) will all become content with benchmarks $\bar{a}_i$, $u_i(\bar{a})$. Call this state z''. Since $\bar{a}$ was played twice in succession on the path $z \to z' \to z''$, every hopeful player in z' has become content in z'', every content player in z' is still content, and by construction all the discontent players have become content. Furthermore all of the content players in z'' have benchmarks $\bar{a}_i$, $u_i(\bar{a})$. There remains the possibility that someone who was watchful in z' has just become discontent in z''. However, *in one more transition,* $\bar{a}$ will be played again and *everyone* will become content with the benchmarks $\bar{a}_i$, $u_i(\bar{a})$, all with probability $\approx \varepsilon^0$. 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, which establishes Claim 1.

Claim 2. If $e = (m, \bar{a}, \bar{u}) \in E^o$ and z has action benchmarks that differ from $\bar{a}$, 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, their actions constitute a pure equilibrium $\bar{a}$, and their benchmark payoffs are aligned with their actions. 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 content 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, and this lowest resistance is either 0 or 1. 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.

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 $(\bar{a}, \bar{u})$ be the benchmarks in state z^1, which are aligned in the sense that $\bar{u}_i = u_i(\bar{a})$ 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 $a_i \neq \bar{a}_i$ such that $u_i(a_i, \bar{a}_{-i}) > u_i(\bar{a}_i, \bar{a}_{-i}) = \bar{u}_i$. The probability that $(a_i, \bar{a}_{-i})$ 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 action benchmark is a_i, $i's$ new payoff benchmark is $u_i(a_i, \bar{a}_{-i})$, 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 $(1-\varepsilon)^{n-1}\varepsilon/(|A_i|-1) \approx \varepsilon$, 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 seen, it is a *monotone increasing path* (with respect to the payoff benchmarks) in the sense that no one's payoff benchmark decreased and someone's strictly increased.

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, because $z^1 \in C^0$ and 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. As before, 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 is 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, an event that has probability $\approx \varepsilon^0$). 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 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 content 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 $\bar{a}^k, \bar{u}^k$ be the benchmark actions and payoffs in state z^k; these are aligned because $z^k \in C^o$ by construction. Let $\bar{a}^{k+1}, \bar{u}^{k+1}$ be the benchmarks in state z^{k+1}. Note that only $i's$ benchmark action and payoff changed between the two states (due to $i's$ successful experiment); 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 $\bar{a}^{k+1}$ are played again. In this case all the watchful agents experience another disappointing payoff and become discontent, while all the other agents become (or stay) 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 partially aligned in the sense that $u_j(\bar{a}^{k+1}) = \bar{u}_j^{\,k+1}$ 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 $(\bar{a}, \bar{u})$. By assumption G is interdependent, hence there exists an agent $j \notin D$ and an action-tuple a'_D such that $u_j(a'_D, \bar{a}_{N-D}) \neq u_j(\bar{a}_D, \bar{a}_{N-D}) = \bar{u}_j$. 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. $u_j(a'_D, \bar{a}_{N-D}) > u_j(\bar{a}_D, \bar{a}_{N-D})$.

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 $\bar{a}_D$, *all the while remaining discontent*. (In each of these periods the players not in D keep playing $\bar{a}_{N-D}$.) This initially raises j's expectations, which are later quashed (the 'goat effect' in reverse). The sequence of transitions and play realizations looks like this:

actions	$(a'_D, \bar{a}_{N-D})$		$(a'_D, \bar{a}_{N-D})$		$(\bar{a}_D, \bar{a}_{N-D})$		$(\bar{a}_D, \bar{a}_{N-D})$	
states	z^{k+2}	$\rightarrow$	z^{k+3}	$\rightarrow$	z^{k+4}	$\rightarrow$	z^{k+5}	$\rightarrow$ z^{k+6}
payoffs			$u_j \uparrow$		$\bar{u}_j \uparrow$		$u_j \downarrow$	
moods			*j hopeful*		*j content*		*j watchful*	*j discontent*

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 $\bar{a}_i$, which

has probability $1-\varepsilon$ if content and probability 1 if watchful or hopeful. These probabilities are 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. $u_j(a'_D, \bar{a}_{N-D}) < u_j(\bar{a}_D, \bar{a}_{N-D})$

In this case it suffices that everyone in D play a'_D and stay discontent, while the others play $\bar{a}_{N-D}$. This makes player j discontent in two steps.

Proceeding in this way, we conclude that there is an easy path from z^{k+2} to a state z^d in which *all* agents are discontent. Given any $e \in E^o$, the probability is at least $(\theta/m)^n$ that $z^d \to e$ 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 $z \notin E^o$, there exists an easy path to some state in E^o. This establishes claim 3.

Recall that, for any state z, $\rho(z)$ is defined to be the *resistance of a least-resistant tree rooted at* z. To establish theorem 1, it therefore suffices to show the following (see the discussion at the end of section 3).

Claim 4. $\forall z \notin E, \exists e \in E^o$ *such that* $\rho(e) < \rho(z)$.

Proof. Let z be in the recurrence class Z_j, and let $\mathcal{T}_z$ 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 $e \in E^o \subset E$. Denote this path by $z \to z^1 \to ... \to z^k = e$, and let $\mathcal{P}$ 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 $\mathcal{J}_z$.

In $\mathcal{J}_z$, 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 $\mathcal{P}$ to the tree $\mathcal{J}_z$; this creates some states with two exiting edges -- one from $\mathcal{P}$ and one from $\mathcal{J}_z$. For each such state (*except* e), remove the exiting edge that comes from $\mathcal{J}_z$. The resulting set of edges $\mathcal{S}$ has one more edge than does $\mathcal{J}_z$; in fact, every state (including e) now has exactly one exiting edge, so it is not a tree.

Let us now compare the total resistance, $r(\mathcal{S})$, summed over all the edges in $\mathcal{S}$, with the total resistance, $r(\mathcal{J}_z)$, summed over all the edges in $\mathcal{J}_z$. Since $\mathcal{P}$ 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 each edge from $\mathcal{P}$ that replaced an edge from $\mathcal{J}_z$ led to a decrease (or at least no increase) in the resistance, that is,

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

Furthermore, the "additional" edge $z \to z^1$ has resistance at most 1, since $\mathcal{P}$ is an easy path. It follows that

$$r(\mathcal{S}) \leq r(\mathcal{J}_z) + 1. \qquad (19)$$

Next let $e \to w^1 \to w^2 \to ... \to w^j$ be the unique path in $\mathcal{J}_z$ (and $\mathcal{S}$) leading from e toward z, where w^j is the *firs*t state on the path such that e and w^j do *not* have the same benchmarks. (There is such a state because e corresponds to an equilibrium and z does not.) 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 $\mathcal{S}$, 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 $\mathcal{T}_e$ that is rooted at e. (See figure 1 for an example.) By construction, each of the states $w^1, w^2, ...w^{j-1}$ has the same benchmarks as does e; they differ from e only in that some agents may not be content. Hence

$$r(w^1 \to e) = r(w^2 \to e) = ... = r(w^{j-1} \to e) = 0 . \qquad (22)$$

Combining (19)-(22) it follows that $r(\mathcal{T}_e) < r(\mathcal{T}_z)$ and hence that $\rho(e) < \rho(z)$. This completes the proof of claim 4 and thereby the proof of theorem 1.

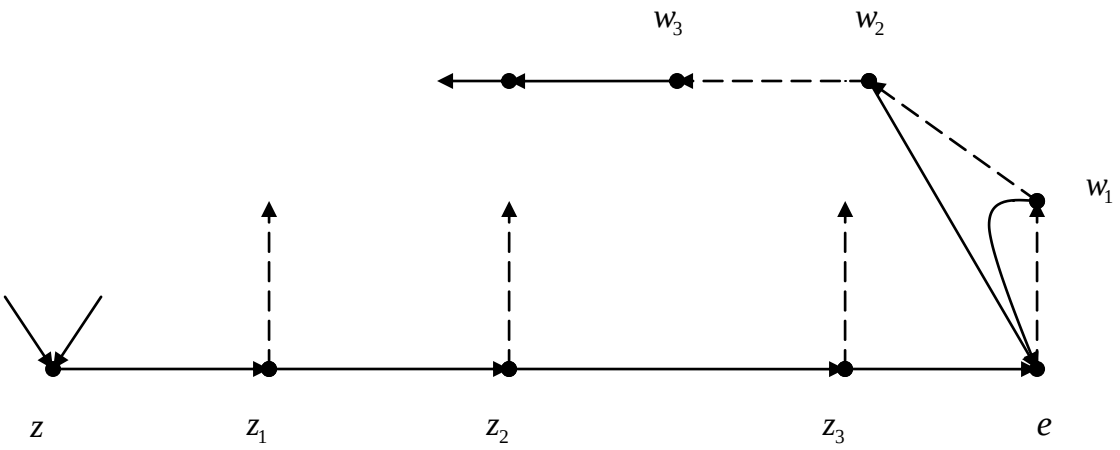


Figure 1. Construction of a tree rooted at e from a tree rooted at z by adding edges (solid) and subtracting edges (dashed).

5. Non-generic payoffs

The interdependence assumption is easy to state, but it is somewhat stronger than necessary. Consider, for example, an n-person game in which the players can be divided into disjoint groups such that the actions of any one group do not affect the payoffs of those outside the group, but the game is interdependent *within* each of these groups. (In effect the game decomposes into two disjoint interdependent games.) If the overall game has a pure equilibrium then so does each of the subgames, and interactive learning will discover it even though the game is not interdependent as a whole.

I shall not attempt to formulate the most general condition under which ITE learning discovers a pure Nash equilibrium; however, *some form of genericity* is needed when there are three or more players (though not when there are two players, as we shall see in theorem 2 below). 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.)

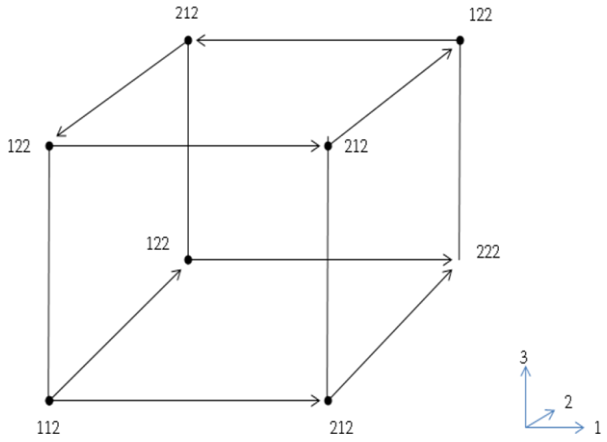


Figure 2. A three-person game with non-generic payoffs in which ITE learning does not necessarily lead to Nash equilibrium play. Arrows indicate best-response transitions.

Similar examples can be constructed when there are more than three players, but this is not the case when there are only two players.

Theorem 2. *Let G be a two-person game on a finite joint action space A that has at least one pure Nash equilibrium. If the players use ITE learning with experimentation*

probability ε *and response function* ϕ*, then for all sufficiently small* ε *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 $a^1 \to a^2 \to ... \to a^m$ 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 $a_{-i}^{k+1} = a_{-i}^k$, $a_i^{k+1} \neq a_i^k$, and a_i^{k+1} is a *strict best response* by i to a_{-i}^k . The sequence is a *best response cycle* if all the states are distinct 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.

Recall that E^o denotes the set of states such that everyone is content, the action benchmarks form a pure Nash equilibrium, and the payoff benchmarks are aligned with the action benchmarks. (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 (hence easy) path to some state $z^1 \in C^o$. If $z^1 \in E^o$ we are done. Otherwise it suffices to show that there exists an easy path from $z^1 \in C^o$ to some state in E^o. Let $(\bar{a}^1, \bar{u}^1)$ 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 $\bar{a}^1$ to some pure Nash equilibrium, say $\bar{a}^1 = a^1 \to a^2 \to ... \to a^m$.

Given such a path, let z^k be the *state* that has action benchmarks a^k, payoff benchmarks $u_i(a^k)$, and everyone is content. (Note that the action and payoff benchmarks are aligned.) We shall construct an easy path to $z^m \in E^o$ that mimics 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 with probability $\approx \varepsilon$. This leads to a state x^2 in which the action benchmarks are a^2 (though some players may be hopeful or watchful). In the next period they play a^2 with probability $\approx \varepsilon^0$. This results in a state y^2 where the watchful players in x^2 (if any) have become discontent, and the action benchmarks are still a^2. In the next period the probability is $\approx \varepsilon^0$ that a^2 is played again and everyone becomes content. (For this to happen each discontent player must play his part in a^2 and become content; by assumption this event has a probability that is bounded away from zero independently of ε.) This results in the all-content state z^2. Thus we have constructed an easy path of form $z^1 \to x^1 \to y^1 \to z^2$, where the first transition has resistance one, and the next two transitions have resistance zero. Repeating the argument we conclude that there is an easy path to the target state $z^m \in E^o$ that mimics the given best response path with the help of some intermediate transitions that have zero resistance.

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

Given that there is no best response path from $\bar{a}^1$ to a pure Nash equilibrium, there must exist a best response path from $\bar{a}^1$ that leads to a best response cycle. Denote such a cycle by $b^0 \to b^1 \to ... \to b^{m-1} \to b^0 ...$ and let the path to it be $\bar{a}^1 = a^1 \to a^2 \to ... \to a^j = b^0$. As in case 1 we can construct an easy path that mimics the best response path up to b^0; we need to show that it can be extended as an easy path to a Nash equilibrium.

Along the *b*-cycle the two players alternate in choosing best responses, say player 1 chooses a strict best response going from b^0 to b^1, player 2 from b^1 to b^2, and so forth, all indexes being modulo m. Since these are strict best responses and the process cycles, each player's payoff must at some stage *decrease*. Proceeding from b^0, let $b^k \to b^{k+1}$ be the *first* transition in 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 $b^k \to b^{k+1}$. Moreover, in the *preceding* transition, $b^{k-1} \to b^k$, player 2's payoff must *strictly increase* because the players alternate in making best responses. We therefore know that

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

We now consider two possibilities.

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

By assumption $b^k \to b^{k+1}$ was the first transition (starting from b^0) in which any decrease occurred, and by assumption it occurred for player 2. Hence $k = 0$ and the hypothesis of case 2a is that $u_1(b^0) < u_1(b^{m-1})$.

As in case 1 we can construct an easy path (in the full state space) that mimics the transitions along the path $\bar{a}^1 = a^1 \to a^2 \to ... \to a^j = b^0$ and then mimics the cycle from b^0 on. Consider the situation when this path *first returns* to b^0, that is, the players play b^0 again after having gone around the cycle once. By construction, the players were all content in the previous state and their benchmarks were aligned. In the transition to b^0, player 1's payoff decreases so he becomes watchful, while player 2's payoff increases so she remains content.

In the next period, the probability is $\approx \varepsilon^0$ that: player 1 plays his current action benchmark b_1^0 again and *becomes discontent*, while player 2 plays action b_2^0 again and remains content. In the next period after that, the probability is $\approx \varepsilon^0$ that player 1 chooses b_1^1 and *remains discontent*, while player 2 does not experiment, chooses $b_2^0 = b_2^1$ again, and remains content. (By assumption, player 1 changed action in the transition $b^0 \to b^1$, hence player 2 *did not* change action, that is, $b_2^1 = b_2^0$.) By (22), player 2's payoff decreases in this transition ($b^0 \to b^1$), so she is now watchful. In the period after that, with probability $\approx \varepsilon^0$ they play b^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 spontaneously become content (with aligned benchmarks), all with probability $\approx \varepsilon^0$. Thus in case 2a we have constructed an easy path to a state in E^o, that is, to an all-content, aligned Nash equilibrium state.

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

In this case let us first construct an easy path (in the full state space) that mimics the transitions along the path $\bar{a}^1 = a^1 \to a^2 \to ... \to a^j = b^0$, and then mimics the cycle up to the point where b^{k+1} is first played. (Recall that this is the first

transition on the cycle where someone's payoff decreases.) At this point player 2 becomes watchful while player 1 remains content. In the next period the probability is $\approx \varepsilon^0$ that b^{k+1} will be played again and that player 2 becomes discontent while player 1 remains content. In the next period after that, the probability is $\approx \varepsilon^0$ that player 2 plays b_2^{k-1} and remains discontent, while player 1 plays b_1^{k+1} . Denote the resulting pair of actions by $\tilde{b} = (b_1^{k+1}, b_2^{k-1})$. Again we may distinguish two cases.

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

In this case player 1 has become watchful in the transition to $\tilde{b}$ while player 2 is still discontent. Hence in one more period the probability is $\approx \varepsilon^0$ that $\tilde{b}$ will be played again and that both players will be discontent. As we have already shown, this leads in one more easy step to an all-content Nash equilibrium, and we are done. It therefore only remains to consider the following.

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

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 b^{k-1} to b^k, player 1 best responded in the previous move. It follows that b_1^{k-1} is 1's best response to b_2^{k-1}, from which we deduce that $u_1(b^{k-1}) \geq u_1(\tilde{b})$. Putting this together with the case 2b″ assumption we obtain

$$u_1(b^k) \geq u_1(b^{k-1}) \geq u_1(\tilde{b}) \geq u_1(b^{k+1}), \qquad (23)$$

which implies that $u_1(b^k) \geq u_1(b^{k+1})$, contrary to (22). This concludes the proof of theorem 2.

6. 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 –* ε *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 (due to sample variability) 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 and intuitive heuristic 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, although these have a fairly special structure. We conclude that there exist simple methods for learning equilibrium even when players know nothing about the structure of the game, who the other players are, or what strategies they are pursuing.

Acknowledgments. I am indebted to Jason Marden, Thomas Norman, Tim Salmon, and Christopher Wallace for helpful comments and suggestions.

References

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

Foster, D. P., Young, H. P., 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38 219-232.

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

Freidlin, M., Wentzell, A., 1984. Random Perturbations of Dynamical Systems. Berlin: Springer-Verlag.

Germano, F., Lugosi, G., 2007. Global convergence of Foster and Young's regret testing. *Games and Economic Behavior* 60, 135-154.

Hart, S., Mas-Colell, A., 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93, 1830-1836.

Hart, S., Mas-Colell, A., 2006. Stochastic uncoupled dynamics and Nash equilibrium. *Games and Economic Behavior* 57, 286-303.

Kandori, M., Mailath, G., Rob, R., 1993. Learning, mutation, and long-run equilibrium in games. *Econometrica* 61, 29-56.

Kirchsteiger, G., Rigotti, L., Rustichini, A., 2006. Your morals are your moods. *Journal of Economic Behavior and Organization* 59, 155-172.

Marden, J., Young, H. P., Arslan, G., Shamma, J. S., 2007. Payoff-based dynamics for multi-player weakly acyclic games. Working Paper, Department of Mechanical and Aerospace Engineering, UCLA.

Smith, K., Dickhaut, J., 2005. Economics and emotion: institutions matter. *Games and Economic Behavior* 52, 316-335.

Young, H. P., 1993. The evolution of conventions. *Econometrica,* 61, 57-84.