

ISSN 1471-0498



**DEPARTMENT OF ECONOMICS
DISCUSSION PAPER SERIES**

**LEARNING EFFICIENT NASH EQUILIBRIA IN
DISTRIBUTED SYSTEMS**

Bary S.R. Pradelski and H. Peyton Young

Number 480
February 2010

Manor Road Building, Oxford OX1 3UQ

Learning Efficient Nash Equilibria in Distributed Systems

Bary S. R. Pradelski and H. Peyton Young*

University of Oxford

June 23, 2011

Abstract. An individual's learning rule is *completely uncoupled* if it does not depend directly on the actions or payoffs of anyone else. We propose a variant of log linear learning that is completely uncoupled and that selects an efficient (welfare-maximizing) pure Nash equilibrium in all generic n -person games that possess at least one pure Nash equilibrium. In games that do not have such an equilibrium, there is a simple formula that expresses the long-run probability of the various disequilibrium states in terms of two factors: i) the sum of payoffs over *all* agents, and ii) the maximum payoff gain that results from a unilateral deviation by *some* agent. This *welfare/stability trade-off criterion* provides a novel framework for analyzing the selection of disequilibrium as well as equilibrium states in n -person games.

JEL: C72, C73

* Correspondence: H. P. Young, Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom. Tel 44 1865 271086. Fax 44 1865 271094
peyton.young@economics.ox.ac.uk

1. Learning equilibrium in complex interactive systems

Game theory has traditionally focussed on situations that involve a small number of players. In these environments it makes sense to assume that players know the structure of the game and can predict the strategic behavior of their opponents. But there are many situations involving huge numbers of players where these assumptions are not particularly persuasive. Commuters in city traffic are engaged in a game because each person's choice of route affects the driving time of many other drivers, yet it is doubtful that anyone 'knows the game' or fully takes into account the strategies of the other players as is usually posited in game theory. Other examples include decentralized procedures for routing data on the internet, and the design of information sharing protocols for distributed sensors that are attempting to locate a target.

These types of games pose novel and challenging questions. Can such systems equilibrate even though agents are unaware of the strategies and behaviors of most (or perhaps all) of the other agents? What kinds of adaptive learning rules make sense in such environments? How long does it take to reach equilibrium assuming it can be reached at all? And what can be said about the welfare properties of the equilibria that result from particular learning rules?

In the last few years the study of these issues has been developing rapidly among computer scientists and distributed control theorists (Papadimitriou, 2001; Roughgarden, 2005; Mannor and Shamma, 2007; Marden and Shamma, 2008; Marden, Arslan, and Shamma, 2009; Marden et al., 2009; Asadpour and Saberi, 2009; Shah and Shin, 2010). Concurrently game theorists have been investigating the question of whether decentralized rules can be devised that converge to Nash equilibrium (or correlated equilibrium) in general n -person games (Hart and Mas-Colell, 2003, 2006; Foster and Young, 2003, 2006; Young, 2009; Hart and Mansour, 2010). A related question is whether decentralized learning procedures can be devised that optimize some overall measure of performance or welfare without necessarily inducing equilibrium (Arieli and Babichenko, 2011; Marden, Young, and Pao, 2011). This is particularly relevant to problems of distributed control, where measures of system performance are given (e.g., the total power generated by a windfarm, the speed of data transmission in a communications network), and the aim is to design local response functions for the components that achieve maximum overall performance.

Much of the recent research on these topics has focussed on potential games, which arise frequently in applications (Marden and Shamma, 2008; Marden, Arslan, and Shamma, 2009). For this class of games there exist extremely simple and intuitively appealing learning procedures that cause the system to equilibrate from any initial conditions. A notable example is logit learning, in which an agent chooses actions with log probabilities that are a linear function of their payoffs. In this case equilibrium occurs at a local or global maximum of the potential function. However, the potential function need not measure the overall welfare of the agents, hence the equilibrium selected may be quite inefficient. This is a well-known problem in congestion games for example. The problem of inefficient equilibrium selection can be overcome by a congestion pricing scheme, but this requires some type of centralized (or at least not fully decentralized) mechanism for determining the price to charge on each route (Sandholm, 1998).

The contribution of this paper is to demonstrate a simple learning rule that incorporates log linear learning as one component, and that selects an efficient equilibrium in any game with generic payoffs that possesses at least one pure Nash equilibrium. (An equilibrium is *efficient* if there is no other equilibrium in which someone is better off and no one is worse off.) By ‘select’ we mean that, starting from arbitrary initial conditions, the process is in an efficient equilibrium in a high proportion of all time periods. Our learning rule is *completely uncoupled*, that is, the updating procedure does not depend on the actions or payoffs of anyone else. Thus it can be implemented even in environments where players know nothing about the game, or even whether they are in a game. All they do is react to the pattern of recent payoffs, much as in reinforcement learning (though the rule differs in certain key respects from reinforcement learning).

Our notion of selection – in equilibrium a high proportion of the time – is crucial for this result. It is not true that the process converges to equilibrium or even that it converges to equilibrium with high probability. Indeed it can be shown that, for general n -person games, there exist no completely uncoupled rules with finite memory that select a Nash equilibrium in this stronger sense (Babichenko, 2010a; see also Hart and Mas-Colell, 2003, 2006).

The learning rule that we propose has a similar architecture to the trial and error learning procedure of Young (2009), and is more distantly related to the ‘learning by sampling’

procedure of Foster and Young (2006) and Germano and Lugosi (2007).¹ An essential feature of these rules is that players have two different search modes: i) deliberate experimentation, which occurs with low probability and leads to a change of strategy only if it results in a higher payoff than the current aspiration level; ii) random search, which leads to a change of strategy that may or may not have a higher payoff. Young (2009) demonstrates a procedure of this type that selects pure Nash equilibria in games where such equilibria exist and payoffs are generic. However this approach does not provide a basis for discriminating between pure equilibria, nor does it characterize the states that are selected when such equilibria do not exist.

In contrast to these earlier papers, the learning rule described here permits a sharp characterization of the equilibrium and disequilibrium states that are favored in the long run. This results from several key features that distinguish our approach from previous ones, including Young (2009). First, we do not assume that agents invariably accept the outcome of an experiment even when it results in a strict payoff improvement: acceptance is probabilistic and is merely increasing in the size of the improvement. Second, players accept the outcome of a random search with a probability that is increasing in its realized level of payoff rather than the gain in payoff. Third, the acceptance functions are assumed to have a log linear format as in Blume (1993, 1995). These assumptions define a learning process that selects efficient pure Nash equilibria whenever pure Nash equilibria exist. Moreover when such equilibria do not exist we obtain a precise characterization of the *disequilibrium* states that have high probability in the long run. These states represent a trade-off between welfare and stability: the most likely disequilibrium states are those that maximize a linear combination of: i) the total welfare (sum of payoffs) across *all* agents and ii) the payoff gain that would result from a deviation by *some* agent, where the first is weighted positively and the second negatively.

¹ Another distant relative is the aspiration-based learning model of 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 if his current payoff falls below his current aspirations. Unlike the present method, this procedure does not necessarily lead to Nash equilibrium behavior even in 2 x 2 games.

2. The learning model

We shall first describe the learning rule informally in order to highlight some of its qualitative features. At any given point in time an agent may be searching in one of two ways depending on his internal state or ‘mood’. In the *content* state an agent occasionally experiments with new strategies, and adopts the new strategy with a probability that increases with the associated gain in payoff. (This is the conventional exploration/exploitation form of search.) In the *discontent* state an agent flails around, trying out randomly chosen strategies every period. The search ends when he spontaneously accepts the strategy he is currently using, where the probability of acceptance is an increasing function of its realized payoff. The key differences between these modes of search are: i) the rate of search (slow for a content agent, fast for a discontent agent); and ii) the probability of accepting the outcome of the search. In the content state the probability of acceptance is determined by the *change* in payoff, whereas in the discontent state the probability of acceptance is determined by the *level* of payoff. The rationale for the latter assumption is that a discontent agent will typically try out many different strategies before settling on any one of them. Over time his previous benchmark payoff fades in importance and his current payoff becomes more salient in deciding whether to accept or reject his current strategy.² This stands in contrast with the more deliberative behaviour of a content agent, who compares the payoff from a single experiment to the benchmark payoff in the immediately preceding period.

A second key feature of the learning process is the mechanism that triggers transitions between content and discontent states. A transition from content (*c*) to discontent (*d*) occurs when an agent’s realized payoff goes down for two periods in succession and he does not experiment during those periods. In other words, a $c \rightarrow d$ transition is triggered by a drop in payoff that was not instigated by the agent, but results from a change of action by someone else. By contrast, a $d \rightarrow c$ transition occurs when an agent tires of searching and accepts his current strategy as ‘good enough.’

To illustrate these ideas in an everyday situation, consider a commuter who ordinarily takes a particular route to work. A ‘content’ commuter will occasionally experiment with new routes just to see if there might be something better out there. If one of these experiments results in

² The algorithm could be modified so that an agent compares his current payoff to a discounted version of his old benchmark. We conjecture that this would lead to similar long-run behavior but it would also significantly complicate the analysis.

lower travel time, he adopts the new route with a probability that is increasing with the gain in payoff (the decrease in travel time), and otherwise he returns to his usual route. However, if the travel time on his usual route worsens over several consecutive days, this will prompt him to become discontent and to start actively looking for a new route. This search eventually comes to an end when he adopts the route he is currently using, where the probability of acceptance increases the more desirable the route is.

It is worth pointing out that these acceptance probabilities depend on cardinal changes in payoffs, and therefore go beyond the standard assumptions of von Neumann Morgenstern utility theory. In particular, the learning model is not invariant to changes of scale or changes of origin in the agent's utility function. The larger the payoff gain from an experiment, the higher the probability that it will be accepted; similarly, the larger the current payoff level from a random search, the higher the probability that the current action will be accepted (and the benchmarks adjusted accordingly). Since agents' payoffs determine their probabilities of making various choices, it should not be surprising that the probability of different states of the system depends on their welfare properties. In other words the learning process implicitly makes interpersonal comparisons of welfare among the agents. Our main theorem will exhibit the precise sense in which the long-run probability of the states is related to both their welfare and their stability.

We now describe the learning algorithm in detail. Let \mathcal{G} be an n -person game on a finite joint action space $A = \prod_i A_i$ and let $u_i(\cdot)$ be i 's utility function. At each point in time the *state* of each player i is a triple $z_i = (m_i, \bar{a}_i, \bar{u}_i)$ consisting of

- i) a mood m_i : a state variable that determines his method of search;
- ii) a benchmark action \bar{a}_i : the action he would take when not searching;
- iii) a benchmark payoff \bar{u}_i : the payoff he expects to get, i.e., his aspiration level.

Each agent has four moods: content (c), discontent (d), watchful (c^-), and hopeful (c^+). We shall assume that each benchmark payoff \bar{u}_i corresponds to a feasible payoff in the game, that is, $\bar{u}_i = u_i(a)$ for some $a \in A$. Since A is finite, it follows that the state space is also finite.

The learning process evolves in discrete time periods $t = 1, 2, 3, \dots$, where the game is played once per period. The agents' exploration rates are governed by a parameter $\varepsilon \in (0, 1)$ that determines the overall level of *noise* in the system, and two real-valued functions $F(u)$ and $G(\Delta u)$, where $\varepsilon^{F(u)}$ is the probability of accepting the outcome of a discontent search when the current payoff is u , and $\varepsilon^{G(\Delta u)}$ is the probability of accepting the outcome of a content search when the payoff gain is $\Delta u > 0$. We shall assume that F and G are *strictly monotone decreasing*, that is, the probability of acceptance is strictly increasing in u and Δu respectively. For simplicity of exposition we shall also assume that ε , F , and G are common across all agents. In section 7 we shall show how to extend our results to the case where agents have heterogeneous learning parameters, including different rates of experimentation.

Fix a particular agent i . Conditional on i being in state z_i at the start of a period, the following rules determine i 's state at the end of the period. (The process is time homogeneous, so there is no need to designate the period in question.)

Content: $z_i = (c, \bar{a}_i, \bar{u}_i)$. At the start of the period i experiments with probability ε and does not experiment with probability $1 - \varepsilon$. An *experiment* involves playing an action $a'_i \neq \bar{a}_i$, where the choice of a'_i is determined by a uniform random draw from the set $A_i - \{\bar{a}_i\}$. Let u'_i be i 's realized payoff during the period whether he experiments or not.

- If i experiments and $u'_i > \bar{u}_i$, i transits to the state (c, a'_i, \bar{u}'_i) with probability $\varepsilon^{G(u'_i - \bar{u}_i)}$, and remains in state $(c, \bar{a}_i, \bar{u}_i)$ with probability $1 - \varepsilon^{G(u'_i - \bar{u}_i)}$.
- If i experiments and $u'_i \leq \bar{u}_i$, i 's state remains unchanged.
- If i does not experiment and $u'_i > \bar{u}_i$, i transits to the state $(c^+, \bar{a}_i, \bar{u}_i)$.
- If i does not experiment and $u'_i < \bar{u}_i$, i transits to the state $(c^-, \bar{a}_i, \bar{u}_i)$.
- If i does not experiment and $u'_i = \bar{u}_i$, i 's state remains unchanged.

Thus if i experiments and the result is a *gain* in payoff, he accepts the outcome (and adopts the corresponding new benchmarks) with a probability that increases with the size of the gain. If i does not experiment and his payoff changes due to the actions of others, he enters a transitional state and keeps his previous benchmarks. The notion here is that it would be premature to change benchmarks based on the experience from a single period, but if the directional change in payoff persists for another period, then i transits to a discontent state or a new content state, as described below. (It is straightforward to modify the algorithm so that an agent waits for k periods before transiting to a new content or discontent state.)

Watchful: $(c^-, \bar{a}_i, \bar{u}_i)$. Agent i plays \bar{a}_i and receives payoff u'_i .

- If $u'_i < \bar{u}_i$, i transits to the state $(d, \bar{a}_i, \bar{u}_i)$.
- If $u'_i > \bar{u}_i$, i transits to the state $(c^+, \bar{a}_i, \bar{u}_i)$
- If $u'_i = \bar{u}_i$, i transits to the state $(c, \bar{a}_i, \bar{u}_i)$

Hopeful: $(c^+, \bar{a}_i, \bar{u}_i)$. Agent i plays \bar{a}_i and receives payoff u'_i .

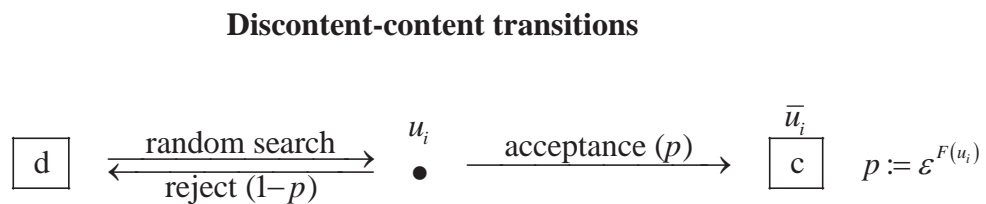
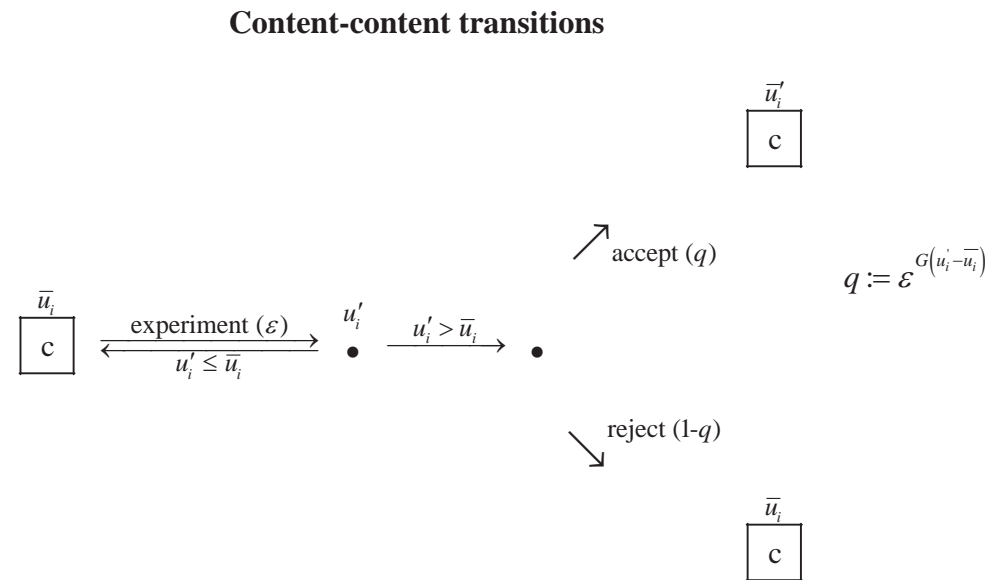
- If $u'_i < \bar{u}_i$, i transits to the state $(c^-, \bar{a}_i, \bar{u}_i)$
- If $u'_i > \bar{u}_i$, i transits to the state $(c, \bar{a}_i, \bar{u}_i)$
- If $u'_i = \bar{u}_i$, i transits to the state $(c, \bar{a}_i, \bar{u}_i)$

The logic of these intermediate states is that if i 's realized payoff *confirms* his current mood, then i becomes discontent (from the watchful state) or content (from the hopeful state). If i 's realized payoff does *not* confirm his current mood, i switches to the opposite intermediate state, or back to the content state if the realized payoff equals his benchmark payoff.

Discontent: $(d, \bar{a}_i, \bar{u}_i)$. Agent i plays an action a'_i drawn uniformly at random from A_i and receives the payoff u'_i .

- i transits to the state (c, a'_i, \bar{u}'_i) with probability $\varepsilon^{F(u'_i)}$; otherwise i remains in the state $(d, \bar{a}_i, \bar{u}_i)$.

Figure 1 illustrates the various transitions diagrammatically.



Intermediate transitions

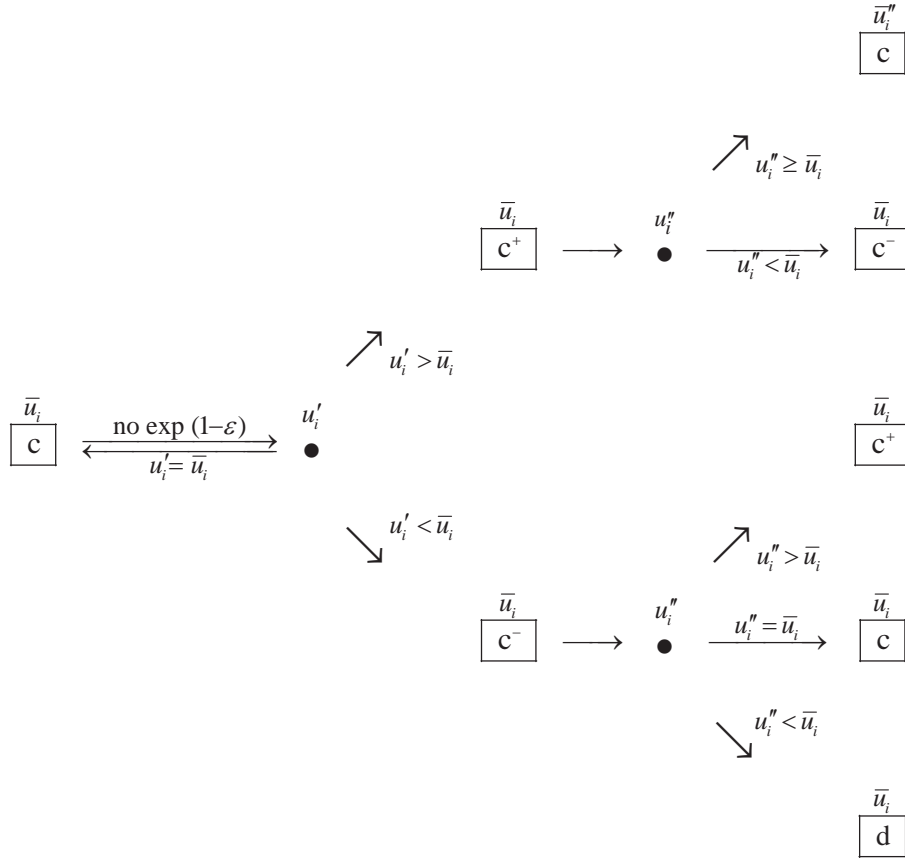


Figure 1. The structure of transitions for a given player i . Payoff benchmarks are above the boxes and current period payoffs are above the dots.

While the details of the learning algorithm are involved, the overall dynamics are qualitatively quite simple. The process oscillates between phases in which search is slow and search is fast.

1. When all players are content they occasionally experiment and search is slow. If the experiments lead to higher payoffs for some and no decrease for anyone, the players become content again with higher payoff benchmarks. This is a hill-climbing phase in which payoffs are monotonically increasing.
2. A sequence of experiments by content players may result in someone's payoff going below his benchmark for two periods in a row. This player becomes discontent and starts searching

every period. With some probability his searching causes other players to become discontent and they start searching. In effect, discontent search by one individual spreads through the population by contagion, creating high variance and leading to fast, wide-area search. This phase ends when the discontent players spontaneously settle down and become content again.

In section 5 we shall walk through an example step-by-step to show how the transition probabilities are computed in a concrete case. Before turning to these calculations, however, we shall discuss how this approach relates to earlier work and then state our main theorem.

3. Discussion

Learning rules that employ fast and slow search have been proposed in a variety of settings. In computer science, for example, there is a procedure known as WoLF (Win or Lose Fast) that has a qualitative flavor similar to the rule proposed above (Bowling and Veloso, 2002). The basic idea is that when a player's payoff is high relative to realized average payoff he updates his strategy slowly, whereas he updates quickly if his payoff is low compared to the realized average. This approach is known to converge to Nash equilibrium in 2×2 games but not in general (Bowling and Veloso, 2002). Similar ideas have been used to model animal foraging behavior (Houston, Kacelnik, and McNamara, 1982; Motro and Shmida, 1995; Thuijsman, Peleg, Amitai, and Shmida, 1995). For example, bees tend to search in the neighborhood of the last visited flower as long as the nectar yield is high, and search widely for an alternative patch otherwise (this is known as a *near-far strategy*). It would not be surprising if human subjects behaved in a similar fashion in complex learning environments, but to our knowledge this proposition has not been tested empirically.

In any event, we make no claim that the rule we have described is accurate from an empirical standpoint. The contribution of the present paper is to demonstrate the existence of simple, completely uncoupled rules that select efficient equilibria in a wide class of games. This addresses an important problem in the application of game theory to distributed control, where the object is to guarantee good system-wide performance using completely decentralized adaptive procedures. The issue of descriptive accuracy does not arise here, because one can build the learning rule into the agents by design. What matters is that the rule is simple to execute and requires little information. In our procedure it suffices to keep track

of just three items – mood, benchmark action, and benchmark payoff – and to compare these with the received payoff each period.

4. Statement of the main result

To state our results in the most transparent form we shall impose certain restrictions on the acceptance functions F and G . In particular we shall assume that they are *strictly decreasing linear functions*

$$F(u) = -\varphi_1 \cdot u + \varphi_2, \text{ where } \varphi_1 > 0, \quad (1)$$

$$G(\Delta u) = -\gamma_1 \cdot \Delta u + \gamma_2, \text{ where } \gamma_1 > 0. \quad (2)$$

The coefficients φ_i, γ_i are chosen so that $F(u)$ and $G(\Delta u)$ are *strictly positive* for all $u \in U$ and $\Delta u \in \Delta U$, and are “small” in the following sense:

$$0 < G(\Delta u) < 1/2 \text{ and } 0 < F(u) < 1/2n. \quad (3)$$

This condition implies that the acceptance probabilities are fairly large relative to the probability of conducting an experiment. In particular the probability of accepting the outcome of an experiment ($\varepsilon^{G(\Delta u)}$) is always greater than $\sqrt{\varepsilon}$, and the probability of accepting the outcome of a random search is greater still (see figure 2). We do not claim that these bounds are best possible, but they are easy to work with and yield sharp analytical predictions. The assumption of linearity will be useful for stating our first theorem, but in fact a more general equilibrium selection result holds even when the functions are nonlinear and differ among agents, as we shall show in section 7.

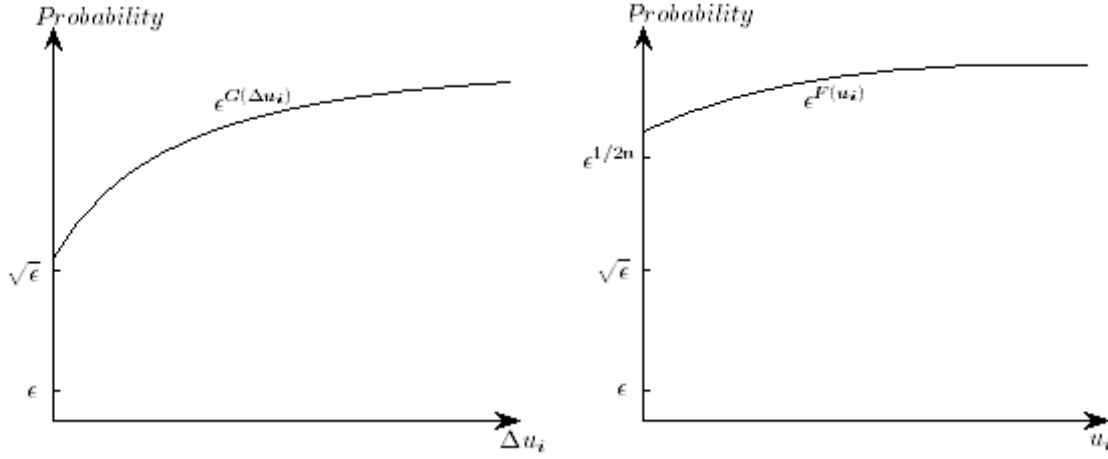


Figure 2. Acceptance probabilities for content agents (left panel) and discontent agents (right panel).

The learning rule described above is a variant of log linear learning that we shall call *log linear trial and error learning*. To see the connection, let $\varepsilon = e^{-\beta}$, where $\beta > 0$. Then the log probability of accepting the outcome of an experiment is

$$\ln P(\text{accept experiment with payoff gain } \Delta u > 0) = -\beta \ln \varepsilon(\gamma_1 \Delta u - \beta \gamma_2). \quad (4)$$

Similarly, the log probability of accepting the outcome of a random search is

$$\ln P(\text{accept search with payoff } u) = -\beta \ln \varepsilon(\varphi_1 u - \beta \varphi_2). \quad (5)$$

Notice that the probability of acceptance depends on the *cardinality* of the payoff gain or payoff level respectively. Hence the players' utilities cannot be rescaled or translated at will, as is the case with von Neumann Morgenstern utility functions. The larger the payoff gain from an experiment, the higher the probability that it will be accepted. Similarly, the larger the payoff level that results from a random search, the higher the probability that the current action will be accepted.

Our equilibrium selection result will be framed in terms of two quantities: the welfare of a state and the stability of a state.

Welfare. The *welfare* of state $z = (m, \bar{a}, \bar{u})$ is the sum of the players' payoffs from their benchmark actions:

$$W(z) = \sum_{i=1}^n u_i(\bar{a}). \quad (6)$$

Stability. An action-tuple $\bar{a} \in A$ is a δ -*equilibrium* for some $\delta \geq 0$ if

$$\forall i, \forall a_i \in A_i, u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a}) \leq \delta. \quad (7)$$

The *stability* of state $z = (m, \bar{a}, \bar{u})$ is the minimum $\delta \geq 0$ such that \bar{a} is a δ -*equilibrium*:

$$S(z) = \min \{ \delta : \text{the benchmark actions constitute a } \delta\text{-equilibrium} \}. \quad (8)$$

Note that the larger δ is, the less stable the state is, i.e., the greater is the incentive for some player to deviate. The following concepts will be needed to state our main result.

Stochastic stability. The set of *stochastically stable states* Z^* of the process $z(t)$ is the minimal subset of states such that, given any small $\alpha > 0$, there is a number $\varepsilon_\alpha > 0$ such that whenever $0 < \varepsilon \leq \varepsilon_\alpha$, $z(t) \in Z^*$ for at least $1 - \alpha$ of all periods t .

Interdependence. An n -person game \mathcal{G} on the finite action space A is *interdependent* if, for every $a \in A$ and every proper subset of players $\emptyset \subset J \subset N$, there exists some player $i \notin J$ and a choice of actions a'_J such that $u_i(a'_J, a_{N-J}) \neq u_i(a_J, a_{N-J})$.

In other words, given any current choice of actions $a \in A$, any proper subset of players J can cause a payoff change for some player not in J by a suitable change in their actions. Note that if a game has generic payoffs (and therefore no payoff ties), interdependence is automatically satisfied. However it is a much weaker condition. Consider, for example, a traffic game in which agents are free to choose any route they wish. There are many payoff ties because a local change of route by one player does not change the payoffs of players who are using completely different routes. But it satisfies the interdependence condition because a

given player, or set of players, can (if they like) switch to a route that is being used by another player and thereby change his payoff.

Aligned. The benchmarks in state $z = (m, \bar{a}, \bar{u})$ are *aligned* if the benchmark payoffs result from playing the benchmark actions, that is, if $\bar{u}_i = u_i(\bar{a})$ for every player i .

Equilibrium state. A state z is an *equilibrium state* if all players are content, their benchmark actions constitute a Nash equilibrium, and their benchmark payoffs are aligned with their benchmark actions.

Theorem 1. *Let \mathcal{G} be an interdependent n -person game on a finite joint action space A . Suppose that all players use log linear trial and error learning with experimentation probability ε and acceptance functions F and G satisfying conditions (1)-(3).*

i) If \mathcal{G} has at least one pure Nash equilibrium, then every stochastically stable state is an equilibrium state that maximizes $W(z)$ among all equilibrium states;

ii) If \mathcal{G} has no pure Nash equilibrium, every stochastically stable state maximizes

$$\varphi_1 W(z) - \gamma_1 S(z). \tag{9}$$

Notice that this result involves a cardinal comparison of the players' utilities. Players with large absolute utility are weighted heavily in the welfare function, and players with large utility differences are weighted heavily in the stability function. If a player's utility function were to be scaled up, he would count more heavily in both the welfare and the stability function. This would improve the likelihood of states in which this player has a high payoff, and decrease the likelihood of states in which this player has an incentive to deviate.

5. Examples

Before turning to the proof we shall consider two simple examples. The first illustrates the result when there is no pure Nash equilibrium. The second illustrates it when there are multiple Nash equilibria. In this case we shall walk through the algorithm to show why the Pareto dominant equilibrium is selected instead of the risk dominant equilibrium.

Example 1. Let \mathcal{G} be a 2×2 game with payoff matrix

	<i>A</i>	<i>B</i>
<i>A</i>	30,30	0,40
<i>B</i>	24,23	10,20

This game has no pure Nash equilibria, so by theorem 1 the learning process selects the combination that maximizes $\varphi_1 \cdot W - \gamma_1 \cdot S$. Figure 3 illustrates the case $\gamma_1 / \varphi_1 = 1$ in which *AA* is selected.

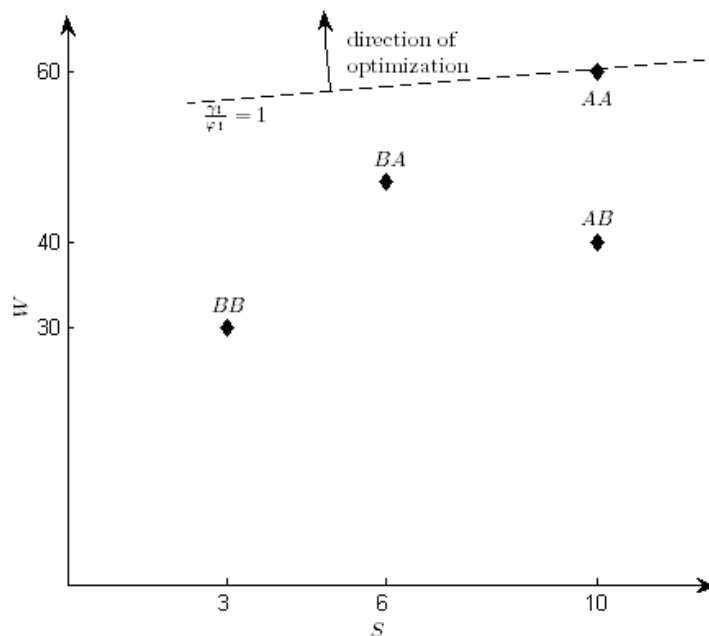


Figure 3. The tradeoff between welfare (W) and stability (S) in the 2×2 game of example 1, which has no pure Nash equilibrium.

In general, the outcome depends on the ratio γ_1 / φ_1 as shown in Figure 4. Note that the welfare maximizing state AA is selected whenever γ_1 / φ_1 is sufficiently small, that is, whenever the marginal change in the rate of acceptance by a discontent player is sufficiently large relative to the marginal change in the rate of acceptance of an experiment.

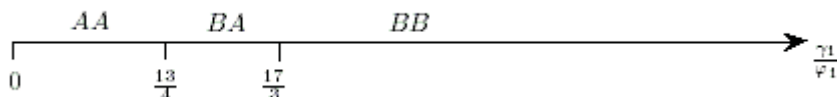


Figure 4. Stochastically stable outcomes for example 1 as a function of γ_1 / φ_1 .

Example 2. Let \mathcal{G} be a symmetric 2×2 coordination game with payoff matrix

	A	B
A	40, 40	30, 10
B	10, 30	50, 50

This game has two pure equilibria: AA and BB . The former is risk-dominant and the latter is Pareto-dominant. By theorem 1 our learning process selects the Pareto-dominant equilibrium over the risk-dominant equilibrium. It therefore differs from many other learning processes in the evolutionary literature (Kandori, Mailath and Rob, 1993; Young, 1993; Blume, 1993, 1995, 2003). It also differs from the algorithm in Young (2009), which provides no clear basis for discriminating between the two equilibria.

It is instructive to walk through the algorithm step-by-step to see how the selection mechanism works. For this purpose we need to consider specific acceptance functions. Let us choose the following:

$$F(u) = -0.001 \cdot u + 0.1, \quad G(\Delta u) = -0.01 \cdot \Delta u + 0.5.$$

It is straightforward to verify that these functions satisfy conditions (1)-(3) for the game in question.

Let us begin by supposing that the learning process is currently in the state AB and that both players are content. In the next period it could happen that the row player experiments by playing action B (probability ε) and the column player does not experiment (probability $1 - \varepsilon$). In this case they play BB , hence the row player experiences a payoff gain of 20 while the column player experiences a payoff gain of 40. Since the row player's experiment led to a payoff gain, he accepts his new action as his benchmark action with probability $\varepsilon^{G(\Delta u)} = \varepsilon^{G(20)} = \varepsilon^{0.3}$. Meanwhile the column player has turned hopeful because he experienced a non-self-triggered payoff gain. In the next period, the content row player continues to play his new benchmark action B with probability $1 - \varepsilon$, and the hopeful column player plays his benchmark action B with probability 1. Therefore the column player experiences a higher payoff than his benchmark for two consecutive periods, which causes him to become content with 50 as his new benchmark payoff and B as his new benchmark action. Overall this sequence of transitions from the all-content state AB to the all-content state BB has probability $\varepsilon \cdot (1 - \varepsilon)^2 \cdot \varepsilon^{0.3}$ (see figure 5).

Next let us consider the probability of exiting the all-content state BB . Note that any such transition involves a payoff decrease for both players. Hence for both of them to become content in another state, they must both accept lower benchmark payoffs. This can only happen if they first become discontent and subsequently become content again. Let D denote the set of states in which both players are discontent. We shall compute the probability of transiting from the all-content state BB to some state in D .

Since BB is a pure Nash equilibrium no unilateral deviation results in a payoff gain, and hence no single player will accept the outcome of an experiment. In fact, to move to D some player must experiment twice in a row. Suppose that next period the row player experiments and the column player does not and that this happens again in the following period. The probability of the first event is $\varepsilon \cdot (1 - \varepsilon)$ while the probability of the second event is ε . (The reason for the latter is that, by the second period, the column player has become watchful and therefore does not experiment.) Since the column player experiences a payoff below his previous benchmark for two successive periods he becomes discontent. Now over the *next two periods* the probability is $(0.5)^2 = 0.25$ that the discontent column player chooses A both times, and that the row player does not experiment. This event has probability $0.25 \cdot (1 - \varepsilon)$

because in the meantime the row player has become watchful. Overall this sequence of transitions from the all-content state BB to the set D has probability $0.25 \cdot \varepsilon^2 \cdot (1 - \varepsilon)^2$.

Next we turn to evaluating the probability of moving from a state in D to an all-content state. By assumption a discontent player spontaneously turns content with probability $\varepsilon^{F(u)}$ in any given period. (Note that this probability is *independent* of his benchmark payoff.) Therefore the probabilities of moving from D to each of the four all-content states are as follows:

$$\begin{aligned} D &\xrightarrow{(\varepsilon^{F(40)})^2 = \varepsilon^{0.12}} AA, & D &\xrightarrow{\varepsilon^{F(30) \cdot F(10)} = \varepsilon^{0.16}} AB \\ D &\xrightarrow{\varepsilon^{F(10) \cdot F(30)} = \varepsilon^{0.16}} BA, & D &\xrightarrow{(\varepsilon^{F(50)})^2 = \varepsilon^{0.10}} BB \end{aligned}$$

Note that among all possible ways of moving out of D , the move to all-content BB has highest probability.

The situation is summarized in Figure 5, which shows the highest probability of transiting between D and each of the four all-content states (possibly via intermediate states, which are not shown).

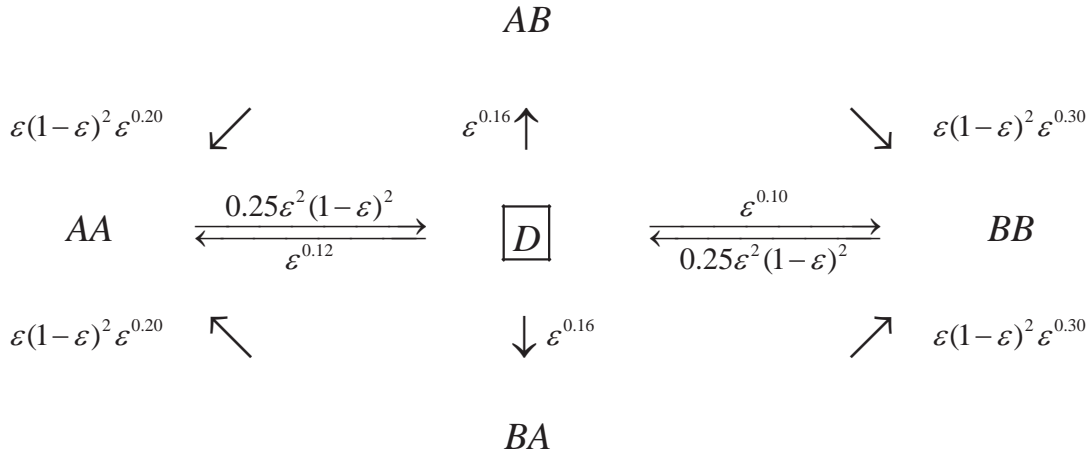


Figure 5. Maximal transition probabilities between the four all-content states and the set D in which everyone is discontent.

The key to the analysis is the relative magnitude of the various transition probabilities when ε is very small. This is determined by the size of the exponent on ε , which is called the *resistance* of the transition. These resistances are shown in figure 6. Note that the two

from any state in R to a state outside of R . The first step in the proof is to characterize the recurrence classes of the unperturbed process P^0 , that is, the process when $\varepsilon = 0$. In this situation, no one experiments and no one converts from being discontent to content.

Notation. Let Z^0 be the subset of states in which everyone's benchmarks are aligned. Let C^0 be the subset of Z^0 in which everyone is content, let E^0 be the subset of C^0 in which the benchmark actions constitute a pure Nash equilibrium and let E^* be the subset of E^0 on which welfare is maximized. Finally, let

$$D = \{\text{all states } z \in Z \text{ in which every player is discontent}\}. \quad (10)$$

Recall that if a player is discontent, he chooses an action according to a distribution that has full support and is independent of the current benchmark actions and payoffs. Moreover, the probability of accepting the outcome of such a search depends only on its realized payoff, and the old benchmarks are discarded. Hence, for any $w \in D$ and any $z \notin D$, the probability of the transition $w \rightarrow z$ is *independent* of w . Next consider a transition of form $z \rightarrow w$ where $w \in D$ and $z \notin D$. In this case the action and payoff benchmarks are the same in the two states, hence there is a *unique* $w \in D$ such that $z \rightarrow w$. We can therefore collapse the set D into a single state \bar{D} and define the transition probabilities as follows:

$$\forall z \in Z - D, P(\bar{D} \rightarrow z) \equiv P(w \rightarrow z) \text{ for all } w \in D. \quad (11)$$

$$\forall z \in Z - D, P(z \rightarrow \bar{D}) \equiv P(z \rightarrow w) \text{ for some } w \in D. \quad (12)$$

Lemma 1. *The recurrence classes of the unperturbed process are \bar{D} and all singletons $\{z\}$ such that $z \in C^0$.*

Proof. First we shall show that every element of C^0 is an absorbing state (a singleton recurrence class) of the unperturbed process ($\varepsilon = 0$). Suppose that $z \in C^0$, that is, the benchmarks are aligned and everyone is content. Since $\varepsilon = 0$, no one experiments and everyone replays his benchmark action next period with probability one. Hence the process remains in state z with probability one.

Next suppose that $z = \bar{D}$, that is, everyone is discontent. The probability that any given player becomes content next period is $\varepsilon^{F(\cdot)} = 0$. (Recall that $F(\cdot)$ is strictly positive on the domain of all payoffs that result from some feasible combination of actions.) Hence \bar{D} is absorbing, i.e. a singleton recurrence class. (Recall that a discontent player chooses each of his available actions with a probability that is bounded away from zero and independent of ε .)

It remains to be shown that there are no other recurrence classes of the unperturbed process. We first establish the following.

Claim. *Given any state z in which at least one player is discontent, there exists a sequence of transitions in the unperturbed process to \bar{D} .*

Consider a state in which some player i is discontent. By interdependence he can choose an action that alters the payoff of someone else, say j . Assume that i plays this action for two periods in a row. If j 's payoff *decreases* then in two periods he will become discontent also. If j 's payoff *increases* then in two periods he will become content again with a higher payoff benchmark. At this point there is a positive probability that the first player, i , will revert to playing his original action for two periods. This causes player j 's payoff to decrease relative to the new benchmark. Thus there is a positive probability that, in four periods or less, i 's behavior will cause j to become discontent. (The argument implicitly assumed that no one except for i and j changed action in the interim, but this event also has positive probability.) It follows that there is a series of transitions to a state where both i and j are discontent. By interdependence there are actions of i and j that cause a third player to become discontent. The argument continues in this manner until the process reaches a state where all players are discontent, which establishes the claim.

An immediate consequence is that \bar{D} is the only recurrent state in which some player is discontent. Thus, to conclude the proof of lemma 1, it suffices to show that if z is a state in which no one is discontent, then there is a finite sequence of transitions to \bar{D} or to C^0 . Suppose that someone in z is in a transient mood (c^+ or c^-). Since no one is discontent and no one experiments, there is no change in the players' actions next period, and therefore no

change in their realized payoffs. Hence everyone in a transient mood switches to c or d in one period. If anyone switches to d there is a series of transitions to \bar{D} . Otherwise everyone becomes content (or already was content) and their benchmarks are now aligned, hence the process has arrived at a state in C^0 . \square

We know from Young (1993, theorem 4) that the computation of the stochastically stable states can be reduced to an analysis of rooted trees on the vertex set R consisting solely of the recurrence classes. By Lemma 1, R consists of the singleton states in C^0 , and also the singleton state $\{\bar{D}\}$. Henceforth we shall omit the set brackets $\{ \}$ to avoid cumbersome notation.

Edge resistance. For every pair of *distinct* recurrence classes w and z , let $r(w \rightarrow z)$ denote the total resistance of the least-resistant path that starts in w and ends in z . We call $w \rightarrow z$ an *edge* and $r(w \rightarrow z)$ the *resistance* of the edge.

r^ -function.* Define the function $r^*(z)$ as follows

$$\forall z \in R, \quad r^*(z) = \min\{r(z \rightarrow w) : w \in R - \{z\}\}. \quad (13)$$

Easy. An *easy edge* from a state z is an edge $z \rightarrow w$, $w \neq z$, such that $r(z \rightarrow w) = r^*(z)$. An *easy path* is a sequence $w^1 \rightarrow w^2 \rightarrow \dots \rightarrow w^m$ in which each edge is easy and all states are distinct. An *easy tree* is a tree all of whose edges are easy.

This idea was illustrated in figure 6, where the easy edges are shown in black and the other edges in grey. Together the four easy edges form an easy tree that is rooted at the state BB . This identifies BB as the stochastically stable state in this particular case. While it is not always possible to determine the stochastically stable state in this way, the identification of the easy edges is a key first step in the analysis. For this purpose we need to evaluate the function r^* on the various recurrence classes. This is the purpose of the next three lemmas.

Lemma 2. $\forall e \in E^0$, $r^*(e) = 2$ and $e \rightarrow \bar{D}$ is an easy edge.

Proof. Let $e = (m, \bar{a}, \bar{u}) \in E^0$, where \bar{a} is a pure Nash equilibrium. Consider any outgoing edge $e \rightarrow z$ where $z \in C^0$ and $z \neq e$. Since e and z are distinct, everyone is content, and the benchmark payoffs and actions are aligned in both states, they must differ in their benchmark actions. Now consider any path from e to z in the full state space Z . Along any such path at least two players must experiment with new actions (an event with probability $O(\varepsilon^2)$), or one player must experiment twice in succession (also an event with probability $O(\varepsilon^2)$) in order for someone's benchmark action or payoff to change. (A single experiment is not accepted by the experimenter because \bar{a} is a Nash equilibrium, so the payoff from an experiment does not lead to a payoff gain. Furthermore, although some other players may temporarily become hopeful or watchful, they revert to being content with their old benchmarks unless a second experiment occurs in the interim.) It follows that $r(e \rightarrow z) \geq 2$.

It remains to be shown that the resistance of the transition $e \rightarrow \bar{D}$ is exactly two, which we shall do by constructing a particular path from e to \bar{D} in the full state space. Choose some player i . By interdependence there exists an action $a_i \neq \bar{a}_i$ and a player $j \neq i$ such that i 's change of action affects j 's payoff: $u_j(a_i, \bar{a}_{-i}) \neq u_j(\bar{a})$. Let player i experiment by playing a_i twice in succession, and suppose that no one else experiments at the same time. This event has probability $O(\varepsilon^2)$, so the associated resistance is two. If $u_j(a_i, \bar{a}_{-i}) > u_j(\bar{a})$, player j 's mood changes to c^+ after the first experiment and to c again after the second experiment. Note that at this point j has a new higher benchmark, namely, $u_j(a_i, \bar{a}_{-i})$. Now with probability $(1 - \varepsilon)^{2n}$ player i reverts to playing \bar{a}_i for the *next* two periods and no one else experiments during these periods. This causes j to become discontent. By the claim in the proof of Lemma 1, there is a zero-resistance path to the all-discontent state. The other case, namely $u_j(a_i, \bar{a}_{-i}) < u_j(\bar{a})$, also leads to an all-discontent state with no further resistance. We have therefore shown that $r(e \rightarrow \bar{D}) = 2$, and hence that $e \rightarrow \bar{D}$ is an easy edge. \square

Lemma 3. $\forall z \in C^0 - E^0$, $r^*(z) = 1 + G(S(z))$, and if $z \rightarrow z'$ is an easy edge with $z' \in C^0$, then $W(z) < W(z')$.

Proof. Let $z = (m, \bar{a}, \bar{u}) \in C^0 - E^0$, in which case \bar{a} is not a Nash equilibrium. Then there exists an agent i and an action $a_i \neq \bar{a}_i$ such that $u_i(a_i, \bar{a}_{-i}) > u_i(\bar{a}) = \bar{u}_i$. Among all such agents i and actions a_i suppose that $\Delta u_i = u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a})$ is a maximum. Let i experiment once with this action and accept the outcome of the experiment, and suppose that no one else experiments at the same time. The probability of this event is $O(\varepsilon^{1+G(\Delta u_i)})$. If the experiment causes everyone else's payoff to stay the same or go up, and if no one experiments in the next period (with the latter event having probability $(1-\varepsilon)^n$), then a state z' is reached after one period in which everyone is content, the benchmarks are aligned ($z' \in C^0$) and $W(z) < W(z')$. The total resistance of this path is $r(z \rightarrow z') = 1 + G(S(z))$. (Recall from (8) that $S(z)$ is the largest $\delta > 0$ such that someone can gain δ by a unilateral deviation.)

The other possibility is that the experiment causes someone else's payoff to *decrease*, in which case there is a zero-resistance series of transitions to the state \bar{D} . Hence in this case we have $r(z \rightarrow \bar{D}) = 1 + G(S(z))$. We conclude that, if there is only one experiment, the process either transits to a recurrence class $z' \in C^0$ satisfying $W(z') > W(z)$, or it transits to \bar{D} . In both cases the resistance of the transition is $1 + G(S(z))$.

If there are two or more experiments, the resistance is at least two. By assumption, however, $G(\cdot) < 1/2$ (condition 3), hence making two experiments has a higher resistance than making one experiment and accepting the outcome (the latter has resistance $1 + G(S(z)) < 1.5$). It follows that $r^*(z) = 1 + G(S(z))$, and if $z \rightarrow z'$ is an easy edge with $z' \neq \bar{D}$, then $W(z) < W(z')$. \square

Lemma 4. $\forall z = (m, \bar{a}, \bar{u}) \in C^0$, $r(\bar{D} \rightarrow z) = \sum_i F(u_i(\bar{a}))$ and $r^*(\bar{D}) = \min_{a \in A} \sum_i F(u_i(a))$.

Proof. Let $\bar{D} \rightarrow z = (m, \bar{a}, \bar{u}) \in C^0$. Recall that the transition probability $P(w \rightarrow z)$ is the same for all $w \in D$, hence take any state $w^1 \in D$ as the starting point. The probability is $O(1)$ that next period every player i chooses \bar{a}_i , in which case their realized payoffs are $\bar{u}_i = u_i(\bar{a})$. They all accept these actions and payoffs as their new benchmarks with probability $\prod_i \varepsilon^{F(u_i(\bar{a}))} = \varepsilon^{\sum_i F(u_i(\bar{a}))}$, hence $r(\bar{D} \rightarrow z) \leq \sum_i F(\bar{u}_i)$.

We claim that in fact $r(\bar{D} \rightarrow z) = \sum_i F(\bar{u}_i)$. Let $z = (m, \bar{a}, \bar{u}) \in C^0$ and consider a least-resistant path $w^1, w^2, \dots, w^m = z$. For each player i there must be some time in the sequence where i was discontent and accepted a benchmark payoff that was \bar{u}_i or less. (There may also have been a time when i was discontent and accepted a benchmark payoff that was strictly more than \bar{u}_i , but in that case there must have been a later time at which he accepted a payoff that was \bar{u}_i or less, which means he must have been discontent.) The probability of such an acceptance is at most $\varepsilon^{F(\bar{u}_i)}$, because the probability of acceptance is increasing in u_i . This reasoning applies to every player, hence the total resistance of this path from \bar{D} to z must be at least $\sum_i F(\bar{u}_i)$. This proves that $r(\bar{D} \rightarrow z) = \sum_i F(\bar{u}_i)$. The claim that $r^*(\bar{D}) = \min_{a \in A} \sum_i F(u_i(a))$ follows from the fact that F is monotone decreasing. \square

w-tree. Identify the recurrence classes R with the nodes of a graph. Given a node w , a collection of directed edges T forms a *w-tree* if from every node $z \neq w$ there is exactly one outgoing edge in T and there is a unique directed path in T from z to w .

Stochastic potential. The *resistance* $r(T)$ of a *w-tree* T is the sum of the resistances of its edges. The *stochastic potential* of w is $\rho(w) = \min\{r(T) : T \text{ is a } w\text{-tree}\}$.

The stochastically stable states are precisely those states where ρ achieves its minimum (Young, 1993, theorem 4). In the next two lemmas we compute the stochastic potential of each type of recurrence class. From these computations theorem 1 will follow.

Lemma 5. *There exists a \bar{D} -tree $T_{\bar{D}}^*$ that is easy.*

Proof. We shall conduct the proof on transitions between recurrence classes, each of which forms a node of the graph. Choose a node $z \neq \bar{D}$ and consider an easy outgoing edge $z \rightarrow \cdot$. If there are several such edges choose one that points to \bar{D} , that is, choose $z \rightarrow \bar{D}$ if it is easy. This implies in particular that for every $e \in E^0$ we select the edge $e \rightarrow \bar{D}$. (This follows from Lemma 2.)

We claim that the collection of all such edges forms a \bar{D} -tree. To establish this it suffices to show that there are no cycles. Suppose by way of contradiction that $z^1 \rightarrow z^2 \rightarrow \dots \rightarrow z^m \rightarrow z^1$ is a shortest cycle. This cycle cannot involve any node in E^0 , because by construction the outgoing edge from any such edge points towards \bar{D} , which has no outgoing edge. Therefore all $z^k \in C^0 - E^0$. Since all of the edges $z^k \rightarrow z^{k+1}$ are easy, Lemma 3 implies that $W(z^k) < W(z^{k+1})$. From this we conclude $W(z^1) < W(z^m) < W(z^1)$, which is impossible. \square

Let
$$\rho^* = \rho(\bar{D}) = r(T_{\bar{D}}^*). \quad (14)$$

Lemma 6. *For every $z \in C^0$ let $z \rightarrow w_z$ be the unique outgoing edge from z in $T_{\bar{D}}^*$ and define*

$$T_z^* = T_{\bar{D}}^* \text{ with } z \rightarrow w_z \text{ removed and } \bar{D} \rightarrow z \text{ added} . \quad (15)$$

T_z^* is a z -tree of least resistance and

$$\rho(z) = \rho^* - r(z \rightarrow w_z) + r(\bar{D} \rightarrow z). \quad (16)$$

Proof. Plainly, the tree T_z^* defined in (15) is a z -tree. Furthermore all of its edges are easy except possibly for the edge $\bar{D} \rightarrow z$. Hence it is a least-resistant z -tree among all z -trees that contain the edge $\bar{D} \rightarrow z$. We shall show that in fact it minimizes resistance among all z -trees.

Let T_z be some z -tree with minimum resistance, and suppose that it does not contain the edge $\bar{D} \rightarrow z$. Since it is a spanning tree it must contain some outgoing edge from \bar{D} , say $\bar{D} \rightarrow z'$. We can assume that $r(\bar{D} \rightarrow z') < r(\bar{D} \rightarrow z)$, for otherwise we could simply take out the edge $\bar{D} \rightarrow z'$ and put in the edge $\bar{D} \rightarrow z$ to obtain the desired result.

Let u_1, u_2, \dots, u_n be the benchmark payoffs in z and let u'_1, u'_2, \dots, u'_n be the benchmark payoffs in z' . By lemma 4,

$$r(\bar{D} \rightarrow z) = \sum_i F(u_i) \quad \text{and} \quad r(\bar{D} \rightarrow z') = \sum_i F(u'_i). \quad (17)$$

Since $r(\bar{D} \rightarrow z') < r(\bar{D} \rightarrow z)$ and $F(u_i)$ is by assumption monotone decreasing in u_i , there must be some i such that $u'_i > u_i$. Let $I = \{i : u'_i > u_i\}$. Consider the unique path in T_z that goes from z' to z , say $z' = z^1, z^2, \dots, z^m = z$, where each z^k is in R and they are distinct. Each edge $z^k \rightarrow z^{k+1}$ corresponds to a sequence of transitions in the full state space Z , and the union of all these transitions constitutes a path in Z from z' to z . Along this path, each player $i \in I$ must eventually lower his payoff benchmark because his starting benchmark u'_i is greater than his ending benchmark u_i . Thus at some point i must become discontent and adopt a new benchmark that is u_i or lower. Assume that this happens in the transition from z^{k_i} to z^{k_i+1} . Since $F(u_i)$ is strictly decreasing, the probability of adopting a benchmark that is u_i or lower is *at most* $\varepsilon^{F(u_i)}$. Hence, along the path from z^{k_i} to z^{k_i+1} someone experiments and accepts, and at some stage player i becomes discontent and accepts a payoff that is u_i or lower. The first event has resistance at least $1 + G(S(z^{k_i}))$ and the second has resistance at least $F(u_i)$. Therefore

$$r(z^{k_i} \rightarrow z^{k_i+1}) \geq 1 + G(S(z^{k_i})) + F(u_i). \quad (18)$$

We know from Lemma 3 that the least resistance of a path out of z^{k_i} is $r^*(z^{k_i}) = 1 + G(S(z^{k_i}))$. Hence $r(z^{k_i} \rightarrow z^{k_i+1}) \geq F(u_i) + r^*(z^{k_i})$. It follows that the total resistance along the sequence $z' = z^1, z^2, \dots, z^m = z$ satisfies

$$\sum_{k=1}^{m-1} r(z^k \rightarrow z^{k+1}) \geq \sum_{i \in I} F(u_i) + \sum_{k=1}^{m-1} r^*(z^k). \quad (19)$$

The resistance of the edge $\bar{D} \rightarrow z'$ satisfies

$$r(\bar{D} \rightarrow z') = \sum_i F(u'_i) > \sum_{i \in I} F(u'_i) \geq \sum_{i \in I} F(u_i). \quad (20)$$

Hence in the tree T_z the outgoing edges from the nodes $\{\bar{D}, z^1, z^2, \dots, z^{m-1}\}$ have a total resistance that is strictly greater than $\sum_i F(u_i) + \sum_{k=1}^{m-1} r^*(z^k)$. But in the tree T_z^* the edges from these nodes have a total resistance equal to $\sum_i F(u_i) + \sum_{k=1}^{m-1} r^*(z^k)$. Furthermore, at every other node the resistance of the outgoing edge in T_z is at least as great as it is in T_z^* (because the latter consists of easy edges). We conclude that T_z^* must minimize resistance among all z -trees, which completes the proof of lemma 6. \square

To complete the proof of theorem 1, we shall first prove the following chain of inequalities:

$$\forall e^* \in E^*, \forall e \in E^0 - E^*, \forall z \in C^0 - E^0, \quad \rho(e^*) \stackrel{(i)}{<} \rho(e) \stackrel{(ii)}{<} \rho(z) \stackrel{(iii)}{<} \rho(\bar{D}) = \rho^*. \quad (21)$$

(Recall that E^* consists of those equilibrium states $e \in E^0$ that maximize $W(e)$.) Let $e = (m, \bar{a}, \bar{u}) \in E^0$. By construction $e \rightarrow \bar{D}$ is an edge in the easy tree $T_{\bar{D}}^*$ (see the beginning of the proof of Lemma 5), so (16) implies that

$$\rho(e) = \rho^* - r(e \rightarrow \bar{D}) + r(\bar{D} \rightarrow e). \quad (22)$$

From Lemma 2 we know that $r(e \rightarrow \bar{D}) = 2$. From Lemma 4 we know that

$$r(\bar{D} \rightarrow e) = \sum_i F(u_i(\bar{a})) = -\varphi_1 W(e) + n\varphi_2. \quad (23)$$

From this and (22) we conclude that

$$\rho(e) = \rho^* - 2 - \varphi_1 W(e) + n\varphi_2. \quad (24)$$

Now suppose that $e^* \in E^*$ and $e \in E^0 - E^*$. Then $W(e^*) > W(e)$, so from (24) we conclude that $\rho(e^*) < \rho(e)$. This establishes (i).

To prove (ii), let $e \in E^0 - E^*$ and $z \in C^0 - E^0$. Recall from (16) that $\rho(z) = \rho^* - r(z \rightarrow w_z) + r(\bar{D} \rightarrow z)$, where $z \rightarrow w_z$ is an easy edge. Since $z \in C^0 - E^0$, we know from Lemma 3 that $r(z \rightarrow w_z) = 1 + G(S(z))$ and from Lemma 4 that $r(\bar{D} \rightarrow z) = -\varphi_1 W(z) + n\varphi_2$. Hence

$$\forall z \in C^0 - E^0: \rho(z) = \rho^* - 1 - G(S(z)) - \varphi_1 W(z) + n\varphi_2. \quad (25)$$

Since $e \in E^0$, (24) implies that

$$\rho(e) = \rho^* - 2 - \varphi_1 W(e) + n\varphi_2. \quad (26)$$

Comparing (25) and (26) we see that $\rho(e) < \rho(z)$ provided

$$\varphi_1 [W(z) - W(e)] < 1 - G(S(z)). \quad (27)$$

The right-hand side of (27) is greater than $1/2$ because we assumed that $G(S(z)) < 1/2$ for all z (see condition (3)). The left-hand side is the sum of n differences $\sum_i F(u_i) - F(u'_i)$, which is smaller than $1/2$ because we assumed that $0 < F(\cdot) < 1/2n$ (condition (3) again). Hence (27) holds and therefore $\rho(e) < \rho(z)$, which proves (ii).

To prove (iii), observe that condition (3) implies

$$-\varphi_1 W(z) + n\varphi_2 \leq n \cdot \max_u F(u) < 1/2. \quad (28)$$

From this and (25) it follows that $\rho(z) < \rho^*$. This completes the proof of (21).

We shall now turn to the two cases of the theorem: a pure Nash equilibrium exists and a pure Nash equilibrium does not exist.

Case 1. A pure Nash equilibrium exists ($E^0 \neq \emptyset$).

By (21) we know that the welfare maximizing equilibrium states $e^* \in E^*$ minimize stochastic potential among all recurrence classes, hence they constitute the stochastically stable states. This establishes statement (i) of Theorem 1.

Case 2. No pure equilibrium exists ($E^0 = \emptyset$).

In this case the chain of inequalities (21) reduces to (iii), because there are no equilibrium states. In particular, \bar{D} cannot be stochastically stable. It follows from (25) that the stochastically stable states are precisely those $z \in C^0$ that minimize

$$\begin{aligned} \rho(z) &= \rho^* - 1 - G(S(z)) - \varphi_1 W(z) + n\varphi_2 \\ &= \gamma_1 S(z) - \varphi_1 W(z) + (n\varphi_2 + \rho^* - 1) \end{aligned} \quad (29)$$

which is equivalent to maximizing $\varphi_1 W(z) - \gamma_1 S(z)$. This concludes the proof of theorem 1.

7. Heterogeneous agents

Thus far we have assumed that ε, F , and G are common to all agents and that F and G are linear. In this section we generalize theorem 1 to the situation where the agents have heterogeneous learning parameters and the acceptance functions may be nonlinear. First let us consider what happens when agents have different rates of experimentation. Suppose that

agent i has experimentation rate $\varepsilon_i = \lambda_i \varepsilon$ where $\lambda_i > 0$ is idiosyncratic to i and $\varepsilon > 0$ is common across agents. Varying ε changes the overall amount of noise in the search process while holding the ratio of the search rates fixed. Assume that agent i accepts the outcome of a random search with probability $\varepsilon_i^{F(u_i)} = (\lambda_i \varepsilon)^{F(u_i)}$, and accepts the outcome of an experiment with probability $\varepsilon_i^{G(\Delta u_i)} = (\lambda_i \varepsilon)^{G(\Delta u_i)}$. The introduction of the λ_i 's does not alter the resistances, which are determined by the exponents of ε in the transition probabilities. Hence the stochastically stable states remain exactly as before, and theorem 1 holds as stated.

Now suppose that i 's probability of accepting the outcome of a random search is governed by some function $F_i(u_i)$, and i 's probability of accepting the outcome of an experiment is governed by $G_i(\Delta u_i)$. Assume that these acceptance functions are strictly monotone decreasing but not necessarily linear, and that they satisfy the bounds in (3), that is,

$$0 < G_i(\Delta u_i) < 1/2 \text{ and } 0 < F_i(u_i) < 1/2n \text{ for all } i. \quad (30)$$

Given a state $z = (m, \bar{a}, \bar{u})$ let us redefine the welfare function as follows:

$$\tilde{W}(z) = -\sum_i F_i(u_i(\bar{a})). \quad (31)$$

Recall that an equilibrium is *efficient* (Pareto undominated) if there is no other equilibrium in which someone is better off and no one is worse off. Since the functions F_i are monotone decreasing, $\tilde{W}(z)$ is monotone increasing in each player's utility $u_i(\bar{a})$. It follows that, among the equilibrium states, $\tilde{W}(\cdot)$ is maximized at an efficient equilibrium. (If the F_i are linear, $\tilde{W}(\cdot)$ maximizes a weighted sum of the agents' utilities.)

Next let us define

$$S_i(z) = S_i(m, \bar{a}, \bar{u}) = \max_{a_i \in A_i} \{u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a}) : u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a}) > 0\}. \quad (32)$$

Further, let $\tilde{S}_i(z) = -G_i(S_i(z))$ and define the stability of a state z to be

$$\tilde{S}(z) = \max_i \{ \tilde{S}_i(z) \}. \quad (33)$$

The larger $\tilde{S}(z)$ is, the more likely it is that some player will accept the outcome of an experiment, hence the more prone the state is to being destabilized. A straightforward modification in the proof of theorem 1 leads to the following:

Theorem 2. *Let \mathcal{G} be an interdependent n -person game on a finite joint action space A . Suppose that each player i uses a learning rule with experimentation probability $\varepsilon_i = \lambda_i \varepsilon$ and monotone decreasing acceptance functions F_i and G_i satisfying the bounds in (30).*

(i) *If the game has a pure Nash equilibrium then every stochastically stable state is an equilibrium state that maximizes $\tilde{W}(z)$ among all equilibrium states, and hence is efficient;*

(ii) *If the game has no pure Nash equilibrium, the stochastically stable states maximize $\tilde{W}(z) - \tilde{S}(z)$ among all $z \in C^0$.*

8. Concluding remarks

In this paper we have identified a completely uncoupled learning rule that selects an efficient pure equilibrium in any n -person game with generic payoffs that possesses at least one pure equilibrium. This provides a solution to an important problem in the application of game theory to distributed control, where the object is to design a system of autonomous interacting agents that optimizes some criterion of system-wide performance using simple feedback rules that require no information about the overall state of the system. The preceding analysis shows that this can be accomplished by a variant of log linear learning and two different search modes – fast and slow. Theorem 1 shows that by choosing the probabilities of experimentation and acceptance within an appropriate range, the process is in an efficient equilibrium a high proportion of the time. This allows one to reduce the price of anarchy substantially, because one need only compare the maximum welfare state to the maximum welfare equilibrium state. As an extra dividend we obtain a simple criterion for the selection

of disequilibrium states. This criterion weights total welfare positively and the incentive to deviate negatively. As we have shown by example this concept differs from risk dominance.

It is, of course, legitimate to ask how long it takes for the learning rule to reach an efficient equilibrium from arbitrary initial conditions. We know in general that it can take an exponentially long time for an uncoupled learning process to converge to Nash equilibrium for games of arbitrary size (Hart and Mansour, 2010); it can also take an exponentially long time to converge to a Pareto efficient, individually rational state (Babichenko, 2010b). The method that we used to prove theorem 1 involved taking the experimentation rate to zero, in which case it takes a very long time in expectation for the process to reach the stochastically stable state(s). However, this does not rule out the possibility that the learning process *strongly favors* the stochastically stable states even when the experimentation probability is large; indeed this has been shown for 2 x 2 games played by large populations of players who make mistakes independently (Young, 1998, Chapter 4). Thus it is possible that the learning algorithm with *intermediate* levels of experimentation would lead quite quickly to an approximate equilibrium, at least for some classes of games. This is a challenging open problem that we shall not attempt to tackle here.

Finally, we should note that the type of learning rules we have proposed are not meant to be empirically descriptive of how humans actually behave in large decentralized environments. Our aim has been to show that it is theoretically possible to achieve efficient equilibrium selection using simple, completely uncoupled rules. Nevertheless, it is conceivable that certain qualitative features of these rules are reflected in actual behavior. At the micro level, for example, one could test whether agents engage in different types of search (fast and slow) depending on their recent payoff history. One could also estimate the probability that they accept the outcome of a search as a function of their realized payoffs. At the macro level, one could examine whether agents playing a congestion game converge to a local maximum of the potential function, to a Pareto optimal equilibrium, or fail to come close to equilibrium in any reasonable amount of time. Whether or not our learning model proves to have features that are descriptively accurate for human agents, the approach does suggest some testable questions that to the best of our knowledge have not been examined before.

Acknowledgements

We thank Gabriel Kreindler and Jason Marden for suggesting a number of improvements to an earlier draft. This research was supported by the Office of Naval Research, grant N00014-09-1-0751.

References

Arieli, Itai, and Yakov Babichenko, 2011, "Average testing and the efficient boundary," Working Paper, University of Oxford and the Center for Rationality, Hebrew University.

Asadpour, Arash and Saberi, Amin, 2009, "On the inefficiency ratio of stable equilibria in congestion games", *5th Workshop on Internet and Networks Economics*, 545-552.

Babichenko, Yakov, 2010a, "Completely uncoupled dynamics and Nash equilibria," Working Paper 529, Center for the Study of Rationality, Hebrew University.

Babichenko, Yakov, 2010b, "How long to Pareto efficiency?", Working Paper 562, Center for Rationality, Hebrew University.

Blume, Lawrence E., 1993, "The statistical mechanics of strategic interaction," *Games and Economic Behavior*, 4, 387-424.

Blume, Lawrence E., 1995, "The statistical mechanics of best-response strategy revision," *Games and Economic Behavior*, 11, 111-145.

Blume, Lawrence E., 2003, "How noise matters," *Games and Economic Behavior*, 44, 251-271.

Bowling, Michael, and Manuel Veloso, 2002, "Multi-agent learning with a variable learning rate," *Artificial Intelligence*, 136, 215-250.

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

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

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

Hart, Sergiu, and Yishay Mansour, 2010, "How Long to Equilibrium? The Communication Complexity of Uncoupled Equilibrium Procedures," *Games and Economic Behavior*, 69, 107-126.

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

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

Houston, A. I., Alex Kacelnik, and John M. McNamara, 1982, "Some learning rules for acquiring information," in *Functional Ontogeny*, D. J. McFarland, ed., New York: Pitman.

Kandori, Michihiro, George Mailath, and Rafael Rob, 1993, " Learning, mutation, and long-run equilibrium in games," *Econometrica*, 61, 29-56.

Karandikar, Rajeeva, Dilip Mookherjee, Debraj Ray, and Fernando Vega-Redondo, 1998, "Evolving aspirations and cooperation," *Journal of Economic Theory*, 80, 292-331.

Mannor, Shie, and Jeff S. Shamma, 2007, "Multi-agent learning for engineers," *Artificial Intelligence*, 171, 417-422.

Marden, Jason R., and Jeff S. Shamma, 2008, "Revisiting log-linear learning: asynchrony, completeness and a payoff-based interpretation," Working Paper, University of Colorado.

Marden, Jason R, Gurdal Arslan, and Jeff S. Shamma, 2009, "Cooperative control and potential games," *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*.

Marden, Jason R., H. Peyton Young, Gurdal Arslan, and Jeff S. Shamma, 2009, "Payoff-based dynamics for multiplayer weakly acyclic games", *SIAM Journal on Control and Optimization*, 48, No. 1, 373-396.

Marden, Jason, H. Peyton Young, and Lucy Y. Pao, 2011, "Achieving Pareto optimality through distributed learning," Discussion Paper, University of Colorado and University of Oxford.

Motro, Uzi, and Avi Shmida, 1995, "Near-far search: an evolutionarily stable foraging strategy," *Journal of Theoretical Biology*, 173, 15-22.

Papadimitriou, Christos, 2001, "Algorithms, games and the internet", *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, 749-753.

Roughgarden, Tim, 2005, *Selfish Routing and the Price of Anarchy*, Cambridge Mass: MIT Press.

Sandholm, W. H., 2002, "Evolutionary implementation and congestion pricing," *Review of Economic Studies*, 69, 667-689.

Shah, Devavrat and Shin, Jinwoo, 2010, "Dynamics in Congestion Games", *ACM SIGMETRICS (preliminary version)*.

Thuijsman, F., Bezael Peleg, M. Amitai, and Avi Shmida, 1995, "Automata, matching, and foraging behavior in bees," *Journal of Theoretical Biology*, 175, 305-316.

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

Young, H. Peyton, 2009, "Learning by trial and error", *Games and Economic Behavior*, 65, 626-643.