Learning Efficient Nash Equilibria in Distributed Systems

Bary S. R. Pradelski and H. Peyton Young

University of Oxford

September 3, 2010

Abstract. An individual's learning rule is *completely uncoupled* if it does not depend 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 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

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 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). Among control theorists and computer scientists, the issue is not whether a given learning rule is descriptively accurate as a model of human behavior, but whether it leads to good system-wide performance when agents are endowed with this behavior; in other words the learning procedure and the agents' payoffs are treated as design elements of the system.

To date the main focus of attention has been on potential games, since these 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, 2010; see also Hart and Mas-Colell, 2003, 2006).

The learning rule that we propose is related to the trial and error learning procedure of Young (2009), and more distantly related to the 'learning by sampling' procedure of Foster and Young (2006) and Germano and Lugosi (2007).[1] An essential feature of the rule is that a player has two different search modes: i) deliberate experimentation, which occurs with low

[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

probability and leads to a change of strategy only if the latter has a higher payoff; ii) random search, which leads to a change of strategy that may or may not have a higher payoff (the probability of acceptance is merely biased towards strategies with high payoffs).

A crucial difference between our procedure and trial and error learning is that a player does not always accept the outcome of an experiment even when it does result in higher payoffs: acceptance is probabilistic. Our rule also differs in that the probabilities of accepting the outcome of an experiment (or of a random search) can be expressed as a log linear function of the payoff gain (in the case of an experiment) or the payoff level (in the case of a random search). The advantage of this approach is that it leads to a simple formula for computing the stochastically stable states of the process. This formula shows that the learning process selects an efficient pure Nash equilibrium whenever a pure Nash equilibrium exists, and identifies the *disequilibrium* states that are selected when a pure Nash equilibrium does not exist. In the latter case there is a tradeoff between welfare and stability: the most likely disequilibrium state is one that maximizes 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.

2. The learning model

We first describe the learning rule informally, then give a detailed definition. A key feature of the rule is that an agent can search in one of two ways. In 'quiet' search he occasionally experiments with new strategies and adopts a new one with a probability that increases with the realized gain in payoff compared to his previous strategy. In 'noisy' search he frequently tries out new strategies and adopts a new one with a probability that increases in the realized level of payoff. These two forms of search can be associated with different 'psychological states'. A *content* agent is not strongly motivated to search but occasionally does so anyway (quiet search). A *discontent* agent flails around trying out new things frequently (noisy search). In the first type of search, the probability of acceptance is determined by the change in payoff, whereas in the second situation the probability of acceptance is determined by the level of payoff. The rationale is that in noisy search an agent tries out many different

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.

strategies before settling on one of them, hence the payoff level seems more salient than the change in payoff from the previous period.[2]

A second feature of the learning process is the mechanism that triggers transitions between content and discontent states. The essential idea is that a transition from content (*c*) to discontent (*d*) occurs when an agent's realized payoff goes down for several periods in succession and he did not search during those periods. In other words, a $c \to d$ transition is triggered by a (negative) change in payoff that was not instigated by the agent, but rather by a change of strategy by someone else (whom the agent cannot necessarily see). By contrast, a $d \to c$ transition occurs when an agent tires of searching and accepts his current strategy as 'good enough'.

To illustrate these ideas in a concrete case, consider a commuter who ordinarily takes a certain route to work. The realized cost is the journey time (the negative of the payoff), which depends on the routes taken by other commuters. Suppose that congestion on his usual route worsens and the journey time increases. If this persists he may become discontent and start looking actively for a different route. We hypothesize that the probability of settling on a new route depends on how short it turns out to be. This is the logic of the $c \to d \to c$ transitions. But even if our commuter is reasonably content, he may not be completely content, and thus may occasionally try out new ways to go to work. Our hypothesis is that he adopts such 'experiments' with a probability that is monotone increasing in the improved travel time.

We now describe the learning rule 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. The *state* of player i is a triplet $z_i = (m_i, \bar{a}_i, \bar{u}_i)$, where $\bar{a}_i \in A_i$ is the player's *benchmark action* (the action he normally plays in this state), $\bar{u}_i$ is his *benchmark payoff* (what he currently expects to get), and m_i is his *mood* (which determines his current search behavior). There are four distinct moods: content (*c*), discontent (*d*), watchful (c^-), and hopeful (c^+). The *state of the process* is denoted by a triple $z = (m, \bar{a}, \bar{u})$ where m is an *n*-vector of the moods of the various players, $\bar{a}$ is an *n*-vector of their benchmark actions, and $\bar{u}$ is an *n*-vector of their benchmark payoffs. For

[2] Learning rules with high and low rates of exploration have been studied in a variety of settings, including computer science (Bowling and Veloso, 2002) and biology (Thuijsman et al., 1995).

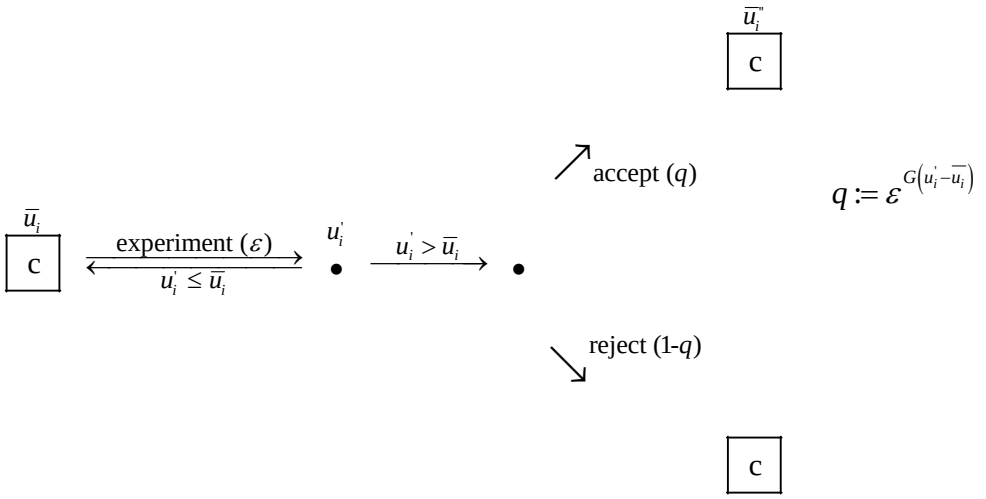
simplicity we assume that each $\bar{u}_i$ is consistent with some payoff in the game, that is, $\bar{u}_i = u_i(a)$ for some $a \in A$. Since A is finite, it follows that the state space Z is also finite. The process evolves in discrete time periods $t = 1, 2, 3, \ldots$, where $z(t)$ denotes the state of the process at time *t*.

Figure 1 depicts the possible one-period transitions for a specific player i. Other transitions may occur during the same period for other players, and the combination of all of these transitions determines the transition function between states. However, we can omit specific reference to the actions and payoffs of players other than i, because the only effect they have on i's transitions is through i's realized payoffs.

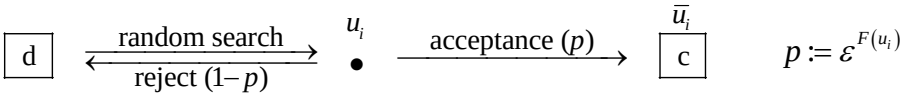
We explain the various cases in turn. A *content-content transition* occurs when player i decides to experiment with a new action. This event has probability $\varepsilon > 0$. Call the resulting payoff u_i'. If u_i' is higher than i's current benchmark payoff $\bar{u}_i$, then with probability $q = \varepsilon^{G(u_i' - \bar{u}_i)}$ he adopts the experimental action and the experienced payoff as his new benchmarks. (The actions are not shown to keep the figure uncluttered.)

The second case is a *discontent-content transition.* When a player i is discontent he chooses an action at random each period, according to some fixed distribution that is independent of ε and has full support on A_i. Player i's current search ends when he spontaneously accepts his current action and its associated payoff as his new benchmarks. The probability of this event is $p = \varepsilon^{F(u_i)}$. (In Young (2009) this probability was assumed to be bounded away from zero and independent of ε. This change in the model has important implications for its equilibrium selection properties and requires a different proof.)

Content-content transitions



Discontent-content transitions



Indirect transitions

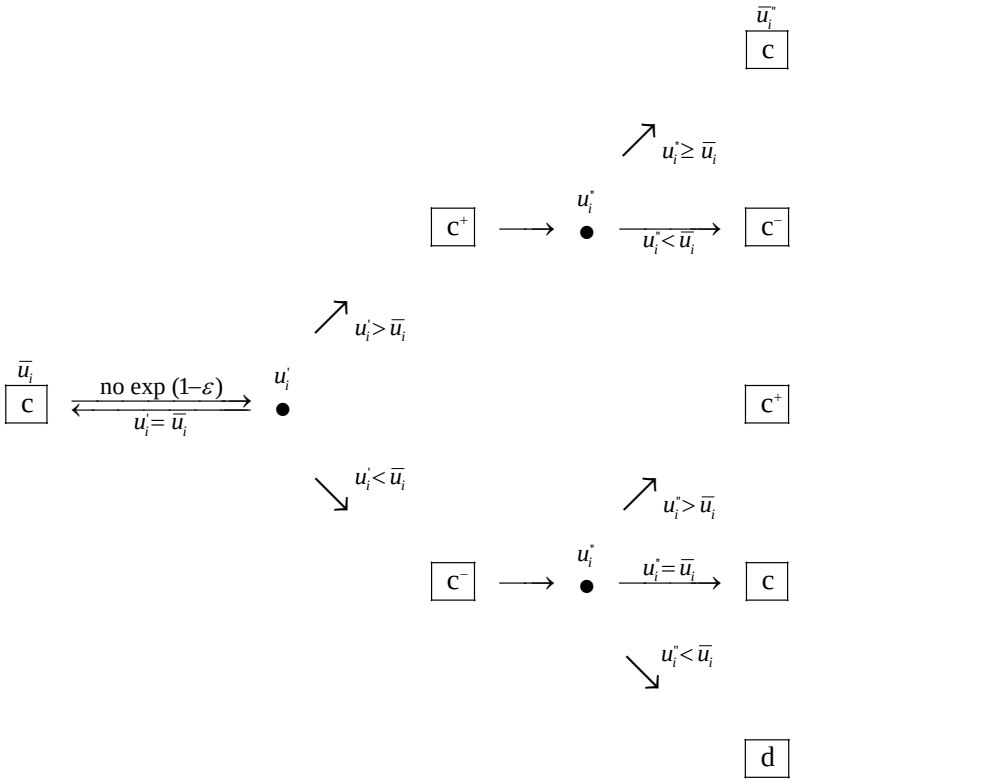


Figure 1. The structure of transitions for a given player *i*.

An indirect transition occurs when player i's payoff changes and he does *not* experiment. Thus the change must have been caused by some other player (whom he may not be able to observe). We call this a *passive change* in payoff. It causes player i to go on the alert: if the new payoff u_i' is higher than his current benchmark $\overline{u}_i$ his mood changes to c^+, whereas if the new payoff is lower than his current benchmark his mood changes to c^-. In the ensuing period he moves from c^- to d if his payoff stays below $\overline{u}_i$, and moves from c^+ to c if his payoff is above or equal to $\overline{u}_i$. (Note that i's payoff in the first period (u_i') may differ from his payoff in the second (u_i'') because of changes in the behavior of other players. Note also that payoff reversals relative to the benchmark cause player i to flip-flop between c^- and c^+, hence these transient states can persist for several periods.

In summary, the parameters of the learning process consist of the *experimentation probability* ε and the two *acceptance functions* $F(u)$ and $G(\Delta u)$. Note that the domain of F is the *set of payoff levels* $U = \cup_i \{u_i(a) : a \in A\}$, whereas the domain of G is the *set of nonnegative payoff differences* $\Delta U = \cup_i \{[u_i(a_i', a_{-i}) - u_i(a_i, a_{-i})]_+ : a_i, a_i' \in A_i, a_{-i} \in A_{-i}\}$.

We shall begin by assuming that F and G are *nonnegative, strictly decreasing, linear functions*, and that they are the same for all players. Later we shall show how to extend our results to the situation where the acceptance functions differ among players and are non-linear. Let

$$F(u) = -f_1 \cdot u + f_2, \text{ where } f_1 > 0, \tag{1}$$

$$G(\Delta u) = -g_1 \cdot \Delta u + g_2, \text{ where } g_1 > 0. \tag{2}$$

We assume that the coefficients f_i, g_i are such that $F(u)$ and $G(\Delta u)$ are *strictly positive* for all $u \in U$ and $\Delta u \in \Delta U$. We shall also assume that F and G are small in the following sense:

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

This means that the probabilities of acceptance are substantially greater than the probability of experimentation. In particular, the probability $\varepsilon^{G(\Delta u)}$ of accepting the outcome of an experiment is greater than $\sqrt{\varepsilon}$; moreover if all agents are discontent then the probability that

they all become content again ($\varepsilon^{nF(u)}$) is also greater than $\sqrt{\varepsilon}$. We do not claim that these bounds are best possible but they are easy to work with and yield sharp analytical predictions.

This rule can be viewed as 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

$$P(\text{accept experiment with payoff gain } \Delta u > 0) = \beta g_1 \Delta u - \beta g_2 . \qquad (4)$$

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

$$P(\text{accept search with payoff } u) = \beta f_1 \Delta u - \beta f_2 . \qquad (5)$$

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 current payoff level that results from a random search, the higher the probability that the current action will be accepted. It follows that interpersonal comparisons of utility can be made, because they imply different acceptance rates among the players.

3. Discussion

Learning rules that employ slow and fast search have been discussed 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 x 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

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:

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

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

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

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

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

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 $\varnothing \subset J \subset N$, 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 $z = (m, \overline{a}, \overline{u})$ are *aligned* if the benchmark payoffs result from playing the benchmark actions, that is, if $\overline{u}_i = u_i(\overline{a})$ 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, and hence is efficient;

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

$$f_1 W(z) - g_1 S(z). \tag{9}$$

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 a large incentive to deviate.

5. Examples

Before turning to the proof we illustrate the result with two simple examples.

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

	A	B
A	a,a	c,d
B	d,c	b,b

Assume that the equilibrium AA is strictly risk-dominant, that is $a-d>b-c>0$. Let us also assume that the equilibrium BB is Pareto optimal, that is, $b>a$. By theorem 1 the learning process selects BB, that is, the benchmark actions are BB a very large proportion of the time (and hence these actions are played a very large proportion of the time). This contrasts with many other adaptive learning procedures – including ordinary log linear learning – that select the risk dominant equilibrium in 2 x 2 games (Kandori, Mailath and Rob, 1993; Young, 1993; Blume, 1993, 1995, 2003).

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

	A	B
A	$30,30$	$0,40$
B	$24,23$	$10,20$

This game has no pure Nash equilibria, so by theorem 1 the learning process selects the combination that maximizes $f_1 \cdot W - g_1 \cdot S$. Figure 2 illustrates the case $f_1 / g_1 = 1$ in which AA is selected. In general the outcome depends on the ratio f_1 / g_1 as shown in Figure 3. Note that the welfare maximizing state AA is selected whenever f_1 / g_1 is sufficiently large, that is, whenever the marginal change in the rate of acceptance by a discontent player (for a given small change in experienced payoff), is sufficiently large relative to the marginal change in the rate of acceptance of an experiment (for a given small change in the gain in payoff).

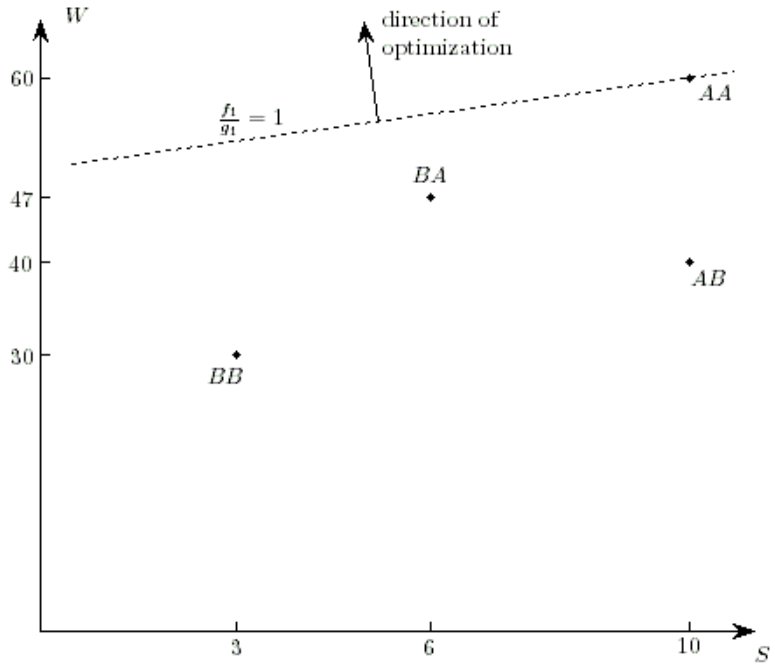


Figure 2. The tradeoff between welfare (W) and stability (S) when no equilibrium exists.

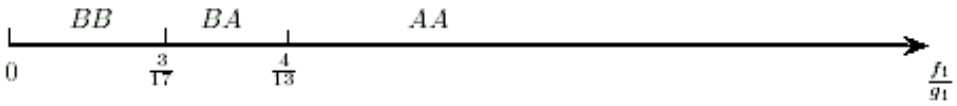


Figure 3. Stochastically stable outcomes as a function of f_1 / g_1

6. Proof of theorem 1.

Before turning to the details of the proof we provide an intuitive outline of the argument. When everyone is content, the learning process moves through a series of steps in which various players experiment and adopt new actions with probabilities that depend on the realized payoff gains. There are two possibilities.

1. A sequence of experiments leads to increasing payoffs for some and no decrease for anyone. In this case everyone eventually becomes content with higher payoff benchmarks.

2. A sequence of experiments ends with someone's payoff going down and staying down for two periods in a row. This person becomes discontent and starts searching widely. With some probability his searching causes other players to become discontent and they start searching. A key part of the argument is that, once a single player becomes discontent, there is a positive probability that all players will become discontent where this probability is bounded away from zero independently of ε.

The process only re-enters a content state once everyone has settled down again. By assumption the probability that everyone settles on a *particular* combination of actions $a \in A$ is proportional to $\prod_i \varepsilon^{F(u_i(a))}$, which is proportional to $\varepsilon^{-f_1 \sum_i u_i(a)}$. Thus high welfare states have an advantage in the sense that the process flows into them with higher probability than to other states. However, one must also consider the probability that the process exits from any given state. Here the equilibria are advantaged because it requires at least two experiments to exit to another (non-transient) state, whereas to exit from a disequilibrium state requires at most one experiment. The essence of the proof is to show that when pure Nash equilibria exist, stability takes precedence over welfare, whereas if no pure equilibrium exists there is an explicit *tradeoff* between welfare and stability.

We now turn to the details of the argument. The learning process defines a finite Markov chain on the state space Z. For every two states $z, z' \in Z$ there is a probability (possibly zero) of transiting from z to z' in one period. We shall write this one-period transition probability as a function of the experimentation rate ε: $P^{\varepsilon}_{zz'}$. The transition is *feasible* if $P^{\varepsilon}_{zz'} > 0$ whenever $\varepsilon > 0$. The *resistance of a feasible transition* $z \to z'$ is the unique real number $r(z, z') \geq 0$ such that $0 < \lim_{\varepsilon \to 0^+} P^{\varepsilon}_{zz'} / \varepsilon^{r(z,z')} < \infty$. The *resistance of an infeasible transition* is defined to be $r(z, z') = \infty$. If the probability of a transition is bounded away from zero when ε is small, we shall say that the transition has *zero resistance* or equivalently has probability of order $O(1)$. Similarly a transition with resistance *r* has probability of order $O(\varepsilon^r)$.

In general, a *recurrence class* of a finite Markov chain is a nonempty subset of states $R \subseteq Z$ such that: i) for every two distinct states $z, z' \in R$ there is a positive probability path from z to z' and a positive probability path from z' to z; ii) there is no positive probability path 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

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

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 $w \to z$ is *independent* of w. Next consider a transition of form $z \to w$ 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 $z \to w$. We can therefore collapse the set D into a single state $\bar{D}$ and define the transition probabilities as follows:

$$\forall z \in Z - D, \;\; P(\bar{D} \to z) \equiv P(w \to z) \text{ for all } w \in D \tag{11}$$

$$\forall z \in Z - D, \;\; P(z \to \bar{D}) \equiv P(z \to w) \text{ for some } w \in D \tag{12}$$

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.) 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. □

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 $r(w \to z)$ denote the total resistance of the least-resistant path that starts in w and ends in z. We call $w \to z$ an *edge* and $r(w \to z)$ the *resistance* of the edge.

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

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

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

The next three lemmas evaluate the function r^* on the various types of recurrence classes.

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

Proof. Let $e = (m, \overline{a}, \overline{u}) \in E^0$, where $\overline{a}$ is a pure Nash equilibrium. Consider any outgoing edge $e \to z$ 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 $\overline{a}$ 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 $r(e \to z) \geq 2$.

It remains to be shown that the resistance of the transition $e \to \overline{D}$ is exactly two, which we shall do by constructing a particular path from e to $\overline{D}$ in the full state space. Choose some player i. By interdependence there exists an action $a_i \neq \overline{a}_i$ and a player $j \neq i$ such that i's change of action affects j's payoff: $u_j(a_i, \overline{a}_{-i}) \neq u_j(\overline{a})$. 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 $u_j(a_i, \overline{a}_{-i}) > u_j(\overline{a})$, 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 $r(e \to \bar{D}) = 2$, and hence that $e \to \bar{D}$ is an easy edge. □

Lemma 3. $\forall z \in C^0 - E^0$, $r^*(z) = 1 + G(S(z))$, *and if* $z \to z'$ *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 $r(z \to z') = 1 + G(S(z))$. 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 $r(z \to \bar{D}) = 1 + G(S(z))$. 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, $G(\cdot) < 1/2$, so 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 $z \to z'$ is an easy edge with $z' \neq \bar{D}$, then $W(z) < W(z')$. □

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

Proof. Let $\bar{D} \to z = (m, \bar{a}, \bar{u}) \in C^0$. Recall that the transition probability $P(w \to z)$ 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 $r(\bar{D} \to z) \leq \sum_i F(\bar{u}_i)$.

We claim that in fact $r(\bar{D} \to z) = \sum_i F(\bar{u}_i)$. Let $z = (m, \bar{a}, \bar{u}) \in C^0$ and consider a least-resistant path $w^1, w^2, ..., w^m = z$. 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 $r(\bar{D} \to z) = \sum_i F(\bar{u}_i)$. 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. □

*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 $\rho(w) = \min\{r(T) : T \text{ is a } w\text{-tree}\}$.

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 $z \to \cdot$. If here are several such edges choose one that points to $\bar{D}$, that is, choose $z \to \bar{D}$ if it is easy. This implies in particular that for every $e \in E^0$ we select the edge $e \to \bar{D}$. (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 $z^1 \to z^2 \to ... \to z^m \to z^1$ 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 $z^k \to z^{k+1}$ 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. □

Let
$$\rho^* = \rho(\bar{D}) = r(T^*_{\bar{D}}). \tag{14}$$

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

$$T^*_z = T^*_{\bar{D}} \text{ with } z \to w_z \text{ removed and } \bar{D} \to z \text{ added } . \tag{15}$$

T^*_z *is a* z*-tree of least resistance and*

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

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 $\bar{D} \to z$. Hence it is a least-resistant *z*-tree among all *z*-trees that contain the edge $\bar{D} \to z$. 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 $\bar{D} \to z$. Since it is a spanning tree it must contain some outgoing edge from $\bar{D}$, say $\bar{D} \to z'$. We can assume that $r(\bar{D} \to z') < r(\bar{D} \to z)$, for otherwise we could simply take out the edge $\bar{D} \to z'$ and put in the edge $\bar{D} \to z$ to obtain the desired result.

Let $u_1, u_2, ..., u_n$ be the benchmark payoffs in z and let $u_1', u_2', ..., u_n'$ be the benchmark payoffs in z'. By lemma 4,

$$r(\bar{D} \to z) = \sum_i F(u_i) \text{ and } r(\bar{D} \to z') = \sum_i F(u_i'). \tag{17}$$

Since $r(\bar{D} \to z') < r(\bar{D} \to z)$ and $F(u_i)$ is 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 $z' = z^1, z^2, ..., z^m = z$, where each z^k is in R and they are distinct. Each edge $z^k \to z^{k+1}$ 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

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

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 $r(z^{k_i} \to z^{k_i+1}) \geq F(u_i) + r^*(z^{k_i})$. It follows that the total resistance along the sequence $z' = z^1, z^2, ..., z^m = z$ satisfies

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

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

$$r(\bar{D} \to z') = \sum_i F(u_i') > \sum_{i \notin I} F(u_i') \geq \sum_{i \notin I} F(u_i). \qquad (20)$$

Hence in the tree T_z the outgoing edges from the nodes $\{\bar{D}, z^1, z^2, ..., z^{m-1}\}$ 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. □

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

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

(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 $e \to \bar{D}$ is an edge in the easy tree $T_{\bar{D}}^*$ (see the beginning of the proof of Lemma 5), so (16) implies that

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

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

$$r(\bar{D} \to e) = \sum_i F(u_i(\bar{a})) = -f_1 W(e) + nf_2. \qquad (23)$$

From this and (22) we conclude that

$$\rho(e) = \rho^* - 2 - f_1 W(e) + n f_2 . \tag{24}$$

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 $\rho(z) = \rho^* - r(z \to w_z) + r(\bar{D} \to z)$, where $z \to w_z$ is an easy edge. Since $z \in C^0 - E^0$, we know from Lemma 3 that $r(z \to w_z) = 1 + G(S(z))$ and from Lemma 4 that $r(\bar{D} \to z) = -f_1 W(z) + n f_2$. Hence

$$\forall z \in C^0 - E^0 : \rho(z) = \rho^* - 1 - G(S(z)) - f_1 W(z) + n f_2 . \tag{25}$$

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

$$\rho(e) = \rho^* - 2 - f_1 W(e) + n f_2 . \tag{26}$$

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

$$f_1 [W(z) - W(e)] < 1 - G(S(z)) . \tag{27}$$

The right-hand side of (27) is greater than ½ 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 ½ 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

$$-f_1 W(z) + n f_2 \le n \cdot \max_u F(u) < 1/2 \quad \text{and} \quad G(\Delta u) > 0 . \tag{28}$$

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 ($E^0 \neq \varnothing$).

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 ($E^0 = \varnothing$).

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

$$\begin{aligned} \rho(z) &= \rho^* - 1 - G(S(z)) - f_1 W(z) + nf_2 \\ &= g_1 S(z) - f_1 W(z) + (nf_2 + \rho^* - 1) \end{aligned} \tag{29}$$

which is equivalent to maximizing $f_1 W(z) - g_1 S(z)$. This concludes the proof of theorem 1.

7. Heterogeneous and nonlinear acceptance functions

The proof is constructed in a way that allows us to see the impact of relaxing our conditions on the acceptance functions F and G. In particular, let us suppose that these functions differ among players and may be nonlinear. Suppose that *i*'s probability of accepting the outcome of an experiment is governed by the function $G_i(\Delta u_i)$, and that *i*'s probability of stopping a search is governed by the function $F_i(u_i)$. As before we shall assume that these acceptance functions are strictly decreasing. These assumptions are not very restrictive. They amount to saying that each agent accepts the outcome of an experiment with a probability that increases

with the payoff gain from the experiment, and accepts the outcome of a random search with a probability that increases with the payoff level of his current choice.

To obtain tight analytical results, we need to assume (as before) that the probabilities of acceptance are large relative to the probability of experimentation. In particular it suffices to assume that

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

Let $z = (m, \bar{a}, \bar{u})$ and redefine the welfare function as follows:

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

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. Define

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

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

$$\tilde{S}(z) = \max_i\left\{\tilde{S}_i(z)\right\}. \tag{33}$$

The larger $\tilde{S}(z)$ is, the more likely it is that some player will accept the outcome of an experiment, hence it is a measure of instability. 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 ε and acceptance functions F_i and G_i satisfying conditions (1)-(3).*

(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 optimize some criterion of system-wide performance using simple feedback rules and 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 – deliberate experimentation and flailing around. Theorem 1 shows that by choosing the probabilities of experimentation and acceptance within an appropriate range, the process is in a welfare-maximizing 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. Prior work would lead one to expect that it may take an exponentially long time to learn Nash equilibrium as a function of the number of players, though this has only been shown for procedures that actually converge to Nash equilibrium (Hart and Mansour, 2010). Faster learning may be possible if one only insists on coming close to equilibrium with high probability, but this is a complex issue that we shall not attempt to address here.

Finally, we should note that the kind of learning rules we have described 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 for suggesting a number of improvements to an earlier draft, which was entitled "Efficiency and Equilibrium in Trial and Error Learning." This research was supported by grants from the Office of Naval Research (#N00014-09-1-0751) and the Air Force Office of Scientific Research (FA9550-09-1-0538).

References

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, 2010, "Completely uncoupled dynamics and Nash equilibria," Working Paper, Center for the Study of 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.

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., Bezalel 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.