Learning by Trial and Error

H. Peyton Young

University of Oxford
The Brookings Institution

March, 2009

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. We introduce a learning rule in which behavior is conditional on whether a player experiences an error of the first or second type. This rule, called *interactive trial and error learning*, implements Nash equilibrium behavior in any game with generic payoffs and at least one pure Nash equilibrium.

JEL Classification: C72, D83

Keywords: learning, adaptive dynamics, 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 other 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 require any information about the opponents' actions or payoffs. Such a rule is said to be *completely uncoupled*.[1] This paper introduces a simple, completely uncoupled learning rule such that, when used by all players in a game, period-by-period play comes close to pure Nash equilibrium play a high proportion of the time, provided that the game has such an equilibrium and the payoffs are generic.

This rule, called *interactive trial and error learning*, has two key ingredients: i) players occasionally experiment with alternative strategies, keeping the new

[1] Foster and Young (2006) use the term *radically uncoupled*. A learning rule is *uncoupled* if it does not require information about the opponents' payoffs, though it may depend on their actions (Hart and Mas-Colell, 2003).

strategy if and only if it leads to a strict increase in payoff; ii) if someone experiences a payoff *decrease* due to a strategy change by *someone else*, he starts a random search for a new strategy, eventually settling on one with a probability that increases monotonically with its realized payoff. A novel feature of the process is that different search procedures are triggered by different psychological states or *moods*, where mood changes are induced by the relationship between a player's realized payoffs and his current payoff expectations.

2. Related literature

Before defining this procedure in detail, I shall briefly outline its relationship with other learning rules in the literature. Perhaps the closest is a recent proposal of Marden, Young, Arslan, and Shamma (2007). In this procedure, hereafter abbreviated MYAS, each player experiments with a small probability in each period, and adopts the experimental action if and only if it results in a higher payoff. It can be shown that in any potential game – in fact in any weakly acyclic game[2] -- this rule implements Nash equilibrium behavior in the sense that a pure Nash equilibrium will be played a high proportion of the time provided that the experimentation probability is sufficiently small. The principal difference between this approach and interactive trial and error learning is that the latter has an additional search phase that is triggered by decreases in payoff caused by someone else. This feature guarantees that in any finite game with generic payoffs and at least one pure Nash equilibrium, such an equilibrium will be played a high proportion of the time proved that the experimentation probability is sufficiently small.

[2] A game is *weakly acyclic* if for every joint action-tuple there exists a sequence of best replies -- one player moving at a time -- that ends at a pure Nash equilibrium (Young, 1993). Potential games and congestion games are special cases.

Another closely related learning rule is *regret testing* (Foster and Young, 2006). In this procedure a player computes his average per period payoff over a long sequence of plays, and compares this with the average payoff he receives from occasional (random) experiments with alternative strategies. If one of these alternative strategies generates a *sufficiently larger average payoff* than the average payoff from his current strategy, he chooses a new strategy *at random*. (In other words, the strategy with the higher payoff is not necessarily chosen; it merely signals to the player that he is not playing the right strategy.) Foster and Young show that, for all finite two-person games, this rule comes close to Nash equilibrium behavior a large proportion of the time. Subsequently, Germano and Lugosi (2007) showed that a slight variant of the procedure comes close to Nash equilibrium behavior in any finite n-person game with generic payoffs. Interactive trial and error learning differs from regret testing in that search is more directed and the rule requires no statistical estimation; however, it only leads to equilibrium behavior in games that have pure equilibria.

A third learning rule that bears some resemblance to the present proposal is due to Karandikar et al. (1998). In this procedure each player has an endogenously generated aspiration level that is based on a smoothed average of his prior payoffs. He changes strategy (with positive probability) when his current payoff falls below his current aspirations. This rule is simple, intuitive, and completely uncoupled. Unlike interactive trial and error learning there is no experimentation per se; rather, the aspiration levels are subjected to small random perturbations. These trembles occasionally cause the players to switch strategies even though the resulting payoffs are *lower* than before. The overall effect is that play transits among strategy-tuples in a way that depends on the rate at which aspirations are updated and also on the probability distribution of the trembles. Unlike the

present method, however, this procedure does not necessarily lead to Nash equilibrium behavior even in 2 x 2 games.

Another, more distantly related, family of learning rules are those based on regret minimization. In general, a player has *ex post regret* if he could have realized a higher average payoff by playing some strategy s' in all those periods when he in fact played s. There exist quite simple learning procedures that minimize ex post regret (Foster and Vohra, 1999; Hart and Mas-Colell, 2000, 2001); moreover they can be cast in a form that is completely uncoupled (Hart and Mas-Colell, 2000). However, unlike interactive trial and error learning, there is no guarantee that behaviors come close to Nash equilibrium most of the time. What can be shown is that regret minimizing rules cause the empirical frequency distribution of play to converge to the set of correlated equilibria (Hart and Mas-Colell, 2000; Foster and Vohra, 1999). This set includes the Nash equilibria as extreme points but is frequently much larger.

There are a number of learning rules that have a similar *stochastic structure* to the present proposal, in the sense that small trembles in behavior (or perceptions of others' behavior) cause play to shift among alternative strategy combinations. When these trembles are small the probability is high that play is concentrated on particular strategy combinations. The rule of Karandikar et al. has this structure, as do the model-based learning rules proposed by Jehiel (1998) and Foster and Young (2003). A key difference between the latter two approaches and the present one is that model-based learning requires *observability* of the opponents' play, whereas interactive trial and learning does not.

Before examining the properties of interactive trial and error learning in detail, a remark is in order about the sense in which it "implements" equilibrium behavior. We have repeatedly said that interactive trial and error learning cause

behaviors to come close to Nash equilibrium a high proportion of the time. Why not just say that behaviors *converge* to Nash equilibrium? Because typically they do *not* converge. In fact, there are very severe limits to what can be achieved if one insists on convergence to Nash equilibrium. To be specific, suppose that a learning rule has the following properties: i) it is uncoupled, ii) each player's choice of action depends solely on the frequency distribution of past play (as in fictitious play), and iii) each player's choice of action, conditional on the state, is deterministic. Hart and Mas-Colell (2003) show that for a large class of games, no such rule causes the players' period-by-period behavior to *converge* to Nash equilibrium.

Matters are not quite so bad when stochastic choice is allowed. In this case there exist simple, uncoupled rules that *converge almost surely* to pure Nash equilibrium behavior in games that have such an equilibrium (Hart and Mas-Colell, 2006; Babichenko, 2007).[3] The approach taken in the present paper shows that one can achieve something similar *even when learning is completely uncoupled*, provided that convergence is weakened to 'close most of the time.'

3. Interactive trial and error learning

Interactive trial and error learning is a modification of ordinary trial and error learning that takes into account the interactive nature of the learning environment. In ordinary trial and error learning, agents occasionally try out new things and accept them if 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

[3] Hart and Mas-Colell (2006) show that the following rule suffices: 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.

than it was before. The former are *active errors* whereas the latter are *passive errors*. We posit that the learning process is conditioned on whether an agent experiences an active or a passive error. Specifically, we propose that these situations trigger different psychological states or *moods*.

In the rule proposed here, an agent can have four different 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 thrashing 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. (The assumption of a one-period waiting time is purely for convenience; it can be any fixed number of periods.)

The proposition that an agent's behavior may be conditional on his emotional state has been examined in a number of experimental papers (Capra, 2004; Smith and Dickhaut, 2005; Kirchsteiger, Rigotti, and Rustichini, 2006). Here I employ the term 'mood' in a more abstract sense: it is simply a state variable that

determines how an agent responds to recent payoff history given the agent's current expectations.[5] The names of these states are meant to be suggestive but should not be taken too literally. In particular, I make no claim that people's search behavior actually does change in the manner prescribed by the rule (though it is certainly possible that different payoff histories induce different types of search).

Why does this process lead to equilibrium? The intuitive idea is that active search leads the players toward progressively higher payoffs and higher aspiration levels until one of two things happens: i) an equilibrium is reached, or ii) someone's aspirations are disappointed before an equilibrium is reached. In the latter case the disappointed player starts searching at random, which causes the other players to become disappointed with positive probability, which leads to a full-scale random search by everyone. This phase concludes when everyone calms down and they start building a new monotone-payoff path. It can be shown that, when the probability of calming down is sufficiently large relative to the probability of experimentation, the process is in a pure Nash equilibrium state much more often than in a disequilibrium state (assuming the game has a pure Nash equilibrium). Of course, there are many variants of the method proposed here that have similar properties, but it would take us too far afield to attempt to formulate the most general such method.

Let us now consider the model in greater 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 $z_i = (m_i, \bar{a}_i, \bar{u}_i)$, 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

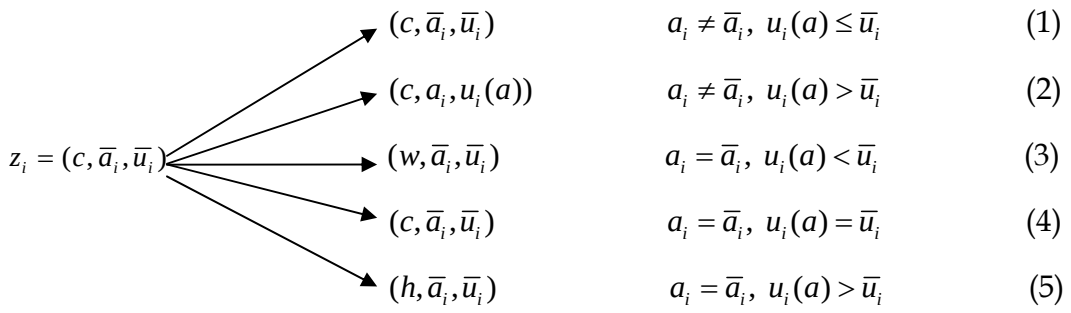
[5] See Compte and Postelwaite (2007) for another setting in which psychological states affect learning behavior.

(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). One way to do this is by an automaton diagram showing the various transitions among states, but this turns out to be somewhat cumbersome. Instead we shall first describe the nature of the shifts in qualitative terms, then give them in detail. Let us focus on a particular player who is currently content. With small probability he can move (in two periods) to a content state with a higher benchmark payoff, assuming such an improvement is possible. Alternatively, he could move (in two periods) to a discontent state, which happens when someone else changes strategy and makes his payoff go down. From a discontent state he eventually moves to a new content state and corresponding new benchmark payoffs; the key feature here is that each of the new content states has positive probability of occurring.

Next we give the transitions in detail. Fix a particular player *i*. There are four cases to consider, depending on the player's current mood.

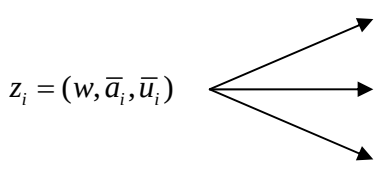
Content: $z_i = (c, \bar{a}_i, \bar{u}_i)$. Player i experiments next period with probability ε, and does not experiment with probability 1 – ε. Denote *i*'s choice of action next period by a_i. (Obviously a_i differs from $\bar{a}_i$ only if i experimented.) The possible transitions are:



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, where any change in i's payoff must be triggered by a change of strategy by someone else.

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:

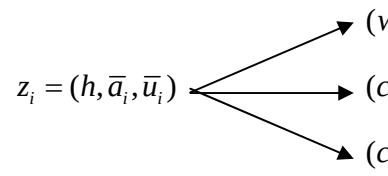


$$(d, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) < \bar{u}_i \tag{6}$$

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

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

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.



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

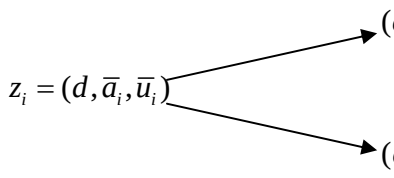
$$(c, \bar{a}_i, \bar{u}_i) \qquad a_i = \bar{a}_i,\ u_i(a) = \bar{u}_i \tag{10}$$

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

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.[7] 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 $\theta > 0$.[8] 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.

[7] 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.

[8] 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.



$$\text{with prob } \phi(u_i(a), \bar{u}_i) \tag{12}$$

$$\text{with prob } 1-\phi(u_i(a), \bar{u}_i) \tag{13}$$

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.[9]

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,

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

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

[9] 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.

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. Given $\delta > 0$, if the players use ITE learning with response function ϕ and sufficiently small experimentation probability ε, then a pure Nash equilibrium is played at least $1 - \delta$ of the time.* .

The assumption of interdependence is not needed when there are only two players, as we shall show later on in theorem 2, but when there are more than two players some form of non-genericity is needed, as we shall show by example in section 5.

4. Proof of theorem 1: preliminaries

Before formally proving theorem 1 let us briefly outline the argument. We begin by observing that states in which someone is not content are inherently unstable: any given player will leaves a discontent, hopeful, or watchful state and enter a content state with a probability that is high relative to the experimentation probability ε. Next suppose that the process is in an all-content state but that the benchmark actions do not constitute a Nash equilibrium. Then 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. If, however, the process is in an all-content state in which the benchmark actions *do* constitute a Nash equilibrium, then it takes at least two experiments (together or in close succession) to cause the benchmarks to change. In other words, the process either transits temporarily to a state with the same benchmarks and quickly reverts to the equilibrium state, or it transits to a state with new benchmarks, where the latter case has 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).[10] 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 $\overline{a}$, and suppose that the vector a is realized next period, resulting in the state z'. If

[10] 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.

$a \neq \bar{a}$, 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. 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.

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 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.. 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. The intuitive idea of the proof is as follows. One by one the players experiment and find actions that yield payoff improvements and correspondingly higher aspirations (benchmark payoffs). This process continues until a state in E^o is reached, or some player's aspirations are disappointed. This triggers a sequence in which one player becomes discontent and his flailing around causes all the others to become discontent (this is the step where the interdependence property is invoked). Then with positive probability they simultaneously jump to a Nash equilibrium and become content again. It can be shown that all of these transitions occur with least resistance, that is, they generate easy paths to E^o.

We now give the argument in detail. 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 action benchmark is a_i, $i's$ new payoff benchmark is $u_i(a_i, \overline{a}_{-i})$, 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 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			$u_j \uparrow$		$\overline{u}_j \uparrow$		$u_j \downarrow$	
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 $\mathcal{T}_z$.

In $\mathcal{T}_z$, 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 $\mathcal{P}$ to the tree $\mathcal{T}_z$; this creates some states with two exiting edges -- one from $\mathcal{P}$ and one from $\mathcal{T}_z$. For each such state (*except* e), remove the exiting edge that comes from $\mathcal{T}_z$. The resulting set of edges $\mathcal{S}$ has one more edge than does $\mathcal{T}_z$; 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, $r(\mathcal{T}_z)$, summed over all the edges in $\mathcal{T}_z$. 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 $\mathcal{T}_z$ led to a decrease (or at least no increase) in the resistance, that is,

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

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

$$r(\mathcal{S}) \le r(\mathcal{T}_z) + 1. \tag{19}$$

Next let $e \to w^1 \to w^2 \to \ldots \to w^j$ be the unique path in $\mathcal{T}_z$ (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

$$r(e \to w^1) + r(w^1 \to w^2) + \ldots + r(w^{j-1} \to w^j) \ge 2. \tag{20}$$

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

$$w^1 \to e, w^2 \to e, \ldots, w^{j-1} \to e. \tag{21}$$

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 $w^1, w^2, \ldots w^{j-1}$ has the

same benchmarks as does e; they differ from e only in that some agents may not be content. Hence

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

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.

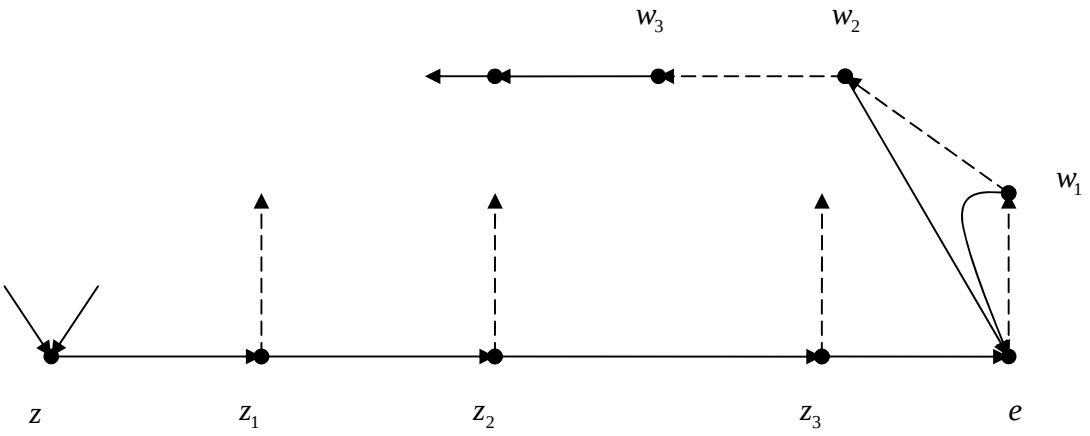


Figure 1. Construction of a tree rooted at e from a tree rooted at z by adding edges (solid) and subtracting edges (dashed).

6. 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. Note that player 3's payoffs remain unchanged no matter what the other players do.

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. *Hence, once player 3 begins in a content state, she remains content and never changes her benchmark 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, so they keep moving around in a best-response cycle. It follows that there are 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.)

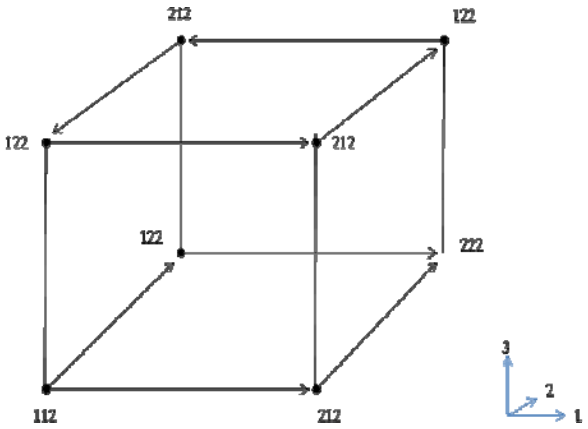


Figure 2. A three-person game with non-generic payoffs in which ITE learning does not necessarily lead to Nash equilibrium play. Each triple represents the payoffs to players one, two, and three respectively. Arrows indicate best-response transitions.

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. Given $\delta > 0$, if the players use ITE learning with response function ϕ and sufficiently small experimentation probability ε, then a pure Nash equilibrium is played at least $1 - \delta$ 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 $a^1 \to a^2 \to \ldots \to a^m$ 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. The key idea is that if the players do not continue up a payoff-monotone path, then they fall into a cycle that causes both to become discontent. (When there are more than two players interdependence assures the existence of such a cycle.)

As before, let E^o denote 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 both players are 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, suppose the relevant player experiments and chooses a best reply to the opponent's current action,

which the opponent continues to play next period. , In other words, the action-tuple a^2 is played next period with probability $\approx \varepsilon$. With probability $\approx \varepsilon^0$ the action-tuple a^2 is played in each of the next two periods and both players become content. Call this state z^2. In this manner we construct an easy path to the 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 $\overline{a}^1 = a^1 \to a^2 \to ... \to a^j = b^0$ 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

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

which implies that $u_1(b^k) \geq u_1(b^{k+1})$, 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 an arbitrarily high proportion of the time when* 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.

One of the issues that we have not dealt with is how long it takes (in expectation) for the learning process to reach an equilibrium from an arbitrary initial state. The proof of theorems 1 and 2, which relies on the theory of large deviations in Markov chains, is not very informative on this point. One can compute a rough upper bound on the waiting time by observing that from any state there exists a sequence of transitions, each having probability at most ε, such that the sequence either ends at an equilibrium, or in an all-discontent state from which the process

jumps to an equilibrium with probability at most ε^n, n being the number of players. To estimate the expected waiting time more precisely requires knowing how long these sequences are, which depends on the payoff structure of the game. This poses an interesting open problem that we shall not pursue 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, Christopher Wallace, the referees, and the Associate Editor for helpful comments and suggestions.

References

Babichenko, Y. 2007. Uncoupled automata and pure Nash equilibria. Discussion Paper #459. Center for the Study of Rationality, Hebrew University.

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

Compte, O., Postelwaite, A. 2007. Repeated games and limited information processing. Mimeo, University of Pennsylvania.

Foster, D. P., and R. Vohra, 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29, 7-35.

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

Foster, D.P., Young, H.P. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45, 73-96.

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. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 1127-50.

Hart, S., Mas-Colell, A. 2001. A general class of adaptive strategies. *Journal of Economic Theory* 98, 26-54.

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.

Jehiel, P. 1998. Learning to play limited forecast equilibria. *Games and Economic Behavior* 22,274-98.

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

Karandikar, R.,Mookherjee, D., Ray, D., Vega-Redondo, F. 1998. Evolving aspirations and cooperation. *Journal of Economic Theory* 80, 292-331.

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.